Report - Data Mining Project

Daniel Vahos

Lucas Sousa

Crimes and living situation in each french department

SD201 - Data Mining

Télécom Paris

Palaiseau, France

12/2021

I. Introduction

Economic inequality is the root of many social problems, and it can present itself in many ways: inequality due to gender, age, geographic location, company hierarchy, among others. It's essential to find out where and how inequality is most present in our society, so that the government can direct its efforts towards the situations where it'll help the most. This report analyzes statistical data on employment in different French towns, addressing the inequality between different regions of France.

For any society, the safety of its citizens is one of the most important factors for quality of life. Without security, there is no peace of mind for the inhabitants. This is why it is essential to identify the parameters that can determine or affect the security. France, despite being a much safer country than others in the world, is showing increasing crime rates in recent years, generating indisposition in its inhabitants.

In this report we analyze statistical data on crime in France and we try to answer questions, in the French context, such as: is there a real correlation between crime and inequality? Can we predict the crime rate of a given location using inequality metrics? What else can be related to the number of crimes committed in a given place?

II. Datasets

Our investigation led us to a dataset uploaded to Kaggle by Étienne LQ which contains information on number and size of companies, average salary in different categories, geographic information and demographics per French town [1]. This dataset compiles information from several reports by Insee (Institut national de la statistique et des études économiques, or National Institute of Statistics and Economical Studies), an official French institute that gathers data for several studies about the country, but the data acquired describes only the year of 2014.

Since our main question concerns crime and inequality, we searched as well for crime reports in France. We found a table from Politologue.com containing information on total number of crimes per French department per year [2] (with data provided by data.gouv.fr, the open platform for public french data) and we were able to scrap it using jQuery. For future analysis, it's possible to get the typification of crimes per department from this same source, but that would require a little more of scraping and, although it'd be interesting to check how

inequality is related to different types of crimes (the more violent ones, for instance), it's not something necessary in the scope of this report.

Furthermore, at a given point we had to guide our efforts towards a slightly different analysis, for which we could only find data from the year of 2015. For this reason, we got some data directly from the Insee website: "Structure et distribution des revenus, inégalité des niveaux de vie en 2015" (Structure and distribution of incomes, inequalities and standards of living in 2015) [3] and "Populations légales 2015" (Legal population 2015) [4].

III. Methods

III.i) Salary difference between different categories of workers

At first, we used the two first datasets aforementioned to get data on crimes per department and salary per town in France. Originally the crime reports table described the department as a string in the format "Name of the department (Department Code)", therefore we used Python's regular expressions (regex) library to separate this two information. Since we had information about salary for each town, we grouped this table by department and computed the average salary per department for 4 different categories of salaried employees: executives, middle managers, employees and workers.

In order to analyze inequality in each department, it was necessary to subtract the average of the maximum salary and the average of the minimum salaries for each department. In this case, the maximum salary always belonged to the executive category salary, but the minimum salary could be either the employees' salary or the workers' salary, so the search for the minimum value was performed to ensure the correct analysis of inequality. We called this relation SIC, for Salary Inequality Coefficient.

$$S.I.C. = \frac{\text{Max salary between categories} - \text{Min salary between categories}}{\text{Min salary between categories}}$$

This formula is equivalent to the relative error formula, and it tells us about the difference between the category with the best salaries and the one with the worst salaries in one department, in relation to the latter.

It is clear that in order to have an objective analysis of crime in each of the departments, the population of each of them must be analyzed, since it is clear that

departments with a greater number of people tend to have a greater number of crimes. For this reason, we merged the crime reports table with the population demographics and computed the rate of crimes per 100,000 inhabitants (since crime rate is usually reported per 1000 or 100,000 people).

As can be seen in the tables, salary data is not available for all departments; there are a few departments that do not have this information. Therefore, to make the union with the crime table, only the departments that are in both tables will be analyzed. This filter is done automatically by pandas' merge function.

We can see on Table 1 below the results commented so far.

	Department Name	Department	Crimes in 2014	Population	Crimes per 100000 inhabitants	Normalized Salary Difference
0	Ain	01	24605	600387	4098.190001	127.407699
1	Aisne	02	23309	514949	4526.467670	125.503356
2	Allier	03	12149	325421	3733.317764	126.555556
3	Alpes-de-Haute-Provence	04	7735	154221	5015.529662	126.920552
4	Alpes-Maritimes	06	80627	1052182	7662.837798	126.680706
...	...	...	...	...	...	...

Table 1 - Information on crime and salary difference per department

We plotted a scatter graphic to visualize how crimes and salary differences are spread between departments, as we can see below on Figure 1.

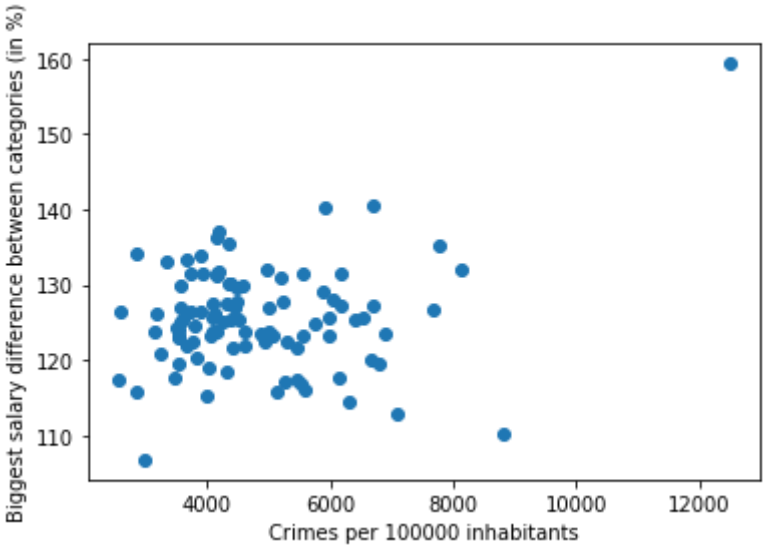

Figure 1 - Crime and salary difference distribution for different French departments

In order to graph and analyze the data it is important to first perform the Min-Max Scaler process, which seeks to scale the features to have zero mean and standard deviation of one, to convert it with properties as a standard normal distribution.

$$\text{New}X_i = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

In this case, characteristic X would refer to Crimes per 100,000 inhabitants and the inequality value per department (this process is used in each of the different analyses made on this project). For each feature, the maximum and minimum values are found, and each value is recalculated so that they are normalized between 0 and 1.

We decided to try and divide this data in clusters using the k-means algorithm to see if there are clusters that could give us insights on the crime rate and salary difference relation. To do so, it is critical to identify the number K of clusters created for a proper analysis. In order to identify the number of clusters, the elbow method was used, in which the number of inertia (sum of squares of distance of points to cluster center) is plotted against the number of possible clusters and a value for K is chosen over the "elbow" of the curve; this way, we can have enough clusters to get a low inertia value but we maintain a reasonable number of samples in each cluster. We can see the inertia plotted against the number of clusters on Figure 2 below. According to this result, an appropriate number of clusters for the analysis is considered to be 5, since the curve is steeper at this point.

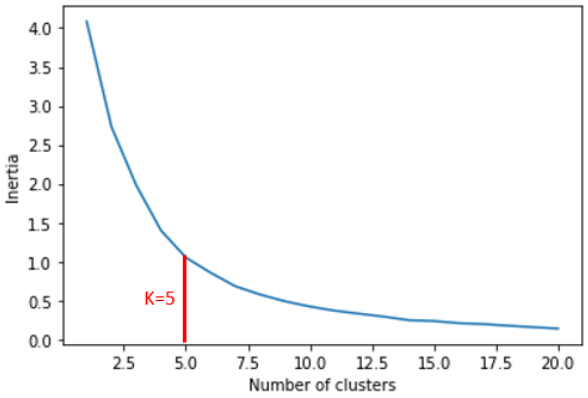

Figure 2 - Inertia curve for several numbers of clusters K

Figure 3 below shows the result of the clustering. As we could see in Figure 1, there is a clear outlier in the top right corner of the plot, which in this case is evidenced in green, well separated from all other values.

It is important to note that MaxMinScaler's scaling does not perform optimally when outliers are found in the data, which affect the scaling and therefore the analysis. Because of this, once identified, we proceed to eliminate-omit it from the study.

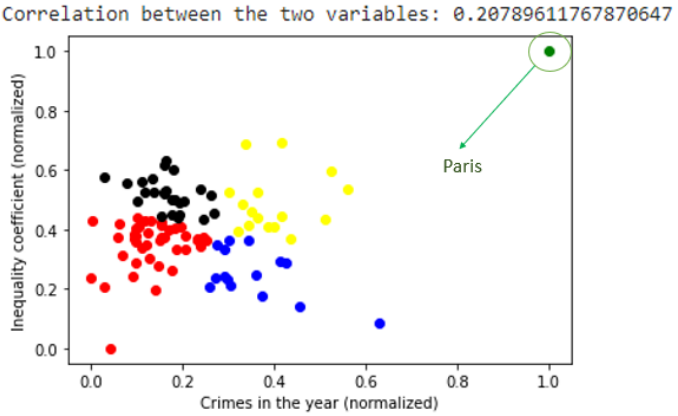

Figure 3 - Five clusters of departments, with Paris as an outlier

Upon further inspection of the data, we could see this outlier is the department of Paris. We seek to ignore the outliers to perform a deeper analysis with the rest of the data. Once this outlier is eliminated, the clustering analysis is taken into account, scaled and grouped into 5 different clusters.

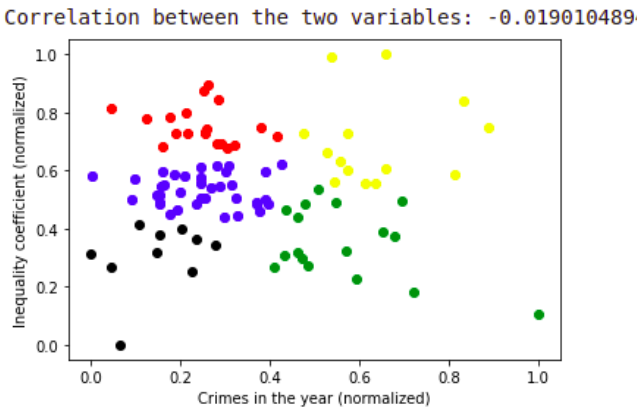

Figure 4 - Five clusters of departments, with the outlier removed

We see that after removing the outlier, the correlation drops from approximately 0.21 to roughly -0.02, which at first sight seems to reject our hypothesis of relation between crime and income inequality. Accordingly, it could be concluded that there is no prediction or modeling of crime cases with respect to inequality in different regions that can be made. However, we also do not guarantee that inequality is not related to the amount of crime in the sector.

At this point, we had to criticize our own methodology of choosing the salary difference between different categories of workers as a metric for inequality. Even though it seemed a good metric at first, it doesn't take into account the number of people with executive, manager, employee or worker salary. In other words, if there are more employees in one of these groups than in another, this would affect the result of inequity in the department, something that is totally disregarded in this previous analysis.

III.ii) Gini Coefficient

Given the failure of the methodology perceived in the previous section, we decided to take a different approach on the inequality metric, so more information was sought from similar sources. Thanks to the INSEE (Institut National de la Statistique et des Études Économiques), we were able to obtain a great deal of information on the fiscal situation in each french department. The data we found corresponds to the year of 2015, thus we got another dataset from INSEE with population information for 2015 as well.

One of the information contained in this fiscal information dataset is the Gini coefficient. As defined by INSEE (2015), "The Gini index (or coefficient) is a synthetic indicator of the level of inequality for a given variable and population. It varies between 0 (perfect equality) and 1 (extreme inequality). Between 0 and 1, the higher the Gini index, the greater the inequality". The Gini coefficient is calculated as a ratio of the area between the "line of perfect equality" and the Lorenz curve over the total area below the "line of perfect equality". In this graph, the x-axis is equivalent to the Cumulative proportion of the population variable and the y-axis to the Cumulative proportion of the income variable. The Lorenz curve, in turn, shows the percentiles of population (x-axis) according to the income (y-axis).

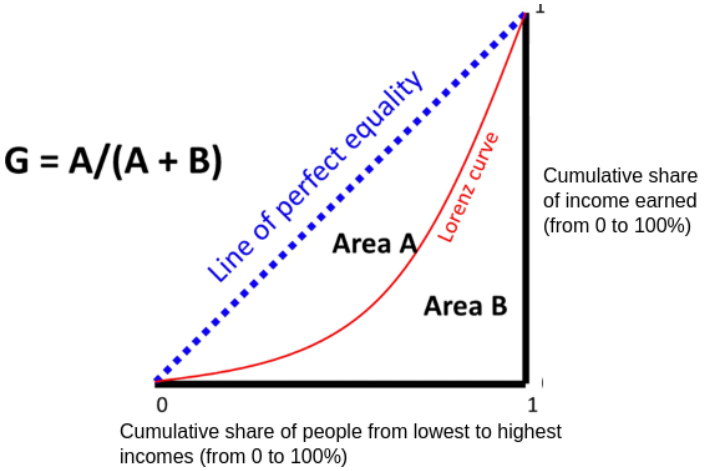

Figure 5 - Gini coefficient graph, obtained from Nature.com (2019), edited by us.

	Department Name	Department	Crimes per 100000 inhabitants	GI15
0	Ain	01	3969.264051	0.354273
1	Aisne	02	4246.835913	0.353915
2	Allier	03	3464.192068	0.329146
3	Alpes-de-Haute-Provence	04	4798.511717	0.337994

Table 2 - Information on crime and Gini Coefficient per department

Once the data was organized, separated and joined, the elbow method is performed again to identify the number of clusters suitable for the analysis. Once the process is done, the following result is obtained.

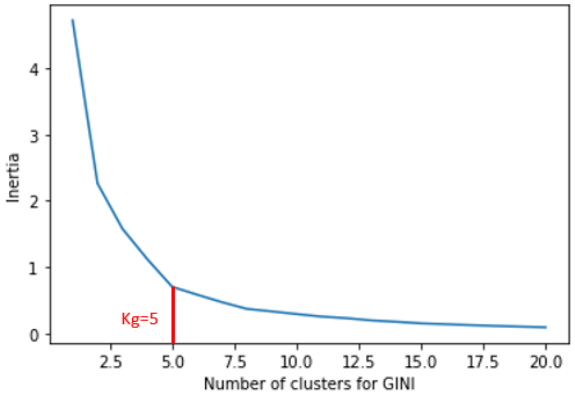

Figure 6. Elbow method for Gini analysis

As can be seen, it is determined that the appropriate number of clusters is again 5, so the clustering procedure is performed, obtaining the following result.

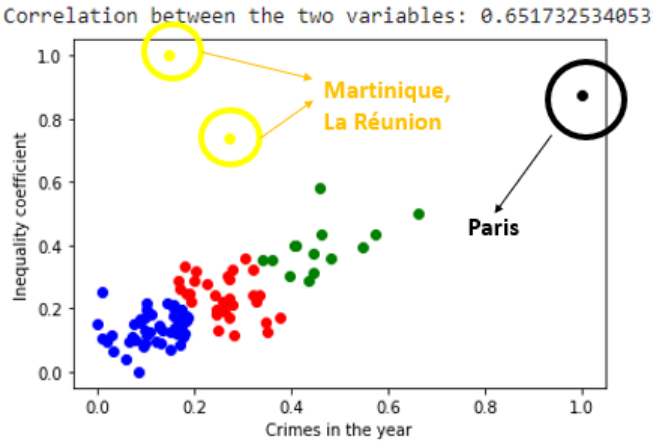

Figure 7. Five clusters with Gini coefficient and outliers

As can be seen, in this new clustering there are 3 outliers. These departments were identified in order to omit them later, and they are: Paris, which was already an outlier before; Martinique, an island in the Caribbean; and La Réunion, an island near Madagascar. These last two are islands far away from the French mainland, so it can be understood that the economy and life conditions are often very different.

We performed the elbow method again for this new data and got a value of K=?. Once again, the clustering analysis was performed omitting the outliers and the following result was obtained.

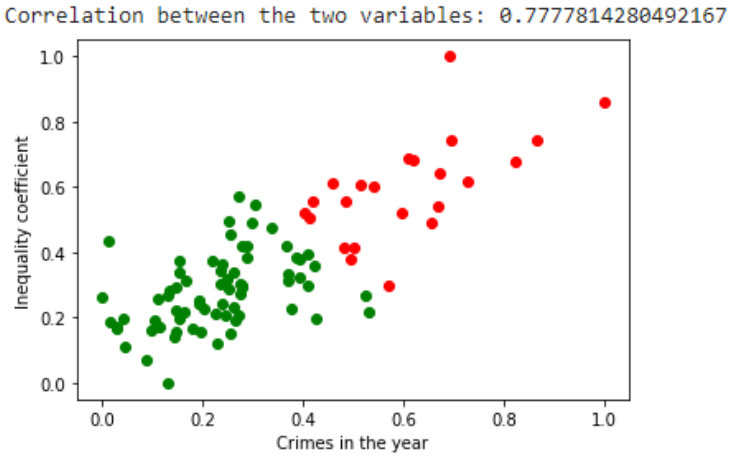

Figure 8. Cluster with Gini coefficient without outliers

As can be observed, the correlation coefficient is almost 0.8, indicating that there is a considerably strong relationship between the Gini coefficient and the crime rate in French departments, but the clusters don't seem to give us good insights about this relation.

III.iii) Predicting crime risk in departments

Knowing that the income inequality (measured by the Gini Coefficient) is strongly correlated with the crimes, we've split the French departments we analyzed in 3 different classes according to their crime rate in 2015, which we called low risk for departments with less than 4000 crimes per 100,000 inhabitants, medium risk for departments with 4000 to 6000 crimes per 100,000 inhabitants, and high risk for departments with a crime rate above the latter. These are arbitrary threshold values, but for the sake of comparison we can see that the first threshold is closer to the crime rate in South America while the latter is closer to the crime rate in Europe. This gives us an unbalanced set, since we have 43 departments in the lower risk class, 41 departments in the medium-risk class and only 12 departments in the highest risk class. We've used a naïve Bayes classifier to try and predict to which class of risk

a department belongs to. We've plotted the histogram for the crime rate and therefore decided to use a Gaussian Naïve Bayes.

At first, we chose as input for our classifier solely the Gini coefficient for each department and we ran our algorithm 1000 times to get an average accuracy, and we used 50% of the data for training and the other 50% for testing. We got a total accuracy of 60.0%, and the recall score for the classes aforementioned were 65.3%, 65.0% and 33.9%, respectively. As expected, we have a terrible prediction for the highest risk class, since it corresponds to only one eighth of our data (while each of the other classes correspond to more than 40% each).

As stated in the introduction of this report, we sought to find different possible relationships of the number of crimes in France. Thus, we decided to try and improve the accuracy of our classifier using other kinds of data. At first, we tried verifying if there's a correlation between geographical localization and the Gini coefficient. For all the departments, we've computed the correlation between inequality and latitude, inequality and longitude, and inequality and distance to Paris. All those three parameters showed weak or close-to-none correlation with the inequality, so it wouldn't really help improve the accuracy of the classifier.

Moreover, another interesting parameter to analyze and predict crime rate in a given region is the age distribution. The Pennsylvania State University professors Jeffery Ulmer and Darrell Steffensmeier's article on "The Age and Crime Relationship" points several sources claiming a relation between crimes and age of the perpetrators:

"Today, the peak age-crime involvement (the age group with the highest age-specific arrest rate) is younger than 25 for all crimes reported in the FBI's UCR (Uniform Crime Report) program except gambling, and rates begin to decline in the late teenage years for more than half of the UCR crimes. [...] In fact, a significant portion of U.S. national crime rate trends over time can be explained by fluctuations in the proportion of the population in the crime-prone age group of 15-to-24-years-old (Steffensmeier & Harer, 1987, 1999)."

We then defined the crime prone age group from 15 to 24 years old, computed the proportion of individuals from each age group in each French department and combined it with the Gini Coefficient information to feed our Naïve Bayes Classifier. We've got an average accuracy of 69.5%, while the recall score for the three classes, in order of increasing risk, were 72.2%, 69.4% and 63.6%. This indicates a very positive evidence pointing towards a certain determination between the number of crimes and the age of the people in each department.

It's important to note that we weren't able to find data on demographics for the year of 2015 (we're using 2015 data for crime rate and Gini coefficient), so we decided to use the demography information of 2014. This generates some uncertainty to our results, since the number of people from each age group defined varies from one year to the other, but we assume that this difference is small enough in one year so that we can consider the distribution roughly the same for 2014 and 2015.

IV. Conclusion

We found out that the salary difference between different categories of workers isn't a good inequality indicator, since it doesn't take into account how many people fall into each category, and this indicator has no correlation to crimes in a region. The Gini coefficient, on the other hand, takes the number of people in each income percentile into account, and therefore is much better as a metric for inequality in a region, and we were able to use it to conclude that there's indeed a strong correlation between inequality and crime rate in the departments of France. We concluded that, given the nature of our data, clustering can't provide much insight about the relations between crime and inequality; a Naïve Bayes Classifier, however, can determine with a much better than random accuracy if a department belongs to a low, medium or high risk category, but will make mistakes around 4 out of 10 for our unbalanced dataset.

We were able to find as well that there's no correlation between inequality and geographical position variables in the French departments (other than in the cases of the oversea departments), and we reproduced Ulmer & Steffensmeier conclusion that there's a crime-prone age group so that demographic characteristics of a region can help predict the crime rate in this region. It is known that crime-rate is related to many variables, but we can conclude that age and inequality are very significant factors.

V. Workload distribution

Although responsibilities were attributed to both members, we did most of the work together at face-to-face meetings at CRDN or in audio conferences. Lucas was responsible for gathering the data and the Naïve Bayes classification part, while Daniel was responsible for processing the data and the clustering part. Throughout the work, both parties shared feedbacks to help improve code comprehension and performance, as well as to stay on the same page regarding the work that was being done.

Bibliography

[1] Kaggle. (2014). *French Employment by Town*. Obtained from: https://www.kaggle.com/etiennelq/french-employment-by-town

[2] Politologue. (2019). *Crimes et délits*. Obtained from: https://crimes.politologue.com/departements/

[3] INSEE. (2015). *Inégalité des niveaux de vie*. Obtained from: https://www.insee.fr/fr/statistiques/3560118

[4] INSEE. (2014). *Statistiques*. Obtained from: https://www.insee.fr/fr/statistiques/3545833?sommaire=3292701

[5] Jeffery T. Ulmer, Darrell Steffensmeier. (s.f.). *The Age and Crime Relationship*. Obtained from: https://www.sagepub.com/sites/default/files/upm-binaries/60294_Chapter_23.pdf

[6] Wikipedia. (s.f.). Obtained from: https://fr.wikipedia.org/wiki/Coefficient_de_Gini