Final Report

Julien DE VOS, Esilv and Uvic student

Mathieu COWAN, Esilv and Uvic student

Enki MILLET, Esilv and Uvic student

1) INTRODUCTION

Nowadays, the population is increasingly trying to make their travels easier, quicker, and more efficient. It is a daily occurrence for individuals in active life to move around, and this often involves the use of cars or other vehicles. People have various reasons for choosing to travel by car, but this also comes with the risk of accidents. Every year, there are numerous accidents, and they vary in terms of their causes, severity, the type of journey, the region, and information about the driver. As it is a concerning issue, we have decided to conduct a study on accidents in France. Specifically, through this project, we aim to address the following question: what environmental factors influence French road accidents and how do they impact them, and what can we do to diminish as much as we can the impact of those environmental factors, whether it is weather, road quality or even visibility.

2) MOTIVATION (WHY):

The impetus for this project arises from a critical need to elevate road safety standards. Our endeavor is to meticulously explore the root causes of severe accidents, providing invaluable insights that can steer the development of targeted preventive measures. Recognizing high-risk locations for severe accidents becomes imperative for optimizing resource allocation and strategically implementing interventions. Moreover, the pursuit of predicting accident severity holds transformative potential, offering a paradigm shift in accident response strategies and bolstering the evolution of advanced driver assistance systems.

In the realm of vehicular accidents, a myriad of studies has been conducted. However, our focus extends beyond the general exploration of accidents to a specific inquiry into the influence of environmental factors on French road accidents. We seek to dissect the impact of various elements, such as weather conditions, road conditions, and visibility, on accident severity.

To underline the significance of our inquiry, we refer to two articles [1][2] providing statistics on the percentage of severe accidents attributed to adverse weather conditions, and another article who give statistics about specific roads impact on accident gravity. These articles not only reinforce the relevance of our investigation but also underscore the real-world consequences of environmental factors on road safety.

To substantiate our research, we will leverage the dataset available at Kaggle [3] - Accidents in France (2005-2016). This dataset encompasses a comprehensive record of accidents and associated variables, providing a robust foundation for our exploration.

3) OBJECTIVES (WHAT):

Concerning the causes of severe accidents, we have as objective to identify key factors which contribute to severe accidents. In a general standpoint, we can affirm that we want to principally analyze factors which are in range of institutional action and prevention.

In the first part, we want to establish a comparison between the potential correlations between the factors and the accidents so that we can analyze which elements are the most needed to be treated.

Furthermore, we want to determine geographic areas with a high incidence of severe accidents. We can then provide actionable insights for targeted safety interventions in identified hotspots.

Finally, in the aspect of predicting accident severity, a good objective would be to assess the feasibility of predicting accident severity using machine learning models.

Hypothesis we may issue is than in addition to what seems trivial factors such as the non-respect of security rules linked to assets such as seat belt, harsh weather, especially rain should be a prominent factor concerning severity of accidents.

4) APPROACHES (HOW):

Introduction: Data plot

- Recuperation of the data
- Data sorting
- Data plot
- Study of the number of accidents per year

1st part: What causes the severity of the accidents?

Through regressions we can study:

- Correlation matrix to see the correlations between the different features in the data
- Principal Component Analysis (PCA)

2nd part: Where do the accidents (with high severity) happen the more often?

- Multiple maps to observe the data
- Clusters concerning the accidents' location (KMeans and DBSCAN clusters)

3rd part: Can we predict the severity of an accident?

- Comparison of multiple classification models to determine which one is the most efficient in predicting the severity of an accident.

In conclusion, this methodological framework is strategically crafted to foster a comprehensive comprehension of severe road accidents, intricately navigating through their causative factors, spatial distribution, and delving into the prospects of predictive modeling for accident severity. The interdisciplinary synergy of statistical analysis, geospatial tools, and machine learning techniques is harnessed to yield actionable insights, constituting a significant stride toward enhancing road safety.

5) FIRST RESULTS

DATA FILTERING

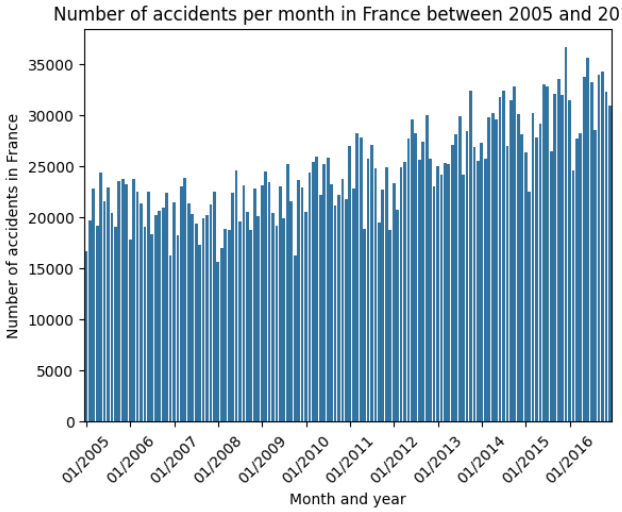
Before using the data and doing our first analysis, we have imported it in the code and filtered it. As we have different csv files containing the data, the first thing to do was to merge them together. After that, we filtered it by selecting only the columns we are interested in. In this way, here is the table we obtained with the 12 features that we kept:

	Year	Month	User category	Severity	Sex	Year of birth	Trip purpose	Securiy	Luminosity	Weather	Type of road	Road surface
0	16	2	1	1	2	1983.0	0.0	11.0	1	8.0	3.0	1.0
1	16	2	1	1	2	1983.0	0.0	11.0	1	8.0	3.0	1.0
2	16	2	1	3	1	2001.0	9.0	21.0	1	8.0	3.0	1.0
3	16	2	1	3	1	2001.0	9.0	21.0	1	8.0	3.0	1.0
4	16	3	1	3	1	1960.0	5.0	11.0	1	1.0	3.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...
3553971	5	12	1	4	1	1990.0	5.0	23.0	1	2.0	4.0	1.0
3553972	5	12	1	4	1	1990.0	5.0	23.0	1	2.0	4.0	1.0
3553973	5	12	1	4	1	1990.0	5.0	23.0	1	2.0	4.0	1.0
3553974	5	12	1	4	1	1951.0	0.0	13.0	5	2.0	2.0	2.0
3553975	5	12	2	4	2	1946.0	0.0	13.0	5	2.0	2.0	2.0

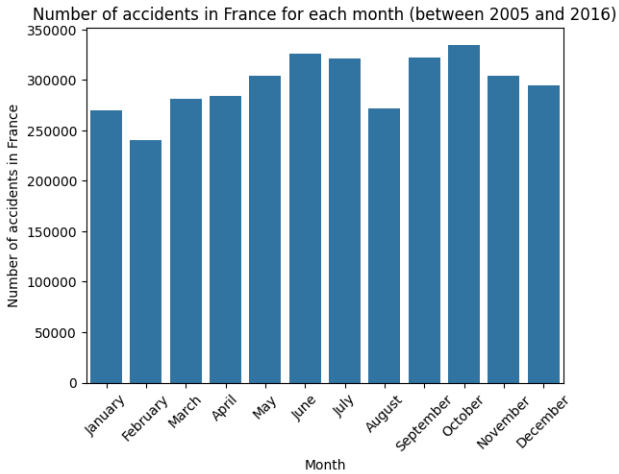
As an introductory work, we decided to have a look at the data. Therefore, we plotted the number of accidents that occurred in France along time.

PLOTS

First, we plot the number of accidents that occurred each month from January 2005 to December 2016:

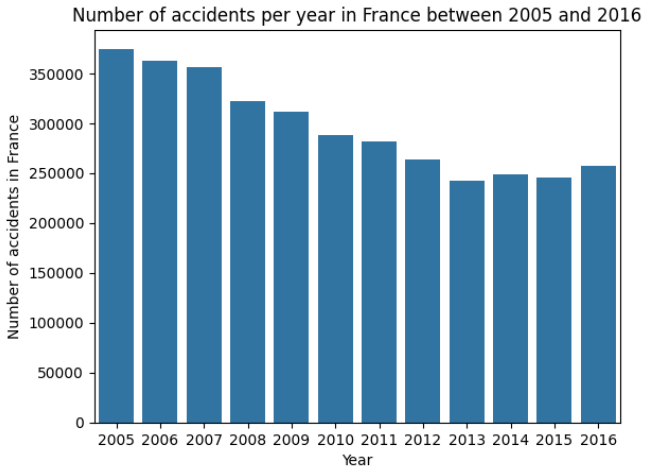


We can observe that the number of accidents in France tend to increase along time. To verify this hypothesis, let's plot the number of accidents through years:



Paradoxically, we observe that the number of accidents per year decreases: from above 350000 in 2005 to approximately 250000 in 2016. If we look again at the first plot, we can observe that during the first years, the number of accidents through months is more regular than in the last years (where there are a lot of peaks). Then, maybe this contradiction is because in 2016 there are months during which the number of accidents is way bigger than during the others, so we have the impression that in the first plot the numbers of accidents increases whereas in reality it decreases through years.

Let's now look at the number of accidents for each month:



We could have thought about several hypothesis such as the fact that there will be a lot of accidents during the summer holidays (July and August) because more people are using the car to go in vacation, or the fact that in winter there are much more cars on the roads so it leads to more accidents, but we can see that these hypotheses are false. We observe that the peaks of accidents are reached in June, July, and October while the lowest numbers of accidents are reached in February and August.

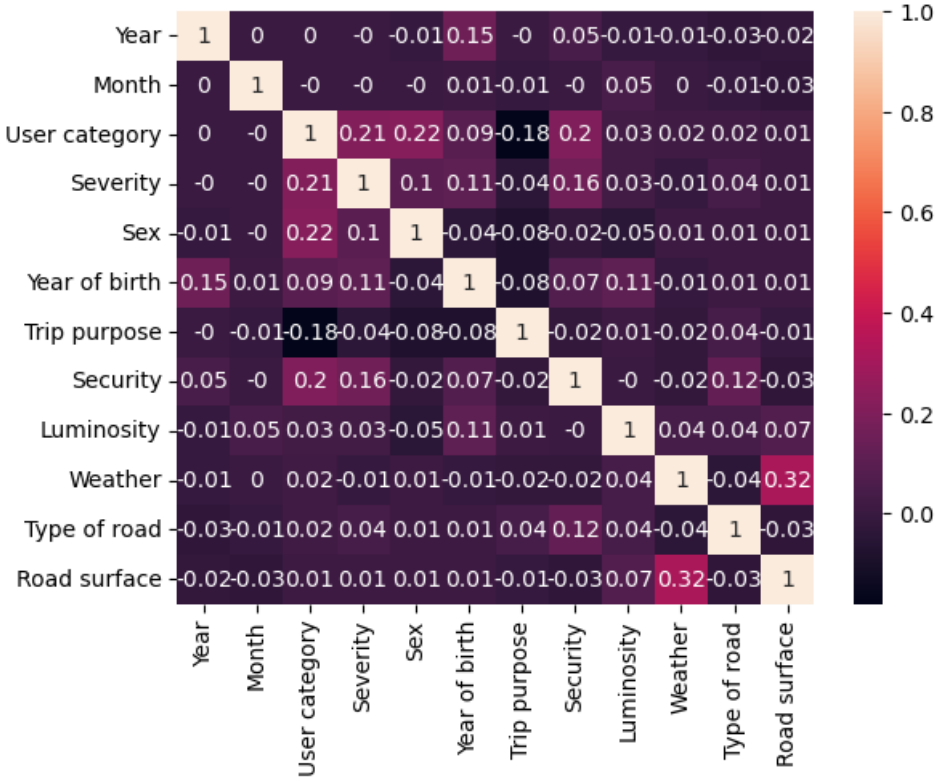
In the next parts of this project, we will try to explain more precisely all these values and try to see if natural condition (like weather, road surface, luminosity, etc.) have an influence on them.

PART 1 – CORRELATIONS

In this part, we will try to study the correlations between the weather and the severity of the accident, the luminosity and the severity of the accident, the age and the severity of the accident, the sex of the conductor and the severity of the accident.

Using the columns of the dataset that interest us and that are connected to an accident and its severity (Year, Month, User category, Severity, Sex, Year of birth, Trip purpose, Security, Luminosity, Weather, Type of road, Road surface), we make our analysis.

CORRELATION MATRIX

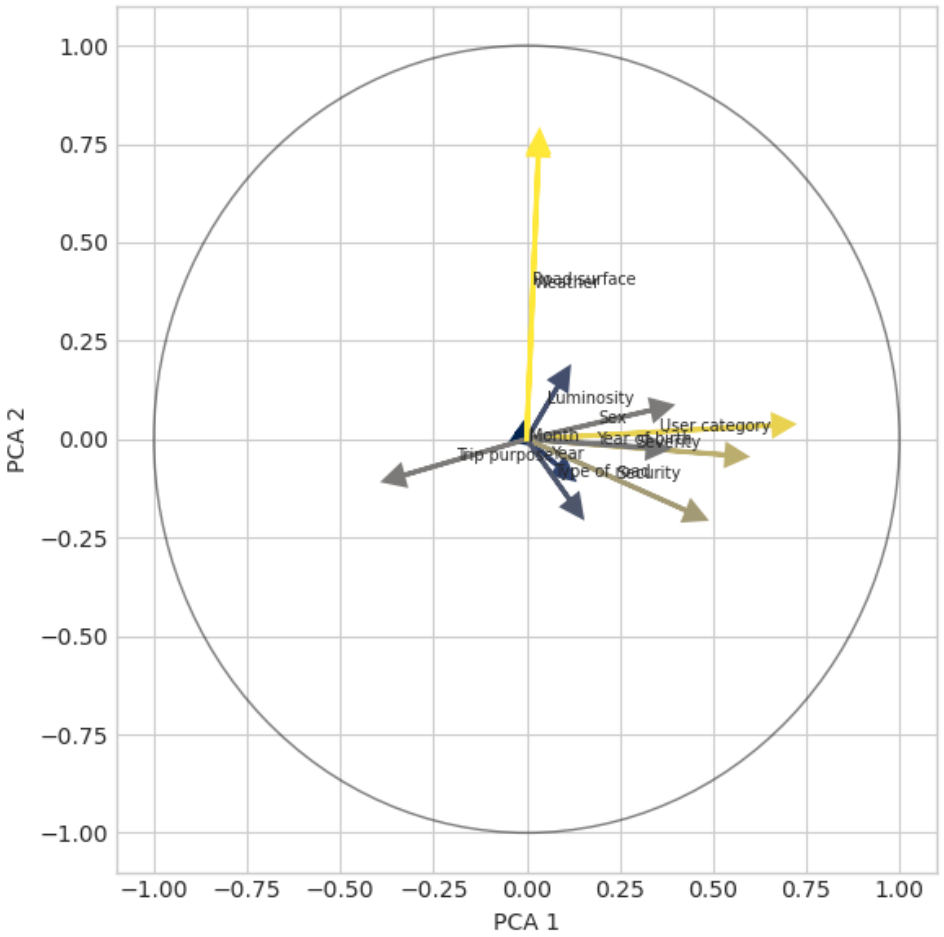


We worked with a correlation matrix to first make an initial study of the dataset to try to answer our problematics. We thus obtain:

We see here that the correlations coefficient aren't very high, however we still see notable positive correlation between the user category and the severity of the accident, the sex of the driver and the severity of the accident, the year of birth and the driver and the severity of the accident, the security precautions and the severity of the accident, as well as a negative correlation between the severity of the accident and the purpose of the trip.

We see other notable correlation between the weather and the road surface, which seems logical. Surprisingly, we don't see a direct linear correlation between the weather or the luminosity with the severity of the car accident. Let's apply the PCA to have a more detailed analysis. Let's work on the PCA part to make a more detailed analysis.

PRINCIPAL COMPONENT ANALYSIS (PCA)



We make a projection on a plan, which means that for some data we lose some information. Here, the more yellow it is, the less we will lose some info. At the inverse, the bluer it is, the more we will lose some info. What we do is that we make a projection on the plan. Therefore, we have the distance and the color for each arrow, and we know could identify and quantify the correlations because the colors indicate in a graduating way how much the variables are dependable.

We see that we obtain the same observations for the correlation matrix is that the severity of the accident is dependent and positively correlated to the sex of the driver, the security that is in place and the year of birth. However, it is negatively correlated to the trip purpose. It has approximatively the same correlation (confirmed on the matrix) because it is a "neighbor color" to the color for the arrow linked to the severity of the accident.

We can make another observation: we see that the road surface and the weather have their arrows which are confounded, which implies that they are directly dependent the one to the other, which makes sense when we think of real situation.

We could also make the same observations of the negative correlation between the scale of the trip purpose and the year of birth as well as for the security put in place for example.

PART 2 – CLUSTERS

In this part, we will first display maps representing the accidents, then we'll perform various clustering.

DATA FILTERING

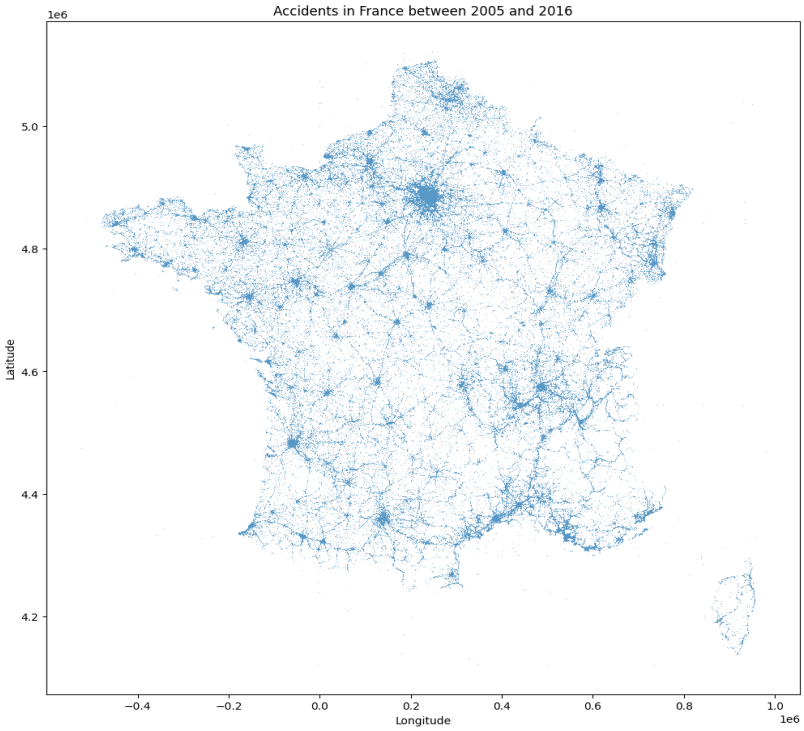
First, we filter the data to only keep the feature we are interested in: the date, the severity and the geographical information. Then, we delete wrong values in the 'Latitude' and 'Longitude' columns, and we apply conditions to only keep coordinates that correspond to accidents in mainland France. We obtain this table:

	Year	Month	Severity	Latitude	Longitude
0	16	4	1	5084579.0	226407.0
1	16	4	1	5084579.0	226407.0
2	16	4	4	5084579.0	226407.0
3	16	4	4	5084579.0	226407.0
4	16	4	1	5084579.0	226407.0
...	...	...	...	...	...
949184	5	12	2	4820100.0	-170700.0
949185	5	12	2	4820100.0	-170700.0
949186	5	12	2	4820100.0	-170700.0
949187	5	12	2	4820100.0	-170700.0
949188	5	12	2	4820100.0	-170700.0

949189 rows × 5 columns

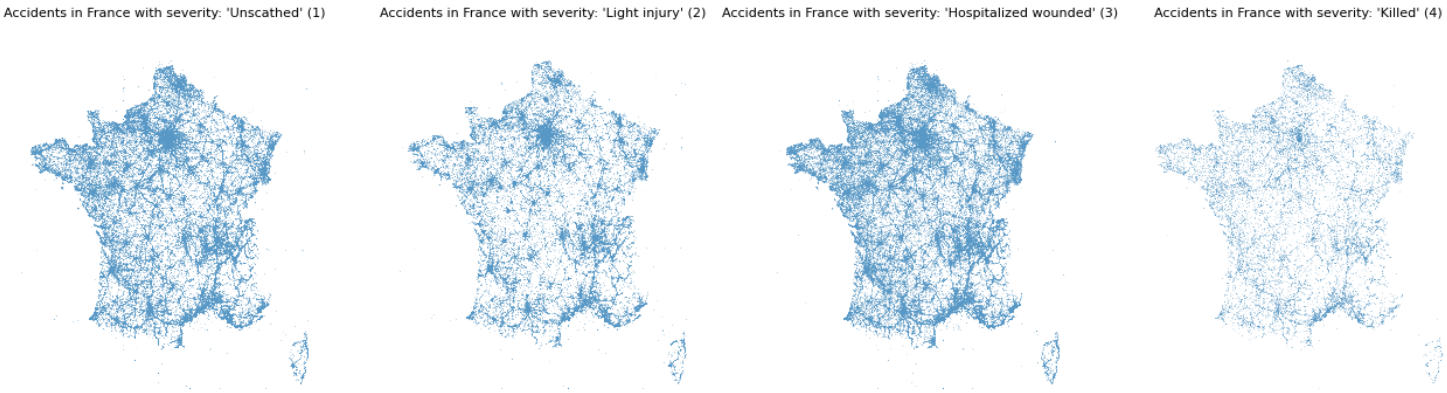
MAPS

Let's first display in simple map of the all the accidents between 2005 and 2016:



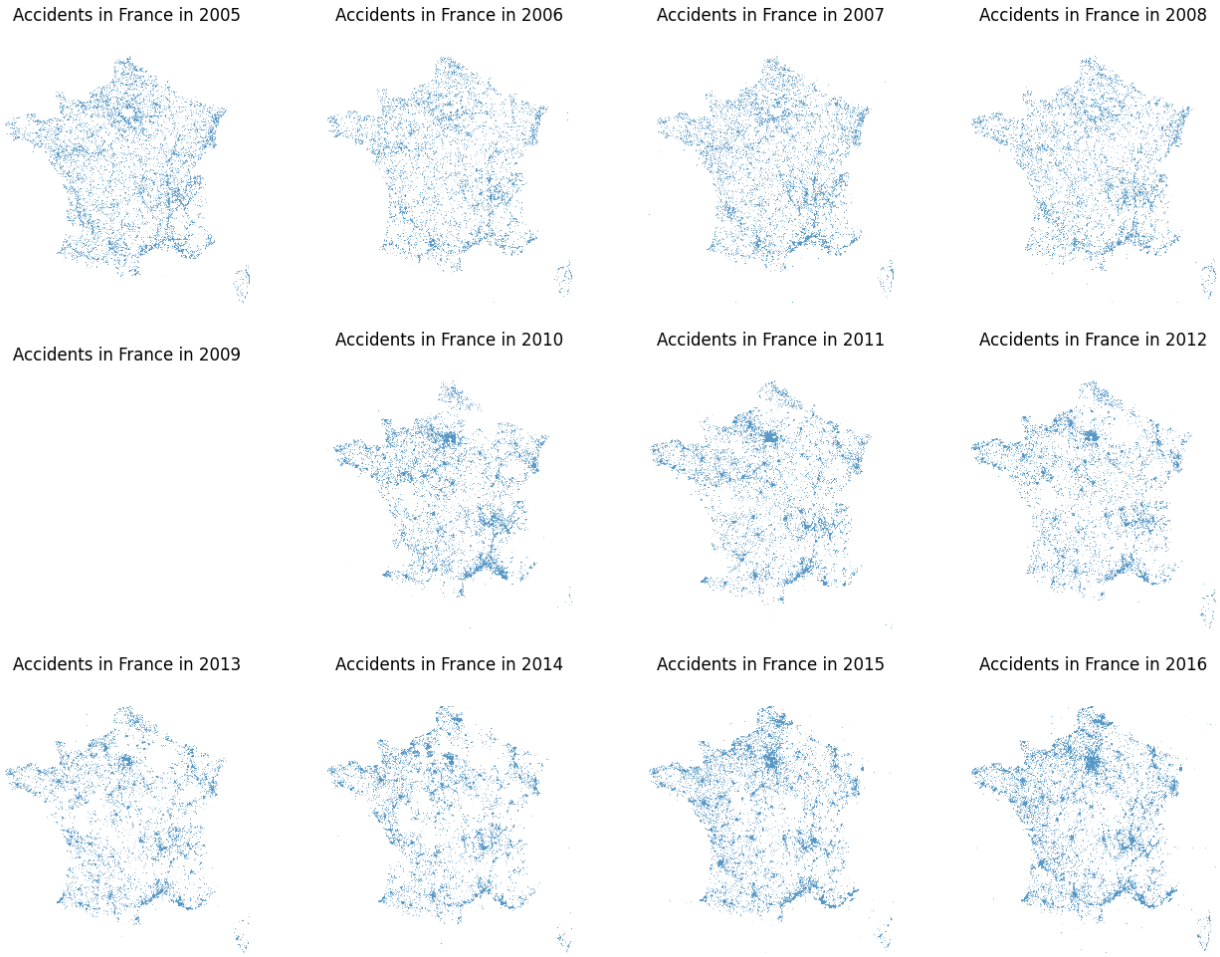
We note that a lot of accidents occur in the biggest cities: Paris, Marseille, Lyon...

Now, let's display a different map for each severity coefficient. This way, we will see if the accidents' location is different according to the severity.



We observe that the location is not linked to the severity: accidents both severe and non-severe accidents occur in cities and on roads.

Let's try to display a map of the accidents for each year between 2005 and 2016. This way, we will see if the accidents' locations change over years.

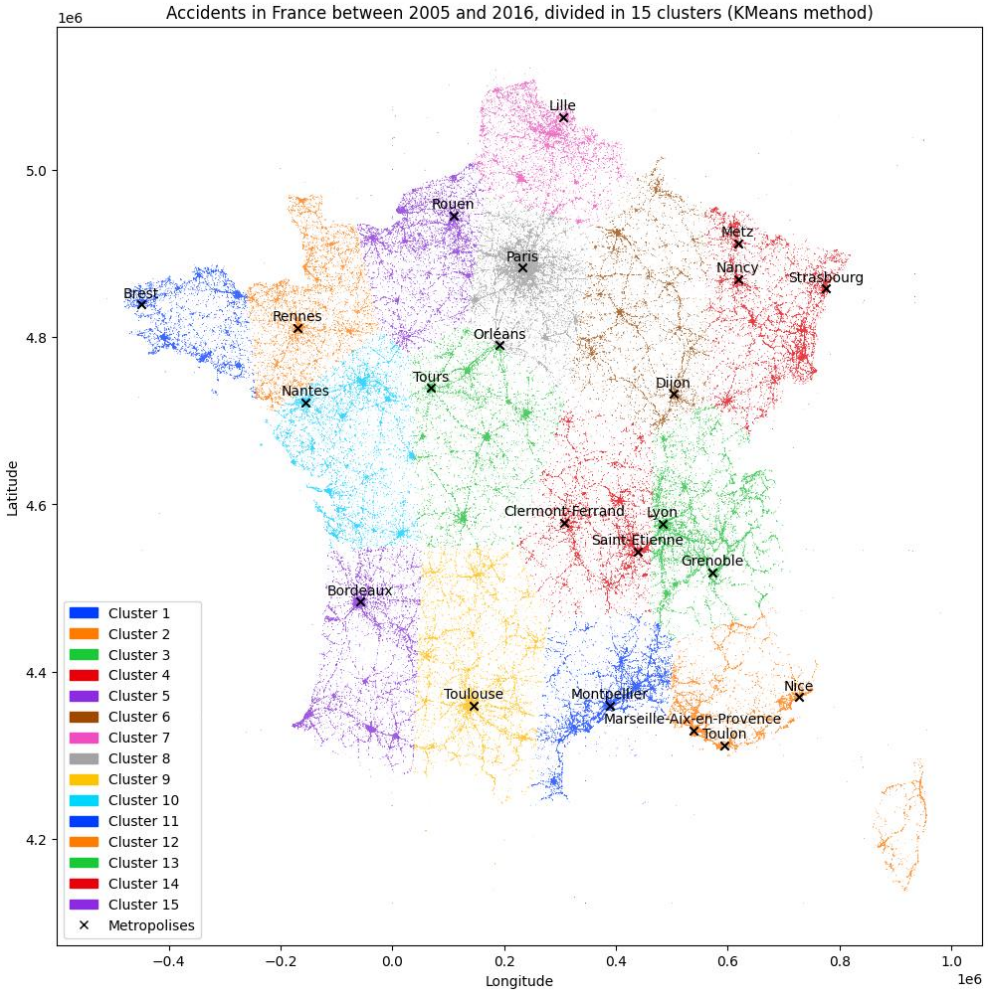


At first glance, we can think that there are more accidents in the cities over years, but we must be careful about the fact that we have not the same amount of data per year. For example, the 2009's accidents have no information about location, so we have no map. Therefore, the fact that there are more accidents in the cities is maybe because we have more data in the most recent years.

In any case, we can note from the previous two blocks of code that the accidents are grouped around France's biggest cities. We will try to confirm this hypothesis thanks to clusters.

CLUSTERS

To verify our hypothesis, we will first use KMeans clustering method to divide the accidents into 15 clusters. In addition to that, we will plot on the map the 22 French metropolises.

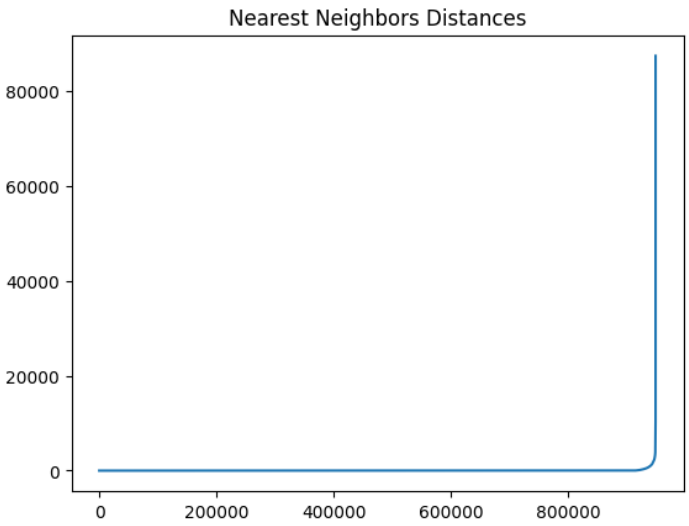


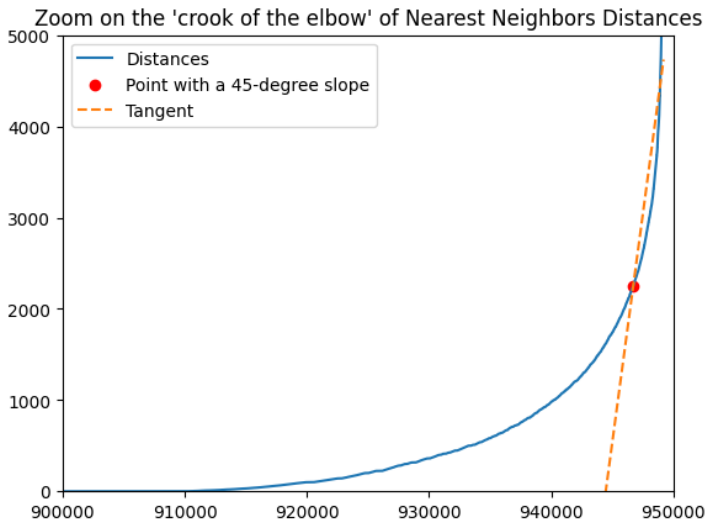
Here, we clearly see that the accidents are grouped around the different metropolises. However, as we didn't know what number of clusters was optimal, we chose 15 arbitrary. Indeed, the KMeans method needs the number of clusters as input to work. To be more rigorous, we can choose a method that determines by itself the optimal number of clusters, such as the DBSCAN method.

The DBSCAN uses different parameters than the KMeans method. The two main parameters are epsilon and the minimum samples. The epsilon corresponds to the maximum distance between two points for them to be considered part of the same cluster. The minimum samples value is the minimum number of points that are needed to form a cluster. As we want the clusters to be cities, we fix it to 300 points.

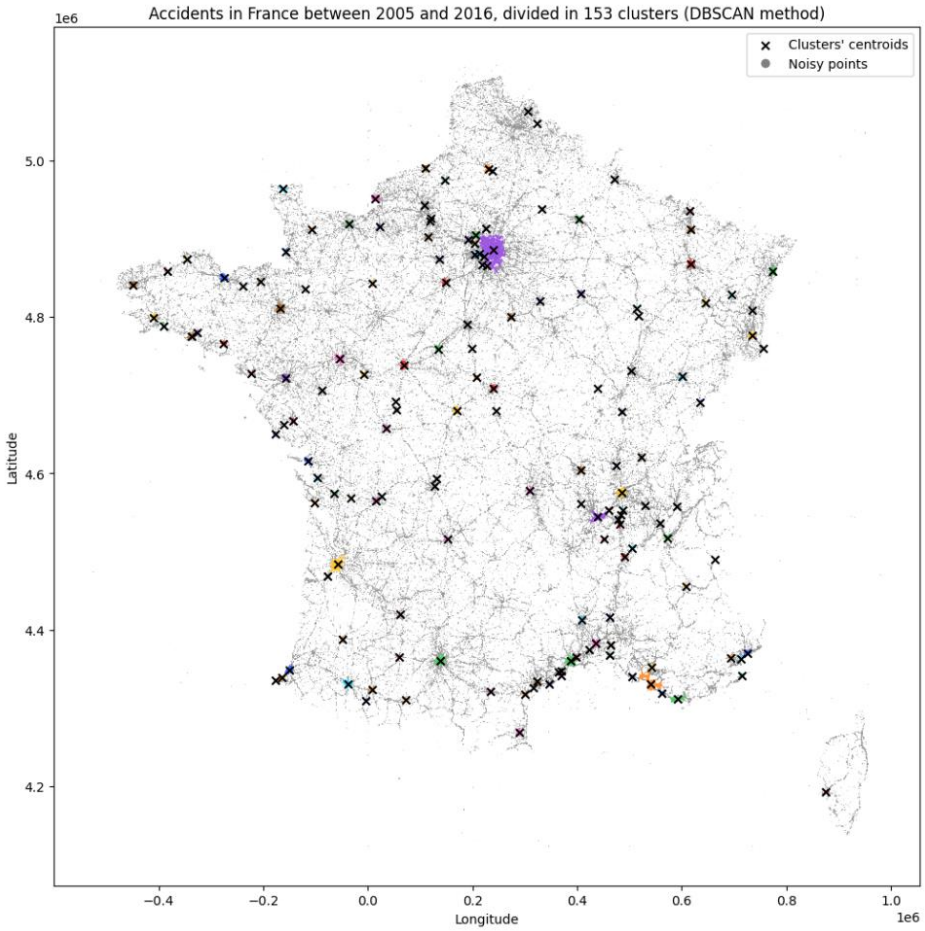
To find the optimal epsilon value, we use the method described in this article. First, we use the Nearest Neighbors class to calculate the average distance between each point in the dataset and its nearest

neighbors. Then, we sort it in ascending order, and we plot it. The ideal value will be equal to the distance value at the “crook of the elbow”, or the point of maximum curvature.





Now that we have the optimal epsilon value, we can apply the DBSCAN clustering method:



This map confirms our hypothesis: the clusters' centroids are located on the biggest French cities, so this is where most accidents occur, particularly around Paris and on the mediterranean coast.

PART 3 – PREDICTIONS

In this part, we will compare several classification models to determine which one can predict the best the severity of an accident.

DATA FILTERING

	User category	Sex	Year of birth	Security	Luminosity	Weather	Type of road	Road surface	Severity
0	1	2	1983.0	11.0	1	8.0	3.0	1.0	1
1	1	2	1983.0	11.0	1	8.0	3.0	1.0	1
2	1	1	2001.0	21.0	1	8.0	3.0	1.0	3
3	1	1	2001.0	21.0	1	8.0	3.0	1.0	3
4	1	1	1960.0	11.0	1	1.0	3.0	1.0	3
...	...	...	...	...	...	...	...	...	...
3499067	1	1	1990.0	23.0	1	2.0	4.0	1.0	2
3499068	1	1	1990.0	23.0	1	2.0	4.0	1.0	2
3499069	1	1	1990.0	23.0	1	2.0	4.0	1.0	2
3499070	1	1	1951.0	13.0	5	2.0	2.0	2.0	2
3499071	2	2	1946.0	13.0	5	2.0	2.0	2.0	2

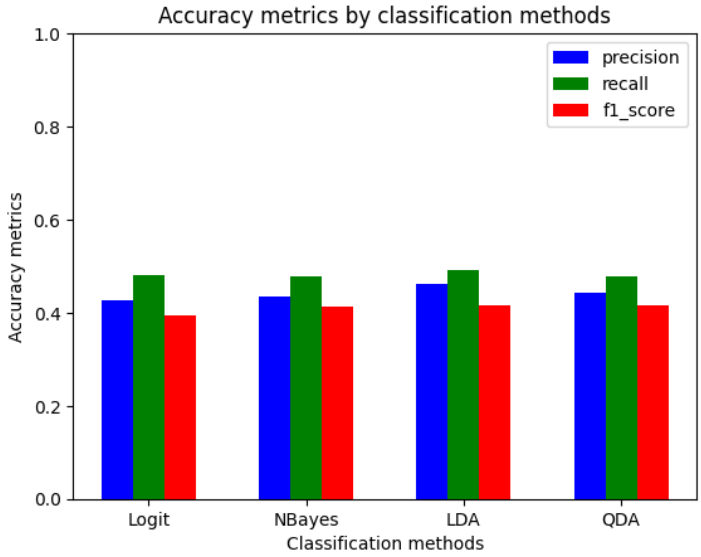
Using the correlation matrix of the **PART 1 – CORRELATIONS**, we retain the features that are the most correlated to the accident's severity. We obtain the following table:

COMPARISON OF THE DIFFERENT CLASSIFICATION METHOD

We will now divide the table into two parts: one for training the models and another for testing them. As we have a very large dataset, we choose the following distribution: 60% for training and 40% for testing. In our case, we get 2099443 rows dedicated to training and 1399629 dedicated to testing.

Now that the data is separated, we test 4 different classification models: Logit, NBayes, LDA and QDA.

Then, we plot the accuracy metrics of each method.



Name	Accuracy	Precision	Recall	F1_score
Logit	0.481	0.427	0.481	0.394
NBayes	0.477	0.434	0.477	0.413
LDA	0.492	0.462	0.492	0.415
QDA	0.477	0.444	0.477	0.417

Thus, we observe that it is difficult to predict the severity of an accident because the correlations between the severity and the features we chose are too weak. However, the best classification model is Linear Discriminant Analysis (LDA), with a success rate of almost 50%.

CONCLUSION

In this first part of our study, we tried to study the correlations between the weather and the severity of the accident, the luminosity and the severity of the accident, the age and the severity of the accident, the sex of the conductor and the severity of the accident. To do so, we used the correlation matrix and the PCA. We concluded for the problematics that we had that the severity of the accident is indeed positively correlated to the sex of the driver, the security that is in place and the year of birth. However, it is negatively correlated to the trip purpose, and we didn't see direct linear correlation between the weather or the luminosity with the severity of the car accident.

For our second part, we first displayed maps representing the accidents, then we'll perform various clustering. We came to the hypothesis that the accidents are indeed grouped around France's biggest cities. We will try to confirm this hypothesis thanks to clusters. To be precise, we used KMeans clustering method to divide the accidents into 15 clusters. At the end we realized that the clusters' centroids are located on the biggest French cities, so this is where most accidents occur, particularly around Paris and on the mediterranean coast.

For the last part, we compared several classification models to determine which one can predict the best the severity of an accident. For the comparison, we divided the table into two parts: one for training the

models and another for testing them (60% for training and 40% for testing). We tested 4 different classification models. Even though due to the weakness between the severity and the features we chose, it is difficult to predict the severity of an accident, we manage to affirm that the best classification model is Linear Discrimination Analysis (LDA).

Contributions

Introduction, Motivation, Objectives and Approach: Enki Millet and Mathieu Cowan

First Results: Julien De Vos

Part 1 – Correlations and PCA: Mathieu Cowan

Part 2 – Clusters : Julien De Vos

Part 3 – Prediction : Julien De Vos

Conclusion : Mathieu Cowan

Poster : Enki Millet

References:

[1] - SOUS LA PLUIE, SUR LES ROUTES: DEUX FOIS PLUS DE RISQUE D'ACCIDENTS]

[2] - L'état des routes, cet ennemi de la sécurité routière qui coûte de plus en plus cher]

[3] Dataset Reference: Kaggle - Accidents in France (2005-2016)

https://www.coordonnees-gps.fr/conversion-coordonnees-gps

https://fr.mapsofworld.com/lat_long/france-lat-long.html

https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd

https://iopscience.iop.org/article/10.1088/1755-1315/31/1/012012/pdf

https://scikit-learn.org

https://ryanwingate.com/intro-to-machine-learning/unsupervised/hierarchical-and-density-based-clustering/

GitHub repository

https://github.com/JulienML/AccidentsInFrance