Milestone 1

Dataset

We are planning on using the GDPR dataset, found on Kaggle. The GDPR dataset consists of violation reports for the years 2018, 2019 and 2020 with format (fine, country, date of decision, ETid, controller/processor, quoted article(s), type, source). There are 437 entries.
The quality of the dataset is high: it is collected from an official department and we can assume that each entry is trusted to be correct and we do not need to do data cleaning. For pre-processing, we need to account for format inconsistencies: some of the columns have different formats that we need to uniformize before tackling visualization; and there are some missing values: some fields are marked as unknown, for example the date of decision or the controller.

Problematic

We want to explore the correlation between features (columns) of the dataset. For example, our visualization could show:

- Relationships between different GDPR articles: is there a subset of articles that are always applied at the same time?
    - Can we process and analyse the articles' text to find similarities?
- What are the most common violations and the associated amounts?
- Are there companies that are consistently paying a fine, which could suggest that this is a strategic choice? If so, in which sector or country?
    - Do they have periodic patterns?
- How have fines evolved over time?
    - (tentative) Are the patterns pre-revision or post-revision? This would require extra data, namely the history of revisions of GDPR articles.

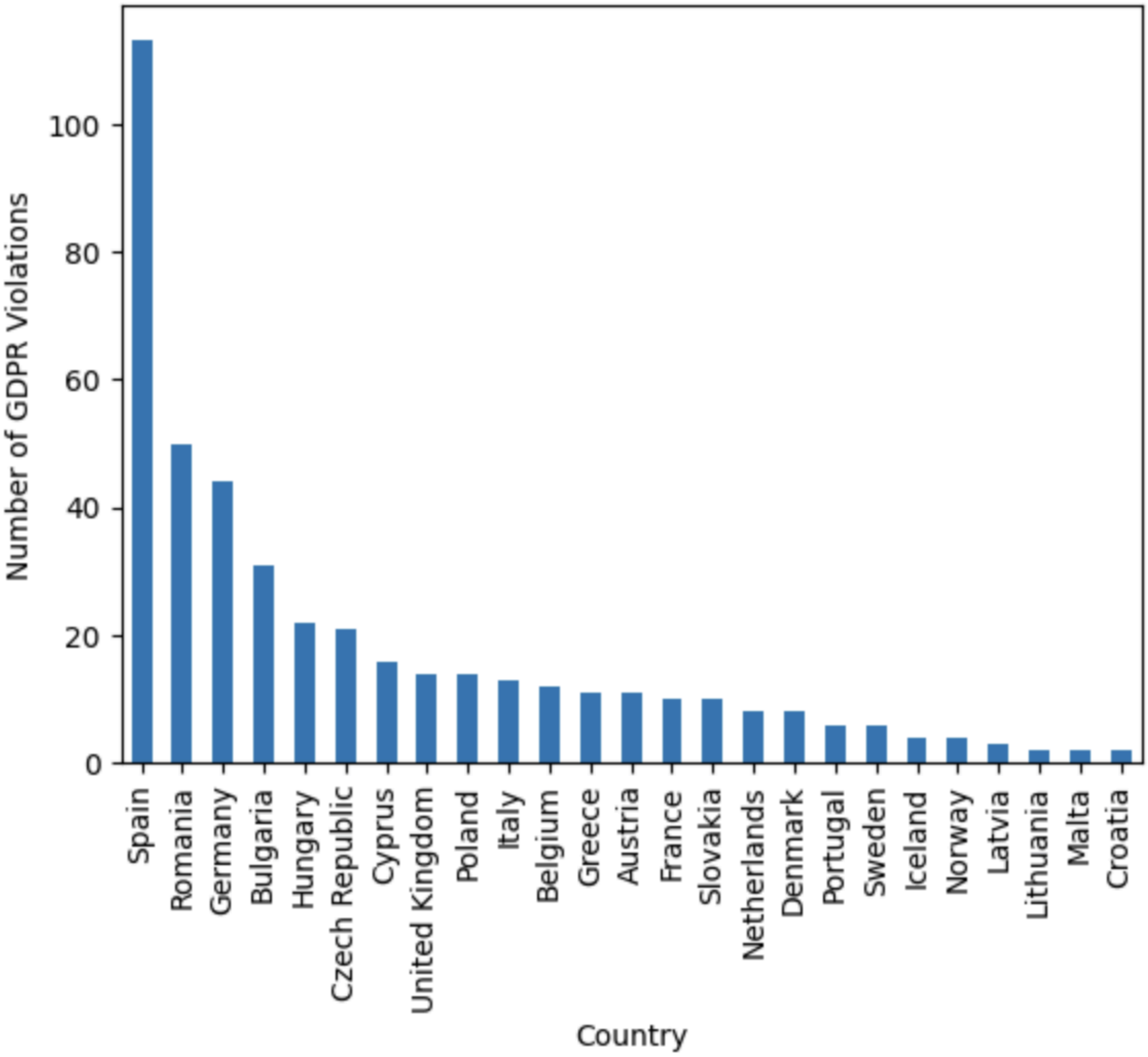
Given the consequent amount of violations, there is something wrong with the current state of GDPR. Through visually highlighting some patterns, we are trying to show how the GDPR is implemented in practice and if it is porous, e.g., if we can find patterns in the fines that suggest that some companies prefer to pay the fine rather than implementing the mitigation; and to find evidence to whether the GDPR law could be improved, and if so, find avenues for improvement, e.g., if articles are redundant, how can they be simplified? Consequently, the target audience is policymakers, so that they can make informed decisions in the future about improving the law.

Exploratory Data Analysis

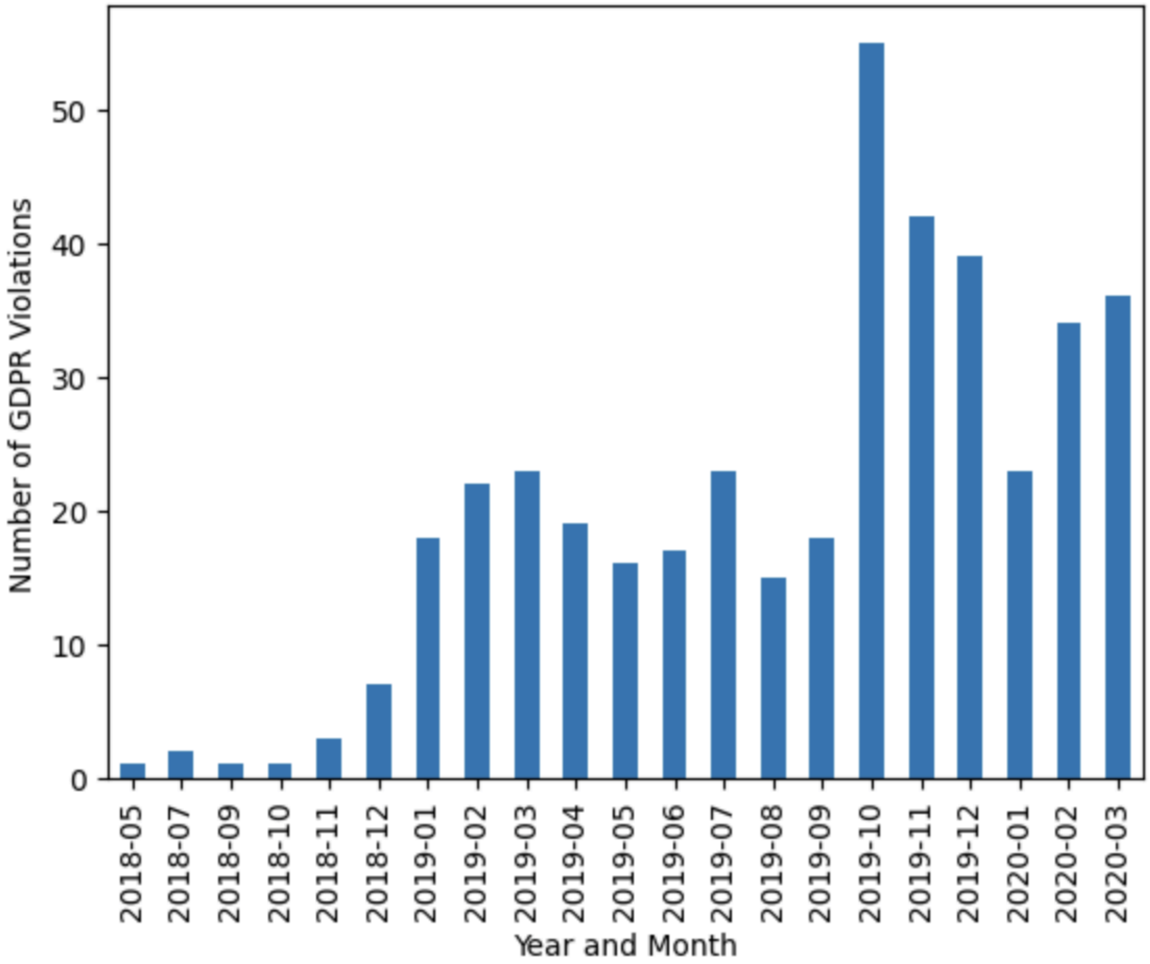
The data has the following headers:
-Country
-Fine
-Date of decision
-Controller/processor
-Quoted article(s)

-Type
-Source
-Summary of the violation

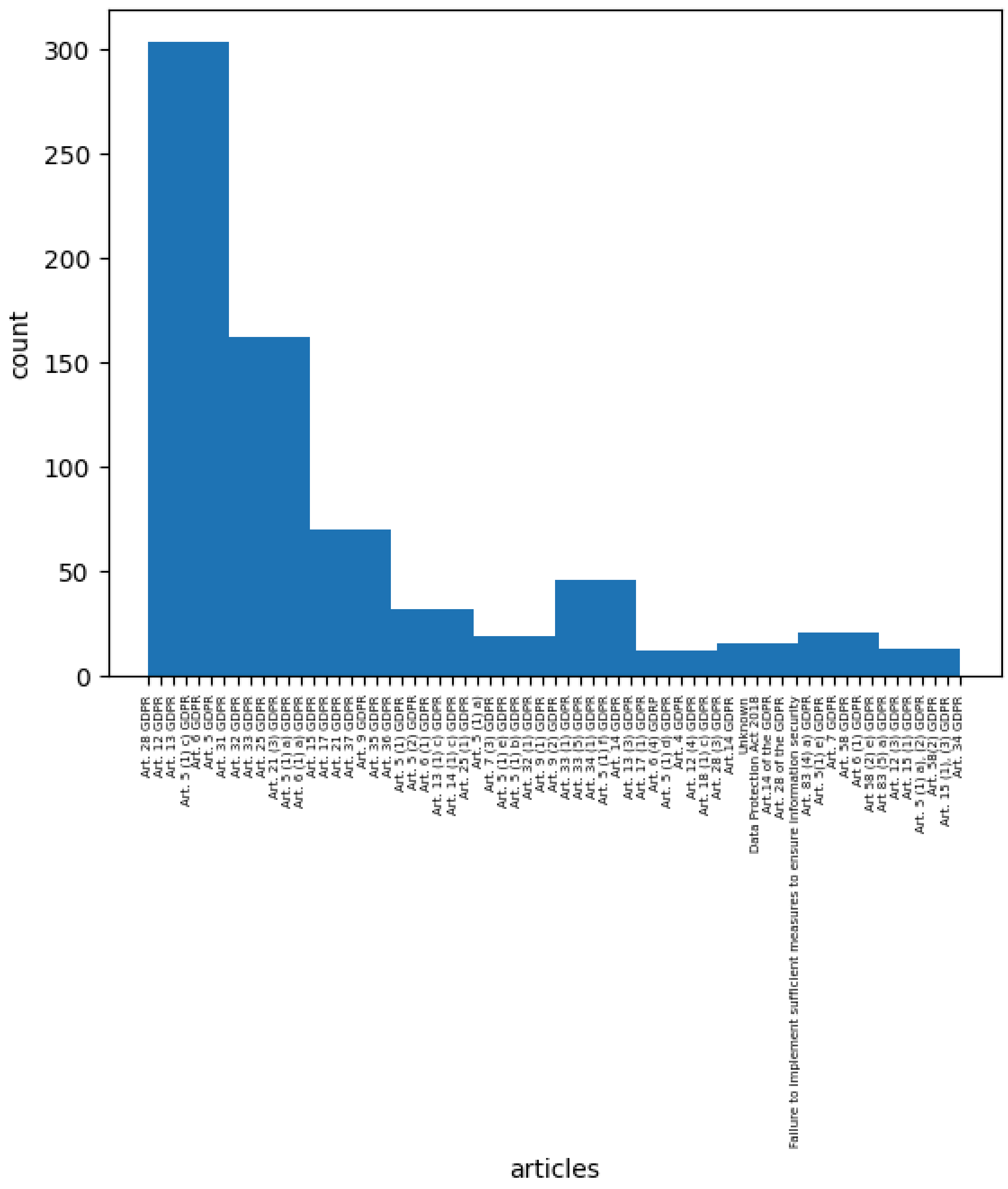
We produced a few basic statistics that show the motivation behind this project:



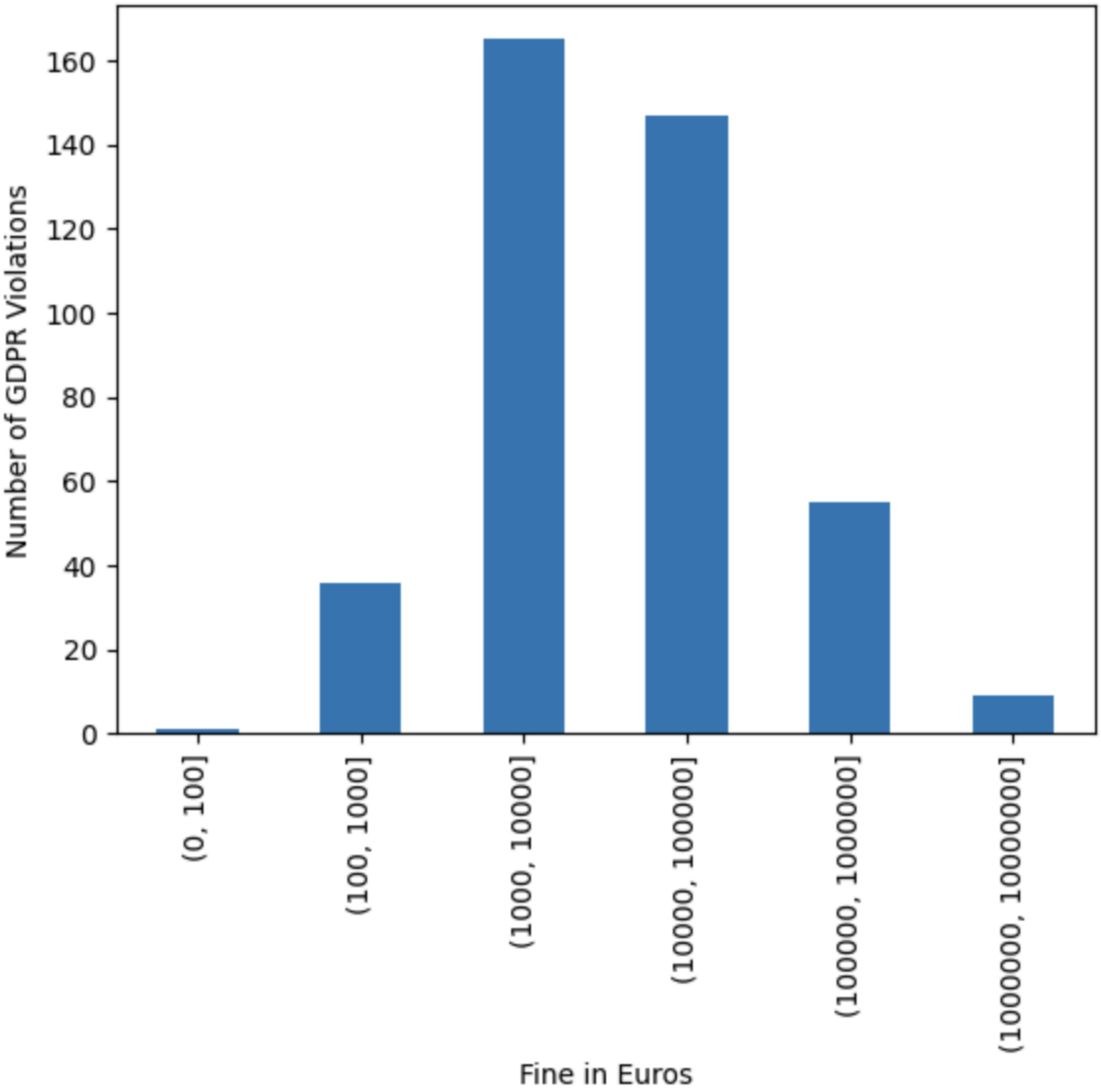
First we look at the number of GDPR violations per country, which shows at least 10 countries having 10 or more violations filed, with Spain reaching 113 violations. We are interested in studying if the trends in types of GDPR violations and articles cited vary across countries.



Then we have plotted the number of violations having been filed per month between May 2018 and March 2020. We will explore the evolution in time of the types of violations, and which countries or controllers were reported.



We additionally look at the redistribution of articles mentioned in the violations. To be noted that 176 entries mention more than one article.



Finally, we have plotted the number of GDPR violations according to the price of the fine (in log scale). We are interested in studying the correlations between the price of the fines, and the type of violations and articles cited.

Related work

Most data visualization efforts focus on first-order statistics and rudimentary figures about the costs and counts of GDPR fines. We detail the following key visualizations we found:

GDPR Enforcement Tracker [1]

- cumulative and non-cumulative fine count and sum over time
- histogram of total fines (counts and sums) per country
- pie chart for fines by violation type
- pie chart for fines by sector

CMS Law [2]

- cumulative fine count and sum over time
- list of top 10 fines
- histogram of average fine cost and total number of fines by sector
- histogram of total fine counts, average, and total fine cost per country
- histogram for fines by violation type

Kaggle [3]

- exploratory data analysis with descriptive statistics and figures about columns
- basic feature correlation matrix
- feature scatter plots

Forms.app [4]

- descriptive statistics (percentages, averages, and totals) across different breakdowns and combined with other sources of statistics.

Statista [5]

- histograms about the breakdown of violations per country and sector

Our Original Contributions

Our approach is original because we look at correlations about multiple features of the data, whereas existing visualizations only look at features in isolation, or at best plot a quantitative metric (e.g., fine amount) against a categorical feature (e.g., country).

(Optional) Do some per capita statistics because they are missing from related work.

Inspiration

correlation circle and correlation matrix (handbook of data visualization)

References

[1] https://www.enforcementtracker.com/?insights
[2] https://cms.law/en/int/publication/gdpr-enforcement-tracker-report/numbers-and-figures
[3] https://www.kaggle.com/code/kerneler/starter-gdpr-violations-bea8ab88-2
[4] https://forms.app/en/blog/gdpr-statistics
[5] https://www.statista.com/topics/9651/tech-regulations-in-europe/#statisticChapter