

**Zurich University of Applied Sciences**

School of Management and Law

MSc Business Information Systems

Spring Semester 2023

*Master's Thesis*

# **How does the introduction of leniency influence cartel activity?**

Evaluation of Legal Texts using Natural Language Processing

**Rebecca Baumann**



W.MA.WIN.21HS.a

Matriculation no: 15530751

Supervisor: Nicole Bellert

Co-Supervisor: Maria Pelli

Date: 31 May 2023

*“If we knew what it was we were doing,  
it would not be called research, would it?”*

– Stedman & Beckley (2007)

## MANAGEMENT SUMMARY

This master's thesis looks at the impact of the European leniency program on detected cartel activity. The starting point is a study by Harrington and Chang, in which the authors conclude that while the introduction of a leniency program aims to reduce cartel activity by increasing the benefit of self-reporting, it can either reduce or increase the cartel rate depending on whether the program diverts resources from non-lenieny enforcement to leniency cases, leading to a crowding-out effect. Derived from this, the following research question is explored: “*How did the introduction of the leniency program of the European Commission in 1996 affect the cartel activity in the European Union?*” Thereby, the change regarding non-lenieny enforcement is also investigated.

The research methodology of design science research is applied, following the three-cycle concept of design science proposed by Hevner, which consists of the cycles of relevance, rigour, and design. Primary data for the study is collected through web-scraping of publicly available resources from the European Commission and secondary data is accessed from an existing manually compiled dataset. Concerning data extraction directly from the scraped prohibition decisions, different Natural Language Processing techniques are employed, and their accuracy rates analysed. Applying an empirical, quantitative cross-sectional survey approach including longitudinal observations, the data collected allows for a comprehensive analysis of temporal trends and patterns across the different cartel entities. Statistical data analysis including panel and multiple linear regression is used to test the hypotheses derived from the research question and its sub-questions.

The results show a nuanced impact of the European Commission’s leniency program on cartel activity. After the introduction of the leniency program in 1996, there was a modest but statistically significant decline in the detection of cartel activity. This decline was also observed in cases detected through non-lenieny enforcement after the program's introduction. Nevertheless, detection rates after 1996 through leniency enforcement were not significantly lower than detection rates for non-lenieny enforcement prior to 1996, suggesting a likely reallocation of resources from non-lenieny enforcement to prosecution of cases based on leniency applications. However, the introduction of the leniency program did not have a significant impact on the duration of detected cartels. Some data even suggest a potential increase in cartel duration, but this finding is not statistically supported. The relationship between the average amount of the imposed fines and the duration of cartels is also not found to be significant. Overall, the results paint a complex

picture of the impact of the European Commission's leniency program on cartel detection and duration.

Analysing various Natural Language Processing techniques showed variable effectiveness based on the type of data extracted. Regular expressions demonstrated a strong performance in identifying case numbers and decision dates, with an impressive 97.87 % and 100 % accuracy rate respectively, though this was heavily dependent on customizing patterns to match the distinct document formats presented by the prohibition decisions. Keyword matching proved to be efficient in detecting instances in which there was an application for leniency, achieving a 90.43 % accuracy rate. Combining Named Entity Recognition, keyword matching and regular expressions delivered mixed results, especially in pinpointing the start and end dates of cartels, which constituted the most difficult data extraction task. Overall, careful selection and combination of Natural Language Processing techniques is vital to meet specific data extraction needs.

Future research should expand the scope of the research conducted in this thesis to include cases published in other languages than English, which would help to mitigate selection bias and offer a more comprehensive view of the impact of the leniency program on cartel activity. Further work should also explore different methods or additional data sources to address the limitations of this study. To enhance analysis, Natural Language Processing techniques could be refined and advanced models, such as BERT or Transformers, could be evaluated. This could improve data extraction, especially for important information like formal decision of the European Commission, the names of cartel members and sector information, which would lead to a deeper understanding of the impact of the European Commission's leniency program regarding possible differential impacts in dissimilar industries and could provide insights on repeat offenders.

## TABLE OF CONTENTS

Management Summary.....	III
Table of Figures.....	VIII
Table of Tables.....	X
Table of Abbreviations.....	XI
1 Introduction.....	1
1.1 Starting Point.....	1
1.2 Problem Definition.....	1
1.3 Objective.....	4
1.4 Research Question.....	5
1.5 Field of Application.....	5
1.6 Artefact Requirements.....	6
1.7 Structure of the Thesis.....	9
2 Methodical Approach.....	10
3 Related Work.....	14
3.1 Leniency Programs and Cartel Deterrence.....	14
3.2 Reduction in Imposed Fines.....	17
3.3 Investigation and Prosecution by Cartel Authority.....	18
3.4 Managing Disclosure Incentives and Abuse Risks.....	19
3.5 Challenges and Research Gap.....	20
4 Methodology.....	22
4.1 Data Acquisition.....	22
4.2 Data Extraction with NLP.....	27
4.2.1 Case Number.....	31
4.2.2 Decision Date.....	34
4.2.3 Cartel Start, End and Duration.....	37
4.2.4 Report Route and Route Indicator.....	42

4.2.5	Leniency .....	44
4.3	Data Preparation .....	45
4.4	Data Analysis .....	48
4.4.1	General Analysis .....	50
4.4.2	Non-leniency Enforcement.....	60
4.4.3	Non-leniency Enforcement before 1996 versus Leniency after 1996 .....	66
4.4.4	Cartel Duration .....	70
5	Results .....	73
5.1	Influence of Leniency on Cartel Activity .....	73
5.1.1	General Analysis .....	73
5.1.2	Non-leniency Enforcement.....	74
5.1.3	Non-leniency Enforcement before 1996 versus Leniency after 1996 .....	75
5.1.4	Cartel Duration .....	76
5.1.5	Limitations.....	77
5.2	Analysis of NLP Techniques for Data Extraction .....	78
5.2.1	Data Extraction Results .....	78
5.2.2	Limitations.....	79
5.3	Evaluation of Artefacts .....	80
5.3.1	Scraping Scripts .....	80
5.3.2	NLP Scripts .....	82
5.3.3	Data Analysis Scripts .....	84
6	Discussion .....	86
6.1	Effectivity of Leniency Programs.....	86
6.2	Investigation by Cartel Authority through Non-leniency Enforcement .....	88
6.3	Cartel Duration and Fines .....	89
7	Conclusion and Future Work .....	91
	Acknowledgements .....	XII

References .....	XIII
Declaration of Authorship .....	XIX

## TABLE OF FIGURES

Figure 1: <i>Conceptional model regarding influence factors contributing to overall cartel activity (own illustration)</i> .....	4
Figure 2: <i>DSR cycles based on Hevner et al. (2004), Hevner (2007)</i> .....	10
Figure 3: <i>DSR method model, Peffers et al. (2007)</i> .....	12
Figure 4: <i>Cartel cases in the year 1964, screenshot of the EC's website</i> .....	23
Figure 5: <i>Search mask on the website of the EC, screenshot of the EC's website</i> .....	24
Figure 6: <i>Example of a table structure where the PDF file to the prohibition decision is provided in the cell located next to the term "Prohibition Decision", screenshot of the EC's website</i> .....	25
Figure 7: <i>Example of a table structure where the prohibition decision is found on another page, Screenshot of the EC's website</i> .....	26
Figure 8: <i>Example of table structure where there is a link to the English PDF without the ".pdf" in the link, Screenshot of the EC's website</i> .....	26
Figure 9: <i>Plot of active cartels per year (own illustration)</i> .....	49
Figure 10: <i>PanelOLS Estimation Summary regarding the impact of the leniency program of the EC on overall detected cartel activity (own illustration)</i> .....	51
Figure 11: <i>Scatterplot of residuals vs. fitted values (1) and dependent variable vs independent variables (2 and 3) (own illustration)</i> .....	52
Figure 12: <i>Line plot of active cartels over time with vertical line in 1996 (own illustration)</i> .....	53
Figure 13: <i>Correlation matrix for the independent variables (own illustration)</i> .....	54
Figure 14: <i>Scatterplot of residuals against unassigned row numbers (own illustration)</i> .....	55
Figure 15: <i>Scatterplot of residuals against independent variables (own illustration)</i> ...	55
Figure 16: <i>PanelOLS Estimation Summary where heteroskedasticity and serial correlation have been accounted for (own illustration)</i> .....	59
Figure 17: <i>Total number of cases per report route comparison (own illustration)</i> .....	61
Figure 18: <i>Report routes in percentage for both periods (own illustration)</i> .....	62
Figure 19: <i>Cartel activity discovered by non-lenieny enforcement (own illustration)</i> .	63
Figure 20: <i>PanelOLS Estimation Summary for filtered data only including cartel activity discovered by non-lenieny enforcement (own illustration)</i> .....	64



Figure 21: <i>PanelOLS Estimation Summary for non-leniency enforcement where serial correlation and heteroskedasticity have been accounted for (own illustration)</i> .....	65
Figure 22: <i>Cartel activity discovered by non-leniency enforcement before 1996 vs. leniency afterwards (own illustration)</i> .....	67
Figure 23: <i>PanelOLS Estimation Summary on non-leniency cases before 1996 vs. leniency cases from 1996 onwards (own illustration)</i> .....	68
Figure 24: <i>PanelOLS Estimation Summary for non-leniency enforcement before 1996 vs. leniency afterwards where serial correlation and heteroskedasticity have been accounted for (own illustration)</i> .....	69
Figure 25: <i>Multiple linear regression regarding impact of leniency on cartel duration (own illustration)</i> .....	71

## TABLE OF TABLES

Table 1: <i>Data needed for comprehensive analysis of the effect the introduction of leniency had on the detected cartel activity (own illustration)</i> .....	27
Table 2: <i>Data needed to perform linear regression with interrupted time series for evaluating the impact of the introduction of leniency on the detected cartel activity (own illustration)</i> .....	30
Table 3: <i>Case number formats from 2010 to 2021 (own illustration)</i> .....	31
Table 4: <i>Decision date formats from 2010 to 2021 (own illustration)</i> .....	34
Table 5: <i>Accuracy rates regarding extraction of cartel start date and cartel end date for versions 2d_a and 2d_b (own illustration)</i> .....	40
Table 6: <i>Success rates regarding extraction of cartel start date and cartel end date for version 2e (own illustration)</i> .....	41
Table 7: <i>Correlation matrix between independent variables (own illustration)</i> .....	54
Table 8: <i>Accuracy rates regarding information extraction using different NLP techniques (own illustration)</i> .....	78
Table 9: <i>Evaluation of the artefact "scraping scripts" (own illustration)</i> .....	80
Table 10: <i>Evaluation of the artefact "NLP scripts" (own illustration)</i> .....	82
Table 11: <i>Evaluation of the artefact "data analysis scripts" (own illustration)</i> .....	84

## TABLE OF ABBREVIATIONS

API	Application Programming Interface
DSR	Design Science Research
EC	European Commission
EU	European Union
GPT	Generative Pretraining Transformer
NER	Named Entity Recognition
NIIC	No-immunity-for-instigators clause
NLP	Natural Language Processing
OLS	Ordinary least squares
Regex	Regular expressions
US	United States
VIF	Variance Inflation Factor

# 1 INTRODUCTION

## 1.1 Starting Point

The leniency program of the European Commission (EC) offers the companies involved in a cartel either complete or partial immunity from fines if they self-report and hand over evidence (EC, 2023). It was introduced in 1996, following the surge in amnesty applications in the wake of the 1993 revision of the Corporate Leniency Program of the United States' (US) Department of Justice's Antitrust Division. Under the 1996 Leniency Notice, the first company to inform the EC about the existence of a secret cartel thereby benefited from either a reduction of at least 75 % of the fine or even from a total exemption from the fine. Other cartel members that cooperated after the EC had already initiated the investigation also benefited from a reduction of the fine, ranging from 10 % to 75 %, depending on the circumstances (EC, 1996). Reports from various implemented leniency programs show that such programs led to numerous applications. However, despite the clear increase in leniency applications, the question poses itself as to whether the programs were also successful in the sense that the actual cartel rate in those countries declined (Harrington Jr. & Chang, 2015; Jochem et al., 2020). The leniency program of the EC is still in force today, although a new Leniency Notice has been introduced in 2002 and again in 2006 (EC, 2006, 2002). The overall concept remains the same, even though one alteration concerns the amount of the fine reduction which has changed slightly for cartel members cooperating after the initiation of proceedings. Under the 2006 Leniency Notice in force at the time of writing this thesis, the first company that provides added value to the proceedings benefits from a reduction between 30 % to 50 %, the second one from 20 % to 30 % and the others from up to 20 % (EC, 2006).

## 1.2 Problem Definition

According to Harrington and Chang (2015), the assessment of leniency programs is generally based on the assumption that they do not affect non-leniency enforcement. Non-leniency enforcement involves the discovery of cartels, their prosecution, and conviction by the authorities without prior notification by one of the cartel members (Harrington & Chang, 2015, p. 418).

Since cartel members only have the incentive to file a voluntary report if the probability of being caught is high, there is a correlation between non-leniency enforcement and the number of reports through the leniency program. Indeed, if there is less non-leniency

---

---

enforcement, the likelihood of being caught and convicted is rather low. Accordingly, fewer cartels are uncovered by the leniency program. The non-lenieny enforcement thus has a direct effect on the number of reports received from the leniency program (Harrington & Chang, 2015, p. 418 et seq.).

In contrast, the introduction of the leniency program also has an impact on non-lenieny enforcement. However, it is not entirely clear in which direction that impact goes. There are two possibilities. Either, because more complaints are received from the leniency program, the authorities have fewer resources to deal with non-lenieny enforcement, which is weakened. Or, because more voluntary reports are received, there are generally fewer cartels in the economic area concerned, decreasing numbers in non-lenieny enforcement altogether (Harrington & Chang, 2015, p. 419).

Thus, Harrington and Chang (2015) assume that while non-lenieny enforcement can be expected to change with the introduction of the leniency program, it is not clear whether it will be strengthened or weakened. In their study, the authors addressed this issue (p. 419).

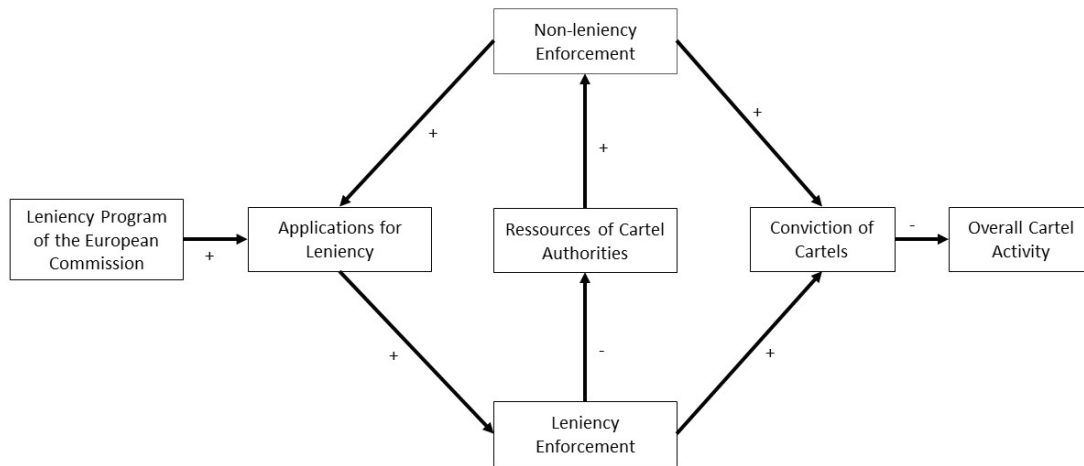
They deduced that, given a fixed non-lenieny enforcement rate, the introduction of a leniency program leads to a reduction in cartel activity (Harrington & Chang, 2015, p. 432). The subsequent findings form the basis for this inference. A leniency program increases the advantage of fraud because a business can now lower its fine by simultaneously applying for leniency. This reduces the range of market conditions in which collusion is stable, thus shortening the anticipated life of a cartel and decreasing the value of collusion. A leniency scheme further reduces the value of collusion. This is because, in the case of a collapsing cartel, the cartel members fight to be the first to apply for leniency. This leads to an increase in expected penalties. A cartel either doesn't develop or lasts less time as a result of the decreased value of collusion, which then results in a lower total cartel rate (Harrington & Chang, 2015, p. 433).

Contrary to the assumption of a given non-lenieny enforcement rate, assuming endogenized non-lenieny enforcement the new program can either increase or decrease the cartel rate. Which of those two outcomes produces itself depends on whether the program shifts resources from non-lenieny cases to leniency cases. Such a shift is likely to happen when the penalties are not harsh enough and not enough prosecutorial resources can be saved by a leniency program (Harrington & Chang, 2015, p. 434). Non-lenieny

enforcement usually aims at prosecuting active cartels. If a leniency program is introduced, the cartels that are about to collapse will seek to self-report. This in turn shifts resources from exposing active cartels to prosecuting cartels that are already coming to an end anyway. This then again creates more work for the authorities, who, instead of focusing on a well-functioning cartel, may now focus on a cartel that is already dying. In this case, the leniency program crowds out non-lenieny enforcement. For this crowding-out effect to occur, it is evidently important that non-lenieny enforcement existed before the introduction of the leniency program (Harrington & Chang, 2015, p. 435 et seq.).

What also must be considered, according to Harrington and Chang (2015), is the fact that a leniency program has a differential impact on the cartel rate depending on the industry (p. 439). The difference arises depending on whether an industry has stable or unstable cartels. Unstable cartels occur when breaching the collusion leads to higher profits. Such instability can be indicated by more firms involved or by higher price elasticity of the firm's demand function. The differential effect is driven by the dying cartels that use the leniency program. In industries with less stable cartels, the rate of dying cartels is higher. The authors proved in their study that in the case of weakened non-lenieny enforcement, the unstable cartels are harmed, and the stable cartels are benefited from a leniency program. This is because the unstable cartels make use of the leniency program, while the stable cartels are no longer prosecuted because there are no longer enough prosecutorial resources. However, if the non-lenieny program is strengthened by the introduction of a leniency program, the average duration of a stable cartel and with that, the cartel rate, declines. The highly stable cartels are not interested in the higher penalties resulting from a leniency race – because a race to apply for leniency is unlikely for these cartels – but in whether they are more or less likely to be prosecuted and convicted outside the leniency program (Harrington & Chang, 2015, p. 442 et seq.).

The conceptual model in Figure 1 can be roughly derived from the aforementioned elaborations.



**Figure 1:** *Conceptual model regarding influence factors contributing to overall cartel activity (own illustration)*

If the increase of a construct has a positive effect (same-pole effect) on the connected construct, this is marked with a plus, if the increase of a construct has a negative effect (opposite-pole effect) on the connected construct, it is marked with a minus. It is important to note that the final impact on overall cartel activity depends on the unknown factors mentioned, such as a possible shift of resources and industry.

In conclusion, the introduction of a leniency program may increase the expected penalties and consequently shorten the duration of cartels in industries where collusion is least stable (or prevent cartels from forming in the first place), while it may lengthen the duration of cartels in industries where collusion is most stable because non-leniency enforcement is weaker. The results of Harrington’s and Chang’s research can be used to examine whether leniency weakens the enforcement of non-leniency, which is a crucial prerequisite for leniency to increase the cartel rate (Harrington & Chang, 2015).

### 1.3 Objective

Interdisciplinarity between different research domains often paves the way for new insights. This study employs techniques of business information systems and data science to illuminate a question in the business law domain. The aim of this work is to evaluate the impact the introduction of the leniency program of the EC had on the overall cartel activity within the material scope of the European Union (EU).

A conclusion on this question will be drawn using regression analysis with time series based on existing data collected manually from 1964 to 2010. In addition to the information provided by an existing dataset, new data from the prohibition decisions of the EC

published in English language from 2010 to 2023 will be included in the regression model. Based on this, it will also be determined what non-leniency enforcement looked like before the introduction of the leniency program and how it has changed with its introduction. At the same time, the average life span of the cartels before and after the introduction of the leniency program are subjected to a comparison.

To facilitate future research in this field, Jupyter Notebook scripts are developed to automatically download the EC's prohibition decisions, conduct data extraction by employing NLP techniques and finally carry out a regression analysis. These resulting scripts constitute the artefacts of this thesis.

#### **1.4 Research Question**

The master's thesis aims to answer the following research question:

*How did the introduction of the leniency program of the European Commission in 1996 affect the cartel activity in the European Union?*

To approach this main research question, the following sub-questions are addressed:

1. How does the level of non-leniency enforcement before the introduction of the leniency program compare to the level of non-leniency enforcement after its introduction?
2. How does the level of non-leniency enforcement before the introduction of the leniency program compare to the level of leniency enforcement after its introduction?
3. How has the average lifetime of cartels changed as a result of the introduction of the leniency program?

The answers to each of these sub-questions serve to answer the main research question, forming a comprehensive conclusion.

#### **1.5 Field of Application**

The scope of this master's thesis is limited to cases under the jurisdiction of the EC. The published decisions date back to 1964. For the analysis, only cases published in the English language are considered.

Regarding cases from 2010 onwards, where no pre-existing data is available, only cases for which the corresponding prohibition decisions have been published as PDF files that

---



contain text that can be converted into machine-readable format are utilised. The cases where the decision is not published in English or the prohibition decisions have been scanned, making them difficult to convert for further use, are excluded from the study.

Furthermore, this research does not differentiate between industries with stable and industries with unstable cartels.

## 1.6 Artefact Requirements

This thesis will produce the following artefacts: 1) two Jupyter Notebook scripts with which the prohibition decisions in English language can be scraped from the website of the EC (referred to as scraping scripts in the following), 2) several Jupyter Notebook scripts in which different NLP methods to extract information from the downloaded PDF files are being explored (further referred to as NLP scripts) and 3) several Jupyter Notebook scripts which conduct different regression analyses with time series based on existing data as well as on additionally collected data (referred to as data analysis scripts henceforth). All the scripts are available on GitHub under [https://github.com/baumareb/cartel\\_analysis](https://github.com/baumareb/cartel_analysis).

The requirements for the artefacts are loosely based on the five defined domains of Prat et al. (2014) for artefact assessment, although the artefacts are not analysed in as much detail, since how the evaluation can be carried out and the artefacts can finally be assessed always depends on the corresponding application area (Prat et al., 2014). In the following, the requirements that shall be met for the respective Jupyter Notebook scripts are presented.

- 1) Web scraping scripts:
  - a. Goal:
    - i. There shall be two scripts: 1) cases from 1964 to 1998 and 2) cases from 1999 to today's date.
    - ii. For every case number, only one PDF file shall be downloaded.
    - iii. The PDF file that will be downloaded by the scripts shall be the latest prohibition decision on the case, i.e., where there are more than one PDF files named "Prohibition Decision", the one with the more recent date shall be downloaded.

- iv. The cases from 1964 to 1998 shall be saved in a newly created folder “cases\_until\_1998” and the cases from 1999 to the current date shall be saved in a newly created folder “cases\_from\_1999”.
  - b. Environment:
    - i. The scripts shall be compatible with the Anaconda environment and should execute successfully in Jupyter Notebook.
    - ii. The scripts shall be written in the Python programming language.
  - c. Structure:
    - i. The scripts shall include sufficiently clear documentation in English language so the code can be understood and used by other researchers for future studies.
  - d. Activity:
    - i. The scripts shall include all the necessary install statements that can be run if the necessary libraries are not yet installed on the end user’s device.
  - e. Evolution:
    - i. The scripts should include error handling capabilities to manage exceptions, such as missing data, inaccessible files, or unsupported file formats.
- 2) NLP scripts:
- a. Goal:
    - i. The scripts must be capable of extracting the following information from the previously collected PDF files:
      1. Case number
      2. Decision date
      3. Cartel start date and cartel end date, and, accordingly, cartel duration
      4. Report route and report route indicators
      5. Whether there was an application for leniency
    - ii. Different NLP methods shall be tested to read out the specified data; where multiple techniques are employed to extract the same data, they should be implemented in separate Jupyter Notebook scripts to facilitate the comparison of results.
-

- iii. The extracted data must be stored in an Excel file, which should include all cases and their corresponding data.
  - b. Environment:
    - i. The scripts shall be compatible with the Anaconda environment and should execute successfully in Jupyter Notebook.
    - ii. The scripts shall be written in the Python programming language.
  - c. Structure:
    - i. The scripts shall include sufficiently clear documentation in English language so the code can be understood and used by other researchers for future studies.
  - d. Activity:
    - i. The scripts shall include all the necessary install statements that can be run if the necessary libraries are not yet installed on the end user's device.
  - e. Evolution:
    - i. The scripts should include error handling capabilities to manage exceptions, such as missing data, inaccessible files, or unsupported file formats.
- 3) Data analysis scripts:
- a. Goal:
    - i. The scripts combine both the manually collected data until 2010 that is accessed in the folder "Data-until-2010" from the existing file "Cartels1964-2010.xls" as well as data after 2010, whereas case number and decision date are taken from the output of the NLP scripts and the missing information is added manually.
    - ii. For every sub-question to the research question, the scripts include the corresponding regression model as well as assumption testing.
    - iii. A separate Jupyter Notebook script shall be created for each sub-question in order to distinguish between the different aspects of the research question.
  - b. Environment:
    - i. The scripts shall be compatible with the Anaconda environment and should execute successfully in Jupyter Notebook.
    - ii. The scripts shall be written in the Python programming language.
-

- c. Structure:
  - i. The scripts shall include sufficiently clear documentation in English language so the code can be understood and used by other researchers for future studies.
- d. Activity:
  - i. The scripts shall include all the necessary install statements that can be run if the necessary libraries are not yet installed on the end user's device.
- e. Evolution:
  - i. The scripts should include error handling capabilities to manage exceptions, such as missing data, inaccessible files, or unsupported file formats.

The following prerequisites for the Jupyter notebooks need to be met by the end user:

1. Anaconda Environment must be installed.
2. Jupyter Notebook must be installed.
3. Pip Installs Packages must be installed.
4. Google Chrome must be installed.

## **1.7 Structure of the Thesis**

The remainder of this thesis is structured as follows: Section 2 Methodical Approach describes the systematic approach followed during the research process. Section 3 Related Work reviews existing literature and studies on leniency programs and cartel deterrence. It also identifies the research gap that this research aims to contribute to filling. Section 4 Methodology explains in detail the methods used for data acquisition, extraction, preparation, and analysis. It outlines the role of NLP in the extraction of data and presents the statistical models used for analysing it. Section 5 Results presents the findings from the data analysis. It assesses the influence of leniency on cartel activity, evaluates the effectiveness of the NLP techniques used for data extraction, and reviews the performance of the developed artefacts. In Section 6 Discussion, the results are interpreted and their implications for both theory and practice discussed. Section 7 Conclusion and Future Work summarises the research, draws conclusions based on the findings, and suggests potential directions for future research to address the limitations of the study.

## 2 METHODOICAL APPROACH

The research design used in this thesis is the Design Science Research (DSR) methodology, which is an important research paradigm in the field of information systems for conducting applicable but still rigorous research. Its strength lies in its practicality (Peffer et al., 2006, p. 84). This approach has been chosen for this thesis because the main part consists of the development of an artefact. This artefact is composed of different Jupyter Notebook scripts (see Section 1.6 Artefact Requirements).

The methodology that will be applied is thus roughly based on the three-cycle view of DSR proposed by Hevner (2007), as illustrated in Figure 2 below.

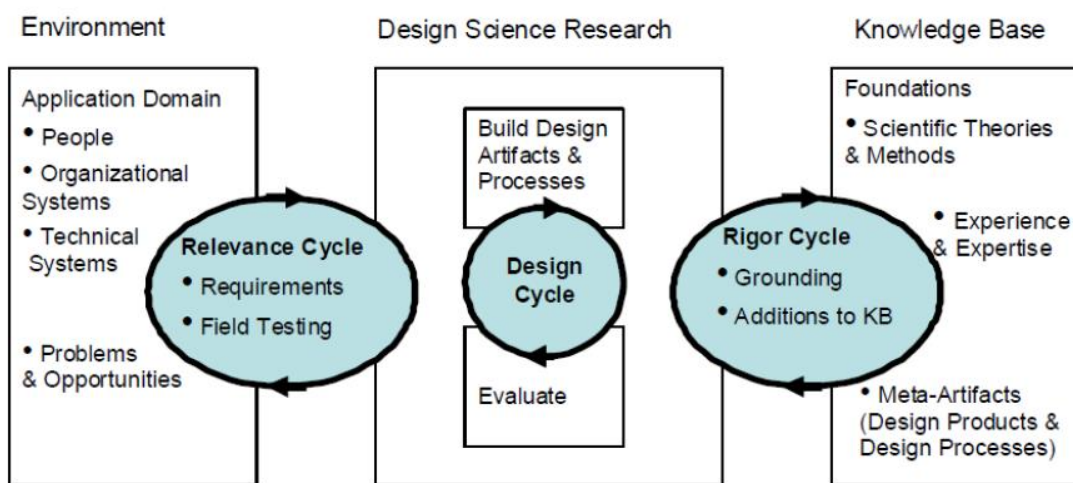


Figure 2: DSR cycles based on Hevner et al. (2004), Hevner (2007)

These cycles consist of 1) the relevance cycle, in which the relevance of answering the research question is stated through the formulation of the starting point, the problem definition, and the objective of the research, 2) the rigor cycle, in which research that has already been done is analysed and put into context with the planned research, and 3) the design cycle, in which the practical research is carried out (Hevner, 2007, p. 88 et seq.).

The relevance cycle establishes the connection between the contextual environment of the research project and the activities that are carried out during the DSR on that project (Hevner, 2007, p. 89). The relevance cycle is illustrated in this thesis in Section 1 Introduction. The problem definition together with the formulation of the objectives of the study serve to put this research into context with the practice-oriented environment. Furthermore, Section 1.6 Artefact Requirements defines the requirements for the artefact to be created and presents the evaluation criteria, based on which a decision is made at the

end of the project whether a further iteration of the relevance cycle is necessary (Hevner, 2007, p. 89).

The rigor cycle connects the DSR activities with the theoretical knowledge base, which is built on scientific theories and methods, experience and expertise, and meta-artefacts (Hevner, 2007, p. 89 et seq.). In this cycle, research done up to date is analysed. Starting with the research from Harrington and Chang (2015), an in-depth literature review is conducted in Section 3 Related Work. The thesis deals with a mature topic on which there is already an extensive body of research that needs to be analysed and summarised. For this reason, a rather thorough literature review according to Webster and Watson (2002) on related work is conducted. This represents the knowledge base. Based on the current state of the art, the research gap is defined, which the creation of the artefact is intended to contribute to closing, aiming to expand existing research (Webster & Watson, 2002, p. 14). The finished master's thesis then contributes to the knowledge base by gaining new insights through the created artefact. This contribution to existent research is discussed in Section 6 Discussion.

Finally, the centrally located design cycle revolves around the creation of design artefacts and processes and their ongoing evaluation (Hevner, 2007, p. 90 et seq.). This process is mapped in Section 4 Methodology. This is where the creation of the artefacts, the main part of this master's thesis, is described. In Section 5.3 Evaluation of Artefacts, where the results of the thesis are presented, the created artefacts are assessed and suggestions for their further improvement in a new DSR iteration are made based on the evaluation criteria described by Prat et al. (2014).

The described procedure is also in line with the DSR process according to Peffers et al. (2007), which is displayed in Figure 3.

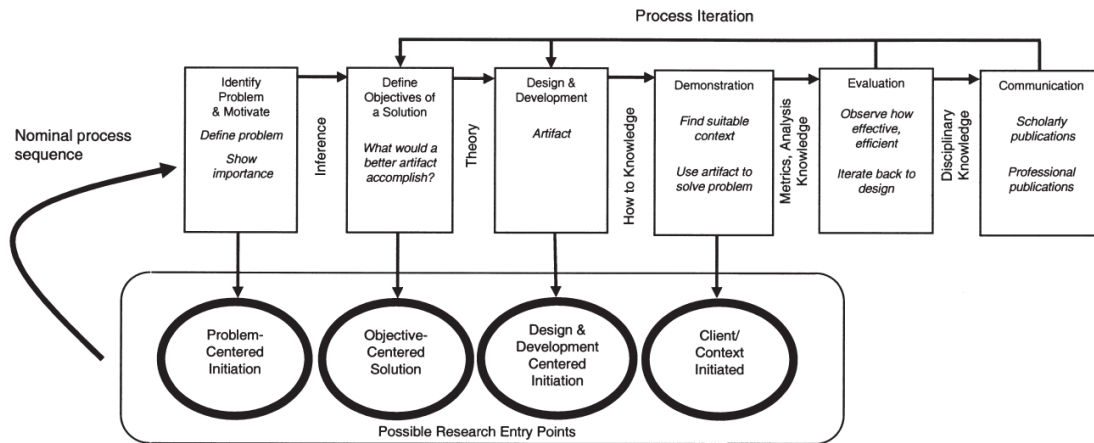


Figure 3: DSR method model, Peffers et al. (2007)

As this thesis starts with the formulation of a research question, it qualifies as a problem-centred initiation (Peffers et al., 2006, p. 54). It also needs to be noted that due to the limited timeframe of this thesis, it is not feasible to complete all steps or to iterate multiple times during the development of the artefact, as would be envisaged by the model of Peffers et al. (2007).

The practical research component of this study uses a mix of primary and secondary data. While web scraping methods are applied to obtain first-hand data directly from the publicly available resources of the EC, a significant part of the data is also derived from an existing dataset put together by previous researchers. This fusion of data sources allows for a comprehensive empirical and quantitative analysis. In particular, the data contain both longitudinal observations that capture a sequence of events over time, allowing the study of temporal trends and patterns and cross-sectional data, as multiple data points which refer to various cartels as different entities are analysed simultaneously. The practical research can thus be qualified as an empirical, quantitative cross-sectional survey based on systematic observation, where longitudinal observations are included (Wilde & Hess, 2006, p. 8). The data is collected at one specific point in time, and the collection is not repeated during this master's thesis (Alavi & Carlson, 1992, p. 47 et seq.).

Usually, a linear and highly structured research process is realised in the quantitative paradigm. This process generally starts with theory work and the development of hypotheses, collects numerical data as representative of a sample as feasible, and ultimately results in statistical data analysis for hypothesis testing (Döring & Bortz, 2016, p. 52 et seq.). It is precisely this process that is followed in this thesis as well.

Based on the findings presented by Harrington and Chang (2015), a central research question is formulated, which is divided into sub-questions for further specification. For each of these sub-questions, a corresponding null hypothesis is derived, which serves as a theoretical construct for the subsequent investigation. The hypotheses are then subjected to rigorous testing using the data collected. Methods of statistical data analysis, especially regression analyses, are used to perform these tests. Panel regression is used for the time series data, while multiple linear regression is used to examine the questions where time is not a factor. Based on the results of these analyses, conclusions are drawn regarding the research question posed in Section 1.4 Research Question that not only confirm or reject the formulated hypotheses, but also contribute new insights to the existing body of knowledge.



---

### 3 RELATED WORK

#### 3.1 Leniency Programs and Cartel Deterrence

The paper of Harrington and Chang (2015) serves as a starting point for this thesis, which builds on the findings of the authors. In order to comprehend when leniency programs are likely to be successful in lowering the prevalence of cartels, the authors constructed and examined a theoretical framework. They derived logical explanations for why a leniency program might even lead to the growth of more cartels instead of a reduction of cartel activity, as outlined in Section 1.2 Problem Definition (Harrington & Chang, 2015). This was not the first time Harrington and Chang had tackled the issue of leniency. Already in a study that was published in 2007, Harrington, together with Chen, dealt with the question of what effect a corporate leniency program has on the formation of cartels and on the cartel price path. The authors demonstrated that maximum leniency always makes collusion more challenging (Chen & Harrington, 2007, p. 18). However, they also found that partial leniency programs can facilitate collusion when leniency is not offered (Chen & Harrington, 2007, p. 12). Although the works of Harrington and Chang are fundamental to this master's thesis, there exist many additional studies that merit acknowledgment. Numerous other scholars have engaged with leniency programs, and their contributions to the field are considerable. While some of the contributions refer to leniency programs other than the one of the EC, the insights are relevant to the present thesis in order to obtain a holistic picture with regard to the impact such policies have on overall cartel activities, and are therefore included in this study.

Motta and Polo (2003) and Aubert et al. (2006) both conducted research on competition policy enforcement strategies, specifically leniency programs and whistleblowing rewards, in the context of deterring collusion. Motta and Polo (2003) demonstrated that leniency programs, which grant lower fines to firms providing information to antitrust authorities, can enhance enforcement effectiveness, particularly when the cartel authority has limited resources (p. 375). However, they also acknowledged that such programs can inadvertently increase collusion by lowering the expected costs of misconduct (Motta & Polo, 2003, p. 349). Despite this potential drawback, the authors show that under an optimal policy, the former effect predominates, arguing in favour of leniency programs when the cartel authority has limited resources (Motta & Polo, 2003, p. 375). Aubert et al. (2006), on the other hand, compared the effects of reduced fines and positive rewards for individuals, including company employees, in preventing collusion. They argued that

---

rewarding individuals is a more effective tool than leniency programs for deterring cartels, thus taking a more sceptical view of leniency programmes (Aubert et al., 2006, p. 36).

Hinloopen and Soetevent (2008) noted that the number of cartels uncovered in the US and Europe had increased significantly since the introduction of their respective leniency programs. However, according to the authors this does not necessarily have to be the result of a successful cartel policy but could also be attributed to increased cartel activity (Hinloopen & Soetevent, 2008, p. 607). The authors investigated the effects of leniency programs on pricing and cartel activity by conducting an experiment. They found that, on the one hand, fewer cartels are formed in the laboratory when a leniency program is in place and, on the other hand, cartels that nevertheless exist are less successful when they charge prices above the static Nash equilibrium price and have poorer survival rates (Hinloopen & Soetevent, 2008, p. 611 et seq.).

Miller (2009) devoted himself to antitrust enforcement in the US. The author expressed concerns about the ambiguity surrounding leniency in the expanding game-theoretic literature at the time. To address this, he developed a theoretical model of cartel behaviour that provides empirical predictions (Miller, 2009, p. 751 et seq.). Using statistical tests, he confirmed the assumption that leniency increases deterrence and detection (Miller, 2009, p. 759 et seq.). In particular, Miller showed that the number of cartel detections increased around the time of leniency introduction and then fell below pre-leniency levels, arguing that this pattern is consistent with improved cartel detection and deterrence capabilities (Miller, 2009, p. 761 et seq.). However, the author warned that the results should be interpreted with caution due to the lack of cross-sectional variation in the data (Miller, 2009, p. 765).

Sauvagnat (2010), Bigoni et al. (2012), Nicolau (2015) and Pinha et al. (2016) all delved into the effectiveness of leniency programs in deterring cartels, but from different perspectives. Sauvagnat (2010) presented a model where the cartel authority is privately informed about the strength of the case against a particular cartel. In this setting, the cartel authority can obtain confessions from cartel members, even when it opens an investigation without the possibility of finding valid evidence (Sauvagnat, 2010, p. 5 et seq.). The author with his research showed that leniency programs increase the conviction rate, subsequently strengthening cartel abstinence and deterrence (Sauvagnat, 2010, p. 22 et seq.). Bigoni et al. (2012) conducted an experiment to examine the influence of fines and

---

---

leniency programs on cartel formation and prices (p. 372 et seq.). They noted that while a leniency program stabilises surviving cartels, well-managed leniency programs still possess a strong cartel deterrence potential (Bigoni et al., 2012, p. 386 et seq.). Nicolau (2015), on the other hand, focused on an economic analysis of the EC's leniency program. He concluded that the program effectively induces voluntary declarations, leading to an increased detection of cartels and higher fines imposed (Nicolau, 2015, p. 34). The leniency program, according to the author, enables the detection of long-running cartels and reduces their size (Nicolau, 2015, p. 27 et seq.). However, the author also observed areas for improvement, such as the considerable number of repeat offenders and the lack of significant reduction in the duration of investigations (Nicolau, 2015, p. 29). Finally, Pinha et al. (2016) used the Brazilian example to explain leniency programs in the context of applied economics. They conclude that, some preconditions given, leniency programs are effective instruments (Pinha et al., 2016, p. 148). Together, these studies highlight the overall effectiveness of leniency programs in deterring cartels while also pointing out the need for continuous improvement in program implementation and management.

Emons (2018), Dijkstra and Frisch (2018), Borrell et al. (2022), and Jochem et al. (2020) added to the discussion by examining various aspects of leniency programs, providing insights into the factors that influence the success of leniency programs in deterring cartels and highlighted areas for improvement. Emons (2018) concluded that leniency is ineffective if firms are sufficiently patient. He also established that increasing the likelihood of investigation at a high level reduces collusion, but never eliminates it completely (Emons, 2018, p. 15). Dijkstra and Frisch (2018) used an empirical study of the Dutch leniency program to investigate the impact of the introduction of sanctions and leniency programs on the number of cartel disclosures. They found that the program in the Netherlands did not lead to more cartels being abandoned, as the number of cartel disclosures decreased over time (Dijkstra & Frisch, 2018, p. 121 et seq.). An analysis of the characteristics of cartel members conducted by the authors showed that enforcement was tougher after the revision, suggesting that the decrease in cartel detections is associated with higher deterrence of cartels (Dijkstra & Frisch, 2018, p. 130 et seq.). Borrell et al. (2022) confirmed this view, suggesting that leniency programs have strong and clear short-term cartel destabilizing and long-term cartel deterrent effects (p. 32). Jochem et al. (2020) analysed whether the 2002 EU leniency reform, which moved the EU leniency regime considerably closer to that of the US, improved the EC's ability to destabilise

---

cartels while making law enforcement more efficient. They conclude that the 2002 reform reduced the duration of cartels by about 87 % but did not significantly affect the other outcome variables. The authors thus conclude that the 2002 reform improved the cartel-stabilizing effect of EU leniency, but without increasing its effectiveness in prosecuting cartels (Jochem et al., 2020, p. 15 et seq.).

### **3.2 Reduction in Imposed Fines**

Broos et al. (2016), Borrell et al. (2022) and Motchenkova (2004) explored the impact of leniency programs on fines imposed by competition authorities. While Broos et al. (2016) focused on cartels fined by the EC since May 2004, Borrell et al. (2022) examined the Spanish leniency program in comparison to the EC leniency program. Broos et al. (2016) presented various statistical findings related to cartel design and enforcement, noting that over half of the companies applied for leniency and benefited from an average fine reduction of 37 % (Broos et al., 2016, p. 86). According to the study authors, the statistics presented suggest a longer duration of infringement for large cartels as well as for cartels whose structure changes over time (Broos et al., 2016, p. 92 et seq.). On the other hand, Borrell et al. (2022) investigated the theoretically and empirically unresolved question of leniency programs' effects on cartel duration, fines, and investigation duration. They did this by comparing the Spanish leniency program with the EC leniency program (Borrell et al., 2022, p. 3). The comparison revealed that such programs destabilise existing cartels in the short term and discourage new cartel formation in the long term. The authors found that the deterrence effects empirically dominated in the long run. They noted that fines per firm increased significantly after leniency introduction, even though firms providing information received partial or full exemptions from being fined (Borrell et al., 2022, p. 32 et seq.). Motchenkova (2004) found that the effects of leniency programs on cartel stability depend on the structure of fines and the confidentiality level of the leniency application. According to the author, while leniency programs often shorten the duration of cartels, this is not always the case (Motchenkova, 2004, p. 3). When leniency programs are not overly strict and fines are proportionate to the accumulated illegal profits from price fixing, self-disclosure and immunity from fines under strict antitrust enforcement encourage firms to stop colluding, which shortens the duration of the cartel. However, if sanctions and prosecution are too lax, the introduction of leniency may paradoxically facilitate collusion and, accordingly, prolong the duration of cartels (Motchenkova, 2004, p. 20).

### 3.3 Investigation and Prosecution by Cartel Authority

There are various studies that examine the impact of leniency programs on the behaviour of cartels and the effectiveness of cartel authorities' enforcement policies, thereby highlighting the connection between an enforcement authorities' available resources and the effectiveness of leniency programs. Those studies include the research from Motta and Polo (1999), Harrington and Chang (2008), Brenner (2009) and Harrington (2013).

As early as 1999, Motta and Polo examined the enforcement of competition policy against collusion under leniency programs. They analysed the optimal policy under alternative rules and addressed the problem of limited resources of the cartel authority (Motta et al., 1999, p. 2). They refer mainly to the US leniency program, since at the time of the study, the EC's leniency program had only been applied in very few cases (Motta et al., 1999, p. 20). In another study from 2008, Chang and Harrington examined the effectiveness of a leniency program (Chang & Harrington, 2008, p. 3). They described that, due to an implicit resource constraint, the probability of a cartel being convicted once it has been uncovered depends inversely on the cartel authority's caseload (Chang & Harrington, 2008, p. 6). As in the 2015 study, they conclude that in the case of an unchanged enforcement policy of the cartel authority, a leniency program reduces the cartel frequency. At the same time, they illustrate that the additional number of cases provided by the leniency program leads the cartel authority to prosecute a lower proportion of cartel cases identified outside the program (Chang & Harrington, 2008, p. 19). Accordingly, the authors argue that with a less aggressive enforcement policy, it is possible that the cartel rate is higher when there is a leniency program (Chang & Harrington, 2008, p. 23).

In an empirical study, Brenner (2009) finds strong evidence that a leniency program induces disclosure of information about criminal activities in the sense that authorities are better informed about cartel behaviour than they would be without the program. Brenner (2009) concludes that investigations and prosecutions are accelerated by about 1.5 years as a result of the leniency program (p. 640). However, he postulates that the savings in investigation and prosecution costs are rather modest, calling into question the effectiveness of the program (Brenner, 2009, p. 641). Contrary to Brenner (2009), Borrell et al. (2022) found that the duration of investigations increases with the introduction of leniency programs (p. 33). This finding, even suggesting longer investigation durations than shorter ones, lends further support to Brenner's scepticism regarding the efficiency of leniency programs.

---

Finally, Harrington (2013) examined the incentives to apply for leniency if each cartel member has access to private information about the probability that the competition authority is able to convict them without a cooperating firm. He suggested some measures the competition authority can take to increase the cartel members' fears of a prior conviction and thereby achieve a greater use of the leniency program (Harrington, 2013, p. 25 et seq.).

### **3.4 Managing Disclosure Incentives and Abuse Risks**

Chen and Rey (2013) developed a model that depicts the conflict between the destabilisation of collusion and the potential abuse of leniency. The model puts optimal leniency in relation to the effectiveness of investigations (Chen & Rey, 2013, p. 922 et seq.). The authors show that it is inherently desirable to grant some leniency before an investigation is initiated. Moreover, they conclude that it is also optimal to grant some leniency once an investigation is underway if the investigation is unlikely to lead to the detection of cartels unless self-reporting occurs. Finally, they conclude that it makes sense to limit leniency to the first informant only; in contrast, they do not argue for a ban on leniency for repeat offenders (Chen & Rey, 2013, p. 946 et seq.).

In 2014, the issue of optimal leniency was addressed again by Sauvagnat (2014). He created a simple model of collusion in which the competition authority grants leniency depending on the number of companies reporting information (Sauvagnat, 2014, p. 323). According to the author, the optimal leniency is the so-called "single-informant rule", i.e., leniency should only be granted if a single undertaking reports information. This rule allows to increase the expected sanctions compared to the first informant rule, which improves the overall deterrence of the cartel (Sauvagnat, 2014, p. 325).

Further exploring leniency program intricacies, Chen et al. (2015) examined the inclusion of no-immunity-for-instigators clauses (NIICs). These clauses deny leniency to parties who instigate or act as leaders of a cartel (Chen et al., 2015, p. 20). The authors show that NIICs can lead to both an increase or decrease in cartel behaviour. By removing the benefit that the instigator derives from cooperation with the authorities, a NIIC, according to the study, cancels out part of the destabilizing benefit of leniency and thus promotes the stability of the cartel. On the other hand, according to the authors, the instigator is punished asymmetrically harshly under a NIIC, which can reduce the incentive to instigate (Chen et al., 2015, p. 28 et seq.).

### 3.5 Challenges and Research Gap

The field of leniency program research has faced numerous challenges and developments over time, as scholars seek to determine the actual effectiveness of such programs in combating cartels. One of the earlier concerns was raised by Miller (2009), who critiqued the ambiguity surrounding the concept of leniency within the burgeoning game-theoretic literature at the time (p. 750).

In an effort to explore the foundations of leniency programs, Spagnolo (2006) provided an overview of their development in the US and the EU. He highlighted the still ongoing debate over the effectiveness of these programs, noting that, while their efficacy was widely assumed, there was no absolute certainty about their actual impact (Spagnolo, 2006, p. 8).

A rather striking challenge in this field of research that has been raised by a number of scholars is the fact that the total cartel population is not known. Zhou (2016) was one of the scholars to point out that particular problem of possible sample selection bias. Cartels hide their activities due to their illegal nature and only revealed cartels are observable (Zhou, 2016, p. 17). Since revealed cartels might be a small and characteristically unrepresentative sample of the cartel population, one cannot infer the impact of the cartel investigation on the cartel population from the information obtained from revealed cartels without making additional assumptions that, according to the author, may be correct or incorrect (Zhou, 2016, p. 17). Addressing the challenge of the unobservable cartel population, Harrington and Chang (2009) developed a model of cartel formation and resolution that endogenously derives the population of cartels and detected cartels. The authors paid special attention to changes in the duration of detected cartels, which according to their research is illuminating for assessing whether a new regulation affects the latent cartel rate (Harrington & Chang, 2009, p. 1419 et seq.). Echoing the concerns of previous researchers, Pinha et al. (2016) also point out the challenge of proving the actual effectiveness of leniency programs, as the total population of cartels is not known (p. 149).

Harrington and Chang (2012) further expanded on their previous research by examining the impact of leniency programs on individual industries. Their findings suggest that the impact can vary greatly depending on the industry, and they caution against measuring the performance of a leniency program solely by the number of leniency applications, as a such a program can decrease the cartel rate while no applications are made and increase

the cartel rate while many applications are made (Harrington & Chang, 2012, p. 29 et seq.).

To summarise, the challenges and developments in leniency program research revolve around the ongoing quest to establish their true effectiveness in combating cartels. Key issues include addressing ambiguity in the concept of leniency, examining the impact of leniency programs on individual industries, and grappling with the challenge of the unobservable cartel population.

Building on the existing body of work discussed above, it is clear that one of the key challenges in the field of leniency program research is the issue of the unobservable cartel population. This issue presents a significant gap in the current understanding and ability to accurately gauge the true effectiveness of leniency programs. In essence, it is difficult to measure the impact of these programs without a comprehensive understanding of the population they intend to address.

In order to bridge this gap, it is important to understand how the cartel population has changed with the introduction of the leniency program. This also requires assessing whether there has been a shift from non-lenieny enforcement to leniency enforcement. This is an aspect that will be examined in this paper. Furthermore, an approach that leverages advancements in NLP techniques to automatically extract and collect important data from the EC's prohibition decisions is introduced, constituting a computational contribution to filling the current research gap by providing an efficient way to gather and analyse the wealth of data contained in these legal texts, which is difficult and time-consuming to process manually, and prone to errors. In addition, this study proposes regression models to analyse extracted data, offering fresh insights into how the introduction of the EC's leniency program impacts cartel activity.

By focusing on these areas, this thesis not only addresses the identified research gap but also contributes valuable methodologies and tools that can potentially facilitate future research in the field. The insights gained from this work will, therefore, provide a more nuanced understanding of the effectiveness of the EC's leniency program and its influence on cartel activities in the EU.



## 4 METHODOLOGY

The methodology to address the research question comprises four phases: 1) data collection, 2) data extraction, 3) data preparation and 4) data analysis.

In the first phase, primary data is gathered from the EC's publicly accessible prohibition decisions on cartel cases. Then, in the second step, the data required for the analysis are to be extracted from the PDF files. Considering the time constraints of this master's thesis, the primary focus is on analysing secondary data collected manually by previous researchers, covering cases from 1964 to 2010, which thus already provide a solid basis for effectively answering the research question using linear regression with time series. The missing data from 2010 to the present time is supplemented in the process. Furthermore, this study takes a first step towards implementing automatic data collection by applying NLP techniques. While these preliminary results demonstrate the potential of NLP in this research area, a comprehensive exploration of these methods is beyond the scope of this paper. However, the groundwork laid in this study serves as a valuable starting point for future research efforts. The goal thereby is to create an artefact that can be further developed and refined in subsequent studies, allowing for deeper analysis and the inclusion of more sophisticated NLP techniques. By using existing data and taking the first steps towards automatic data collection, this study lays the foundation for future advances in the field. Finally, in the third phase, the combined primary and secondary data is processed accordingly and then analysed by employing statistical methods in the fourth and last step of the process.

### 4.1 Data Acquisition

For future research, it is important that new decisions can be downloaded efficiently from the EC's website. Thus, the first step in obtaining the needed data is to scrape the prohibition decisions from the EC's website. These are provided as PDF files. Since it is planned to use NLP techniques for further processing, only files available in English will be used. In the context of this time-constrained project, translating decisions from other languages into English and subsequently processing them would not yield a satisfactory return on effort. Addressing this limitation should be prioritised in future research projects though, as restricting the analysis to English decisions skews the data and potentially distorts the results.

At the time of the data acquisition, the cases on the EC's website were divided into two time periods: Cases from 1964 to 1998 could be accessed via the archive. The cases from 1999 onwards could be found via the official search function.<sup>1</sup> For this reason, and because the web scraping involves a relatively long execution time, the required PDF files are downloaded in two separate scraping scripts: script 1a for the cases from 1964 to 1998, and script 1b for the cases from 1999 onwards.

For the cases from 1964 to 1998, a folder is created in the current working directory, which the scraped prohibition decisions are downloaded into. The cases up to 1999 can be accessed through the link <http://ec.europa.eu/competition/antitrust/closed/en/>, with the individual years added at the end of the URL. For instance, for the year 1964, the completed URL would be <http://ec.europa.eu/competition/antitrust/closed/en/1964.html>. On those pages, all the cases in the specific years are listed as can be seen in Figure 4 below, including a link leading to a separate landing page for each case.

#### Decisions Art. 81/82 (ex 85/86) : Year 1964

*Last page update 24.03.00*

**Deca** 22.10.1964 Negative clearance Art.81(1) [ex 85(1)]

Official Journal : L - 31/10/1964 Page : 2761 Celex No. : [364D05 99](#) - IV/71

**Grundig-Consten** 23.09.1964 Infringement Art.81 [ex 85]

Official Journal : L - 20/10/1964 Page : 2545 Celex No. : [364D05 66](#) - IV/3344 IV/4

**Nicholas Freres + Vitapro** 30.07.1964 Negative clearance Art.81(1) [ex 85(1)]

Official Journal : L - 26/08/1964 Page : 2287 Celex No. : [364D05 02](#) - IV/95

**Bendix + Mertens and Straat** 01.06.1964 Negative clearance Art.81(1) [ex 85(1)]

Official Journal : L - 10/06/1964 Page : 1426 Celex No. : [364D03 44](#) - IV/12868

**Grosfillex + Fillistorf** 11.03.1964 Negative clearance Art.81(1) [ex 85(1)]

Official Journal : L - 9/04/1964 Page : 915 Celex No. : [364D02 33](#) - IV/61

**Figure 4:** Cartel cases in the year 1964, screenshot of the EC's website

Using Selenium and BeautifulSoup, all pages from 1964 to 1998 are iterated through and the links to the cases are crawled. After that, each one of the collected links is accessed separately. On those landing pages, the prohibition decisions are provided in several languages. Again, with Selenium, all the links that contain “*legal-content/EN/TXT*” are searched for. Finally, those links are called upon and the PDF files are downloaded into the current working directory. All in all, when the search was conducted on 23 April 2023, 342 cases have been downloaded, whereas it needs mention that regarding the cases

<sup>1</sup> In the meantime, the structure of the EC's website has slightly changed. Those changes are not accounted for in the present thesis. Both scraping scripts are still fully functional at the time of submission of this thesis.

published in the initial years, the EC does not provide any documents in English language. The first case with a decision available in English dates from 1973. However, it should be noted that no distinction is made between antitrust and cartel cases in this procedure. This may explain the difference between the number of prohibition decisions downloaded through the scraping script 1a and the data from 1964 to 2010 provided in the Excel file. Should the code be used for future research work, the cartel cases would have to be filtered out – preferably using NLP techniques – for further processing. Since the present work henceforth concentrates on dealing with the more recent cases for which no data are yet available, this aspect will not be elaborated in more detail here.

The scraping script for downloading the cases from 1973 to 1998 is included in this master thesis for the sake of completeness, neither the script nor the PDF files it downloaded are further used in the remaining part of this thesis. The script is part of the first artefact, the corresponding file is named *1a\_scraping-cases-until-1998.ipynb*.

For cases from 1999 onwards, a new folder titled "cases\_from\_1999" is created in the current working directory. Selenium is then used to access the EC's search function via Chrome Webdriver, which can be accessed through the following link: [https://ec.europa.eu/competition/elojade/isef/index.cfm?clear=1&policy\\_area\\_id=1%2C2%2C3](https://ec.europa.eu/competition/elojade/isef/index.cfm?clear=1&policy_area_id=1%2C2%2C3).

The search mask looks as displayed in Figure 5 below.

**Figure 5:** Search mask on the website of the EC, screenshot of the EC's website

The following restrictions are made in the search: 1) The *Policy Area* is limited to *Cartels* and 2) for *Decision Date*, the time period from 01/01/1999 to the current date of running the script is selected. When the search was conducted, on 23 April 2023, it returned 127 hits.

The page with the results is then scraped through to collect all the listed case numbers the search has given back. For this purpose, the XPath and the class name are searched for. The case numbers are written into a text file, which is then used to call up the cases individually. To do this, the case number for each case is added at the end of the following URL: [https://ec.europa.eu/competition/elojade/isef/case\\_details.cfm?proc\\_code=1\\_](https://ec.europa.eu/competition/elojade/isef/case_details.cfm?proc_code=1_).









On the pages accessed this way, the accessible information concerning the cases are embedded in a table format. With Selenium, the table is searched through to find the term “*Prohibition Decision*”. If that text is found in a cell, the cell located next to it is crawled through to get the link to the corresponding prohibition decision. There are three possibilities: 1) either there is a PDF file available in the table itself, as illustrated in Figure 6, 2) there is a link to another page where the prohibition decision is available in different languages, or 3) there is a link to another landing page on which the decision is available in English language only.

#### AT.38238 Raw Tobacco (ES)

Companies: [ACOTAB](#) | [ANETAB](#) | [ASAJA](#) | [Agroexpansion](#) | [CCAE](#) | [COAG](#) | [Cetarsa](#) | [Deltafina](#) | [Dimon Inc.](#) | [FNCT](#) | [Intabex Netherlands](#) | [Standard Commercial](#) | [TABARES](#) | [TCLT](#) | [Taes](#) | [UPA](#) | [Universal Corp.](#) | [Universal Leaf Tobac.](#) | [WWTE](#)

Economic Activity: [C.12.00](#) - Manufacture of tobacco products  
[G.46.21](#) - Wholesale of grain, unmanufactured tobacco, seeds and animal feeds  
[G.46.35](#) - Wholesale of tobacco products

Events:

Date	Document Type	Document
16.06.2017	Memo	 <a href="#">EN</a> published on 20.06.2017
16.06.2017	Prohibition Decision (Art. 101 Ex 81)	 <a href="#">EN</a>  <a href="#">ES</a> published on 15.12.2017
16.06.2017	Summary Decision	Summary Decision : <a href="#">Multilingual</a>
31.05.2017	Report of the Hearing Officer	Hearing Officer report : <a href="#">Multilingual</a>
30.05.2017	Opinion of the Advisory Committee	Opinion of Advisory Committee : <a href="#">Multilingual</a>
19.04.2007	Opinion of the Advisory Committee	Official Journal C 85, 19.4.2007, p. 14–14 : <a href="#">Multilingual</a> Official Journal C 85, 19.4.2007, p. 15–15 : <a href="#">Multilingual</a>
19.04.2007	Report of the Hearing Officer	Official Journal C 85, 19.4.2007, p. 16–16 : <a href="#">Multilingual</a>
19.04.2007	Summary Decision	Official Journal L 102, 19.4.2007, p. 14–14 : <a href="#">Multilingual</a>
20.10.2004	Press Release	Ammende della Commissione a società del settore del tabacco grezzo in Spagna : <a href="#">EN</a>
20.10.2004	Prohibition Decision (Art. 101 & 102 Ex 81 & 82)	 <a href="#">DE</a>  <a href="#">EN</a>  <a href="#">ES</a>  <a href="#">FR</a>  <a href="#">IT</a>

*These publications are for information purposes only and should not be considered as an official publication*

**Figure 6:** Example of a table structure where the PDF file to the prohibition decision is provided in the cell located next to the term “*Prohibition Decision*”, screenshot of the EC’s website

On each page accessed in this way, Selenium first searches for PDF files that are available in English. If a PDF file in English language is found, the links to the PDF file is written to a new text file and then downloaded into the current directory. The example that is illustrated in Figure 6 also shows another challenge that needs to be overcome; in some

cases, there are two different prohibition decisions available, one dating further back than the other. For further data processing, only the newer one of the decisions shall be used. Therefore, the first PDF file that meets the specified criteria is downloaded, and the process is then halted for the corresponding case. As the more recent decision is always placed higher in the table, only the newest decision is considered.

If instead of a PDF file, Selenium finds a link to another landing page, as is the case in the example shown in Figure 7 below, those links will be added to a list that will then again be accessed individually by Selenium. All those pages are again searched through for a PDF file with the title attribute “*PDF English*”, if such a file is found, it is also downloaded into the current directory.

#### AT.36604 Citric acid

<b>Companies:</b>	<a href="#">ADM</a>   <a href="#">Bayer AG</a>   <a href="#">Cerestar Bioproducts</a>   <a href="#">Citrique Belge NV</a>   <a href="#">F. Hoffmann-La Roche</a>   <a href="#">Haarmann &amp; Reimer</a>   <a href="#">Jungbunzlauer AG</a>										
<b>Economic Activity:</b>	<a href="#">C.20.14</a> - Manufacture of other organic basic chemicals										
<b>Events:</b>	<table border="1"> <thead> <tr> <th>Date</th> <th>Document Type</th> <th>Document</th> </tr> </thead> <tbody> <tr> <td>05.12.2001</td> <td>Press Release</td> <td>Commission fines five companies in citric acid cartel : <a href="#">EN</a></td> </tr> <tr> <td>05.12.2001</td> <td>Prohibition Decision (Art. 81 &amp; 82)</td> <td>Official Journal L 239, 2002, P. 0018 - 0065 : <a href="#">Multilingual</a></td> </tr> </tbody> </table>		Date	Document Type	Document	05.12.2001	Press Release	Commission fines five companies in citric acid cartel : <a href="#">EN</a>	05.12.2001	Prohibition Decision (Art. 81 & 82)	Official Journal L 239, 2002, P. 0018 - 0065 : <a href="#">Multilingual</a>
Date	Document Type	Document									
05.12.2001	Press Release	Commission fines five companies in citric acid cartel : <a href="#">EN</a>									
05.12.2001	Prohibition Decision (Art. 81 & 82)	Official Journal L 239, 2002, P. 0018 - 0065 : <a href="#">Multilingual</a>									
<small>These publications are for information purposes only and should not be considered as an official publication</small>											

**Figure 7:** Example of a table structure where the prohibition decision is found on another page, Screenshot of the EC's website

The last of the three beforementioned versions, the one where there is a link to another landing page on which the decision is available only in English language, only concerns one case, which is depicted in Figure 8 below. This specific case – and thus all future cases that may have the same structure – is also accounted for in the scraping script.

#### AT.35860 Seamless steel tubes

<b>Companies:</b>	<a href="#">British Steel plc</a>   <a href="#">Corus UK</a>   <a href="#">Dalmine SpA</a>   <a href="#">Europipe</a>   <a href="#">ILVA Lamiere Tubi</a>   <a href="#">JFE Engineering Corp</a>   <a href="#">JFE Steel Corp</a>   <a href="#">Mannesmannröhrenwerk</a>   <a href="#">NSC</a>   <a href="#">Sumitomo Metal Ind.</a>   <a href="#">Usinor Sacilor</a>   <a href="#">Vallourec Industries</a>										
<b>Economic Activity:</b>	<a href="#">C.24.20</a> - Manufacture of tubes, pipes, hollow profiles and related fittings, of steel										
<b>Events:</b>	<table border="1"> <thead> <tr> <th>Date</th> <th>Document Type</th> <th>Document</th> </tr> </thead> <tbody> <tr> <td>25.01.2007</td> <td>Press Release</td> <td>Competition: Commission welcomes judgments of the European Court of Justice in Seamless steel tubes cartel case : <a href="#">EN</a></td> </tr> <tr> <td>08.12.1999</td> <td>Prohibition Decision (Art. 81 &amp; 82)</td> <td>Official Journal L 140, 2003, P. 0001 - 0029 : <a href="#">EN</a></td> </tr> </tbody> </table>		Date	Document Type	Document	25.01.2007	Press Release	Competition: Commission welcomes judgments of the European Court of Justice in Seamless steel tubes cartel case : <a href="#">EN</a>	08.12.1999	Prohibition Decision (Art. 81 & 82)	Official Journal L 140, 2003, P. 0001 - 0029 : <a href="#">EN</a>
Date	Document Type	Document									
25.01.2007	Press Release	Competition: Commission welcomes judgments of the European Court of Justice in Seamless steel tubes cartel case : <a href="#">EN</a>									
08.12.1999	Prohibition Decision (Art. 81 & 82)	Official Journal L 140, 2003, P. 0001 - 0029 : <a href="#">EN</a>									
<small>These publications are for information purposes only and should not be considered as an official publication</small>											

**Figure 8:** Example of table structure where there is a link to the English PDF without the ".pdf" in the link, Screenshot of the EC's website

At the end of this procedure, 111 PDF files have been downloaded into the folder “*cases\_from\_1999*”. This means that out of the 127 search results for cases from 1999

onwards, at the time of the script being run, on 23 April 2023, for 16 cases there has been no decision available in English. Those cases are thus not included in the further work for the present research.

Among the downloaded files, there are some PDF files that pertain to the same case. A visual inspection shows that these are mainly amending decisions. Since this issue will be addressed again in Section 4.2 Data Extraction with NLP, it will not be elaborated upon here.

The script scraping the cases from 1999 to the current date is the second part of the first artefact, the file is named *1b\_scraping-cases-from-1999.ipynb*.

In total, 342 files from 1973 to 1998 and 111 files from 1999 onwards have been found and downloaded on 23 April 2023.

## 4.2 Data Extraction with NLP

Legal documents differ in structure, vocabulary, ambiguity, citations, and size from other texts such as news articles, blog entries or scientific texts (Zadgaonkar & Agrawal, 2021, p. 5452; Kanapala et al., 2019, p. 372 et seq.). Because of that, the structure and semantics of legal texts need to be understood for the application of techniques used for information extraction (Zadgaonkar & Agrawal, 2021, p. 5452 et seq.). The text extracted from the downloaded prohibition decisions is unstructured. To be able to work with that unstructured information and later use it to conduct an analysis, the data need to be extracted from that unstructured text and put into a more structured format, i.e., into an Excel file.

To perform a comprehensive analysis of the data from the prohibition decisions, the data outlined in Table 1 below would be needed. This data is already available for the cases from 1964 to 2010, having been manually collected by previous researchers. Thus, it would be desirable to extend the dataset with the corresponding information for the total of all cartel cases.

**Table 1:** *Data needed for comprehensive analysis of the effect the introduction of leniency had on the detected cartel activity (own illustration)*

Label	Description
<b>case_number</b>	Identification of an instance of a case handled by the EC
<b>cartel_name</b>	Name/title of the cartel case

---

<b>party_name</b>	Name of the involved cartel members, can be used for identifying repeat offenders
<b>decision_date</b>	Date when the EC took the final decision
<b>party_address</b>	Addresses of the involved cartel members
<b>party_country</b>	Countries in which the involved cartel members are based
<b>country_indicator</b>	Indicator for the different countries in which the involved cartel members are based
<b>european_union</b>	Binary value indicating whether the cartel member is based in a country that is part of the EU or not
<b>parties_count</b>	Number of all involved parties in the cartel
<b>cartel_start</b>	Start date of the cartel, needed to compare the average cartel duration before the introduction of leniency with its average duration afterwards
<b>cartel_end</b>	End date of the cartel, needed to compare the average cartel duration before the introduction of leniency with its average duration afterwards
<b>cartel_duration</b>	Duration of the cartel, calculated from <i>cartel_start</i> and <i>cartel_end</i>
<b>decision_text</b>	Complete text of the decision*
<b>points</b>	Number of paragraphs in the decision text
<b>oecd_sector</b>	Integer per OECD sector, which shall be used to gain insights in differences between industries with stable cartels and industries with unstable cartels
<b>manufacturing_sub</b>	Integer per manufacturing subdivision
<b>report_route</b>	Route through which the case came to the EC
<b>route_indicator</b>	Numerical indicator for the report route (assigned integer per report route: 1) notification, 2) complaint, 3) Commission's own initiative, 4) leniency application)

---

<b>leniency</b>	Binary value indicating whether leniency was applied or not
<b>formal_decision</b>	Formal decision taken by the EC regarding whether there was an infringement
<b>decision_indicator</b>	Numerical indicator for the formal decision, shall be used to compare the success rate of convicting offenders before the introduction of leniency and after the introduction of leniency
<b>conduct_nature</b>	Nature of conduct
<b>conduct_indicator</b>	Numerical indicator regarding the nature of conduct
<b>minimum_fine</b>	Minimum possible fine
<b>maximum_fine</b>	Maximum possible fine
<b>fine</b>	Actual fine imposed on the cartel member, shall be used to assess whether there was a change in fines imposed after leniency has been introduced
<b>total_fines</b>	Sum of the fines imposed on all involved parties in the cartel, shall be used to assess whether there was a change in fines imposed after leniency has been introduced
<b>turnover</b>	Turnover of the involved parties in the cartel
<b>commissioner</b>	Name of the responsible commissioner

*\* Note: The complete decision text cannot be stored in an Excel file since it exceeds the limit of 32'767 characters per cell. Instead, it can be put into a txt File and the name of the file could be mentioned in the Excel file instead.*

Nevertheless, emphasis is placed on the data which is most crucial to conduct a regression analysis with time series to evaluate the impact the introduction of leniency had on the detected cartel activity falling under the EC's jurisdiction. The most critical data for doing that is consolidated in Table 2 below. Subsequently, NLP methods that can be used for extracting that data from the previously scraped PDF files are analysed in more detail, alongside an assessment of their accuracy rates regarding successful information extraction.



**Table 2:** Data needed to perform linear regression with interrupted time series for evaluating the impact of the introduction of leniency on the detected cartel activity (own illustration)

Label	Description
<b>case_number</b>	Identification of an instance of a case handled by the EC
<b>decision_date</b>	Date when the EC took the final decision
<b>cartel_start</b>	Start date of the cartel, needed to compare the average cartel duration before the introduction of leniency with its average duration afterwards
<b>cartel_end</b>	End date of the cartel, needed to compare the average cartel duration before the introduction of leniency with its average duration afterwards
<b>cartel_duration</b>	Duration of the cartel, calculated from <i>cartel_start</i> and <i>cartel_end</i>
<b>report_route</b>	Route through which the case came to the EC
<b>route_indicator</b>	Numerical indicator for the report route (assigned integer per report route: 1) notification, 2) complaint, 3) Commission's own initiative, 4) leniency application)
<b>leniency</b>	Binary value indicating whether leniency was applied or not

Due to varying data formats and time constraints, the NLP methods that can be used for extracting the information defined in Table 2 are examined only in context of the prohibition decisions downloaded from 1999 onwards. This restriction ensures the processing of a sufficiently large amount of data and avoids unnecessary loss of time, while at the same time focusing on the cases for which no data has been manually collected yet.

Out of the initial 111 files, twelve have been duplicates (case13.pdf, case27.pdf, case34.pdf, case55.pdf, case65.pdf, case76.pdf, case86.pdf, case103.pdf, case106.pdf, case108.pdf, case109.pdf and case110.pdf), three are unreadable because they have been scanned as an image (case24.pdf, case58.pdf and case59.pdf), one file has been wrongly labelled as English by the EC but actually is German and can thus not be processed (case16.pdf), and one file is broken (case87.pdf). Following the removal of duplicates after extracting case numbers from the acquired PDF files, 94 distinct values remain.

These 94 cases represent a comprehensive dataset, and the accuracy rate for the data extraction is calculated based on this total in the following.

A perfectly successful data extraction for individual data points could not be attained. Nonetheless, the below elaborations together with the scripts that are accessible on GitHub provide a valuable direction for further development and refinement of the code in the future by other researchers that will be dealing with the influence the introduction of leniency had on cartel activity.

#### 4.2.1 Case Number

##### 4.2.1.1 Regular expressions

The method applied to extract the case number from the prohibition decisions are regular expressions (regex). Regex are used for pattern matching in text strings. Using this method, sequences of patterns can be matched, which is particularly beneficial for cases where the information to be extracted is presented as a combination of strings with unique and specific structures (Bird & Klein, 2006, p. 1 et seq.).

Given the significance of precise regex pattern structures in this approach, the sole data cleaning performed involves eliminating redundant spaces. Looking at the case number, there are various ways in which it is formatted. Most important are the newer cases from 2010, since the manually collected data only encompasses the cases until that year. Formats in which the case numbers are presented from 2010 to 2021 are listed in Table 3 below, with one example presented per format for each year for better illustration.

**Table 3:** Case number formats from 2010 to 2021 (own illustration)

Year	Case number formats	Regex patterns
2010	Case COMP/39092	<i>COMP\(\d{5}\)</i>
	COMP/38.344	<i>COMP\(\d{2}\)\.\(\d{3}\)</i>
2011	COMP/39579	<i>COMP\(\d{5}\)</i>
2012	CASE AT.39437	<i>AT\.\(\d{5}\)</i>
2013	CASE AT.39633	<i>AT\.\(\d{5}\)</i>
2014	Case AT.39574	<i>AT\.\(\d{5}\)</i>
	CASE AT.39610	<i>COMP\(\d{2}\)\.\(\d{3}\)</i>

	CASE COMP/39922	
	AT.39965	
<b>2015</b>	CASE AT.39563	<i>AT\.(d{5})</i>
	AT.39861	
<b>2016</b>	CASE AT.38589	<i>AT\.(d{5})</i>
	Case AT.39965	
<b>2017</b>	CASE AT.39258	<i>AT\.(d{5})</i>
	AT.39780	<i>COMPV(d{2})\.(d{3})</i>
	CASE AT.AT.39881	
	COMP/38.238	
<b>2018</b>	CASE AT.39920	<i>AT\.(d{5})</i>
<b>2019*</b>	CASE AT. 40127	<i>AT\.(s\d{5})</i>
<b>2020</b>	CASE AT.39563	<i>AT\.(d{5})</i>
<b>2021*</b>	CASE AT.40178	<i>AT\.(d{5})</i>

\* Years in which only one decision is available in the dataset (as of 23 April 2023)

The purpose of the list in Table 3 is to determine whether there has been any recent standardisation or harmonisation in the formatting of the prohibition decisions. If such standardisation has occurred, it would significantly simplify the process of extracting information in the future. However, as of 2017, there is no clear evidence of standardisation. Although Table 3 may suggest the presence of standardisation starting from 2018, it is important to note that only two decisions were published in 2018, followed by just one decision per year in the subsequent years, except for the year 2020, which saw the publication of two decisions. Consequently, it is not possible to draw any reliable conclusions on this matter.

Based on the list in Table 3, which is exhaustive for the scraped cases from 2010 to 2021 and which are in English readable language format, four general patterns can be observed: 1) COMP/XXXXX format: *COMPV(d{5})*, 2) COMP/XX.XXX format: *COMPV(d{2})\.(d{3})*, 3) AT.XXXXX format (without a space after the dot): *AT\.(d{5})* and 4) AT. XXXXX format (with a space after the dot): *AT\.(s\d{5})*. The

digits are grouped in brackets for further processing of the case numbers. For cases before 2010, there are additional patterns which will not be elaborated upon here. With the four defined regex patterns above, 72 case numbers are correctly extracted from a total of 94 values from 1999 to 2021, which corresponds to a success rate of 67.68 %. Therefore, to find all the case numbers, the patterns have been further expanded and adapted in the corresponding NLP script 2a. For instance, some regex patterns have been combined into single regex patterns with optional parts. Due to various variations and challenges while running the code through the individual PDF files (i.e., patterns not being matched if they are directly followed by an alphanumeric character, or false matches in case they are), the four defined regex patterns are no longer represented as such in the code, but they do form the starting point for the further development of different patterns. The final patterns can be found in the NLP script *2a\_data-extraction-case-number-and-decision-date-regex.ipynb*.

Running the NLP script 2a, out of a total of 94 unique cases, in 92 of them, the case number has been extracted correctly. This is a success rate of 97.87 %. The two cases in which the case number has not been extracted fully successfully both refer to a case where there are two case numbers and only the first one has been read out of the file. Given the good result, no further NLP techniques have been tried out for extracting the case numbers.

In certain exceptional instances, multiple case numbers are present, such as in 1) (*COMP/D/32.448 and 32/450 - Compagnie Maritime Belge*), 2) (*Case COMP.D.2 37.444 — SAS Maersk Air and Case COMP.D.2 37.386 — Sun-Air versus SAS and Maersk Air*) and 3) (*Case COMP/E - 1/37.919 (ex 37.391) Bank charges for exchanging euro-zone currencies Germany*). While the first case is addressed in the final script, the latter ones require further examination and incorporation into the code should these formats recur in future case files. Attempts were made to retrieve multiple matches, but this led to incorrect case numbers being returned for other cases. This issue arose when the code identified pattern matches intended as references to other cases, resulting in inaccurate data. However, with the current code, the first case number is extracted in situations where two case numbers exist. This meets the requirement of identifying entities by their respective case numbers. Consequently, a more detailed investigation of this issue was not carried out.

### 4.2.2 Decision Date

The decision date is important because it helps identifying the cases that were decided before leniency was introduced compared to the rulings that followed its introduction. Similar to the case number, dates do have very specific structures. The decision date is usually the very first date in the document, which helps identify it. Also, from a pure logical perspective, in most cases it should be the most recent date that can be found in the entire document. This is an assumption that has not been confirmed for all cases and there might be exceptions to the rule.

#### 4.2.2.1 Regex

Considering the unique and specific structure of a date string, regex have been used for extracting the decision dates from the prohibition decision. Again, the decision date comes in a variety of different formats. In Table 4, the formats from 2010 to 2021 are illustrated. The dates are either preceded by “*COMMISSION DECISION of*” or “*Brussels,*”, which is taken into account in the NLP script 2a since this further restriction is helpful in avoiding false positives, but for the sake of simplicity is not included in the date formats in Table 4.

**Table 4:** Decision date formats from 2010 to 2021 (own illustration)

Year	Decision date formats	Regex patterns
2010	30.6.2010	$(\d{2})\d{1}\d{4}$
	8.12.2010	$(\d{1})\d{2}\d{4}$
	23 June 2010	$(\d{2})\s(?:January February March April May June July August September October November December)\s\d{4}$
2011	13.4.2011	$(\d{2})\d{1}\d{4}$
	12.10.2011	$(\d{2})\d{2}\d{4}$
2012	27/03/2012	$(\d{2})/(\d{2})/(\d{4})$
	27.6.2012	$(\d{2})\d{1}\d{4}$
2013	2 7.11.2013	$(\d{1})\s\d{1}\d{2}\d{4}$
	10/07/2013	$(\d{2})/(\d{2})/(\d{4})$

<b>2014</b>	3.9.2014	$(\backslash d\{1\})\backslash(\backslash d\{1\})\backslash(\backslash d\{4\})$
	19.3.2014	$(\backslash d\{2\})\backslash(\backslash d\{1\})\backslash(\backslash d\{4\})$
	21.10.2014	$(\backslash d\{2\})\backslash(\backslash d\{2\})\backslash(\backslash d\{4\})$
	02/04/2014	$(\backslash d\{2\})/(\backslash d\{2\})/(\backslash d\{4\})$
<b>2015</b>	24/06/2015	$(\backslash d\{2\})/(\backslash d\{2\})/(\backslash d\{4\})$
	4.2.2015	$(\backslash d\{1\})\backslash(\backslash d\{1\})\backslash(\backslash d\{4\})$
<b>2016</b>	29/06/2016	$(\backslash d\{2\})/(\backslash d\{2\})/(\backslash d\{4\})$
	6.4.2016	$(\backslash d\{1\})\backslash(\backslash d\{1\})\backslash(\backslash d\{4\})$
<b>2017</b>	17/03/2017	$(\backslash d\{2\})/(\backslash d\{2\})/(\backslash d\{4\})$
	16.6.2017	$(\backslash d\{2\})\backslash(\backslash d\{1\})\backslash(\backslash d\{4\})$
	8.2.2017	$(\backslash d\{1\})\backslash(\backslash d\{1\})\backslash(\backslash d\{4\})$
<b>2018</b>	21/02/2018	$(\backslash d\{2\})/(\backslash d\{2\})/(\backslash d\{4\})$
<b>2019*</b>	27/09/2019	$(\backslash d\{2\})/(\backslash d\{2\})/(\backslash d\{4\})$
<b>2020</b>	17/12/2020	$(\backslash d\{2\})/(\backslash d\{2\})/(\backslash d\{4\})$
<b>2021*</b>	08/07/2021	$(\backslash d\{2\})/(\backslash d\{2\})/(\backslash d\{4\})$

\* Years in which only one decision is available in the dataset (as of 23 April 2023)

The data in Table 4 which encompasses an exhaustive list of formats from 2010 to 2021 also suggests that some standardisation has taken place as of 2018. However, since too few cases have been published since then, it is not possible to draw a definitive conclusion.

Looking at the regex patterns presented in Table 4, there are a few patterns that can be combined:  $(\backslash d\{1\})\backslash(\backslash d\{1\})\backslash(\backslash d\{4\})$ ,  $(\backslash d\{1\})\backslash(\backslash d\{2\})\backslash(\backslash d\{4\})$ ,  $(\backslash d\{2\})\backslash(\backslash d\{1\})\backslash(\backslash d\{4\})$  and  $(\backslash d\{2\})\backslash(\backslash d\{2\})\backslash(\backslash d\{4\})$  i.e. become  $(\backslash d\{1,2\})\backslash(\backslash d\{1,2\})\backslash(\backslash d\{4\})$ . Thus, for the cases from 2010 to 2021, there are the following three patterns: 1)  $(\backslash d\{1,2\})\backslash(\backslash d\{1,2\})\backslash(\backslash d\{4\})$ , 2)  $(\backslash d\{2\})\backslash(?:(January|February|March|April|May|June|July|August|September|October|November|December))\backslash(\backslash d\{4\})$ , 3)  $(\backslash d\{2\})/(\backslash d\{2\})/(\backslash d\{4\})$  and as a special case, 4)  $(\backslash d\{1\})\backslash(\backslash d\{1\})\backslash(\backslash d\{2\})\backslash(\backslash d\{4\})$ . Again, other cases from 1999 onwards contain more formats than that, which is the reason why in the final NLP script 2a, the patterns are slightly different to those just listed above. Special characters from before 2010 include roman

numerals indicating the months, and there are also cases in which only the latter two digits for the year are displayed.

Running the NLP script `2a_data-extraction-case-number-and-decision-date-regex.ipynb`, in which the decision dates are extracted using regex, out of 94 files, 94 dates get extracted correctly, which results in a success rate of 100 %. However, regex are very inflexible, even small changes in the format can lead to the date no longer being recognised. For instance, this could be the case if there is suddenly a whitespace character in the year, or, as in the example of “2 7.11.2013”, a whitespace between the two digits that indicate the day, which can lead to errors in the data extraction. For this reason, another NLP technique is explored to extract the decision date.

#### 4.2.2.2 *Named Entity Recognition, Keyword Matching and Regex*

SpaCy is an open-source software library built with the programming language Python and is made for NLP (SpaCy, 2023a). Its linguistic features encompass part-of-speech tagging, morphology, lemmatisation, dependency parsing, Named Entity Recognition (NER), entity linking, tokenisation, merging and splitting, sentence segmentation, rule-based mappings and exceptions, and word vectors and semantic similarity (SpaCy, 2023b). With regard to extracting the decision dates, NER is a technique likely to be successful. SpaCy’s NER system is based on a pre-trained model that labels word entities like persons, locations, or dates (SpaCy, 2023b). Accordingly, it is a more flexible NLP technique than regex and might thus be the better option for the extraction of the decision dates.

For this reason, in a secondary NLP script 2b, the decision date is extracted using SpaCy’s NER model. The first part of the script is the same as in the regex one, but after that, instead of using regex again for the extraction of the decision date, SpaCy’s NER model is used. In a first attempt, the first three pages of the prohibition decisions are transformed into a SpaCy document object and then the first entity that is recognised as a date and is between 1999 and 2023 is returned. The code uses the SpaCy library to tokenise the text. Tokenisation is the process of splitting the text into words, phrases, symbols, or other meaningful elements, called tokens (Menzli, 2023). In this code, tokenisation is done by the `nlp(text)` function, which processes the text and returns a SpaCy document object with tokenised words. Unfortunately, this preliminary approach proved largely ineffective, securing accurate decision dates for only 22 out of 94 files, marking an accuracy rate of just

---

23.41 %. For this reason, this approach has subsequently been further developed, as elaborated in the following.

To improve the performance and enhance accuracy, other NLP techniques are added. Alongside keyword matching and application of the library “datefinder”, which uses regex and heuristics to find and parse dates in the text, 71 decision dates out of 94 are extracted correctly, which is a success rate of 75.53 %. This is already a huge improvement on the previous code. In addition to those 71 correctly extracted dates, in 16 cases (17.02 %), the correct date is initially identified within the text as well. However, while the year is then extracted correctly, the day and month are mistakenly swapped. The code can be found in the NLP script *2b\_data-extraction-decision-date-spacy.ipynb*.

There are several promising possibilities for improving the date extraction process that could be explored in future studies. A first step could be to perform a detailed error analysis of the cases where data was not returned correctly or not returned at all. This would mean examining these cases in detail to understand the common patterns of these errors, which could enable the development of more sophisticated extraction rules or the refinement of existing rules. In addition, the possibility of training SpaCy’s NER model on the specific dataset could be considered. This would involve fine-tuning the existing NER model to the prohibition decision texts, which could lead to improved performance in identifying the decision dates in these documents. Also, the exploration of advanced deep learning algorithms such as BERT could be considered. While SpaCy’s NER model incorporates aspects of such algorithms, a stand-alone implementation of BERT or similar models could provide a more sophisticated understanding of the date extraction task and therefore yield better results. Furthermore, an experimental approach could include an interface to advanced language models such as ChatGPT. This would involve querying the model with text passages containing the decision date and analysing the responses for date information. As ChatGPT has been trained on a variety of internet texts, it could have a robust ability to extract date information from the EC’s prohibition decision texts.

#### 4.2.3 Cartel Start, End and Duration

The cartel start date and end date are usually mentioned in a sub-chapter entitled “Duration of the infringement” (or slightly modified versions thereof, such as “Duration”, “Duration of Infringement”, etc.).



For this reason, in a first step, regex are used to filter for the according subchapter headings. The regex patterns are used to identify specific markers or starting points in the text related to infringement duration. The code then also attempts to find the end of these sections by looking for the beginning of the next chapter with different regex patterns. This step could be improved as it is not as successful yet as finding the starting point. Due to time constraints, this has not been optimised further, but there is certainly great potential for improvement.

The extracted text for each case is then saved as a separate txt file in a newly created folder in the current working directory labelled *dates*. The text files are named after the original PDF files, making it easier to associate the extracted data with the original files later on. That first step, contained in NLP script 2c, is a standalone process. Because the subsequent NLP scripts 2d\_a and 2d\_b both depend on it, this code is isolated in its own script *2c\_extract\_text\_from\_duration\_of\_infringement.ipynb*.

The process is designed to simplify data extraction by substantially reducing the amount of text to be searched through. However, there are several factors that considerably complicate data extraction. First, it would be a natural logical assumption that the date in that section of the text which dates furthest back is the date on which the cartel started, and the most recent date indicates the end date of the cartel. However, this is not the case. Second, cartels with several parties often have different start and end dates listed because the parties joined the cartel at different times, and ceased collusion at different times, too. Third, in the case of different infringements, a different duration might be given for each infringement. Combined with many parties joining and leaving the cartel at different times, numerous dates need to be parsed through. All of this makes the extraction of the cartel start date and end date intricate and complex.

#### *4.2.3.1 SpaCy's NER Model with Keywords*

Extracting the data regarding cartel start date and cartel end date is the most challenging part of the data extraction done in this thesis. In the following, an attempt was made to first select the sentences in which data is found using SpaCy's NER model. For this, two different versions were created and tested.

In the first version 2d\_a, several iterations through the previously created txt files are made. Using SpaCy's NER model, all sentences that contain a date object are selected. Then, according to defined keywords for both cartel start date and cartel end date, those

---

sentences are combed through. If a date object is found in the same sentence as one of the keywords it gets selected, the first match is chosen as cartel start date or cartel end date respectively. In a second iteration, the files for which no cartel end date but a cartel start date has been found are iterated through. This time, another set of keywords is defined. All the dates that match are put into a list and converted into datetime format. Then, they are sorted, and the most recent date is set as cartel end date. Together with the cartel start date, the cartel duration is then calculated and put into the newly created Excel file *cartel\_data\_duration.xlsx* in the folder *dates*. This is done again for the cases with missing values for cartel start date or both cartel start date and cartel end date in a third iteration. The code can be found in the NLP script *2d\_a\_data\_extraction\_duration\_spacy\_v1.ipynb*.

The second version *2d\_b* is very similar to the first one, except that in this case, only one iteration for cartel start date and one iteration for cartel end date is made and the keywords are defined once. Overall, the code is cleaner and better structured, but, looking at the results displayed in Table 5 below, the overall accuracy of extracting the correct dates is lower, even though this version *2d\_b* gets at least the year values correct more often than version *2d\_a*. For this reason, the code is nonetheless included in this thesis. The code for version *2d\_b* can be found in the NLP script *2d\_b\_data\_extraction\_duration\_spacy\_v2.ipynb*. The output of version *2d\_b* is saved in the Excel file *cartel\_data\_duration-v2.xlsx*, which is located in the folder *dates*.

The first version *2d\_a* extracts a total of 86 dates out of 94 data points for the cartel start dates, and a total of 87 dates for cartel end dates. The second version *2d\_b* extracts a total of 81 dates out of 94 data points for the cartel start dates, and a total of 88 dates out of 94 data points for cartel end dates. Whether the dates correspond to the actual cartel start dates and cartel end dates is assessed in the next step.

The extracted values were manually compared on the one hand with the data from the Excel file *Cartels1964-2010.xls* which contains the cartel start dates and cartel end dates for cases up to 2010. On the other hand, for the decisions from 2010 onwards, the values for each case were compared directly with the information from the corresponding prohibition decisions. The following should be clarified about the evaluation; as soon as the year is wrong, the month and day are also automatically indicated as wrong. The same applies to month; as soon as it is incorrect, the day is also automatically evaluated as incorrect. This means that wherever the day is correct, the entire date is correct. In other

---

cases, it may be that only the year or only the year and month are correct. In cases where no day is mentioned in the prohibition decision, the first day of the month is automatically set as the start date and the last day of the month as the end date. In these cases, if the year and month are correct, the day is also considered to be correct. The accuracy rates displayed in Table 5 were calculated.

**Table 5:** Accuracy rates regarding extraction of cartel start date and cartel end date for versions 2d\_a and 2d\_b (own illustration)

	Cartel start date			Cartel end date		
	Year	Month	Day	Year	Month	Day
<b>Version 2d_a</b>	29.79 %	26.60 %	21.28 %	23.40 %	20.21 %	13.83 %
<b>Version 2d_b</b>	30.85 %	22.34 %	19.15 %	34.04 %	22.34 %	10.64 %

Interestingly, the dates that are correctly recognised largely concern different cases in the two versions 2d\_a and 2d\_b. While in version 2d\_b more years are given back correctly, in version 2d\_a more dates are correct as a whole (where the date up to and including day is correct). There is certainly a lot of potential for improvement here. Due to the shortage of time, no detailed error analysis was carried out, but a second approach has been tried, which is again in large parts based on the same process as described for the two previous versions.

Since the success rates are relatively modest for both versions, the process is carried out again in version 2e. Now, however, it is not the text after "Duration of Infringement" that is looked at, but the text after "HAS ADOPTED THIS DECISION". This is because in most cases the duration of infringement is also listed there again under "Article 1". However, this is not the case in all prohibition decisions, especially the earlier ones might not include the duration of infringement again in the text under the announcement of the formal decision. Nonetheless, the later cases most often include that, and thus, this approach is the most promising to build on for future research.

There are some slight changes again in this code with regard to versions 2d\_a and 2d\_b. For instance, there is a limit set with respect to the cartel start date, which cannot be older than 01/10/1946. This date was chosen since it is three years earlier than the earliest

known start date for a cartel that is included in the used dataset. Thus, the code should be able to find all cartel start dates in the known range for the data that is looked at. Also, the cartel start date cannot be more recent than the cartel end date, and the cartel end date cannot be more recent than the decision date. The code this time only iterates through every file once and extracts both cartel start date and cartel end date in one go. This third version 2e extracts a total of 76 dates out of 94 data points for the cartel start dates, and a total of 76 dates out of 94 data points for cartel end dates. The code for version 2e can be found in NLP script *2e\_data\_extraction\_duration\_spacy\_v3.ipynb*.

Clearly, this last version 2e yields the best results overall, as can be seen by the computed accuracy rates depicted in Table 6 below.

**Table 6:** *Success rates regarding extraction of cartel start date and cartel end date for version 2e (own illustration)*

	Cartel start date			Cartel end date		
	Year	Month	Day	Year	Month	Day
<b>Version 2e</b>	67.02 %	56.38 %	55.32 %	62.77 %	55.32 %	27.66 %

What is noticeable and warrants further investigation is that the cartel end dates are recognised significantly less accurately than the cartel start dates. With regard to the cartel start dates, version 2e with 55.32 % accurate to the day date extractions already provides a fairly good value in view of the fact that there was not enough time to conduct a detailed error analysis. Such an analysis should certainly be carried out by future researchers building on that code and could certainly improve the accuracy of the data readout.

Furthermore, several approaches could potentially improve the performance of the data extraction model. First and foremost, a combination of the three approaches used and presented in this section could potentially enhance the data extraction regarding cartel start dates and cartel end dates, since in all three versions, some dates were extracted correctly which the other two versions did not catch. Second, a more sophisticated approach to keyword selection could significantly improve the performance of the data extraction model. Currently, keywords are selected manually, which has its limitations. To improve this process, statistical analysis techniques or advanced NLP methods such as topic modelling or semantic analysis could be used. These techniques can provide a more

sophisticated understanding of the context and reveal keywords that might have been missed in a manual selection process. Third, other models could be employed. The current model uses SpaCy's pre-trained NER model, which is undeniably effective. However, there are other NLP libraries and models which might achieve a higher accuracy regarding date extraction. For example, models such as BERT or Transformers could be explored. Also, training SpaCy's NER model specifically for this task may also improve performance. This would require feeding the model with more cartel-related data to improve its ability to understand and accurately predict the cartel start dates and cartel end dates. This approach could be particularly effective if there is a unique language pattern or terminology specific to the topic. Fourth, an interesting possibility would be to integrate an Application Programming Interface (API) into language models, like ChatGPT. Given the advanced speech understanding capabilities of the Generative Pretraining Transformer (GPT), it could be used, for example, to check the context and verify the identified data, or to fill in missing data.

There are many methods that could be explored further, the necessary data is available and given enough time the success rate can undoubtedly be improved. By exploring these different approaches, future research could certainly achieve more accurate and reliable results in extracting cartel start dates and cartel end dates from the dataset.

#### 4.2.4 Report Route and Route Indicator

There are four different ways in which a cartel can be discovered: 1) notification, 2) complaint, 3) Commission's own initiative and 4) leniency application. How the cartel authorities became aware of the infringement is usually described in the chapter "Procedure" (or similar wording).

A leniency application is usually made under Article 8(a) since the leniency notice of 2002 (EC, 2002, 2006). The article reads as follows:

*“A. IMMUNITY FROM FINES*

*8. The Commission will grant an undertaking immunity from any fine which would otherwise have been imposed if:*

*(a) the undertaking is the first to submit evidence which in the Commission's view may enable it to adopt a decision to carry out an investigation in the sense of Article 14(3) of Regulation No 17 (2) in connection with an alleged cartel affecting the Community; or*

---

*(b) the undertaking is the first to submit evidence which in the Commission's view may enable it to find an infringement of Article 81 EC (3) in connection with an alleged cartel affecting the Community.”*

It is usually referred to as “[...] *under point 8(a) of the Commission Notice [...]*” in the prohibition decisions. This also means that this specification is the most explicit. If this expression is found in the document, it means the investigation was initiated by an application for leniency.

The same applies to a complaint, which is submitted through official channels and thus should be explicitly referred to as “complaint” in the text. This means that if no leniency application has been made, complaints are then searched for, followed by notification and finally the Commission’s own initiative.

#### *4.2.4.1 Regex and Keyword Matching*

So, similar to what was done for extracting the cartel start dates and cartel end dates, in a first step, regex are used to filter the text for the according subchapter headings. This is done by accessing the previously created txt files in the *folder cases\_from\_1999 > commission\_decisions\_text\_files*.

Different keywords are then defined in a dictionary to assign the matches found to the four possible report routes, in the order inferred above. The keywords for leniency were also supplemented by other keywords than the ones referring to “point 8(a)”, as not every decision explicitly refers to the article, especially in the cases in which the application was still submitted on the basis of the 1996 Leniency Notice.

The decisions are then examined in three iterations according to the keywords; in the first iteration, the text is combed through directly after the "Procedure" chapter. Normally, the relevant information should be found there, thus, by only searching through that part of the decision text first, incorrectly assigned report routes should be avoided. If no initiation type could be assigned in the first iteration, a little more text is taken and searched through in the second run. If nothing is found after the second iteration, the entire decision text is combed through. For files where a specific report route could not be discerned, it is assumed that the investigation was initiated by the most common method, which is the Commission's own initiative. Accordingly, the value Commission's own initiative is assigned in these cases. For the present data, this only concerns one case.

The values for *route\_indicator* depend on the values in *report\_route*. Because of this, in a second step, the integers are assigned according to the values in *report\_route*. In addition, where the integer in *route\_indicator* is equal to 4, a 1 is assigned in the *leniency* column, and a 0 in all other cases. The code can be found in the NLP script *2f\_data\_extraction\_report\_route\_leniency.ipynb*.

Out of 94 cases, in 69 cases the correct report routes were assigned. This is an accuracy rate of 73.4 %. Regarding leniency, out of 94 cases, in 74 cases the correct Boolean value was assigned, constituting an accuracy rate of 78.72 %.

Depending on the order of the report route indicators in the dictionary, different results are obtained. This suggests that the keyword selection should be further refined in a next step. In addition, as in the previous processes, the accuracy of the data read out could certainly be improved by implementing further NLP techniques.

#### 4.2.5 Leniency

Information about whether the proceedings were initiated due to an application for leniency can be inferred from the data extraction detailed in Section 4.2.4 Report Route and Route Indicator. Consequently, values derived from that previous process can be applied to determine if the investigation was initiated because of a leniency application. This involves checking the *route\_indicator* column for the presence of the number 4, as has been done in Section 4.2.4.1 Regex and Keyword Matching.

Beyond merely indicating whether the proceedings were initiated due to a leniency application, additional information can also be ascertained. For example, it can be determined whether the cartel members applied for leniency only after the cartel came to the attention of the cartel authorities through another report route, or whether any application for leniency was filed throughout the proceedings, not necessarily under Article 8(a) of the Commission Notice.

##### 4.2.5.1 Keyword Matching

To analyse this, a very simple keyword matching is applied, which searches the complete text of the prohibition decisions for the mention of the leniency notice. If it is found, a value of 1 is set under the column *leniency\_applied*, otherwise a value of 0. The code can also be found in the NLP script *2f\_data\_extraction\_report\_route\_leniency.ipynb*.

Out of 94 cases, the correct Boolean value was assigned in 85 cases, which is a success rate of 90.43 %.

It is a simple process that can be improved further, given the time. In a next step, the information extraction could be improved by combining this procedure with other NLP techniques. Then, by comparing the previously read data, which can be found in the leniency column, it can be found out in which cases Article 8(a) was applied and thus the investigation was started by the application for leniency, and in which cases the cartel members referred to the leniency notice after the procedure was started and provided information to facilitate the investigation by the cartel authorities. Additional guidance on possible further research regarding NLP techniques for data extraction from the prohibition decisions can be found in Section 7 Conclusion and Future Work.

### 4.3 Data Preparation

To answer the main research question, a regression analysis with time series is carried out on the basis of the existing data. This requires collecting and formatting the relevant data in a consistent manner.

The main data were compiled manually by previous researchers and cover cartel cases from 1964 to 2010. For cases from 2010 to 2021, the missing data regarding case number, decision date, cartel start dates and cartel end dates, cartel duration, leniency and report route were added, partly using a data analysis script for already existing data (case number and decision date), partly manually for the data that had not yet been read out at the time of the analysis being conducted.

The data preparation process consists of the following steps:

1. In the first step, the existing data is processed and formatted. This is done with the data analysis script *3a\_data-preparation-regression-part1.ipynb*.
2. In a second step, the missing data is added manually.
3. In the last step, an automated process is used to supplement the dataset which is done with the data analysis script *3b\_data-preparation-regression-part2.ipynb*, which can be found both in the annex and on GitHub.

The first step is to create an Excel file named `new_cartel_data_regression.xlsx` with the columns `year`, `case_number`, `decision_date`, `cartel_start`, `cartel_end`, `leniency_regulation`, `leniency_application`, `report_route` and `fines`. The manually collected data can be found in the Excel file `Cartels1964-2010.xls`, which is stored in the folder `Data_until_2010`. Both this folder and the Excel file can also be found on GitHub. In the following, the case



numbers, the decision dates as well as the cartel start years and cartel end years are taken from the file *Cartels1964-2010.xls*. The same is done for fines, whereas it is assumed that the set values in the fines column depict the average fine imposed per cartel member. This assumption is made because the same amount is usually set for all parties to the cartel in the Excel file, even though it is presumed that not the same fine has been imposed on all members by the EC in the corresponding prohibition decisions. Nevertheless, to mitigate possible mistakes, the average of all the imposed fines per entity is calculated.

Subsequently, for the other case numbers and the decision dates of the cases from 2010 onwards, the Excel file *cleaned\_cartel\_data.xlsx* is accessed, which was created during the data collection for the extraction of these two values in NLP script 2a. Both case number and decision date are read out and appended for the case numbers that are not yet present in the data frame. Any duplicate values are dropped.

After the case numbers and decision dates have been transferred, the cartel start years and cartel end years must be added for the cases from 2010 onwards, as well as the integer indicating report route and leniency application, and the average amount of fines imposed. At the time the analysis was carried out, this information could not be read out automatically from the PDF files. For this reason, this step is carried out manually. Likewise, some of the data that has been read out of the file *Cartels1964-2010.xls* is manually deleted from the file *new\_cartel\_data\_regression.xlsx* because there is neither cartel start year nor cartel end year. For these files, a manual check is made in the original file to see if there is a comment from the researchers who created the data set. In the original Excel file *Cartels1964-2010.xls*, the rows regarding the corresponding case numbers are marked in red and/or labelled "*Mistake!*", which is why they are deleted and not used for further processing. Otherwise, the cartel start years and cartel end years are added manually. Also, in one case concerning case number 38338, the cartel end year had mistakenly been indicated as 199, which is corrected to 1999. After the missing data has been added, the file is saved under the name *new\_cartel\_data\_regression\_complete.xlsx*. This file is then used for further data preparation. It is also available on GitHub.

The value 1 is set under the column *leniency\_application* if the procedure was initiated due to a leniency application. If the value is 1, the report route is automatically set to number 4. However, there are several cases where the investigation was initiated based on one of the other three possible report routes, but at the very beginning of the investigation, one or more companies involved in the cartel agree to cooperate under the 1996

Leniency Notice (or, in later cases, under the 2002 Leniency Notice or the 2006 Leniency Notice). As an illustrative example of such a case, an excerpt from the decision on case number 38238 is reproduced here:

*“(3) On the basis of information to the effect that the Spanish raw tobacco processors and producers had infringed Article 81 of the Treaty, the Commission carried out the following inspections under Article 14(3) of Regulation No 17:*

*[...]*

*(4) By letter of 16 January 2002, the four Spanish processors (the fourth processing firm in Spain is Tabacos españoles, S.L. - "Taes") and their association ANETAB announced that they were committed to cooperating with the Commission in the proceedings. They referred to the 1996 Commission notice on the non-imposition or reduction of fines in cartel cases ("the 1996 leniency notice"). They also informed the Commission that, as from 3 October 2001, they had put an end to the practices concerned by this Decision.*

*[...]*

*(6) By fax of 15 February 2002, Universal Leaf Tobacco Company Inc. ("Universal Leaf"), the parent company of Taes, informed the Commission that it supported its subsidiary's initiative of cooperating with the Commission within the framework of the 1996 leniency notice. It also pointed out that its subsidiary in Italy, Deltafina Spa ("Deltafina"), was cooperating with Taes in drafting the memorandum that the latter was expecting to send to the Commission in the days ahead as part of the announced cooperation and that it hoped that Deltafina could thus also benefit from the advantages flowing from the 1996 leniency notice.*<sup>2</sup>

The first sentence in paragraph three, “*On the basis of information to the effect that the Spanish raw tobacco processors and producers had infringed Article 81 of the Treaty*”, indicates that the EC has received a notification and therefore the investigation has been launched. However, in paragraphs four and six, it is mentioned that some of the involved undertakings cooperate under the 1996 Leniency Notice.

---

<sup>2</sup> Note: Underlining of sentences was done by the author of this thesis.

To account for these cases and since it is not entirely clear how they were handled by the previous researchers that put together the existing dataset from 1964 to 2010, a column "*leniency\_applied*" was added for the additional cases from 2010 to 2021. It employs a Boolean value to indicate whether there was any application for leniency even though the proceedings were initiated through another report route. For cases up to 2010, the values from *leniency\_application* are taken over into this column, despite the prevailing uncertainty regarding how these were managed by prior researchers.

As mentioned above, the report routes distinguish between 1) notification, 2) complaint, 3) Commission's own initiative and 4) leniency application. A notification is understood to mean that the EC is informed by another authority, or a natural person or legal entity, of the existence of a possible cartel. A complaint is classified as such when a formal complaint has been lodged by another legal entity or natural person not involved in the cartel. It is assumed that if the term "*information received*" or similar sentence structures, such as "*on the basis of information to the effect [...]*" (see above) are present in the decision text, a notification has likely been received. Nonetheless, it is rather difficult to distinguish between notification and complaint. It was attempted to assign the report routes as accurately as possible, but there is no 100 % certainty as to their correct assignment.

Once the missing data has been added manually, the data processing can be continued. For this purpose, the file *new\_cartel-data-regression-complete.xlsx* is loaded into a new data frame. For the regression analysis with time series, it is crucial to indicate, for each case number per year, whether the cartel was active or not. To achieve this, the series is duplicated for each case number for the years spanning from 1964 to 2023. The years are matched with cartel start and end year. If the cartel was active in the corresponding year, the number 1 is entered in the cartel column, otherwise, the number 0 is recorded. Finally, the *leniency\_regulation* column is also filled in with a Boolean value; from 1996 onwards, the value 1 is entered, before 1996, the value 0 is set. The final Excel file is again saved under the name *new\_expanded-cartel-data-regression.xlsx*. This file is subsequently used for the regression analysis.

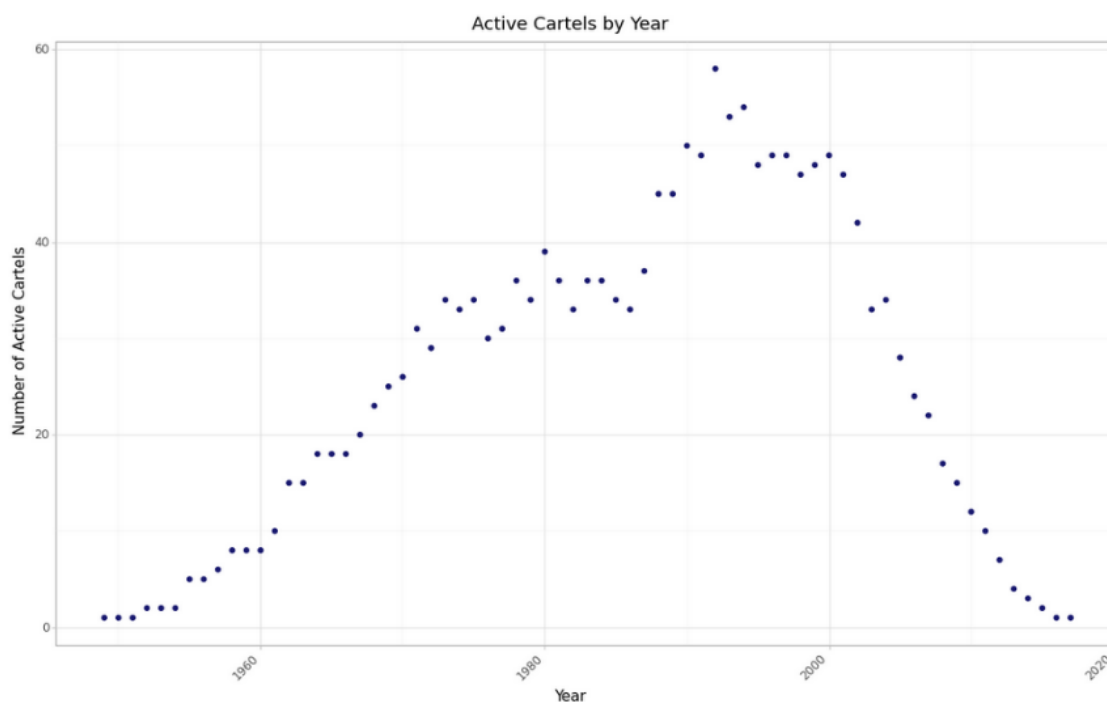
#### 4.4 Data Analysis

For conducting the data analysis, the data structure needs to be examined in a first step. There are 15'124 rows and 9 columns in the Excel file, which contains panel data, that is, data for a total of 199 entities, for each of which datapoints span over a total of 76 years (from 1948 to 2023). The columns *cartel*, *leniency\_regulation* and *leniency\_application*

contain Boolean values, the first indicating the years during which the cartel was active, the second denoting whether the leniency program has been in force, and the last one signifying whether an application for leniency has been made or not. The case number itself identifies the different entities.

Panel regression is a powerful tool when the available data varies both across time (time series, longitudinal data) and across entities (cross-sectional), as it allows to control for variables that change over time as well as variables that are specific to each entity but do not change over time. It is also a tool that can be used over two or more time periods, to conduct a “before and after” comparison (Hanck et al., 2023, p. 231 et seq.; Wooldridge, 2012, p. 459 et seq.). Since the data structure of the collected data matches this approach, panel regression is used for the data analysis to answer the research question.

In a first step, a scatterplot is created that illustrates the yearly sum of active cartels. This plot extends from 1948, continuing until just before 2020. This visual representation, depicted in Figure 9 below, offers a graphical representation of the frequency of cartel activity over the specified timeframe and as such provides an intuitive understanding of cartel activity trends over time.



**Figure 9:** Plot of active cartels per year (own illustration)

A visual inspection of the data suggests the presence of an upward trend until the mid-1990s and a strong downward trend thereafter, with the highest number of active cartels

appearing in 1991. These apparent trends could indicate non-stationarity in the time series. This would potentially violate the assumption of stationarity often required in time series analysis (Hyndman & Athanasopoulos, 2018). However, these initial observations are statistically tested using the Augmented Dickey-Fuller test, a more rigorous method for determining stationarity. The test results for both the *cartel* and *leniency\_regulation* variables strongly suggest that both series are, in fact, stationary, with p-values of 0.000 and test statistics lower than the critical values at the 1 %, 5 %, and 10 % levels. This provides solid grounds to reject the null hypothesis of a unit root, indicating that despite the observed trends, the series can be considered stationary. Consequently, proceeding with the panel regression analysis is deemed statistically appropriate.

#### 4.4.1 General Analysis

To figure out how the implementation of the leniency program in 1996 influenced the detected cartel activity within the EU, a panel regression analysis is performed and the common assumptions on time series are subsequently tested for the model.

The null hypothesis is as follows:

*The introduction of the leniency program in 1996 has not significantly affected the detected cartel activity within the European Union.*

$$H_0: \beta_1 = 0 \text{ in the panel regression model: } \text{cartel}_t = \alpha + \beta_1 \text{leniency\_regulation} + \beta_2 \text{year} + u_t$$

Accordingly, the alternative hypothesis is:

$$H_a: \beta_1 \neq 0 \text{ in the panel regression model } \text{cartel}_t = \alpha + \beta_1 \text{leniency\_regulation} + \beta_2 \text{year} + u_t$$

The panel regression analysis is – as already indicated by the null hypothesis – modelled by the following equation:

$$\text{cartel}_t = \alpha + \beta_1 \text{leniency\_regulation} + \beta_2 \text{year} + u_t$$

In the equation, *cartel<sub>t</sub>* is the dependent variable, which as a metric variable represents the sum of active cartels per year.  $\alpha$  is the intercept,  $\beta_1 \text{leniency\_regulation}$  is an independent dummy variable with values 1 or 0 and thus represents a categorical, nominal variable.

$B_2$ year is another independent variable that can also be considered a control variable and indicates a trend component. The year variable is a metric, interval-level variable. Finally,  $u_t$  denotes the error term. The coefficients  $\beta_1$  and  $\beta_2$  correspond to the effects of the leniency regulation and the year on detected cartel activity, respectively.

Running the model, in which the year variable is included as a fixed effect, the output displayed in Figure 10 is computed.

```

PanelOLS Estimation Summary
=====
Dep. Variable:          cartel    R-squared:                0.0022
Estimator:              PanelOLS  R-squared (Between):      -0.1204
No. Observations:      15124    R-squared (Within):       0.0022
Date:                   Mon, May 15 2023  R-squared (Overall):     -0.0211
Time:                   21:56:47    Log-likelihood            -3585.9
Cov. Estimator:        Unadjusted

                               F-statistic:                32.271
Entities:                199      P-value                   0.0000
Avg Obs:                 76.000   Distribution:              F(1,14924)
Min Obs:                 76.000
Max Obs:                 76.000   F-statistic (robust):     32.271
                               P-value                   0.0000
Time periods:           76      Distribution:              F(1,14924)
Avg Obs:                 199.00
Min Obs:                 199.00
Max Obs:                 199.00

                               Parameter Estimates
=====
                               Parameter  Std. Err.   T-stat   P-value   Lower CI   Upper CI
-----
leniency_regulation      -0.0296    0.0052     -5.6808  0.0000   -0.0398   -0.0194
=====

F-test for Poolability: 6.8109
P-value: 0.0000
Distribution: F(198,14924)

Included effects: Entity

```

**Figure 10:** PanelOLS Estimation Summary regarding the impact of the leniency program of the EC on overall detected cartel activity (own illustration)

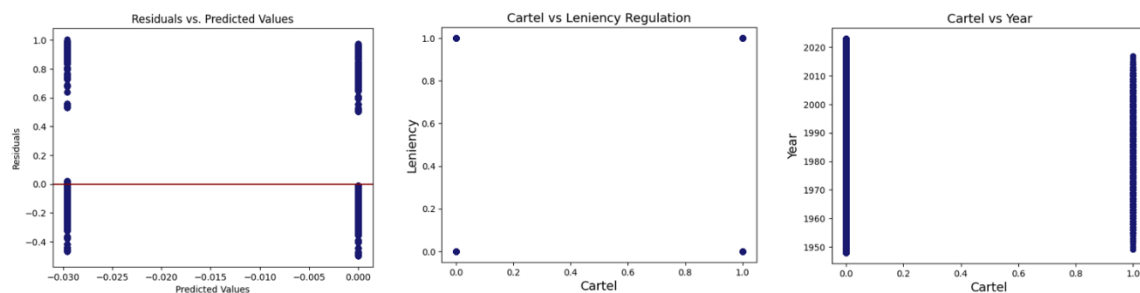
The model has a very low R-squared value, suggesting that it explains only a small fraction of the variance in the dependent variable *cartel*. The negative parameter estimate of -0.0296 for the explanatory variable *leniency\_regulation* suggests that there is a negative relationship between the introduction of the leniency regulation and the detection of cartels. In other words, an increase in leniency regulation is associated with a decrease in the likelihood of cartel activity being detected. The exact p-value of the coefficient for *leniency\_regulation* is  $1.37e-08$ . Together with the t-statistic of -5.6808, this indicates that the null hypothesis can be rejected, and the observed relationship between the introduction of leniency and the detected cartel activity is not due to random chance. However, to

reject the null hypothesis with certainty, certain assumptions regarding time series need to be met.

Hereinafter, the common assumptions for ordinary least squares (OLS) regression in the context time series are tested, which are: 1) the relationship between  $X$  and  $Y$  is linear in parameters, 2) no independent variable in the sample is constant, nor is it a perfect linear combination of the others, 3) the error term is uncorrelated with the explanatory variables  $X$  in every time period  $t$ , 4) there is homoskedasticity, 5) there is no serial correlation between the errors of different time periods and 6) the error term follows a normal distribution with a mean of 0 and a constant variance of  $\sigma^2$  (Wooldridge, 2012, p. 349 et seq.).

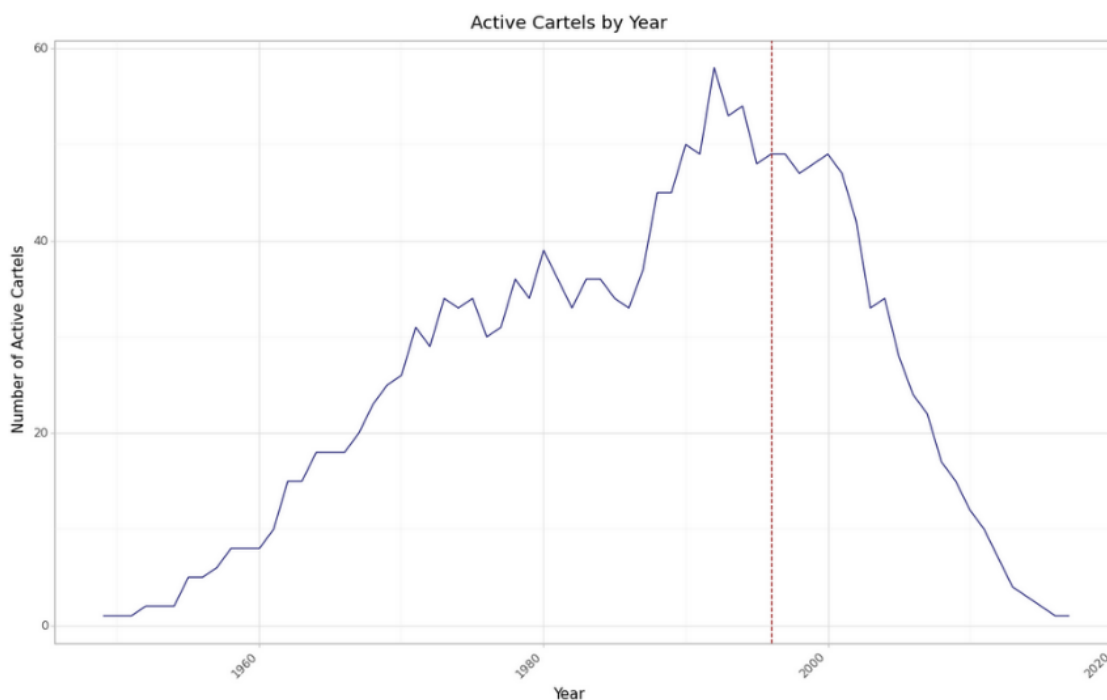
#### *Assumption 1: Linear Relationship*

The first assumption states that the data in the time series follows a model which is linear in its parameters (Wooldridge, 2012, p. 349). To check for linearity, both residuals vs. fitted values and scatterplots for the independent variables are looked at. The corresponding scatterplots are depicted in Figure 11.



**Figure 11:** Scatterplot of residuals vs. fitted values (1) and dependent variable vs independent variables (2 and 3)  
(own illustration)

The visual inspection implies that a linear relationship may not be present, which is consistent with the binary nature of the dependent variable *cartel* for each entity. However, given that the observations stem from time series data, the temporal aspect needs to be accounted for as well. Looking at the total of active cartels over time, the created line plot reveals more of a curve instead of a perfectly linear relationship between  $X$  and  $Y$  over time, as can be seen in Figure 12.



**Figure 12:** Line plot of active cartels over time with vertical line in 1996 (own illustration)

However, the line plot in Figure 12 could also be used to argue in favour of a linear relationship from the beginning of the observations to 1996 and then from 1996 to 2023, but with opposite trend lines (upwards trend before the introduction of the leniency program vs. downwards trend after).

#### *Assumption 2: No Perfect Collinearity*

The second assumption rules out perfect collinearity between independent variables; no independent variable must be constant, nor can it be a perfect linear combination of the others (Wooldridge, 2012, p. 350).

One of the most common ways to check for multicollinearity is to look at the Variance Inflation Factor (VIF). The VIF measures the inflation in the variances of the parameter estimates because of multicollinearity. As a general rule of thumb, only VIFs higher than 5 warrant further investigation, as they indicate that the associated regression coefficients are poorly estimated due to multicollinearity (Shrestha, 2020, p. 40 et seq.; Daoud, 2017, p. 4; Paul, 2006, p. 4 et seq.).

The VIF for *leniency\_regulation* and *year* is computed as 3.3, indicating that the independent variables are moderately correlated and thus that there is some collinearity but not high collinearity between the two variables.



However, plotting a correlation matrix, which gives back the values as listed in Table 7, the correlation between *year* and *leniency\_regulation* is over 0.8, which is a value that suggests high correlation.

**Table 7:** Correlation matrix between independent variables (own illustration)

	leniency_regulation	year
leniency_regulation	1.000000	0.835573
year	0.835573	1.000000

The correlation matrix is also visually displayed in Figure 13 below.



**Figure 13:** Correlation matrix for the independent variables (own illustration)

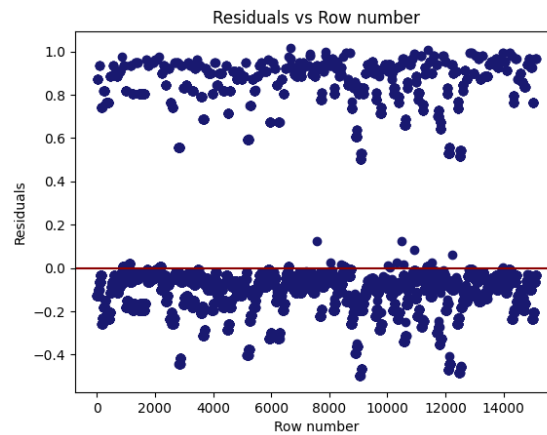
Considering that despite that high correlation there is still no perfect collinearity present, the second assumption is not violated in the model.

#### *Assumption 3: Zero Conditional Mean*

The third assumption states that over all time periods, the expected value of the error term  $u_t$  is zero, given the explanatory variables for all time periods. This implies that the error at any time  $t$ ,  $u_t$ , is uncorrelated with each explanatory variable in every period, which essentially means that the error term should not be predictable from the information available (Wooldridge, 2012, p. 350).

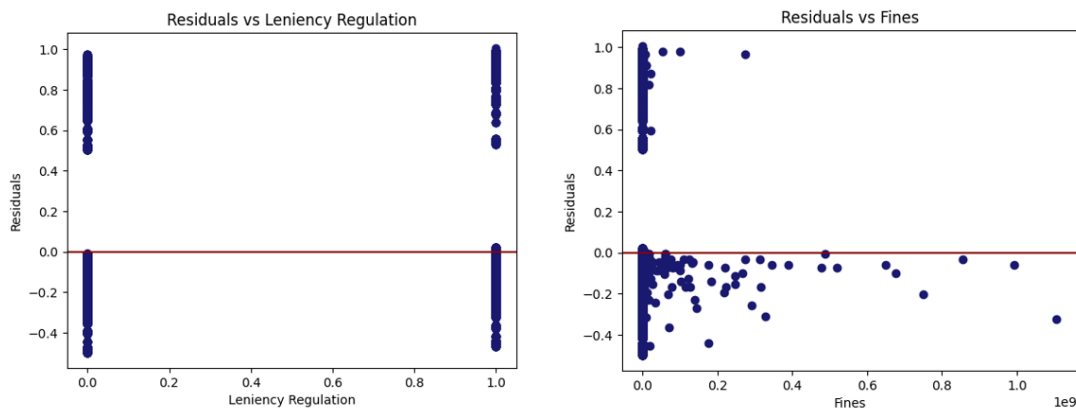
One way to check this is to plot the residuals against unassigned row numbers, which means that the row numbers are not specifically associated with any particular value of the dependent variable. This is because the order of observations in the dataset which is represented by row numbers should not inherently have any correlation with the dependent variable. The residuals should be randomly and symmetrically distributed around

zero. If this is the case, it indicates that there is no correlation between consecutive errors (Shuangyuan, 2021). Doing so, the plot as is displayed in Figure 14 does not show any discernible pattern, which suggests that the third assumption is not violated, either.



**Figure 14:** Scatterplot of residuals against unassigned row numbers (own illustration)

However, to get more content, the residuals are also plotted against the independent variables of the model, to check for any obvious correlations.



**Figure 15:** Scatterplot of residuals against independent variables (own illustration)

As can be seen in Figure 15, no such obvious correlations can be observed. Again, for *leniency\_regulation*, any relationships are not easily distinguished, it being a dummy variable. So far, there is no indication towards the assumption regarding zero conditional mean being violated.

#### *Assumption 4: Homoskedasticity*

This assumption concerns homoskedasticity. It assumes that the conditional on the explanatory variables  $X$ , the variance of  $u_t$ , is the same for all periods  $t$ . Mathematically this is expressed as follows:

$$\text{Var}(u_t|X) = \text{Var}(u_t) = \sigma^2$$

To confirm this assumption,  $\text{Var}(u_t)$  must be constant over time. If the assumption of homoskedasticity does not hold, it means there is heteroskedasticity. So, if the assumption of homoskedasticity is violated, it can lead to inefficient or even biased coefficient estimates (Wooldridge, 2012, p. 352 et seq.).

Before testing for heteroskedasticity, it is crucial to first check for serial correlation in the model's errors. The presence of serial correlation can invalidate the results of a heteroskedasticity test. Once any detected serial correlation has been addressed, heteroskedasticity reliably can be tested for (Wooldridge, 2012, p. 435 et seq.).

After having done this (see elaborations under Assumption 5 below), a Breusch-Pagan test for homoskedasticity is conducted (Wooldridge, 2012, p. 277 et seq.). Running the Breusch-Pagan test on the original model, the output is as follows:

1. *lagrange multiplier statistic = 2337.16 (rounded)*
2. *p-value for the lagrange multiplier test = 0.0*
3. *F-statistic of the hypothesis that the error variance does not depend on the independent variables X = 1381.96 (rounded)*
4. *p-value for the F-statistic = 0.0*

The p-values for both the lagrange multiplier test and the F-test are given as zero and with that are well below the typical significance level of 0.05. This suggests that the null hypothesis of homoskedasticity must be rejected, implying that there is indeed evidence of heteroskedasticity in the model (Wooldridge, 2012, p. 435). The result is not surprising, since the fifth assumption of serial correlation has already been violated in the panel regression model which did not include lagged values for the dependent variable (see elaborations under Assumption 5 below). However, in an adapted model in which lagged values of the dependent variable are included, the results do not change much, though there is some slight change detectable:

1. *lagrange multiplier statistic = 974.58*
2. *p-value for the lagrange multiplier test = 2.35e-212*
3. *F-statistic of the hypothesis that the error variance does not depend on the independent variables X = 347.49*
4. *p-value for the F-statistic = 3.86e-218*

The p-values are still extremely small and thus again well below the significance level of 0.05. This indicates strong evidence against the null hypothesis of homoskedasticity, suggesting that the model's errors are heteroskedastic and thus the variance of the errors changes across observations. Accordingly, this indicates a violation of the fourth assumption for time series. While this is not causing bias or inconsistency in the  $\hat{\beta}_j$ , it invalidates the usual standard errors, t-statistics and F-statistics (Wooldridge, 2012, p. 434 et seq.).

To account for this violation, the parameter `cov_type='robust'` is added to the `mod.fit()` method, which specifies the use of robust standard errors in the regression analysis. The `'robust'` option in the `fit` method uses a method of calculating the standard errors that is robust to heteroskedasticity. It should be noted, however, that this does not fix the problem of heteroskedasticity in itself; it merely provides a way to still draw valid conclusions despite heteroskedasticity being present. Using the `'robust'` parameter increases the standard error, which now is more robust, and the t-value becomes smaller, which means that the results of the analysis are therefore less significant than before.

#### *Assumption 5: No Serial Correlation*

The fifth assumption indicates that the error term  $u_t$  at any given time period  $t$  should not provide any information about the error term  $u_t$  at any other time period  $s$ . Mathematically, it is stated as:

$$\text{Corr}(u_t, u_s | X) = 0, \text{ for all } t \neq s$$

This notation implies that the correlation between the error term at time  $t$ ,  $u_t$ , and the error term at any different time  $s$ ,  $u_s$ , given the explanatory variables  $X$ , in context of regression analyses the independent variables, is zero (Wooldridge, 2012, p. 353).

A common test to check for serial correlation is the Durbin-Watson test, which is based on OLS residuals (Wooldridge, 2012, p. 418 et seq.). A value of 2 suggests there is no autocorrelation detected in the sample, whereas a value  $< 2$  stands for positive serial correlation and a value  $> 2$  stands for negative serial correlation (Wooldridge, 2012, p. 419). The Durbin-Watson test gives back a value of 0.28 (rounded). That value is close to zero and thus strongly indicates a positive autocorrelation, which means that the fifth time series assumption is violated and there is indeed serial correlation present.

To address this issue, the model is run again, this time including lagged values of the dependent variable. This changes the model equation to:

$$cartel_t = \alpha + \beta_1 leniency\_regulation + \beta_2 year + \beta_3 cartel\_lag_{t-1} + u_t$$

After running the new panel regression model with the lagged values of the dependent variable accounted for, the new Durbin-Watson statistic is at 1.90 (rounded). The new value obtained is close to 2, which suggests that now there is very little autocorrelation in the panel data residuals after including the lagged dependent variable. This means that after having addressed the issue, the assumption of no serial correlation is now largely met.

*Assumption 6: Normality*

The sixth assumption refers to the distribution of the errors (the residuals) in the panel regression model. It assumes that these errors are normally distributed. The assumption suggests that the estimated coefficients generated by the OLS method are normally distributed given independent variables. This provides a basis for using statistical tests, such as the t-test for individual independent variable significance and the F-test for joint significance (Wooldridge, 2012, p. 355).

Concerning normal distribution, it is particularly important for small sample sizes. In the case of larger samples, like the one used for the panel regression, it is less imperative. Given the large sample size in this analysis (> 15'000 observations), it is important to consider the implications of the Central Limit Theorem (CLT). The CLT states that, given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution, regardless of that variable's distribution in the population. This holds true even if the underlying population distribution is not normal (Wooldridge, 2012, p. 767 et seq.).

In this case, the large number of observations suggests that the CLT applies. Therefore, even if the individual errors are not normally distributed, their mean should still follow a normal distribution due to the CLT. This means the distribution of the estimated coefficients, which are essentially averages, should still be approximately normal, thus satisfying the conditions for using the t-test and F-test.

To conclude, having tested the common assumptions for time series, the assumption of no serial correlation and homoskedasticity have been violated at first, which can lead to inefficient parameter estimates and can make the standard errors, and thus t-statistics and p-values, unreliable. This in turn could potentially lead to incorrect conclusions about the significance of the coefficients (Wooldridge, 2012, p. 269 et seq., 355). However, both

violations have been accounted for by using a lagged variable of the dependent variable as well as specifying the use of robust standard errors in the regression model.

The final output of the new panel regression model which should now provide valid results is shown in Figure 16:

PanelOLS Estimation Summary						
Dep. Variable:	cartel	R-squared:	0.7415			
Estimator:	PanelOLS	R-squared (Between):	0.9698			
No. Observations:	14925	R-squared (Within):	0.7415			
Date:	Wed, May 17 2023	R-squared (Overall):	0.7853			
Time:	18:35:47	Log-likelihood	6464.2			
Cov. Estimator:	Robust	F-statistic:	2.111e+04			
Entities:	199	P-value	0.0000			
Avg Obs:	75.000	Distribution:	F(2,14724)			
Min Obs:	75.000	F-statistic (robust):	6002.4			
Max Obs:	75.000	P-value	0.0000			
Time periods:	75	Distribution:	F(2,14724)			
Avg Obs:	199.00					
Min Obs:	199.00					
Max Obs:	199.00					
Parameter Estimates						
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
cartel_lag	0.8600	0.0083	103.37	0.0000	0.8437	0.8763
leniency_regulation	-0.0163	0.0027	-5.9873	0.0000	-0.0217	-0.0110
F-test for Poolability: 0.4736						
P-value: 1.0000						
Distribution: F(198,14724)						
Included effects: Entity						

**Figure 16:** PanelOLS Estimation Summary where heteroskedasticity and serial correlation have been accounted for (own illustration)

The R-squared (Within) value has increased substantially from 0.0022 in the first model to 0.7415 in the second model. This suggests that the new model is much better at explaining the variation within the data than the initial model. Also, the overall R-squared value indicates that the model explains approximately 78.53 % of the variance in *cartel*.

The coefficient  $\beta_1$  for *leniency\_regulation* has changed from -0.0296 in the first model to -0.0163 in the second model. Its p-value is even smaller now, 1.02e-09, and high significance is thus given. The t-statistic regarding the significance of the impact the leniency program has on detected cartel activity is -5.9873, having slightly decreased compared to the model without robust standard errors (where the value was -6.1069), but still very clearly indicating that this predictor is also highly statistically significant.

The F-statistic is very high, and the p-value associated with this F-statistic is zero, which indicates that at least one of the predictors in the model is statistically significant in explaining the dependent variable.

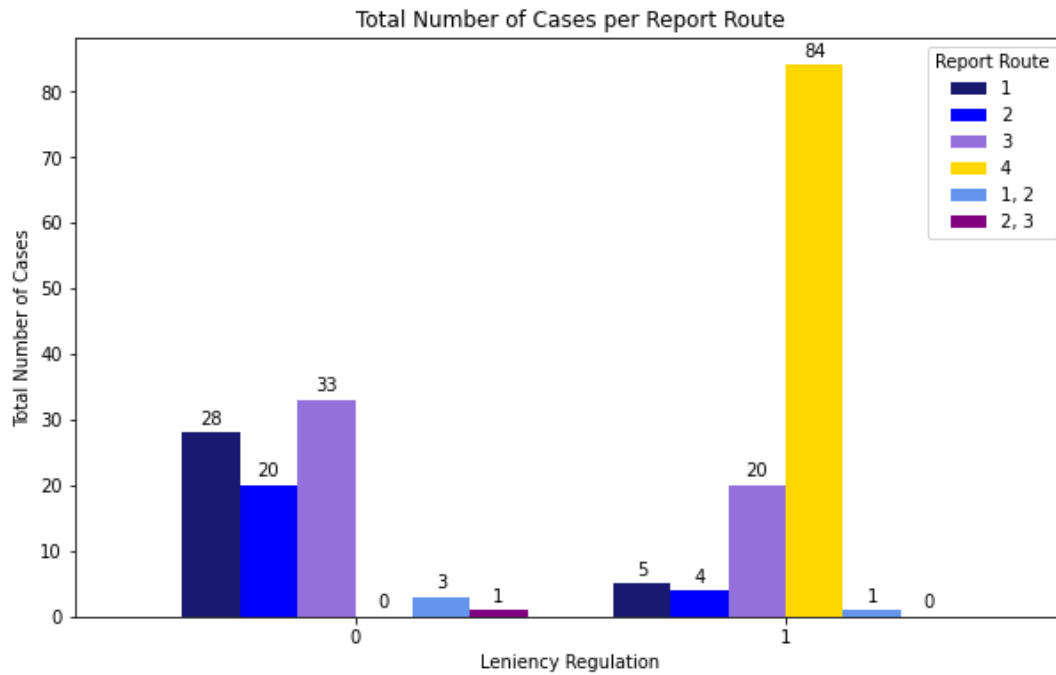
#### 4.4.2 Non-leniency Enforcement

In order to gain a comprehensive understanding of the impact of leniency, it is crucial to analyse the extent of non-leniency enforcement both before and after its introduction. Such an examination not only provides insights into the changing dynamics of cartel prosecution, but also sheds light on the allocation of resources within enforcement agencies. As suggested by Harrington & Chang (2015), the presence of non-leniency enforcement prior to leniency is fundamental to the potential crowding-out effect. By comparing these two periods, significant changes in enforcement patterns are analysed.

At first, a basic visual inspection of the data is made. It should be mentioned here that this constitutes a snapshot of the situation at the time of writing this thesis. At a later point in time, the results will be different, as the time frame before the introduction of the leniency program does not change any more, while the time frame afterwards becomes larger and larger, which lies in the nature of things.

The Excel file *new\_cartel\_data\_regression\_complete\_unique.xlsx* is used for the visual data analysis. The data are supplemented with the values for leniency regulation, whereby the decision date is taken as the basis for the entry of a Boolean value; from 1996 the value is a 1, before that, a 0. The data frame, which is also supplemented with the cartel duration, is saved in a new Excel file *analysis-report-route-duration.xlsx*, which serves as the basis for the subsequent visual data inspection.

To define the level of non-leniency enforcement, the value under *report\_route* is looked at. It includes the following categorical values: 1 = notification, 2 = complaint, 3 = Commission's own initiative, and 4 = leniency application. Some of the values also appear in combination, as they had been set like this by previous researchers. A pivot table compares the report routes before and after the introduction of the leniency program. Figure 17 shows the total number of cases for both periods graphically.



**Figure 17:** Total number of cases per report route comparison (own illustration)

The bars in blue and violet colours represent cases with non-leniency enforcement, while the yellow bar depicts cases where the procedure was initiated due to a leniency application. Upon initial observation, there doesn't appear to be a significant difference in cases initiated by the Commission's own initiative. However, a more pronounced difference is visible in the other categories. To better illustrate the overall picture, percentages are calculated.

Given that the number of cases in the pre-intervention period differs from the number of cases in the post-intervention period, a pie chart offers a more effective visual representation. In Figure 18, the report routes for both periods are shown as percentages.



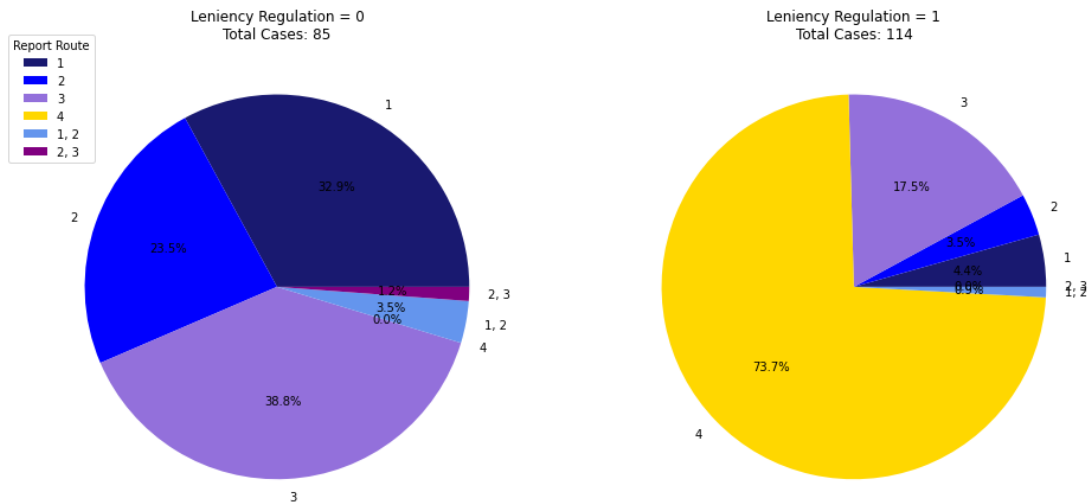


Figure 18: Report routes in percentage for both periods (own illustration)

In order to contribute to answering the research question, the influence of the leniency program on the level of non-lenieny enforcement is analysed in the following, again using the same panel regression model as in Section 4.4.1 General Analysis, based on the data sourced from the Excel file *new\_expanded\_cartel\_data\_regression.xlsx*.

Thus, the null hypothesis is stated as:

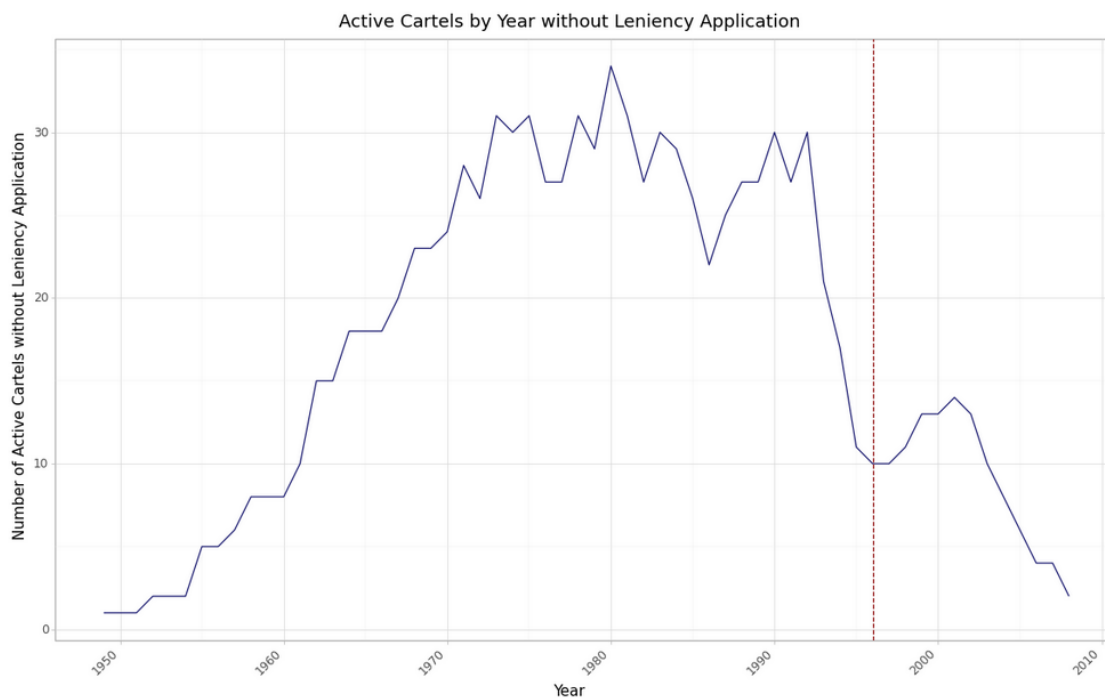
*The introduction of the leniency program in 1996 has not significantly affected the detected cartel activity within the European Union for non-lenieny enforcement cases.*

$H_0: \beta_1 = 0$  in the panel regression model  $cartel_t = \alpha + \beta_1 leniency\_regulation + \beta_2 year + u_t$ , where  $leniency\_application = 0$

And the alternative hypothesis accordingly is:

$H_0: \beta_1 \neq 0$  in the panel regression model  $cartel_t = \alpha + \beta_1 leniency\_regulation + \beta_2 year + u_t$ , where  $leniency\_application = 0$

Looking at the line plot of non-lenieny cartel cases, as depicted in Figure 19, a drop that coincides with the introduction of leniency in 1996 can be observed.



**Figure 19:** Cartel activity discovered by non-lenieny enforcement (own illustration)

After having conducted a first visual exploration of the data, the panel regression analysis modelled by the same equation as in Section 4.4.1 General Analysis is run, however, this time only cartel cases in which the procedure was initiated through non-lenieny enforcement are considered. This is done by applying a filter on the column *leniency\_application*, whereby only data is included in the subset that has the value of 0 in that specific column, because no application for leniency being made to instigate the investigation of the EC automatically means that the procedure was taken up through non-lenieny enforcement. The output of the panel regression (again using the first model) conducted on the accordingly filtered data is displayed in Figure 20.

```

PanelOLS Estimation Summary
=====
Dep. Variable:          cartel    R-squared:              0.0400
Estimator:              PanelOLS  R-squared (Between):    -0.5888
No. Observations:      8740     R-squared (Within):     0.0400
Date:                   Tue, May 16 2023  R-squared (Overall):    -0.0806
Time:                   12:37:04    Log-likelihood          -1926.7
Cov. Estimator:        Unadjusted

                          F-statistic:              359.44
Entities:              115         P-value                 0.0000
Avg Obs:               76.000     Distribution:            F(1,8624)
Min Obs:               76.000
Max Obs:               76.000     F-statistic (robust):   359.44
                          P-value                 0.0000
Time periods:          76         Distribution:            F(1,8624)
Avg Obs:               115.00
Min Obs:               115.00
Max Obs:               115.00

Parameter Estimates
=====
              Parameter  Std. Err.    T-stat    P-value    Lower CI    Upper CI
-----
leniency_regulation  -0.1277    0.0067    -18.959    0.0000    -0.1409    -0.1145
=====

F-test for Poolability: 7.2659
P-value: 0.0000
Distribution: F(114,8624)

Included effects: Entity

```

**Figure 20:** PanelOLS Estimation Summary for filtered data only including cartel activity discovered by non-leniency enforcement (own illustration)

The output suggests a significant negative relationship between the variable *leniency\_regulation* and the dependent variable *cartel*, looking at both the p-value and the t-statistic. With a coefficient of -0.1277, this hints towards the introduction of the leniency regulation being associated with a decrease in the detected cartel activity through non-leniency enforcement. The p-value for the *leniency\_regulation* coefficient is given out as 0.00e+00, literally zero, which calls for further investigation.

Thus, the assumptions for time series are tested again in order to evaluate the model's validity. The tests are conducted in the data analysis script *3d\_Analysis\_2\_panel\_regression\_non-leniency.ipynb* which can be accessed on GitHub and are not elaborated in detail here again, as this has already been done in Section 4.4.1 General Analysis, and the outputs can be retrieved from the data analysis script 3d.

Again, in order to account for serial correlation, the model where the lagged values of the dependent variable *cartel* are included needs to be employed:

$$cartel_t = \alpha + \beta_1 leniency\_regulation + \beta_2 year + \beta_3 cartel\_lag_{t-1} + u_t$$

And, as has been the case before, to counteract the detected heteroskedasticity, the 'robust' option is used for the *model.fit()* method. The output of the second model is as shown in Figure 21:

PanelOLS Estimation Summary						
Dep. Variable:	cartel	R-squared:	0.7443			
Estimator:	PanelOLS	R-squared (Between):	0.9622			
No. Observations:	8625	R-squared (Within):	0.7443			
Date:	Tue, May 23 2023	R-squared (Overall):	0.7866			
Time:	09:47:01	Log-likelihood	3760.0			
Cov. Estimator:	Unadjusted	F-statistic:	1.238e+04			
Entities:	115	P-value	0.0000			
Avg Obs:	75.000	Distribution:	F(2,8508)			
Min Obs:	75.000	F-statistic (robust):	1.238e+04			
Max Obs:	75.000	P-value	0.0000			
Time periods:	75	Distribution:	F(2,8508)			
Avg Obs:	115.00					
Min Obs:	115.00					
Max Obs:	115.00					
Parameter Estimates						
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
cartel_lag	0.8547	0.0056	152.85	0.0000	0.8437	0.8656
leniency_regulation	-0.0237	0.0036	-6.6319	0.0000	-0.0307	-0.0167
F-test for Poolability: 0.5265						
P-value: 1.0000						
Distribution: F(114,8508)						
Included effects: Entity						

**Figure 21:** PanelOLS Estimation Summary for non-lenieny enforcement where serial correlation and heteroskedasticity have been accounted for (own illustration)

After running the new model, a clear improvement in explanatory power is evident when comparing the R-squared (Within) values. The R-squared (Within) value has leaped from a mere 0.0400 in the first model to a much more robust 0.7443 in the second model. This suggests that the revised model is significantly better at explaining the variation within the data than its predecessor. Moreover, the overall R-squared value has increased dramatically, suggesting that the model can now account for approximately 78.66 % of the variance in cartel behaviour, a sizable improvement from the first model's -0.0806.

Turning to the individual coefficients, the  $\beta_1$  coefficient for *leniency\_regulation* has altered from -0.1277 in the first model to -0.0237 in the second model. The p-value remains extremely small, even smaller than before, which continues to signal a high level of statistical significance for this variable. The t-statistic for *leniency\_regulation* has decreased

from -18.959 to -7.5331, indicating a slightly reduced yet still highly significant impact of the leniency program on detected cartel activity.

Lastly, the F-statistic has seen a massive increase from 359.44 in the first model to 1.238e+04 in the second model, reinforcing the overall statistical significance of the new model's predictors. The p-value associated with this F-statistic remains at zero, confirming that at least one of the predictors in the model has a statistically significant relationship with the dependent variable *cartel*. Despite the robust standard errors applied in the second model, the predictors remain highly significant, further solidifying the enhanced explanatory power and reliability of the second model.

#### 4.4.3 Non-lenieny Enforcement before 1996 versus Leniency after 1996

Despite the coexistence of both leniency and non-lenieny enforcement cases since the introduction of the leniency program, the comparison of non-lenieny enforcement before 1996 and leniency enforcement after 1996 is still insightful. This part of the analysis aims to assess whether the leniency program, which represents only a part of cartel activity after 1996, has been able to induce the same or even a higher level of detected cartel activity as the full non-lenieny enforcement regime present before 1996.

The null hypothesis tested to assess that question is as follows:

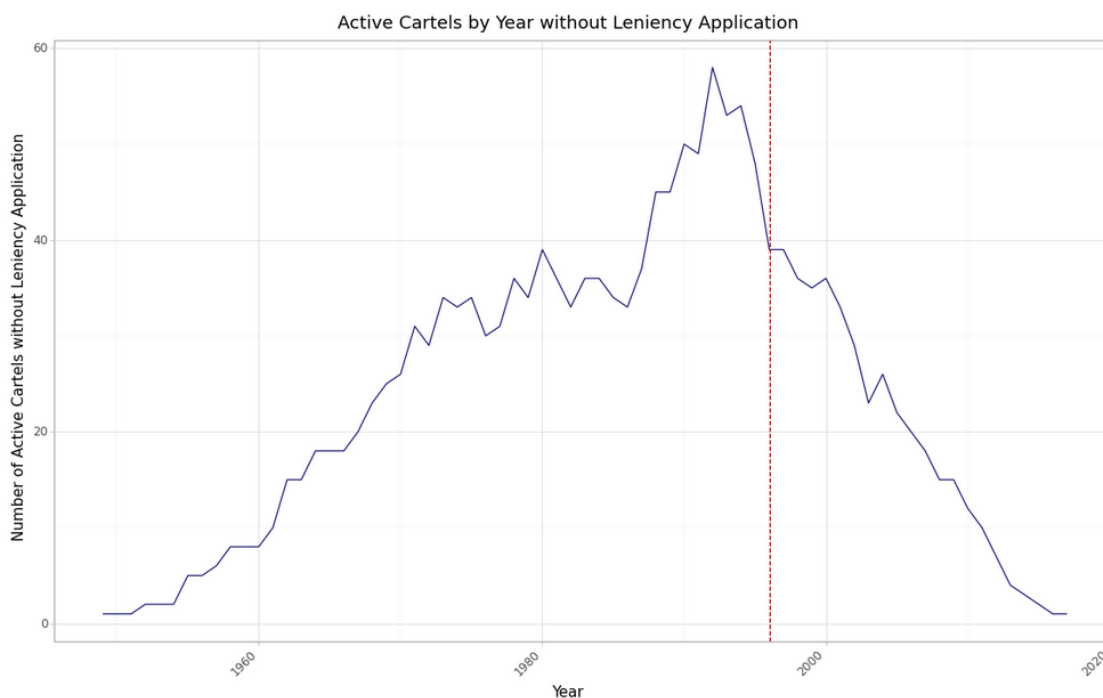
*The level of detected cartel activity where no leniency applications were made before 1996 is the same as the level of detected cartel activity where leniency applications were made after 1996 in the European Union.*

$H_0: \beta_1 = 0$  in the panel regression model  $cartel_t = \alpha + \beta_1 leniency\_regulation + \beta_2 year + u_t$ , where data is divided into pre-1996 (non-lenieny) and post-1996 (leniency) subsets.

And the alternative hypothesis thus is inferred as:

$H_a: \beta_1 \neq 0$  in the panel regression model  $cartel_t = \alpha + \beta_1 leniency\_regulation + \beta_2 year + u_t$ , where data is divided into pre-1996 (non-lenieny) and post-1996 (leniency) subsets.

At first, a basic visual inspection is conducted on this data subset, in which non-lenieny cases before 1996 and only leniency cases from 1996 onwards are included, which is depicted in Figure 22 below.



**Figure 22:** Cartel activity discovered by non-lenient enforcement before 1996 vs. lenient afterwards  
(own illustration)

Similarly to before, the detected cartel activity increases at first (the data before 1996 is the same data as in Sections 4.4.1 General Analysis and 4.4.2 Non-lenient Enforcement) and then drops sharply after 1996, with only the cartels detected through a leniency application being plotted after the red line (these are the cases where the cartel was active after 1996 and where *leniency\_application* = 1).

To test the null hypothesis, the first panel regression model as defined in Section 4.4.1 General Analysis is run on the two combined subsets including non-lenient enforcement cases before 1996 and leniency cases from 1996 onwards. Since there was no leniency program and with that no possibility to apply for leniency before 1996, the first subset includes all detected cartel activity before the introduction of the leniency program, the second subset only includes part of it, as outlined previously. Running the first panel regression model, the results shown in Figure 23 are given back.

PanelOLS Estimation Summary						
Dep. Variable:	cartel	R-squared:	0.0131			
Estimator:	PanelOLS	R-squared (Between):	0.0851			
No. Observations:	11904	R-squared (Within):	0.0131			
Date:	Tue, May 16 2023	R-squared (Overall):	0.0387			
Time:	15:21:13	Log-likelihood	-3353.6			
Cov. Estimator:	Unadjusted	F-statistic:	155.76			
Entities:	199	P-value	0.0000			
Avg Obs:	59.819	Distribution:	F(1,11704)			
Min Obs:	48.000	F-statistic (robust):	155.76			
Max Obs:	76.000	P-value	0.0000			
Time periods:	76	Distribution:	F(1,11704)			
Avg Obs:	156.63					
Min Obs:	84.000					
Max Obs:	199.00					
Parameter Estimates						
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
leniency_regulation	0.1047	0.0084	12.480	0.0000	0.0883	0.1212
F-test for Poolability: 8.9273						
P-value: 0.0000						
Distribution: F(198,11704)						
Included effects: Entity						

**Figure 23:** PanelOLS Estimation Summary on non-lenieny cases before 1996 vs. leniency cases from 1996 onwards  
(own illustration)

The results of running this model seem counterintuitive at first: The *leniency\_regulation* variable has a positive coefficient of 0.1047, suggesting that leniency enforcement has a positive effect on cartel activity. While the positive coefficient may indicate that the leniency program is associated with an increase in detected cartel activity, it does not necessarily mean that it causes an increase in all cartel activity though. However, the p-value for *leniency\_regulation* coefficient is 0.00e+00 and with that warrants further investigations regarding its validity.

To establish the validity of the results, the six time series assumptions are tested in the data analysis script *3e\_Analysis\_3\_panel\_regression\_non-lenieny-vs-lenieny.ipynb*.

After having included lagged values of the dependent variable to counteract the low Durbin-Watson statistic and to thus deal with the detected serial correlation, and having added the 'robust' parameter to the *model.fit()* method to account for heteroskedasticity, the new panel regression model gives out an entirely different coefficient and p-value regarding *leniency\_regulation*, as can be seen in Figure 24.

PanelOLS Estimation Summary						
Dep. Variable:	cartel	R-squared:	0.7268			
Estimator:	PanelOLS	R-squared (Between):	0.9762			
No. Observations:	11705	R-squared (Within):	0.7268			
Date:	Wed, May 17 2023	R-squared (Overall):	0.7888			
Time:	19:42:23	Log-likelihood	4156.4			
Cov. Estimator:	Robust	F-statistic:	1.531e+04			
Entities:	199	P-value	0.0000			
Avg Obs:	58.819	Distribution:	F(2,11504)			
Min Obs:	47.000	F-statistic (robust):	4968.6			
Max Obs:	75.000	P-value	0.0000			
Time periods:	75	Distribution:	F(2,11504)			
Avg Obs:	156.07					
Min Obs:	84.000					
Max Obs:	199.00					
Parameter Estimates						
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
cartel_lag	0.8559	0.0087	97.885	0.0000	0.8388	0.8730
leniency_regulation	-0.0066	0.0045	-1.4686	0.1420	-0.0155	0.0022
F-test for Poolability: 0.5847						
P-value: 1.0000						
Distribution: F(198,11504)						
Included effects: Entity						

**Figure 24:** PanelOLS Estimation Summary for non-lenieny enforcement before 1996 vs. leniency afterwards where serial correlation and heteroskedasticity have been accounted for (own illustration)

Upon executing the new model, the within R-squared value experiences a significant jump from 0.0131 in the first model to 0.7268 in the second model. This indicates that the updated model demonstrates a significantly enhanced ability to account for variation within the data compared to the initial model. In addition, the overall R-squared value in the second model suggests that the model can elucidate approximately 78.88 % of the variation in the dependent variable *cartel*, marking a substantial advancement from the original model's 0.0387.

Examining the coefficient for *leniency\_regulation*,  $\beta_1$  has transitioned from 0.1047 in the first model to -0.0066 in the second model. Interestingly, the p-value for this variable has risen to 0.1420, indicating that *leniency\_regulation* no longer possesses statistical significance in the second model. The t-statistic for *leniency\_regulation* has decreased from 12.480 to -1.4686, suggesting a decrease in the influence of *leniency\_regulation* on detected cartel activity, to an extent where it is not statistically significant anymore.

Lastly, the F-statistic has seen a remarkable rise from 155.76 in the first model to 1.531e+04 in the second model, which reaffirms the overall statistical significance of the



predictors in the updated model. The p-value linked with this F-statistic maintains its position at zero, highlighting that at least one of the predictors in the model holds a statistically significant relationship with the dependent variable *cartel*. However, this probably refers to the newly added variable *cartel\_lag*, which remains highly significant despite the implementation of robust standard errors, and which for this particular analysis is of little importance.

#### 4.4.4 Cartel Duration

Finally, the aspect of cartel duration is analysed. The aim is to investigate the impact of the leniency program on the duration of cartels. Since this consists merely of a comparison of the duration of cartels before and after the introduction of the leniency program, there is no need to control for other time-varying factors, which means in this case, instead of a panel regression analysis, another regression analysis should be conducted (Wooldridge, 2012, p. 68 et seq.). In addition to the leniency regulation, the height of the imposed fines per entity are also considered, thus insinuating the use of a multiple linear regression model, following the below equation

$$cartel\_duration = \alpha + \beta_1 end\_leniency\_regulation + \beta_2 fine + u$$

where *cartel\_duration* is the dependent variable, a metric variable representing the duration of detected cartels.  $\alpha$  is the intercept,  $\beta_1 end\_leniency\_regulation$  is an independent dummy variable with values 1 indicating the cartel ended after the introduction of leniency or 0 indicating the cartel ended before the introduction of leniency and thus represents a categorical, nominal variable.  $\beta_2 fine$  is the second independent variable, a ratio variable, which stands for the average amount of fines imposed per entity.  $u$  denotes the error term. The coefficients  $\beta_1$  and  $\beta_2$  correspond to the effects of the leniency regulation and the fines on the duration of detected cartels, respectively.

The following null hypothesis is tested:

*The introduction of the leniency program has no effect on the duration of detected cartels.*

$H_0: \beta_1 = 0$  in the multiple linear regression model  $cartel\_duration = \alpha + \beta_1 end\_leniency\_regulation + \beta_2 fines + u$

Accordingly, the alternative hypothesis is:

$$H_a: \beta_1 \neq 0 \text{ in the multiple linear regression model } \textit{cartel\_duration} = \alpha + \beta_1 \textit{end\_leniency\_regulation} + \beta_2 \textit{fines} + u$$

To do this, the column *cartel\_duration* was added, the values in which were deducted from the values in the columns *cartel\_start* and *cartel\_end* taken from the previously created Excel file *new\_cartel\_data\_regression\_complete\_unique.xlsx*. The resultant data was then saved in a new Excel file called *new-analysis-cartel-duration.xlsx*. At this point it must be mentioned that only the year values were selected due to some existing gaps in the month and day data collected prior to 2010. Accordingly, the duration is only accurate to the year, as the two values for the start and end dates only reflect the yearly figures. Expending considerable time and effort to manually retrieve this missing data would not have been justified given the expected return. Accordingly, this is a heuristic approach, and the following analysis must be interpreted with caution.

Running the multiple linear regression model, modelled by the equation above, the output as depicted in Figure 25 is displayed.

OLS Regression Results						
Dep. Variable:	cartel_duration		R-squared:	0.010		
Model:	OLS		Adj. R-squared:	0.000		
Method:	Least Squares		F-statistic:	1.016		
Date:	Sun, 14 May 2023		Prob (F-statistic):	0.364		
Time:	14:39:15		Log-Likelihood:	-666.80		
No. Observations:	199		AIC:	1340.		
Df Residuals:	196		BIC:	1349.		
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
constant	7.5344	0.707	10.658	0.000	6.140	8.929
end_leniency_regulation	0.5738	1.039	0.552	0.582	-1.476	2.624
fines	0.5574	0.520	1.073	0.285	-0.467	1.582
Omnibus:	75.602	Durbin-Watson:	1.746			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	188.259			
Skew:	1.715	Prob(JB):	1.32e-41			
Kurtosis:	6.307	Cond. No.	2.78			
Notes:						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						

**Figure 25:** Multiple linear regression regarding impact of leniency on cartel duration (own illustration)

The R-squared value for the model is 0.010, suggesting that the model explains approximately 1 % of the variance in the dependent variable *cartel\_duration*. Given this

---

relatively low R-squared value, the model might not be very effective at accounting for the variation in *cartel\_duration*.

The coefficient for the explanatory variable *end\_leniency\_regulation* is 0.5738, with a standard error of 1.039. The t-statistic for this variable is 0.552, and the corresponding p-value is 0.582. Given the high p-value and the corresponding low t-statistic, this means that *end\_leniency\_regulation* is not statistically significant (as 0.5738 is  $> 0.05$ ) in explaining the duration of cartels. In other words, according to this multiple linear regression model, the estimated effect of the leniency program is an increase in the duration of cartels by about 0.5738 years, but this estimate is not statistically significant, suggesting that this variance might be due to chance. Therefore, there is no strong evidence to conclude that the introduction of the leniency program has had an effect on the duration of cartels.

The coefficient for *finer* is 0.5574, with a standard error of 0.520. The t-statistic for this predictor is 1.073, and the p-value is 0.285. This suggests that *finer*, like *end\_leniency\_regulation*, is not statistically significant in predicting the dependent variable *cartel\_duration*. The F-statistic for the model is 1.016, with an associated p-value of 0.364. This high p-value suggests that there is no sufficient evidence to conclude that at least one of the predictors is statistically significant.

For multiple linear regression models, the following assumptions need to be tested for: 1) linearity in parameters, 2) random sampling, 3) no perfect collinearity, 4) zero conditional mean and 5) homoskedasticity (Wooldridge, 2012, p. 83 et seq.). Those assumptions are tested in the data analysis script *3f\_Analysis\_4\_linear\_regression\_duration.ipynb*, and are not outlined in detail here again, since the tests are either discussed in Section 4.4.1 General Analysis, or more details can be found directly in the data analysis script 3f. Having conducted appropriate tests regarding the five assumptions, no violations are found.

However, there might be selection bias regarding the sample; after all, the data only encompasses cartels that have been detected either through non-lenieny enforcement or leniency applications, which might not be representative of all cartels. Considering the problem of the non-observable cartel population, it is logical to assume that there are many entities in the population that are not considered. This potential selection bias is a limitation of the analysis and could affect the validity of the conclusions.

## 5 RESULTS

### 5.1 Influence of Leniency on Cartel Activity

#### 5.1.1 General Analysis

Looking at the overall cartel activity to gauge the answer to the primary research question “How did the introduction of the leniency program of the European Commission in 1996 affect the cartel activity in the European Union?”, the following null hypothesis has been tested:

$$H_0: \beta_1 = 0 \text{ in the panel regression model } \text{cartel}_t = \alpha + \beta_1 \text{leniency\_regulation} \\ + \beta_2 \text{year} + u_t$$

Running this panel regression model to assess the impact the introduction of the EC’s leniency program had on the detected cartel activity,  $\beta_1$  has been computed as -0.0296, suggesting that an increase in the variable *leniency\_introduction* is related to a decrease in detected cartel activity. The p-value regarding the coefficient for *leniency\_regulation* was computed as 1.37e-08, essentially zero. However, during the testing of the time series assumptions, to counteract the violated assumption of no serial correlation and to account for heteroskedasticity, a new panel regression model including lagged variables of the dependent variable *cartel* was run and robust standard errors were used. Thus, looking at the same question but using the new model, the new null hypothesis  $H_0$  would be as follows:

$$H_0: \beta_1 = 0 \text{ in the panel regression model } \text{cartel}_t = \alpha + \beta_1 \text{leniency\_regulation} \\ + \beta_2 \text{year} + \text{cartel\_lag}_{t-1} + u_t$$

The new coefficient of *leniency\_regulation* with a value of -0.0163 is smaller than in the original model, as is the new p-value of 1.02e-09, which is highly significant. This indicates that the null hypothesis can be rejected in favour of the alternative hypothesis

$$H_a: \beta_1 \neq 0 \text{ in the panel regression model } \text{cartel}_t = \alpha + \beta_1 \text{leniency\_regulation} \\ + \beta_2 \text{year} + \text{cartel\_lag}_{t-1} + u_t$$

and thus, it can be concluded that the introduction of leniency had a highly significant impact on the overall detected cartel activity. The coefficient  $\beta_1$  indicates that a unit increase in *leniency\_regulation* is associated with a decrease in the overall detected cartel activity by 0.0163 units, on average. When this value is applied to 199 entities, the reduction in detected activity is approximately 3.24 cartels per year (0.0163 \* 199). This

suggests that the introduction of the leniency program is associated with approximately three fewer detected cartels per year across all entities, all other factors held constant. However, it is important to interpret this with caution as this decrease is quite small and other variables could also be influencing this outcome.

In essence, these results suggest that the leniency program introduced by the EC in 1996 is associated with a significant, albeit small, reduction in detected cartel activity. This underlines the relevance and effectiveness of the leniency program in reducing cartel activity in the EU, even if the magnitude of the effect is rather modest.

### 5.1.2 Non-leniency Enforcement

To address the sub-question posed in Section 1.4 Research Question regarding the comparison of non-leniency enforcement levels before and after the introduction of the leniency program “*How does the level of non-leniency enforcement before the introduction of the leniency program compare to the level of non-leniency enforcement after its introduction?*”, the following null hypothesis was tested:

$$H_0: \beta_1 = 0 \text{ in the panel regression model } \text{cartel}_t = \alpha + \beta_1 \text{leniency\_regulation} \\ + \beta_2 \text{year} + u_t, \text{ where } \text{leniency\_application} = 0$$

From the regression analysis, the coefficient for leniency regulation,  $\beta_1$ , has been estimated at -0.1277 with a zero p-value. This value suggests that the leniency program's introduction has negatively impacted overall detected cartel activity through non-leniency enforcement within the EU.

However, upon adjusting for serial correlation and heteroskedasticity, the coefficient for leniency regulation changed to -0.0237, although it remained statistically significant. This value still suggests a decrease, albeit smaller, in detected cartel activity through non-leniency enforcement following the introduction of the leniency program. Considering that the entities looked at with regard to non-leniency enforcement encompass a total of 115 cartel cases, this would mean that the number of cartels detected through non-leniency enforcement is expected to decrease by about 2.73 cartels per year ( $0.0237 * 115$ ).

As for the new alternative hypothesis

$$H_a: \beta_1 \neq 0 \text{ in the panel regression model } \text{cartel}_t = \alpha + \beta_1 \text{leniency\_regulation} \\ + \beta_2 \text{year} + \beta_3 \text{cartel\_lag}_{t-1} + u_t, \text{ where } \text{leniency\_application} = 0$$

the results provide strong evidence to support it. Therefore, the level of non-leniency enforcement, as measured by detected cartel activity, appears to have decreased after the introduction of the leniency program. This suggests that the introduction of the leniency program might have shifted the focus of enforcement from non-leniency measures to leniency measures, resulting in fewer detections through non-leniency enforcement methods.

### 5.1.3 Non-leniency Enforcement before 1996 versus Leniency after 1996

This section deals with the sub-question “*How does the level of non-leniency enforcement before the introduction of the leniency program compare to the level of leniency enforcement after its introduction?*”. To answer whether the leniency program itself has been able to induce the same or even a higher level of detected cartel activity as the full non-leniency enforcement regime before 1996, the following null hypothesis has been tested:

$$H_0: \beta_1 = 0 \text{ in the panel regression model } \text{cartel}_i = \alpha + \beta_1 \text{leniency\_regulation} + \beta_2 \text{year} + u_i, \text{ where data is divided into pre-1996 (non-leniency) and post-1996 (leniency) subsets.}$$

Initially, the results from the panel regression model suggested that when the leniency program is in effect, the probability of detecting a cartel increases by approximately 0.1047 (or 10.47 %). However, when adjusting for detected serial correlation and heteroskedasticity, the coefficient of *leniency\_regulation* significantly changes to -0.0066, and its corresponding p-value to 0.1420, losing statistical significance.

This finding implies that, when accounting for past cartel activity through the lagged values and using robust standard errors, the leniency program does not significantly alter the detection of cartels compared to the level of non-leniency enforcement before 1996. Hence, the null hypothesis cannot be rejected. Consequently, it can be tentatively inferred that the detection level of cartel activity under the leniency program after 1996 is not significantly different from the level of cartel activity detected through non-leniency enforcement before the introduction of the leniency program.

In essence, the analysis suggests that the introduction of the leniency program did not result in a statistically significant shift in cartel detection rates compared to the period of non-leniency enforcement before 1996. Based on the available data and the model used, the leniency program's implementation did not markedly increase the detection of cartels compared to previous non-leniency enforcement methods. However, this interpretation

should be approached with caution, as it does not necessarily imply that the leniency program was ineffective. Instead, it highlights that in this particular analysis, the leniency program's measured impact on detection rates was not found to be significant.

#### 5.1.4 Cartel Duration

Regarding the impact of the introduction of leniency on the cartel duration, and the last sub-question from Section 1.4 Research Question, “*How has the average lifetime of cartels changed as a result of the introduction of the leniency program?*”, the following null hypothesis was tested:

$$H_0: \beta_1 = 0 \text{ in the multiple linear regression model } \textit{cartel\_duration} = \alpha + \beta_1 \textit{end\_leniency\_regulation} + \beta_2 \textit{fines} + u$$

The analysis yields a coefficient of 0.5738 for the *end\_leniency\_regulation* variable that indicates the presence of the leniency program. This coefficient suggests that the introduction of the leniency program is associated with an increase in the duration of detected cartels by about 0.5738 years.

However, the p-value for that coefficient with 0.582 is not statistically significant, indicating the null hypothesis cannot be rejected. That is, the observed increase in cartel duration may be due to chance rather than a direct effect of the leniency program. Therefore, according to the multiple linear regression model, there is no strong statistical evidence to suggest that the introduction of the leniency program has had a significant impact on the duration of detected cartels.

Also, looking at the average fines imposed on the cartel members, no statistical significance can be observed. If the fines were statistically significant with a coefficient of 0.5574, it would imply that higher fines lead to longer cartel durations, which is counter-intuitive to what one might expect. This further substantiates the non-significance of the *fines* variable in this context.

Based on the statistical analysis conducted, it can be cautiously concluded that the introduction of the leniency program has not significantly altered the average lifetime of detected cartels. While the model indicates a slight increase in cartel duration in correlation with the introduction of the leniency program, this increase is not statistically significant. Therefore, it cannot confidently be asserted that the leniency program has effectively

decreased the duration of detected cartels. Again, these findings should be interpreted with caution, as other unmeasured factors could also influence cartel duration.

#### 5.1.5 Limitations

Out of the six assumptions for time series regression models, two have been violated. While serial correlation has been accounted for by including lagged variables of the dependent variable, the assumption of homoskedasticity, crucial for the validity of the model, has been addressed by applying the *'robust'* parameter to the *model.fit()* method. However, this merely allows to validate the results of the panel regression model, as the use of robust standard errors provides more reliable coefficient estimates, but heteroskedasticity is still present. The presence of heteroskedasticity suggests that there may be factors not accounted for in the panel regression model, which are influencing the dependent variable in a way that changes over different levels of the independent variables. It also indicates that the error term in the model may not be behaving randomly, which is a fundamental assumption of regression analysis.

With respect to the overall data analysis, it is important to recognise the potential for selection bias in the sample. The dataset exclusively contains cartels that have been detected either through non-leniency enforcement measures or through leniency applications submitted by a cartel member. This subset may not accurately represent the entire spectrum of cartels. Given the inherent difficulty in observing the full population of cartels, it is plausible that a considerable number of entities remain unaccounted for in the dataset. This potential selection bias poses a limitation to the analysis, which could influence the reliability of the findings and conclusions drawn.

Another form of selection bias may result from the language of publication. The analysis conducted in this study is based only on cases published in English. It is noteworthy that more recent cases have been published predominantly in English, thus ensuring their inclusion in the analysis conducted in this thesis. However, some cases, particularly those that occurred at the very beginning of the reporting period, may have been overlooked because of their publication in other languages. In addition, some cases from the period from 1999 onwards were also not included today due to language barriers, as lined out in Section 4.1 Data Acquisition. Therefore, it is important to exercise caution when interpreting the results of this thesis.



## 5.2 Analysis of NLP Techniques for Data Extraction

### 5.2.1 Data Extraction Results

The success of the different NLP techniques applied varied depending on the specific information sought. The results obtained per NLP technique applied for data extraction are shown in Table 8.

**Table 8:** Accuracy rates regarding information extraction using different NLP techniques (own illustration)

Label	Technique employed	Accuracy rate
<b>case_number</b>	Regex	97.87 %
<b>decision_date</b>	Regex	100.00 %
	NER, keyword matching & datefinder	75.53 %
<b>cartel_start</b>	Version 2d_a: Regex + NER, keyword matching	21.28 %
	Version 2d_b: Regex + NER, keyword matching	19.15 %
	Version 2e: Regex + NER, keyword matching	55.32 %
<b>cartel_end</b>	Version 2d_a: Regex + NER, keyword matching	13.83 %
	Version 2d_b: Regex + NER, keyword matching	10.64 %
	Version 2e: Regex + NER, keyword matching	27.66 %
<b>cartel_duration</b>	– <i>(dependent on cartel start date and cartel end date)</i>	
<b>report_route</b>	Regex and keyword matching	73.40 %
<b>route_indicator</b>	– <i>(dependent on report_route)</i>	
<b>leniency</b>	Keyword matching	90.43 %

Although regex had a perfect extraction rate for decision data, such results warrant a closer look. The 100 % success rate was the result of intensive work to refine the regex patterns to perfectly match the specific formats in the unique dataset used. Thus, the effectiveness of regex as a technique is limited by its flexibility, or lack thereof, in handling different formats, especially those that deviate from the standard. It is important to emphasise that the perfect result of regex may not hold in the face of format changes, as possible pattern variations lead to errors and lower accuracy.

Nevertheless, the regex approach proved to be very effective in extracting case numbers and decision dates, with success rates of 97.87 % and 100 %, respectively. The keyword matching technique also performed well in leniency detection, with a success rate of 90.43 %. NER combined with keyword matching and regex showed mixed results, especially in detecting cartel start and end dates. Success rates ranged from 10.64 % to 55.32 %.

Although all the applied methods have their strengths, the most promising results are likely to come from a combination of these NLP techniques. When used together, they can complement each other's weaknesses to provide more comprehensive and accurate data extraction.

### 5.2.2 Limitations

As far as the exploration of different NLP techniques for data extraction are concerned, the primary constraint was time, as the vast potential and complexity of NLP methods necessitates significant time and resources for proper exploration and implementation. Due to the scope of this study, only a fraction of these techniques could be applied, and even then, their usage was limited to the English language.

It is paramount to note that the linguistic restriction poses a significant limitation. The approach employed in this thesis is predominantly suited for English texts, which is the reason why cases published in languages other than English have been excluded from the analysis. Future research in this area should consider extending the NLP techniques to incorporate multiple languages to ensure a more comprehensive and inclusive data set. This would provide a more accurate understanding of the impact of the leniency program of the EC.

### 5.3 Evaluation of Artefacts

In the following, it will be assessed whether the artefact requirements as defined in Section 1.6 Artefact Requirements have been met. This evaluation will take the form of a checklist, and the fulfilment of each requirement will be reported as “met”, “partially met”, or “not met”, with additional comments provided where necessary for further clarification. Although efforts have been made to maintain objectivity in this assessment, it should be noted that certain elements may be subjective, particularly those relating to the clarity and precision of documentation.

#### 5.3.1 Scraping Scripts

The scripts being assessed hereunder are scraping scripts *1a\_scraping\_cases\_until\_1998.ipynb* and *1b\_scraping\_cases\_from\_1999.ipynb*.

**Table 9:** Evaluation of the artefact "scraping scripts" (own illustration)

Domain	Requirement	Evaluation
<b>Goal</b>	There shall be two scripts: 1) cases from 1964 to 1998 and 2) cases from 1999 to today's date.	met
<b>Goal</b>	For every case number, only one PDF file shall be downloaded.	not met <i>Note: Due to amending decisions, for some cases there are several PDF files that are downloaded.</i>
<b>Goal</b>	The PDF file that will be downloaded by the scripts shall be the latest prohibition decision on the case, i.e., where there are more than one PDF files named "Prohibition Decision", the one with the more recent date shall be downloaded.	partially met <i>Note: It is the case for the downloaded PDF files from 1999 onwards. Regarding the first scraping script, the time has not sufficed to conduct an assessment on that matter. Also, as mentioned before, where there are amending decisions present, not only the</i>

		<i>most recent prohibition decision is downloaded.</i>
<b>Goal</b>	The cases from 1964 to 1998 shall be saved in a newly created folder “cases_until_1998” and the cases from 1999 to the current date shall be saved in a newly created folder “cases_from_1999”.	met
<b>Environment</b>	The scripts shall be compatible with the Anaconda environment and should execute successfully in Jupyter Notebook.	met
<b>Environment</b>	The scripts are written in the Python programming language.	met
<b>Structure</b>	The scripts shall include sufficiently clear documentation in English language so the code can be understood and used by other researchers for future studies.	met
<b>Activity</b>	The scripts shall include all the necessary install statements that can be run if the necessary libraries are not yet installed on the end user’s device.	partially met <i>Note: in some instances, some libraries that are usually pre-installed by Python might not be included.</i>
<b>Evolution</b>	The scripts should include error handling capabilities to manage exceptions, such as missing data, inaccessible files, or unsupported file formats.	met

Overall, most of the requirements have been met, although there is room for improvement. Also, the scraping script 1a would need to be analysed in more detail to be able to make accurate statements about it.

### 5.3.2 NLP Scripts

The scripts being assessed hereunder are NLP scripts *2a\_data-extraction-case-number-and-decision-date-regex.ipynb*, *2b\_data-extraction-decision-date-spacy.ipynb*, *2c\_extract\_text\_from\_duration\_of\_infringement.ipynb*, *2d\_a\_data\_extraction\_duration\_spacy\_v1.ipynb*, *2d\_b\_data\_extraction\_duration\_spacy\_v2.ipynb*, *2e\_data\_extraction\_duration\_spacy\_v3.ipynb* and *2f\_data\_extraction\_report\_route\_leniency.ipynb*.

**Table 10:** Evaluation of the artefact "NLP scripts" (own illustration)

Domain	Requirement	Evaluation
<b>Goal</b>	<p>The scripts must be capable of extracting the following information from the previously collected PDF files:</p> <ol style="list-style-type: none"> <li>1. Case number</li> <li>2. Decision date</li> <li>3. Cartel start date and cartel end date, and, accordingly, cartel duration</li> <li>4. Report route and report route indicators</li> <li>5. Whether there was an application for leniency</li> </ol>	<p>partially met</p> <p><i>Note: see detailed evaluation of this requirement in Section 5.2 Analysis of NLP Techniques for Data Extraction.</i></p>
<b>Goal</b>	<p>Different NLP methods shall be tested to read out the specified data, where multiple techniques are employed to extract the same data, they should be implemented in</p>	<p>partially met</p> <p><i>Note: not for every needed information several NLP techniques have been tried, however, in the overall picture,</i></p>

---

	separate Jupyter Notebook scripts for easy comparison of results.	<i>different techniques have been employed.</i>
<b>Goal</b>	The extracted data must be stored in an Excel file, which should include all cases and their corresponding data.	met
<b>Environment</b>	The scripts shall be compatible with the Anaconda environment and should execute successfully in Jupyter Notebook.	met
<b>Environment</b>	The scripts are written in the Python programming language.	met
<b>Structure</b>	The scripts shall include sufficiently clear documentation in English language so the code can be understood and used by other researchers for future studies.	met
<b>Activity</b>	The scripts shall include all the necessary install statements that can be run if the necessary libraries are not yet installed on the end user's device.	partially met <i>Note: in some instances, some libraries that are usually pre-installed by Python might not be included.</i>
<b>Evolution</b>	The scripts should include error handling capabilities to manage exceptions, such as missing data, inaccessible files, or unsupported file formats.	partially met <i>Note: In some instances, error handling is included pretty well, in others, it could definitely be further improved.</i>

---

The applied NLP techniques do not yet achieve the desired results, much could be improved. Also, more NLP techniques should be tried and evaluated in the different use cases in future research. In addition, the error handling aspect should be further optimised.

### 5.3.3 Data Analysis Scripts

The scripts being assessed hereunder are data analysis scripts *3a\_data-preparation-regression-part1.ipynb*, *3b\_data-preparation-regression-part2.ipynb*, *3c\_Analysis\_1\_panel\_regression\_general.ipynb*, *3d\_Analysis\_2\_panel\_regression\_non-leniency.ipynb*, *3e\_Analysis\_3\_panel\_regression\_non-leniency-vs-leniency.ipynb* and *3f\_Analysis\_4\_linear\_regression\_duration.ipynb*.

**Table 11:** Evaluation of the artefact "data analysis scripts" (own illustration)

Domain	Requirement	Evaluation
<b>Goal</b>	The scripts combine both the manually collected data until 2010 that is accessed in the folder "Data-until-2010" from the existing file "Cartels1964-2010.xls" as well as data after 2010, whereas case number and decision date are taken from the output of the NLP scripts and the missing information is added manually.	met
<b>Goal</b>	For every sub-question to the research question, the scripts include the regression model as well as assumption testing.	met
<b>Goal</b>	A separate Jupyter Notebook script shall be created for each sub-question in order to distinguish between the different aspects of the research question.	met
<b>Environment</b>	The scripts shall be compatible with the Anaconda environment and should execute successfully in Jupyter Notebook.	met

---

<b>Environment</b>	The scripts are written in the Python programming language.	met
<b>Structure</b>	The scripts shall include sufficiently clear documentation in English language so the code can be understood and used by other researchers for future studies.	met
<b>Activity</b>	The scripts shall include all the necessary install statements that can be run if the necessary libraries are not yet installed on the end user's device.	partially met <i>Note: in some instances, some libraries that are usually pre-installed by Python might not be included.</i>
<b>Evolution</b>	The scripts should include error handling capabilities to manage exceptions, such as missing data, inaccessible files, or unsupported file formats.	not met

---

Regarding the data analysis scripts, most of the set requirements have been met. However, there is especially room for improvement about error handling. Since relatively few errors occurred during the creation of the data analysis scripts that related to non-existent data, unreadable files or data formats, this aspect was neglected. This should be incorporated if future studies are to be based on the scripts.



---

## 6 DISCUSSION

In the following, a connection is established with the existing literature that has been extensively discussed in Section 3 Related Work. While further research would need to be conducted on the questions where the null hypotheses were not rejected, in the following, the implications that would arise if the null hypotheses were indeed true are discussed.

### 6.1 Effectivity of Leniency Programs

As outlined in Section 3.1 Leniency Programs and Cartel Deterrence, the existing research provides compelling evidence of the success of leniency programs in deterring cartels. However, observations diverge regarding the detected cartel activity prior and subsequent to the leniency program's inception.

For instance, Miller (2009), focusing on the US antitrust program, found an increase in detected cartel activity around the time of the program's introduction. He furthermore described that detected cartel activity then fell below pre-leniency levels (Miller, 2009, p. 761 et seq.). Similar results were presented by Dijkstra and Frisch (2018) after analysing the Dutch leniency program. These authors also observed a decrease in cartels detected after its introduction, although they clearly attributed this to stricter enforcement and thus a stronger deterrence of cartels (Dijkstra & Frisch, 2018, p. 121 et seq.). The research conducted within the framework of this thesis found that most cartels were also detected around the time frame of leniency introduction and the overall level of detected cartel activity fell sharply thereafter and are thus in line with the findings of both Miller (2009) and Dijkstra and Frisch (2018).

At the same time, the findings run contrary to what Nicolau (2015), Hinloopen and Soetevent (2008), and Spagnolo (2004) observed during their research. In all these studies, the authors found that after the introduction of the leniency program, more cartel activity was detected than before (Nicolau, 2015, p. 34 et seq.; Hinloopen & Soetevent, 2008, p. 607 et seq.; Spagnolo, 2004, p. 2). As far as Hinloopen and Soetevent (2008) are concerned, it should be noted that the authors themselves only refer to Spagnolo (2004) in their paper, and it is not clear from the text whether they have analysed the fact again themselves (Hinloopen & Soetevent, 2008, p. 607 et seq.). In the following, their study refers to an experiment that shows that fewer cartels exist when a leniency program is in place, which, should this also be the case in the field, would again support the results of this thesis with

---

regard to the leniency program resulting in fewer detected cartels (Hinloopen & Soetevent, 2008, p. 609 et seq.). Again, it should be noted that only cases published in English have been considered in this thesis and that it is highly likely that judgments concerning the last few years have not yet been published at the time this paper was written.

The results derived from the present thesis indicate a slight but statistically significant decrease in detected cartel activity after the introduction of leniency in 1996. The coefficient on the leniency variable in the employed panel regression model shows an average decrease in detected cartel activity of 0.0163 units for each of the 199 entities. This corresponds to an approximate decrease of about three cartels per year within the sampled data, indicating a subtle but discernible effect of the leniency program. While this modest decrease underscores the role of leniency in suppressing cartel activity, it should be interpreted with caution given the statistical uncertainty, possible unmeasured confounding factors and other limitations of the study.

Furthermore, the possible crowding-out effect regarding the reallocation of resources from non-lenieny enforcement to leniency enforcement must also be examined, which according to Harrington and Chang (2015) can undermine the success of the leniency program. This is further elaborated upon in the following in Section 6.2 Investigation by Cartel Authority through Non-lenieny Enforcement.

In addition to examining the cartel activities detected, it would be crucial to assess how leniency affects the conviction rate in order to better gauge the actual success of leniency. Sauvagnat (2010) assumes an increase, but unfortunately such an investigation could not be considered in the context of this study due to time constraints in data collection. Moreover, the study could not be extended to investigate the differential impact in different industries characterised by weak and strong cartels, so it cannot further contribute to this discourse suggested by Harrington and Chang (2015). Although these aspects are not the focus of this thesis, their importance in assessing the overall effectiveness of leniency programs in deterring cartel activity cannot be underestimated. Therefore, their brief mention in this chapter is considered essential to provide a broader context. A detailed discussion of these factors would add further depth to the findings.

## 6.2 Investigation by Cartel Authority through Non-leniency Enforcement

Harrington and Chang (2015) discuss the possible crowding-out effect of leniency, resulting in resources being diverted from non-leniency enforcement to leniency cases. If this is the case, it could actually increase the cartel rate (Harrington & Chang, 2015).

The authors conclude that leniency reduces cartel activity if the enforcement policy of the cartel authority remains unchanged (Chang & Harrington, 2008). This view is also held by Dijkstra and Frisch (2018), who put enforcement under the microscope after observing declining cartel detections. They found enforcement to be stricter than before and concluded accordingly that the cartel rate actually declined (Dijkstra & Frisch, 2018, p. 121 et seq.). In line with those authors, Motta and Polo (1999) and Brenner (2009) also postulate a connection between the existing resources of the cartel authorities and the effectiveness of the leniency programme.

In light of these existing viewpoints, the results of this thesis offer further insights. Through statistical analysis of the 115 entities representing various cartels that came to light through non-leniency enforcement, this study shows a slight but statistically significant decrease in the detection of cartel activity through non-leniency after the introduction of leniency. The leniency coefficient of -0.0237, while modest, suggests a discernible effect, reflected in an average decrease of about three cartels per year being investigated through non-leniency enforcement.

This result lends some credence to the crowding-out effect hypothesised by Harrington and Chang (2015, 2008), which suggests a shift of resources from non-leniency enforcement to leniency cases following the introduction of the program. The question arises as to whether this reallocation of resources did undermine the overall effectiveness of cartel enforcement or whether it may actually have complemented it.

To provide more context on that, the study looked at whether leniency, which accounts for only a portion of post-1996 cartel activity, was able to produce the same or even higher levels of detected cartel activity as the full enforcement regime without leniency that existed prior to 1996.

The results imply that leniency does not significantly change the cartel detection rate compared to the levels of enforcement without leniency before 1996. The conclusion from these data suggests that the detection rate of cartel activity under leniency after 1996 is not statistically different from the detection rate obtained by non-leniency enforcement

---

before the program was implemented. Together with the observed statistical change in non-leniency enforcement rates, these findings seem to indicate the presence of a crowding-out effect, where resources have potentially been shifted from non-leniency to leniency enforcement.

This verdict calls for a more in-depth examination of the current enforcement policy of the EC, as despite the alleged shift of resources to leniency enforcement, where a strict enforcement policy is applied, the decline in detected cartel activity can still be attributed to a successfully implemented leniency program.

### **6.3 Cartel Duration and Fines**

Finally, the duration of the cartel and the amount of the fine were considered. These two aspects are also important in the context of evaluating the success of a leniency program.

With respect to the lifetime of a cartel, Lefouili and Roux (2012) brought forward the argument that leniency programs can shorten the duration of a cartel by providing an incentive to self-report. In particular, they point to the approach of Amnesty International, in which members of a detected cartel can also benefit from a reduction in fines after the cartel is discovered, provided they self-report and deliver information (Lefouili & Roux, 2012, p. 634 et seq.). This is in line with the view of Harrington and Chang (2015), who shed light on the race between cartel members to be the first to tip off the authorities under leniency programs where only the first whistle-blower is exempt from fines. Nevertheless, both studies highlight the potential of leniency programs to shorten the longevity of cartels.

Contrary to the suggestions made by Lefouili and Roux (2012) and Harrington and Chang (2015), the results of this paper do not significantly support the assertion that the introduction of leniency programs reduces the duration of detected cartels. Rather, a differentiated picture emerges. On the one hand, the data analysis suggests that the introduction of leniency may even be associated with an increase in the duration of detected cartels by about 0.5738 years. On the other hand, this relationship is found to not be statistically robustly evidenced. Thus, this suggests that the introduction of the leniency program did not have a significant impact on the duration of detected cartels at all.

Similar to the issue of cartel duration, the impact of leniency programs on the amount of fines imposed on cartel members is also worthy of examination. For instance, Motta and Polo (2003) demonstrated that leniency programs, which reduce penalties for cartel

members that supply information to antitrust authorities, can increase their efficacy, particularly when the resources of enforcement agencies are constrained. However, they also presented the counterpoint that such programs might inadvertently incentivise the formation of cartels, given the reduced repercussions upon exposure when cooperation is offered (Motta & Polo, 2003, p. 375). Regarding the height of imposed fines per company, Borell et al. (2022) found that they increased significantly after the introduction of the leniency program. Motchenkova (2004) found that when sanctions and prosecution are lax, the introduction of leniency can paradoxically facilitate collusion and prolong the duration of cartels. Only when implemented strictly do higher fines also lead to shortened cartel duration (Motchenkova, 2004, p. 26).

This thesis focused on whether the amount of fines imposed influences the duration of cartels. The present analysis shows no significant statistical relationship between the amount of fines imposed and the duration of the cartels. In particular, a coefficient of 0.5574, which theoretically suggests that higher fines are associated with longer cartel duration, contradicts conventional wisdom. However, the statistically insignificant relationship observed confirms the fine's minimal role in affecting cartel duration in this context.

While this thesis can thus contribute to the current state of research, it is important to reiterate the limitations of the analysis and point out that further research is needed to confirm the findings.

---

## 7 CONCLUSION AND FUTURE WORK

In this thesis, the impact of the EC's leniency program on the cartel rate was analysed and NLP methods for data extraction from EC decisions were examined.

In summary, to answer the research question, there was a slight but statistically significant decrease in detected cartel activity after the introduction of leniency programs in 1996. The study also shows a slight but statistically significant decrease in non-leniency detections after the introduction of leniency programs. Interestingly, detection rates under post-1996 leniency programs were not found to be statistically lower than those of non-leniency enforcement prior to 1996. This suggests a possible shift in enforcement resources from non-leniency methods to those involving leniency. However, leniency does not appear to have a significant effect on the duration of cartels detected, with data suggesting a possible increase in cartel duration, although this could not be clearly demonstrated statistically. The effect of fines on the duration of cartels was also insignificant. Overall, these results provide a nuanced view of the impact of leniency programs on cartel detection and duration and suggest a possible crowding-out effect as resources may have been shifted from non-leniency enforcement to leniency enforcement.

As far as the analysed NLP techniques are concerned, the effectiveness of the different methods applied revealed considerable variance depending on the specific data extracted. Regex showed robust capabilities in extracting case numbers and decision dates, with success rates of 97.87 % and 100 % respectively. Nevertheless, it's crucial to acknowledge that these outcomes relied heavily on thorough adjustments of regex patterns to suit individual document formats, suggesting possible adaptability challenges. Keyword matching stood out in its ability to identify leniency, posting a 90.43 % success rate. The integration of NER, keyword matching, and regex, however, yielded mixed results, particularly in the identification of cartel start dates and cartel end dates. These variations emphasise the multifaceted nature of data extraction and the necessity for careful choice and combination of NLP techniques to align with specific data requirements.

Future research in this area offers numerous opportunities for further exploration. First, it is recommended that the scope of the study be expanded to include cases that were not published in English. This could mitigate selection bias due to language and lead to a more comprehensive understanding of the impact the introduction of the EC's leniency program had on the overall cartel activity. Also, the inherent selection bias of the dataset should be an important consideration. Currently, the dataset includes only detected

---

cartels, which very likely does not represent the entire cartel population. Future studies could explore innovative methods or incorporate additional data sources to address this limitation. Furthermore, careful consideration of the general limitations identified in this study could inform the design and conduct of further research. For example, methods to better address the problem of heteroscedasticity in the regression models could be explored.

The results of the research into NLP techniques employed for data extraction open up an equally large variety of avenues for further exploration. For example, consideration could be given to refining SpaCy's NER model to the dataset used to improve its effectiveness in the specific legal context. Other advanced NLP models could also be evaluated, such as BERT or Transformers. These future efforts could significantly improve the results of the analysis. In addition to this, there would be much more possibilities for data analysis if also the other important key information, which in Section 4.2 Data Extraction with NLP are listed in Table 1, could be extracted using combined NLP techniques. When all these data are available, the impact of the leniency program can be evaluated even better. Specifically, the *formal\_decision*, *party\_name*, and *oecd\_sector* data could provide deeper insights into repeat offenders and the potential differential impact of the leniency program depending on the industry. In addition, it is advisable to extend the analysis of NLP methods to cases prior to 1998. Prohibition decisions for this time period were scraped for in this thesis and are thus available for further research but could not be examined in this study due to time constraints. Also, a linguistic expansion to texts in languages other than English could be achieved either by using multilingual NLP models or by using machine translation techniques to convert non-English texts into English. Such an approach would not only cover a broader range of cases, but also provide a more comprehensive understanding of the EC's leniency program and its broader implications.

## ACKNOWLEDGEMENTS

The writing of this thesis has undoubtedly taken an immense amount of time and fortitude, but every moment and every effort has been immeasurably worthwhile. First and foremost among those who deserve thanks is my supervisor Nicole Bellert. Her willingness to supervise me and her invaluable support when obstacles arose are greatly appreciated.

Gratitude also goes to Merlin Baumann, who carefully reviewed the Python scripts to ensure their compatibility and robustness on other devices than my own (who knew this could actually cause problems!?). His meticulous efforts have successfully fixed many issues, and I sincerely hope that all possible shortcomings have been corrected. Thanks are also due to Isaac Kargar for his timely inspiration and guidance when it came to programming challenges.

Finally, but not least, my heartfelt thanks go to Michelle Wehrli, Céline Bertet and Mirjam Baumann. Their careful review of this thesis has increased both its readability and comprehensibility and their contributions have ensured that reading this master's thesis is as enjoyable and understandable as possible.



## REFERENCES

- Alavi, M., & Carlson, P. (1992). A Review of MIS Research and Disciplinary Development. *Journal of Management Information Systems*, 8(4), 45–62. <https://doi.org/10.1080/07421222.1992.11517938>
- Aubert, C., Rey, P., & Kovacic, W. E. (2006). The impact of leniency and whistle-blowing programs on cartels. *International Journal of Industrial Organization*, 24(6), 1241–1266. <https://doi.org/10.1016/j.ijindorg.2006.04.002>
- Bigoni, M., Fridolfsson, S.-O., Le Coq, C., & Spagnolo, G. (2012). Fines, leniency and rewards in antitrust. *The RAND Journal of Economics*, 43(2), 368–390. <https://doi.org/10.1111/j.1756-2171.2012.00170.x>
- Bird, S., & Klein, E. (2006). *Regular Expressions for Natural Language Processing*.
- Borrell, J.-R., García, C., Jiménez, J. L., & Ordóñez-de-Haro, J. M. (2022). *Cartel destabilization effect of leniency programs*.
- Brenner, S. (2009). An empirical study of the European corporate leniency program. *International Journal of Industrial Organization*, 27(6), 639–645. <https://doi.org/10.1016/j.ijindorg.2009.02.007>
- Broos, S., Gautier, A., Ramos, J. M., & Petit, N. (2016). Analyse statistique des affaires d'ententes dans l'UE (2004-2014): *Revue économique*, Vol. 67(HS1), 79–94. <https://doi.org/10.3917/reco.hs01.0079>
- Chang, M.-H., & Harrington, J. E. (2008). *The impact of a corporate leniency program on antitrust enforcement and cartelization*.
- Chen, J., & Harrington, J. E. (2007). Chapter 3 The Impact of the Corporate Leniency Program on Cartel Formation and the Cartel Price Path. In *Contributions to Economic Analysis* (Vol. 282, pp. 59–80). Elsevier. [https://doi.org/10.1016/S0573-8555\(06\)82003-1](https://doi.org/10.1016/S0573-8555(06)82003-1)

- Chen, Z., Ghosh, S., & Ross, T. W. (2015). Denying leniency to cartel instigators: Costs and benefits. *International Journal of Industrial Organization*, 41, 19–29. <https://doi.org/10.1016/j.ijindorg.2015.04.003>
- Chen, Z., & Rey, P. (2013). On the Design of Leniency Programs. *The Journal of Law and Economics*, 56(4), 917–957. <https://doi.org/10.1086/674011>
- Daoud, J. I. (2017). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series*, 949, 012009. <https://doi.org/10.1088/1742-6596/949/1/012009>
- Dijkstra, P. T., & Frisch, J. (2018). Sanctions and Leniency to Individuals, and its Impact on Cartel Discoveries: Evidence from the Netherlands. *De Economist*, 166(1), 111–134. <https://doi.org/10.1007/s10645-017-9309-4>
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Human- und Sozialwissenschaften* (5th edition). Springer.
- Emons, W. (2018). *The Effectiveness of Leniency Programs when Firms choose the Degree of Collusion* [Application/pdf]. <https://doi.org/10.7892/BORIS.145867>
- European Commission. (1996). Commission Notice on the non-imposition or reduction of fines in cartel cases. *Official Journal of the European Union*. [https://ec.europa.eu/competition/antitrust/legislation/96c207\\_en.html](https://ec.europa.eu/competition/antitrust/legislation/96c207_en.html)
- European Commission. (2002). Commission Notice on Immunity from fines and reduction of fines in cartel cases. *Official Journal of the European Union*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52002XC0219%2802%29&qid=168465551214>
- 5
- European Commission. (2006). Commission Notice on Immunity from fines and reduction of fines in cartel cases. *Official Journal of the European Union*. <https://eur->

- lex.europa.eu/legal-con-  
tent/EN/ALL/?uri=CELEX%3A52006XC1208%2804%29
- European Commission. (2023). Leniency. *Competition Policy*. [https://competition-policy.ec.europa.eu/cartels/leniency\\_en](https://competition-policy.ec.europa.eu/cartels/leniency_en)
- Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2023). *Introduction to Econometrics with R*.
- Harrington, J. E. (2013). Corporate Leniency Programs when Firms have Private Information: The Push of Prosecution and the Pull of Pre-emption: The Push of Prosecution and the Pull of Pre-emption. *The Journal of Industrial Economics*, 61(1), 1–27. <https://doi.org/10.1111/joie.12014>
- Harrington, J. E., & Chang, M.-H. (2009). Modeling the Birth and Death of Cartels with an Application to Evaluating Competition Policy. *Journal of the European Economic Association*, 7(6), 1400–1435. <https://doi.org/10.1162/JEEA.2009.7.6.1400>
- Harrington, J. E., & Chang, M.-H. (2012). *Endogenous Antitrust Enforcement in the Presence of a Corporate Leniency Program*.
- Harrington, J. E., & Chang, M.-H. (2015). When Can We Expect a Corporate Leniency Program to Result in Fewer Cartels? *The Journal of Law and Economics*, 58(2), 417–449. <https://doi.org/10.1086/684041>
- Hevner, A. R. (2007). *A Three Cycle View of Design Science Research*. 19.
- Hinloopen, J., & Soetevent, A. R. (2008). Laboratory evidence on the effectiveness of corporate leniency programs. *The RAND Journal of Economics*, 39(2), 607–616. <https://doi.org/10.1111/j.0741-6261.2008.00030.x>
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). <https://otexts.com/fpp2/>

- Jochem, A., Parrotta, P., & Valletta, G. (2020). The impact of the 2002 reform of the EU leniency program on cartel outcomes. *International Journal of Industrial Organization*, 71, 102640. <https://doi.org/10.1016/j.ijindorg.2020.102640>
- Kanapala, A., Pal, S., & Pamula, R. (2019). Text summarization from legal documents: A survey. *Artificial Intelligence Review*, 51(3), 371–402. <https://doi.org/10.1007/s10462-017-9566-2>
- Lefouili, Y., & Roux, C. (2012). Leniency programs for multimarket firms: The effect of Amnesty Plus on cartel formation. *International Journal of Industrial Organization*, 30(6), 624–640. <https://doi.org/10.1016/j.ijindorg.2012.04.004>
- Menzli, A. (2023, April 25). Tokenization in NLP: Types, Challenges, Examples, Tools. *Neptune.Ai*. <https://neptune.ai/blog/tokenization-in-nlp>
- Miller, N. H. (2009). Strategic Leniency and Cartel Enforcement. *American Economic Review*, 99(3), 750–768. <https://doi.org/10.1257/aer.99.3.750>
- Motchenkova, E. (2004). Effects of Leniency Programs on Cartel Stability. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.617224>
- Motta, M., Fabra, U. P., & Polo, M. (1999). *Leniency Programs and Cartel Prosecution*.
- Motta, M., & Polo, M. (2003). Leniency programs and cartel prosecution. *Int. J. Ind. Organ.*
- Nicolau, J. S. (2015). *The effectiveness of Leniency Programs on Cartel Prosecution*.
- Paul, R. K. (2006). *MULTICOLLINEARITY: CAUSES, EFFECTS AND*.
- Peffer, K., Tuunanen, T., Gengler, C. E., Rossi, M., & Hui, W. (2006). *THE DESIGN SCIENCE RESEARCH PROCESS: A MODEL FOR PRODUCING AND PRESENTING INFORMATION SYSTEMS RESEARCH*. 83–106.
- Pinha, L. C., Braga, M. J., & Oliveira, G. A. S. (2016). *A efetividade dos programas de leniência e o contexto brasileiro*. 4.

- Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). *ARTIFACT EVALUATION IN INFORMATION SYSTEMS DESIGN-SCIENCE RESEARCH – A HOLISTIC VIEW*.
- Sauvagnat, J. (2010). Prosecution and Leniency Programs: A Fool’s Game. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.1826324>
- Sauvagnat, J. (2014). Are leniency programs too generous? *Economics Letters*, 123(3), 323–326. <https://doi.org/10.1016/j.econlet.2014.03.015>
- Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*, 8(2), 39–42. <https://doi.org/10.12691/ajams-8-2-1>
- Shuangyuan, S. W. (2021). Testing the Assumptions of Linear Regression. *Towardsdatascience.Com*. <https://towardsdatascience.com/testing-the-assumptions-of-linear-regression-f38857abc08a>
- SpaCy. (2023a). SpaCy, Facts & Figures [Documentation]. *Facts & Figures*. <https://v2.spacy.io/usage/facts-figures>
- SpaCy. (2023b). SpaCy, Linguistic Features [Documentation]. *Linguistic Features*. <https://spacy.io/usage/linguistic-features>
- Spagnolo, G. (2004). *Divide et Impera: Optimal Leniency Programs*.
- Stedman, R. C., & Beckley, T. M. (2007). “If We Knew What it Was We Were Doing, it Would Not be Called Research, Would it?” *Society & Natural Resources*, 20(10), 939–943. <https://doi.org/10.1080/08941920701561031>
- Webster, J., & Watson, R. T. (2002). *Analyzing the Past to Prepare for the Future: Writing a Literature Review*.
- Wilde, T., & Hess, T. (2006). *Methodenspektrum der Wirtschaftsinformatik: Überblick und Portfoliobildung*. 1–14.
- Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach*.

- Zadgaonkar, A. V., & Agrawal, A. J. (2021). An overview of information extraction techniques for legal document analysis and processing. *International Journal of Electrical and Computer Engineering (IJECE)*, 11(6), 5450. <https://doi.org/10.11591/ijece.v11i6.pp5450-5457>
- Zhou, J. (2016). *The dynamics of leniency application and the knock-on effect of cartel enforcement.*

## DECLARATION OF AUTHORSHIP

I hereby declare that this master's thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented is entirely my own.

■ 30 May 2023

■

Rebecca Baumann