Analyzing GDPR Fines

by Andre Buser and Victor Adafinoaiei

Background

On May 25, 2018, the European Union (EU) “General Data Protection Regulation” (GDPR) became effective. The GDPR is a new data privacy initiative adopted by the EU to provide enhanced protection to EU citizens and their personal data. The penalties violations can result in up to twenty million euros or four percent of the company’s global annual revenue from the previous year, whichever number is higher. In addition, EU legislators impose fines for penalties to enforce data protection compliance.

Objectives *(Project Phase 1)*

The purpose of the project was to analyze GDPR fines issued since 2018 and to

(A) Address the following basic questions:

- Which **industry sectors** were penalized the most per country?
- Which **EU countries** issued the **most** GDPR fines for the healthcare sector?
- Which **GDPR articles** were quoted the most per reported compliance issue in the healthcare sector?
- What are the **average costs** per GDPR compliance issue for the healthcare sector?

The **outcomes** could guide the Group Data Privacy Officer of our healthcare organization in deciding on the **critical compliance areas and regions** and how to **allocate the limited resources** (people and budget).

(B) Verify additional assumptions by considering the following proxy measures: population by country (**POP**), gross domestic product (**GDP**), and corruption perception index (**CPI**):

- A higher GDP could lead to higher fines
- A higher CPI could lead to higher fines
- A higher population could lead to more or higher fines

The **result** will help understand if those proxy measures could be considered in future **prediction models** to improve their accuracy, e.g., in predicting the expected average fine.

Analyzing GDPR fines imposed by the European data protection authorities could reveal the main reasons and focus areas of the authorities for non-compliance and could allow our organization to timely address similar gaps in their data privacy strategy. A correlation with additional proxy measures could help to build future prediction models.

Executive Summary

Project Phase 1

Recommendations *(Project Phase 1)*

- Between 2018 and 2021, Spain imposed the most GDPR fines (352) across all industries sectors.
- Based on the fines imposed in the healthcare sector (total and ratio), it is recommended to review, assess and monitor our data processing activities in Sweden, Italy, Spain, Estonia, and Portugal.
- Based on the distribution of the quoted GDPR articles and the average costs for a compliance issue, it is recommended to review and assess our controls and requirements for Information Security (Art. 32), Legal Basis for Data Processing (Art. 6 and Art. 9) and The General Data Processing Principles (Art 5.)
- The primary GDPR Compliance Issues that account for €11.81m (96.5%) in the healthcare sector are related to (1) insufficient technical and organizational measures to ensure information security and (2) non-compliance with general data processing principles.
- For the data grouped by country, some positive trends were identified showing that countries with a: (A) higher GDP issue higher fines, (B) higher CPI issue higher fines (C) higher population issue higher fines.

Limitations

- Completeness and accuracy: The source for the GDPR fine dataset, enforcementtracker.com, is not a complete representation of all GDPR fines since not all fines are made public.

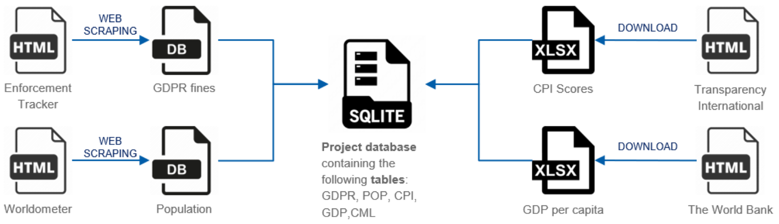
However, we can use this sample data to make an inference or conclusion for the population.

Next Steps *(Project Phase 2)*

- Update and verify results once the actual 2021 values are available for GDP per capita, population, and CPI score.
- Add annual revenues of fined healthcare companies, if available, to calculate the average % of the fines.
- Analyze GDPR fine text summaries (NLP) for additional data mining.
- Refine binning with the use of the Fisher-Jenks algorithm to rank data into natural breaks instead of using quantiles.
- Explore outliers discovered in the datasets and evaluate how to address those.
- Build prediction models (regression and classification).
- Update report on a quarterly or annual basis.

Methodology

Data Sources



The following **data sources** were Considered and consolidated into an SQLite DB:

GDPR The GDPR fines data is the primary dataset. The information is scraped from www.enforcementtracker.com and contains details about the imposed GDPR fines. The information was **scrapped** with the Selenium library.

No. of records:	986 *on 2022-01-13*
No. of attributes:	10
Format:	HTML tables (dynamic)

POP Countries of the world with their population over the years (1955 - 2020). The data is scraped from www.worldometers.info. For the **parsing**, the Beautiful Soup library was used.

No. of records:	4212* *on 2021-11-28*
No. of attributes:	5
Format:	HTML table (static)

CPI The CPI dataset describes the Corruption Perceptions Index (CPI) per country. The CPI scores and ranks countries based on how corrupt a country's **public sector is perceived** to be. The data is manually **downloaded** from www.transparency.org.

No. of records:	180 *on 2021-11-29*
No. of attributes:	34
Format:	XLSX

GDP Gross Domestic Product (GDP) is the monetary value of all finished goods and services made within a country during a specific period. GDP provides an economic snapshot of a country, used to estimate the size of an economy and growth rate. The dataset is manually **downloaded** from data.worldbank.org.

No. of records:	266 *on 2021-11-29*
No. of attributes:	65
Format:	XLS

**Micronesia was excluded due to parsing issues. Micronesia does not have all the attributes compared to the other countries in the Worldmeter dataset. Considering that this country is not relevant for our analysis (no GDPR fines in Micronesia) the project team decided not include that country in the parsing.*

Methodology

Data Manipulation

The following **major data manipulations** were applied (full details available in the referenced notebooks):

	GDPR	POP	CPI	GDP
CREATE	▪ Added mapping column: COUNTRY-YYYY in order to ensure reference key across all tables ▪ Added categories (binning)			
	▪ Created country master list (CML) ▪ Applied label encoding for categorical features		▪ Added two missing countries: "Isle of Man" and "Liechtenstein"	
UPDATE	▪ Fine: Re-labeled missing values ("not assigned") as "unknown" and replaced all non-numerical entries with "np.NAN" ▪ Fine: Changed fine type to float ▪ Article: Cleaned and streamlined article naming convention to "Art. xx (x) GDPR" ▪ Article: Kept "unknown" values as dedicated category ▪ Decision date: Imputed missing values ("unknown") with a forward-fill ▪ Decision date: Kept year (YYYY) only	▪ Aligned country naming with country master list (CML)		
		▪ ! Calculated expected values for 2021 based on **average growth rate** between 2015 and 2020	▪ ! Calculated expected values for 2021 based on the **mean CPI** between 2012 and 2020 ▪ Updated CML with iso3 codes	▪ ! Calculated expected values for 2021 based on the **mean GDP** between 2012 and 2020
DELETE	▪ Removed observations for year 2022 ▪ ! Dropped observation ETid-31 as invalid case ▪ ! Removed non-GDPR article observations	▪ Removed unwanted observations: year < 2018		
	Final Shape: (978,14)	**Final Shape: (124,6)**	**Final Shape: (124,6)**	**Final Shape: (124,6)**

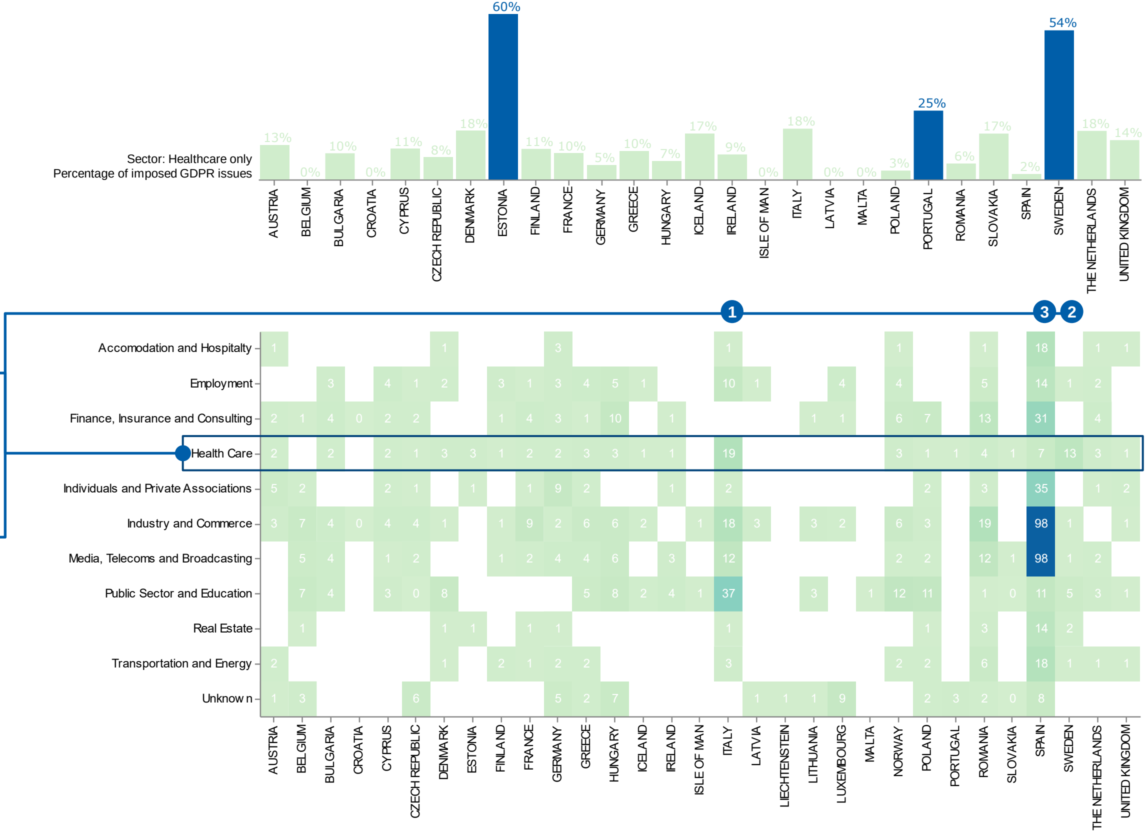
Analysis

GDPR Fines per Sector

Firstly, we evaluated: **(A)** which industry sectors were penalized the most per country and **(B)** in which county was the healthcare sector (HCS) penalizing the most:

- In total, **80** fines were imposed for the healthcare sector.
- Based on **total** imposed fines for the healthcare sector, ❶ **Italy** imposed most of the healthcare-related fines: **19 out of** 103, followed by ❷ **Sweden** with **13 out of** 24 and ❸ **Spain** with **7 out of 352**.
- Based on the **percentage** of imposed GDPR fines for HCS, we also would need to consider **Estonia (60%)**, again, **Sweden (54%)**, and **Portugal (25%)**

Based on the fines imposed in the healthcare sector (total and ratio), ***it is recommended to review, assess and monitor our data processing activities in Sweden, Italy, Spain, Estonia and Portugal***.

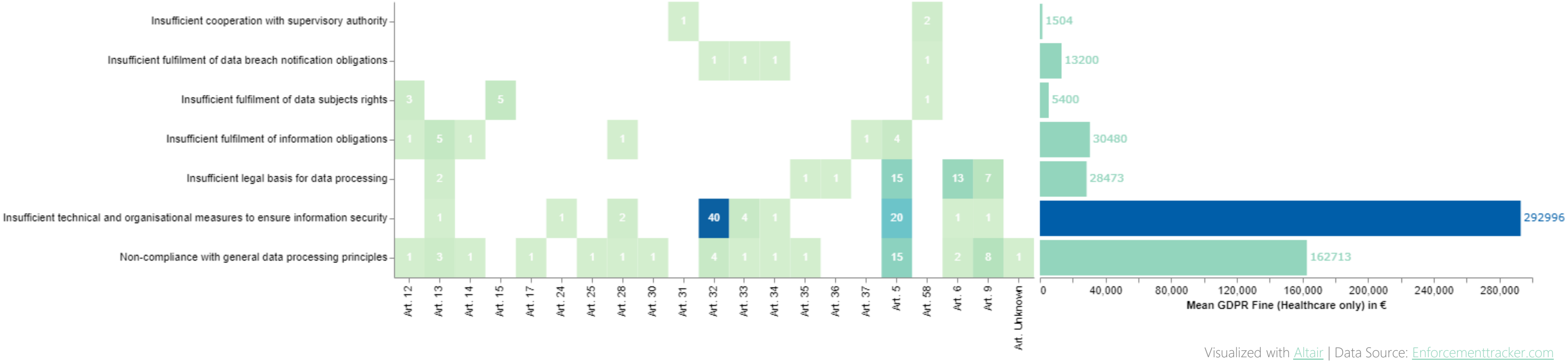


	AUSTRIA	BELGIUM	BULGARIA	CROATIA	CYPRUS	CZECH REPUBLIC	DENMARK	ESTONIA	FINLAND	FRANCE	GERMANY	GREECE	HUNGARY	ICELAND	IRELAND	ISLE OF MAN	ITALY	LATVIA	LIECHTENSTEIN	LITHUANIA	LUXEMBOURG	MALTA	NORWAY	POLAND	PORTUGAL	ROMANIA	SLOVAKIA	SPAIN	SWEDEN	THE NETHERLANDS	UNITED KINGDOM
Accomodation and Hospitalty	1						1				3						1						1			1		18		1	1
Employment			3		4	1	2		3	1	3	4	5	1			10	1			4		4			5		14	1	2	
Finance, Insurance and Consulting	2	1	4	0	2	2			1	4	3	1	10		1					1	1		6	7		13		31		4	
Health Care	2		2		2	1	3	3	1	2	2	3	3	1	1		19						3	1	1	4	1	7	13	3	1
Individuals and Private Associations	5	2			2	1		1		1	9	2			1		2							2		3		35		1	2
Industry and Commerce	3	7	4	0	4	4	1		1	9	2	6	6	2		1	18	3		3	2		6	3		19		98	1		1
Media, Telecoms and Broadcasting		5	4		1	2			1	2	4	4	6		3		12						2	2		12	1	98	1	2	
Public Sector and Education		7	4		3	0	8					5	8	2	4	1	37			3		1	12	11		1	0	11	5	3	1
Real Estate		1					1	1		1	1						1							1		3		14	2		
Transportation and Energy	2						1		2	1	2	2					3						2	2		6		18	1	1	1
Unknow n	1	3				6					5	2	7					1	1	1	9			2	3	2	0	8			

Visualized with Altair | Data Source: Enforcementtracker.com
Raw Data Processed | Reporting Period: 2018 to 2021 | Created: 2022-01-23

Analysis

GDPR Compliance Issues for the Healthcare Sector



Visualized with Altair | Data Source: Enforcementtracker.com
Raw Data Processed | Reporting Period: 2018 to 2021 | Created: 2022-01-23

In the next step, it was evaluated:

(A) What are the **average costs** per GDPR compliance issue for the healthcare sector? Those will provide a risk-based focus.

(B) Which GDPR articles have been quoted the most per reported compliance issue in the healthcare sector? Those will guide us to the relevant requirements.

Based on the distribution of the articles and the average costs for a compliance issue, ***it is recommended to review and assess the controls and requirements*** from: Art. 32 (Security of processing), Art. 5 (Principles relating to the processing of personal data), Art. 6 (Lawfulness of processing) and Art. 9 (Processing of special categories of personal data).

Analysis

GDPR Costs and Financial Risks

€12.51m / $14.20m

in total fines (80) since 2018 across the **healthcare** sector

The primary **GDPR Compliance Issues** that account for €11.81m (96.5%) in the healthcare sector are related to:

- (1) Insufficient technical and organisational measures to ensure information security
- (2) Non-compliance with general data processing principles

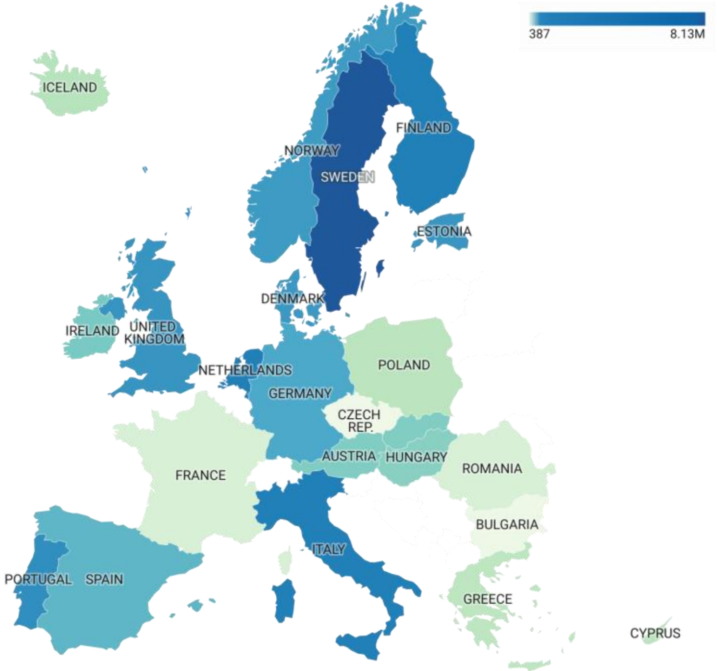
For violation (1), **GDPR Art.32**[1] **and Art.5**[2] account for 99% of the total related fines.

For example, **Capio St. Göran's Hospital AB (Sweden)** was fined **€2.9m** in 2020 for failing to implement adequate technical and organizational measures to ensure information security resulting in unauthorized full access to confidential patient data.

Top 5 fined countries represent ~85% of total fines

Sweden	**€8.13m**
Netherland	**€0.91m**
Italy	**€0.72m**
Finland	**€0.61m**
Portugal	**€0.40m**

Compliance Issue	**Health Care Sector**			**Across All Sectors**		
	Freq.	Avg. Fine	Highest Fine	Freq.	Avg. Fine	Highest Fine
(1)	40.0%	€0.29m	€2.9m	19.9%	€0.36m	€22.0m
(2)	18.8%	€0.16m	€1.2m	20.2%	€3.98m	€746.0m



Visualized with Datawrapper | Data Source: Enforcementtracker.com
Raw Data Processed | Reporting Period: 2018 to 2021 | Created: 2022-01-21

[1] https://gdpr-info.eu/art-32-gdpr/

[2] https://gdpr-info.eu/art-5-gdpr/

Analysis

Correlations: Non-Aggregated Data

Spearman's rank correlation coefficient was used to explore the correlations because **none** of the attributes were **normally distributed**, and we assumed **linear relationships**. Statistical outliers were not removed.

For the **non-aggregated** dataset **weak positive** correlations were found for:

- gdp and fine: 0.15
- cpi_score and fine: 0.28.

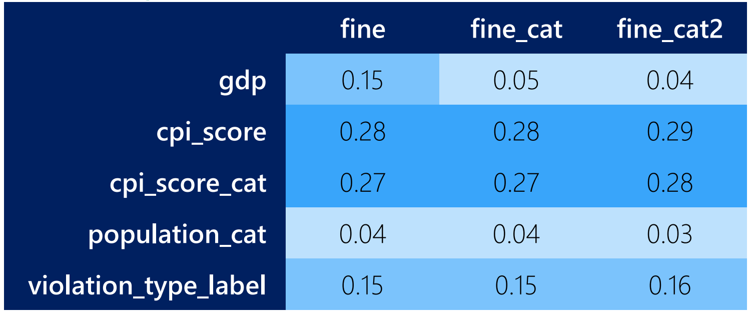
For all correlation coefficients, the p-value was below 1%.

Conclusion

The results show a **weak positive trend** that countries with:

- a higher GDP also issue higher fine
- a higher CPI also issue higher fines

Spearman's rank correlation coefficients



	fine	fine_cat	fine_cat2
gdp	0.15	0.05	0.04
cpi_score	0.28	0.28	0.29
cpi_score_cat	0.27	0.27	0.28
population_cat	0.04	0.04	0.03
violation_type_label	0.15	0.15	0.16

Calculated with Pandas
Data Source: Enforcementtracker.com
Raw Data Processed | Reporting Period: 2018 to 2021
Created: 2022-01-23

_label: Categorical feature was label encoded
_cat: Binning was applied using rounding
_cat2: Binning was applied using quantiles

Analysis

Correlations: Aggregated Data

For the next correlation check, we decided to **group the data by country** and aggregate the important features (mean, median).

Again, **Spearman's rank correlation coefficient** was used, because also after the aggregation the attributes were **not normally distributed**, and we assumed **linear relationship**. Statistical outliers were not removed.

For the **aggregated** dataset **moderate positive** correlations were found for:

- gdp_cat2_median and fine_mean: 0.62
- cpi_score_cat_mean and fine_median: 0.59

Conclusion

The results show **some positive trends** that countries with:

- a higher GDP issue higher fines
- a higher CPI issue higher fines
- a higher population issue higher fines

Spearman's rank correlation coefficients

	fine_mean	fine_median
gdp_mean	0.59	0.30
gdp_cat2_median	0.62	0.26
cpi_score_mean	0.47	0.57
cpi_score_cat_mean	0.43	0.59
population_mean	0.42	0.10
population_cat2_median	0.45	0.12

Calculated with Pandas
Data Source: Enforcementtracker.com
Raw Data Processed | Reporting Period: 2018 to 2021
Created: 2022-01-23 | Data grouped by country

_label: Categorical feature was label encoded
cat: Binning was applied using rounding
cat2: Binning was applied using quantiles

References

Project Phase 1

Data Sources

- www.enforcementtracker.com
- www.worldometers.info
- www.transparency.org
- data.worldbank.org

GDPR Articles

- https://gdpr-info.eu/

Project Structure Template

- https://drivendata.github.io/cookiecutter-data-science/

External Visualization Tool

- https://www.datawrapper.de/

Data Science Ethics Checklist

- https://deon.drivendata.org/

Jupyter Notebooks Web Scrapping

- 01_DCO01_WS_GDPR.ipynb
- 01_DCO02_WS_POP.ipynb

Jupyter Notebooks Data Cleaning

- 02_DCL01_GDPR.ipynb
- 02_DCL02_CPI.ipynb
- 02_DCL03_POP.ipynb
- 02_DCL04_GDP.ipynb

Jupyter Notebooks Data Analysis

- 03_ANA01_EDA.ipynb
- 03_ANA02_basics.ipynb
- 03_ANA03_correlations.ipynb