Team NAN 101
Project report

Team member:
Alexandru Bârsan (No.: 7396261)
Johannes-Jona Detjenn (No.: 6836948)

Our source data consists of datasets from two different sources, first a high quality dataset from Kaggle and an automatically generated dataset from Roel M. Hogervorst a user of Github.
Kaggle dataset:
https://www.kaggle.com/eljailarisuhonen/gdpr-fines-in-eu-20182019-120-rows-8-columns
Github record:
https://github.com/RMHogervorst/scrape_gdpr_fines
Our code and all files that were used are available here:
https://github.com/AlexBarsen/Data_Science

Our records are about the given fines of the EU members.
The General Data Protection Regulation (GDPR) is an EU regulation on data protection and privacy in the European Union and the European Economic Area.
It also deals with the transfer of personal data outside the EU and EEA areas. The main aim of the GDPR is to give individuals control over their personal data and to simplify the regulatory environment for international transactions by harmonising regulation within the EU.
If institutions or companies do not comply with the GDPR rules, they will be fined.

In this project we would like to find out which variables from both dataset relate with each other. For example if there are more articles applied, the value of the fine should also be growing. We assume that the number of articles which are cited have an influence on the height of the fine. In our hypothesis this relation should also apply for the economic wealth of a country measured in Gross domestic product (GDP) and the total eight of fines per country.

The Kaggle dataset is a sample of given information from the site "www.enforcementtracker.com", consisting of 120 rows and 8 columns: *Country, Authority, Date, Fine, Controller / Processor, Quoted Article, Type, Infos*. The sample represents only about 15 to 20% of all given data points from the period 2018-2019, the most recent datapoint being "2019-11-25". The correctness of the dataset is taken for granted by the creator (Elja-Ilari Suhonen) and is not further checked by us. With more time a check for updates and correctness could have been done.

The second dataset of Github was created by the owner of the repository by web scraping the website "https://www.privacyaffairs.com/gdpr-fines/".
The website tracks current procedures to implement the EU regulation on data protection. All entries from the Privacy Affairs site are reported by official authorities, which guarantees their legitimacy.
The relevant information has been extracted from the source code of the website. Afterwards the website was processed as HTML code in R and the data was mostly just formatted and filtered. The dataset we chose is a resulting comma-separated file (.csv), which is derived from the JSON structure of the raw data.

Consequently, this tabularly formatted dataset consists of 250 rows and 11 columns: *id, picture, country, price, authorithy, date, org_fined, article Violated, type, source, summary*, which are manipulated to merge the two datasets.

To edit and cleanse the data, we imported both datasets into the Kaggles (Python) and edited them using functions of the Python framework pandas.
The two .csv files of the datasets are read in as dataframes named *data1* and *data2* using the parser integrated in pandas.
The column names of data2 with the same or similar content as in data1 have been named the same in order to simplify later merging:
The column *Fine [€]* of data1 was renamed to *Fine*
The *Price* column of data2 was changed to Fine
The *Controller/Processor* column of data1 has been changed to *Org_fined*
The column *Quoted Art.* of data1 was changed to *ArticleViolated*

Using the function **concat()** from the pandas framework and the parameter *join="inner"*, an intersection of the two datasets was created and named dataset_clust. If we decide that we want to use clustering algorithms, this will probably the basis for out clustering analysis.
To get not only the intersection of both datasets, but the total amount of both datasets, we executed the function **concat()** with the parameter *join="outer"* and created a dataset named *dataset*.

Further cleansing steps that where done:
Standardize the country names with the function **upper()** capital letters in example "germany" => "GERMANY".
To correct the problem of decimal points in currencies we first convert the *Fine* column to a string and then delete the "," and "." from the Fine column, because the dataset does not contain any decimal point values.
We check if all entries in the Fine column are numbers and notice that we have 6 "Unknown" entries.
Further, because of inconsistency we move the Index column and drop the following columns from the data frame: *Id, Infos, Picture, Source, Summary, index.*
Using a dictionary, we generate a *Country_code* column with the corresponding country code manually. Tricky was the country "Netherlands" which appears in both original datasets differently, we found four time "The Netherlands" instead and had to treat it separately.

Before being processed in Python, our datasets looked like this:

Kaggle Data Set

121 rows — Extensions: Wikidata
Show as: rows records Show: 5 10 25 50 rows — « first ‹ previous 1 - 10 next › las

All		Country	Authority	Date	Fine [â,¬]	Controller/Processor	Quoted Art.	Type	I
	1.	ROMANIA	Romanian National Supervisory Authority for Personal Data Processing (ANSPDCP)	2019-11-25	11,000	Courier Services Company	Art. 32 GDPR	Insufficient technical and organisational measures to ensure information security	link
	2.	ROMANIA	Romanian National Supervisory Authority for Personal Data Processing (ANSPDCP)	2019-11-22	2,000	BNP Paribas Personal Finance S.A.	Art. 12 GDPR, Art. 17 GDPR	Insufficient fulfilment of data subjects rights	link
	3.	SPAIN	Spanish Data Protection Authority (aepd)	2019-11-21	60,000	Viaqua XestiĂłn Integral Augas de Galicia	Art. 6 GDPR	Insufficient legal basis for data processing	link
	4.	FRANCE	French Data Protection Authority (CNIL)	2019-11-21	500,000	Futura Internationale	Art. 5 GDPR, Art. 6 GDPR, Art. 13 GDPR, Art. 14 GDPR, Art. 21 GDPR	Insufficient fulfilment of data subjects rights	link
	5.	SPAIN	Spanish Data Protection Authority (aepd)	2019-11-19	60,000	CorporaciĂłn radiotelevisiĂłn espanola	Art. 32 GDPR	Insufficient technical and organisational measures to ensure information security	link

Github Data Set

250 rows — Extensions: Wikidata
Show as: rows records Show: 5 10 25 50 rows — « first ‹ previous 1 - 10 next › last »

All			id	picture	country	price	authority	date	org_fined	articleViolated	type	source	summary
		1.	1	https://www.privacyaffairs.com/wp-content/uploads/2019/10/republic-of-poland.svg	Poland	9380	Polish National Personal Data Protection Office (UODO)	2019-10-18	Polish Mayor	Art. 28 GDPR	Non-compliance with lawful basis for data processing	https://uodo.gov.pl/decyzje/ZSPU.421.3.2019	No data processing agreement has been concluded with the company whose servers contained the resources of the Public Information Bulletin (BIP) of the Municipal Office in AleksandrĂłw Kujawski. For this reason, a fine of 40.000 PLN (9400 EUR) was imposed on the mayor of the city.
		2.	2	https://www.privacyaffairs.com/wp-content/uploads/2019/10/romania.svg	Romania	2500	Romanian National Supervisory Authority for Personal Data Processing (ANSPDCP)	2019-10-17	UTTIS INDUSTRIES	Art. 12 GDPR, Art. 13 GDPR, Art. 5 (1) c) GDPR, Art. 6 GDPR	Information obligation non-compliance	https://www.dataprotection.ro/?page=A_patra_amenda&lang=ro	A controller was sanctioned because he had unlawfully processed the personal data (CNP), and images of employees obtained through the surveillance system. The disclosure of the CNP in a report for the ISCIR training in 2018 wasnâ€™t legal, as per Art.6 GDPR.
		3.	3	https://www.privacyaffairs.com/wp-content/uploads/2019/10/spain.svg	Spain	60000	Spanish Data Protection Authority (AEPD)	2019-10-16	Xfera Moviles S.A.	Art. 5 GDPR, Art. 6 GDPR	Non-compliance with lawful basis for data	https://www.aepd.es/resoluciones/PS-00262-2019_ORI.pdf	<p>The company had unlawfully processed the personal data despite the subjectâ€™s request to stop doing so.

After processing in Python our dataset looks like this:

Out[31]:

	ArticleViolated	Authority	Country	Date	Fine	Org_fined	Type	Country_code
0	Art. 32 GDPR	Romanian National Supervisory Authority for Pe...	ROMANIA	2019-11-25	11000	Courier Services Company	Insufficient technical and organisational meas...	21
1	Art. 12 GDPR, Art. 17 GDPR	Romanian National Supervisory Authority for Pe...	ROMANIA	2019-11-22	2000	BNP Paribas Personal Finance S.A.	Insufficient fulfilment of data subjects rights	21
2	Art. 6 GDPR	Spanish Data Protection Authority (aepd)	SPAIN	2019-11-21	60000	Viaqua Xestión Integral Augas de Galicia	Insufficient legal basis for data processing	23
3	Art. 5 GDPR, Art. 6 GDPR, Art. 13 GDPR, Art. 1...	French Data Protection Authority (CNIL)	FRANCE	2019-11-21	500000	Futura Internationale	Insufficient fulfilment of data subjects rights	08
4	Art. 32 GDPR	Spanish Data Protection Authority (aepd)	SPAIN	2019-11-19	60000	Corporación radiotelevisión espanola	Insufficient technical and organisational meas...	23
...	...	...	...	...	...	...	...	...
366	Art. 33 GDPR, Art. 34 GDPR	Data Protection Authority of Hamburg	GERMANY	2019-01-01	20000	https://datenschutz-hamburg.de/assets/pdf/28_...	Failure to implement sufficient measures to en...	09
367	Art. 5 GDPR, Art. 32 GDPR	Data Protection Authority of Baden-Wuerttemberg	GERMANY	2019-04-04	80000	Company in the financial sector	Failure to implement sufficient measures to en...	09
368	Art. 5 GDPR, Art. 6 GDPR	Data Protection Authority of Nordrhein-Westfalen	GERMANY	2019-08-05	200	Private person (YouTube-Channel)	Failure to comply with data processing principles	09
369	Art. 5 GDPR, Art. 32 GDPR	Data Protection Authority of Baden-Wuerttemberg	GERMANY	2019-10-24	100000	Food company	Failure to implement sufficient measures to en...	09
370	Art. 5 GDPR, Art. 6 GDPR, Art. 32 GDPR	Hellenic Data Protection Authority (HDPA)	GREECE	2019-12-19	150000	Aegean Marine Petroleum Network Inc.	Failure to comply with data processing principles	10

371 rows × 8 columns

We also exported the dataframe from Kaggle to a CSV, since we continue to clean the dataset in OpenRefine.

We have edited the following in OpenRefine:
Deleted 6 rows that had Fine and Date as "Unknown" values
6 rows which had Fine 0 Value and date 1970-01-01 deleted
4 lines had 2.018 as year specification, so we rewrote them to 2018-01-01
8 lines had 2.019 as year specification, so that we rewrote them to 2019-01-01
9 lines had "Unknown" as year, so we changed it to 2021-01-01
9 lines had the year 1970-01-1, so we rewrote them to 2021-01-01
Furthermore we have converted the fine and country_code lines to number with the "Common Transform" function

After processing the data in OpenRefine, our dataset looks like this:

121 rows — Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 rows — « first ‹ previous 1 - 10 next › las

All		Country	Authority	Date	Fine [â,¬]	Controller/Processor	Quoted Art.	Type	In
	1.	ROMANIA	Romanian National Supervisory Authority for Personal Data Processing (ANSPDCP)	2019-11-25	11,000	Courier Services Company	Art. 32 GDPR	Insufficient technical and organisational measures to ensure information security	link
	2.	ROMANIA	Romanian National Supervisory Authority for Personal Data Processing (ANSPDCP)	2019-11-22	2,000	BNP Paribas Personal Finance S.A.	Art. 12 GDPR, Art. 17 GDPR	Insufficient fulfilment of data subjects rights	link
	3.	SPAIN	Spanish Data Protection Authority (aepd)	2019-11-21	60,000	Viaqua XestiÃn Integral Augas de Galicia	Art. 6 GDPR	Insufficient legal basis for data processing	link
	4.	FRANCE	French Data Protection Authority (CNIL)	2019-11-21	500,000	Futura Internationale	Art. 5 GDPR, Art. 6 GDPR, Art. 13 GDPR, Art. 14 GDPR, Art. 21 GDPR	Insufficient fulfilment of data subjects rights	link
	5.	SPAIN	Spanish Data Protection Authority (aepd)	2019-11-19	60,000	CorporaciÃn radiotelevisiÃn espanola	Art. 32 GDPR	Insufficient technical and organisational measures to ensure information security	link

From the 370 entries, only 358 entries remained after the data was cleaned.

In addition a Fines_applied column is generated which contains the number of broken articles.

Out[86]:

	ArticleViolated	Authority	Country	Date	Fine	Org_fined	Type	Country_code	Fines_applied
0	Art. 32 GDPR	Romanian National Supervisory Authority for Pe...	ROMANIA	2019-11-25	11000	Courier Services Company	Insufficient technical and organisational meas...	21	1
1	Art. 12 GDPR, Art. 17 GDPR	Romanian National Supervisory Authority for Pe...	ROMANIA	2019-11-22	2000	BNP Paribas Personal Finance S.A.	Insufficient fulfilment of data subjects rights	21	2

We added to our existing data the GDP per citizen and GDP per country after processing it in OpenRefine, together with some string manipulation in python.

76 rows

Show as: rows records Show: 5 10 25 50 rows

	unit,na_item,geo\time	2000	2001	2002	2003	2004
1.	CLV10_EUR_HAB,B1GQ,AT	31710	31990	32360	32520	33200
2.	CLV10_EUR_HAB,B1GQ,BE	29890	30110	30490	30680	31640
3.	CLV10_EUR_HAB,B1GQ,BG	3010	3230	3440	3640	3890
4.	CLV10_EUR_HAB,B1GQ,CH	50750	51160	50880	50530	51590

24 rows

Show as: rows records Show: 5 10 25 50 rc

All			Country	GDP_2019
		1.	AUSTRIA	38250
		2.	BELGIUM	35900
		3.	BULGARIA	6800
		4.	CYPRUS	24250

After we have cleaned the data we opened it in Openrefine and converted all values which should be numbers to a numerical value. That includes the columns: “Fine”, “Country_code”, “GDP_pc_2019”, “GDP_2019”, “ Fines_applied” .

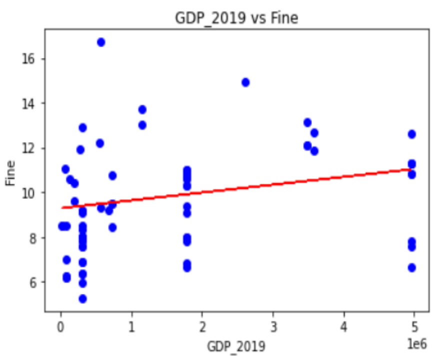
Sources of Datasets used:
GDP per capita:
https://ec.europa.eu/eurostat/web/products-datasets/-/sdg_08_10
https://datacatalog.worldbank.org/dataset/gdp-ranking

Our final cleaned data looks as following:

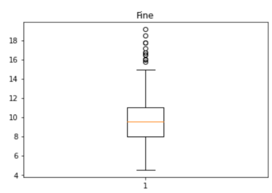
[33]:

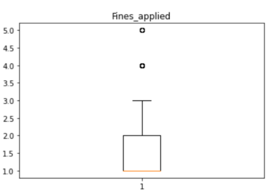
	ArticleViolated	Authority	Country	Date	Fine	Org_fined	Type	Country_code	Fines_applied	GDP_pc_2019	GDP_2019
0	Art. 32 GDPR	Romanian National Supervisory Authority for Pe...	ROMANIA	2019-11-25	11000	Courier Services Company	Insufficient technical and organisational meas...	21	1	9130	268299
1	Art. 12 GDPR, Art. 17 GDPR	Romanian National Supervisory Authority for Pe...	ROMANIA	2019-11-22	2000	BNP Paribas Personal Finance S.A.	Insufficient fulfilment of data subjects rights	21	2	9130	268299

Now when we are assured all the Data is correct we begin to apply both algorithms on our cleaned dataset which we called “data_gdpr_aftercleaning”.
First of all we want to know if our hypothesis is true so we will fit a linear regression model onto the data points which represent the country and the height of fines and also GDP and height of fines.

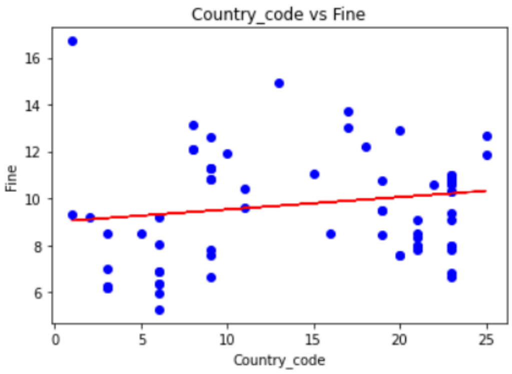


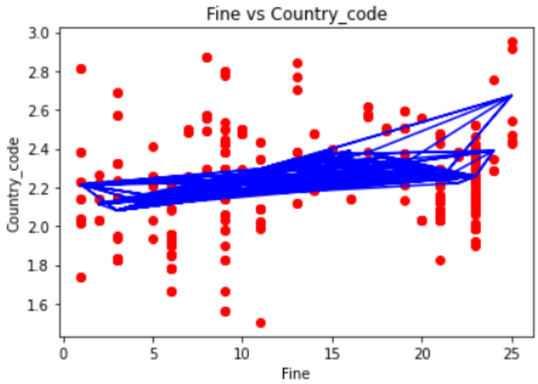
Linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables.
This regression shows a relation between the GDP of a country and the sum of the fine applied for the ones that break the laws. The higher the GDP of a country is, the higher the value of the applied fine.





Furthermore the following box plot indicates another relation between the number of articles applied when giving a fine.
It is clear to be seen that the more articles there are applied for a fine the higher that fine will be.





Our hypothesis do not fit very well. We cannot find real dependencies in the data with linear regression it does not fit as expected on the data. That's why we think both linear and polynomial regression were not the best algorithm which we could have applied. If we would have had more time, we would think about a different hypothesis to test and maybe also a different method to fit onto the data. For a future approach we could think of measuring correlations an take them as a distance for clustering. We have lost a lot of time with the data preprocessing and to get the processing environment running.