Accident Analysis in the "Hauts-de-Seine" department

-

RAMEIX Philippe | Coursera Applied Data Science Capstone | 05/06/2020

Introduction

In high density urban area, doing his daily commute is a real puzzle. People spend a lot of time in public transport or traffic jam. Many residents are looking for alternative mobility to save time or for ecological considerations, but the risk of accident is major factor restraining the adoption of new means of locomotion.

So, Road safety is a matter of concern for this kind of people or for every parent concerned for the safety of their child.

The aim of this project is to analyse road accidents in the “Hauts de Seine” department (in France, in the Paris agglomeration) in order to provide to the resident an overview of:

- the evolution of accident statistics for the last years
- where are located the places with the more accident
- what are the accident categories?

This analysis can give some metrics highlight where road security should be enhanced.

DATA COLLECTION AND CLEANING

Data Sources

This project is based on the analysis of the open data provided by the “Hauts de Seine”. This file contains information about all accident resulting in death or injury in the department from 2006 to 2017.:

I will also use the following GeoJSON file that defines the areas/boundaries of the town to draw choropleth map:

https://github.com/gregoiredavid/france-geojson/blob/master/departements/92-hauts-de-seine/communes-92-hauts-de-seine.geojson

Data Cleaning & Feature Selection

The file « accidents-corporels-de-la-circulation-routiere.csv » contains 139 columns.

It provides a lot of information about:

- the localization of the accident
- the vehicles involved (number, type)
- the context (luminosity, type and state of the road, date and hour etc…)
- accident description (collision type, type of maneuver, attitude of people)
- accident consequences (severity, number of people injured detailed by type of vehicle)

This file is very rich, but some data are missing or gives too much detail for our needs. So, I first simplify this file to target the following main topics:

- The severity of the accident: light/serious/lethal
- Does it occur during the day or the night?
- Type of vehicles involved (including pedestrian)
- Driver was under the influence of alcohol or drug?
- How evolve the number of accident during this period?

This information is given by the following columns:

- DATE_1 : accident datetime (created by concatenation of the DATE_ and HEURE initial columns
- ID_PV : police id for this accident
- LUMINOSITE,'COND_ATMOS':
- TYPE_COLLI: collision type
- TYPE_ACCI: severity of the accident
- ETAT_USA1,'ETAT_USA2': information about the drivers and passengers (drunk for exemple)
- NB_USAGER: number of people
- NB_VEH: number of vehicles
- NB_PIE: number of pedestrians
- NB_VEL: number of cyclists
- NB_CYC: number of moped (scooter for example)
- NB_MOT: number of motorcycles
- NB_VL: number of "light" vehicle (car)
- NB_PL: number of "heavy" vehicles like truck
- NB_TC: number of other vehicles type

And of course, in order to display information on maps I extract the latitude and longitude of each accident and store them in the corresponding columns.

Data transformation:

I use the one hot encoding technique to transform several categorical values in “dummies” column:

- TYPE_ACCI: one hot encoding to create the 'LIGHT','SERIOUS','LETHAL' columns (accident severity)
- TYPE_COLLI: transform the French categorical description in the following columns. It describes the type of collision that occurs.

```
•  SIDE_COLLI          10897
•  OTHER_COLLI          6937
•  REAR_COLLI           4399
•  WITHOUT_COLLI        4077
•  CHAIN_COLLI          1015
•  MULTIPLE_COLLI        960
•  FRONT_COLLI           957
```

Creation of DRUNK_DRUG column from the categorical value ETAT_USA 1 & 2 (driver state). These columns contain a description of the state drivers involved in the accident (tired, drunk, malaise etc). I only keep the drunk/medic/drug values in order to display the localization of this type of accident.

I wanted to make distinction between accident that occurs during the day or the night.

The LUMINOSITE column (means luminosity) gives good information but is not filled for all record. So, I find more reliable to create two columns DAY/NIGHT (value 0/1) by checking the date and hour of each accident.

METHODOLOGY

Evolution and Geographical distribution of the accidents

First, I wanted to see how the accidents are distributed geographically. So, I started by creating a map. Because the important number of item (31725) I had to use the folium `MarkerCluster` function to avoid memory problems.

Then the next topic of my project is: How evolve the security in the "Haut-de-Seine" towns from 2006 to 2017? Are these towns more secure now that ten years before?

To answer that question, I had to create a dataframe that enable us to compute:

- The evolution year by year of the number of accidents for the whole department
- The evolution year by year of the number of accidents for each town
- The total number of accidents for each town during the period
- The number of accidents for the last year (2017) to observe if the ranking is the same that for the whole period

I decided use bar / line plot and to create a choropleth map to answer these questions.

The plot (cf: image in the Results sections) has shown a group of 3 towns that has a more important number of accident than other cities.

By using foursquare I have searched information about these 3 cities to determine if they have the same kind of "profile".

Clustering

The next point is: Is there some hidden groups in this file that share common characteristics.

I have done two clustering operation by using the K-Means algorithm:

- One with the columns describing: the number of vehicles, day/night, the severity of the accident, the types of vehicles involved, the drunk state and the type of collision.
- The second one, simpler, without the type collision

Each I used the shoulder technique to determine the best number of clusters to use.

Additional Maps

Finally, I wanted to see if the drunk or drug usage is responsible of many accidents and if there is specific dangerous localization. So, I decided to draw two maps:

- A map the localize the accident which involve alcohol
- A map of the lethal accidents

RESULTS

Evolution and Geographical distribution of the accidents

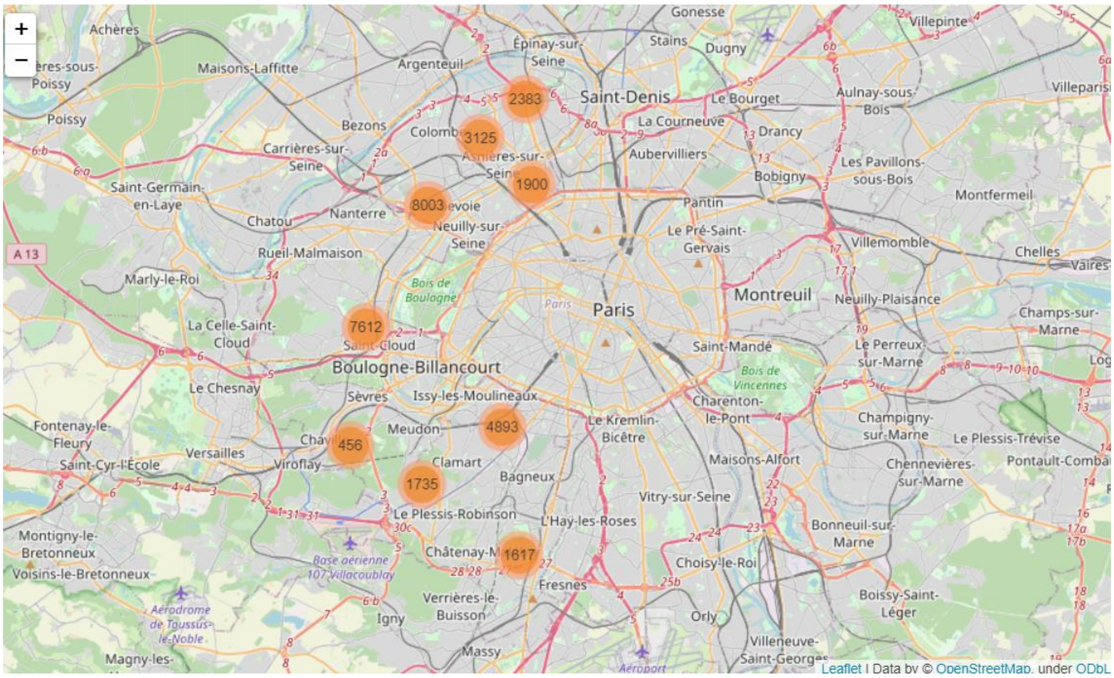
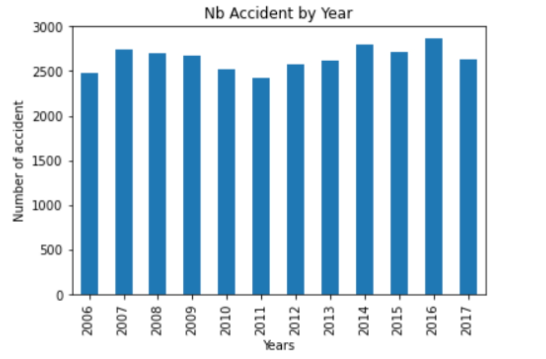


Figure 1Accident localization

	We can observe that between 2006 and 2017 the number of accident alternate increase and decrease period with a final number in 2017 which is quite the same as in 2006.

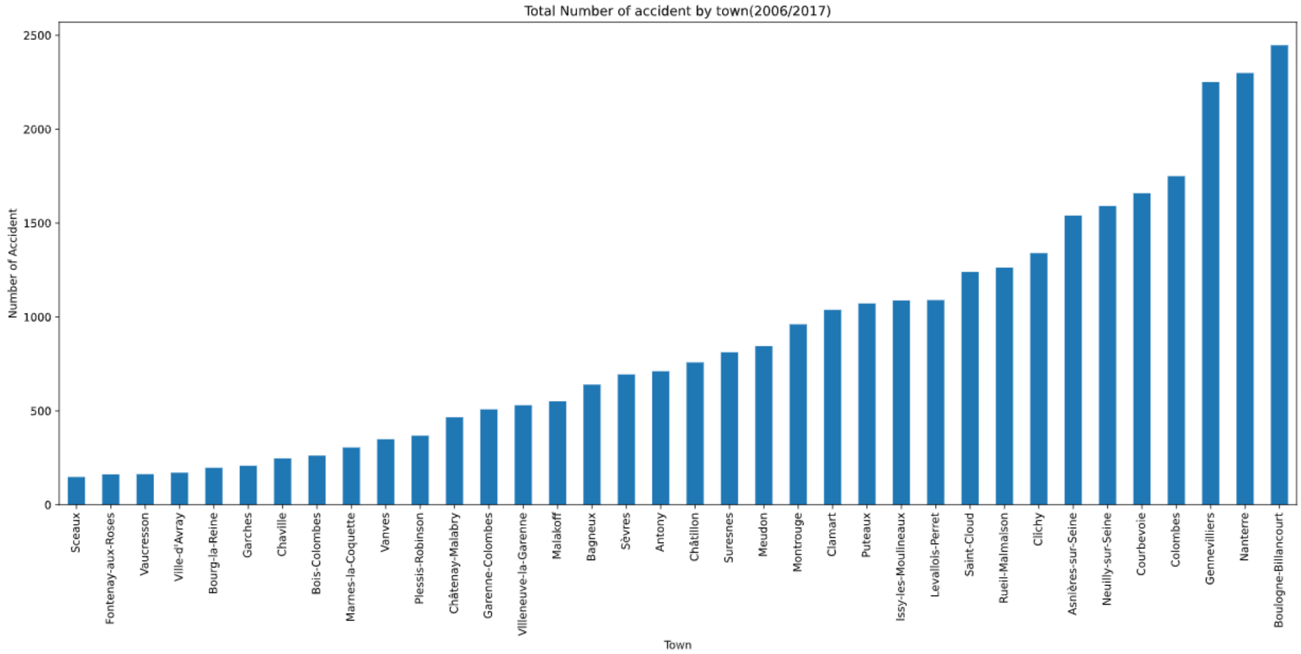


Figure 2 Number of accident by town

A group of 3 town (Boulogne-Billancourt, Nanterre, Gennevilliers) stand out from the other towns with a higher number of accident.

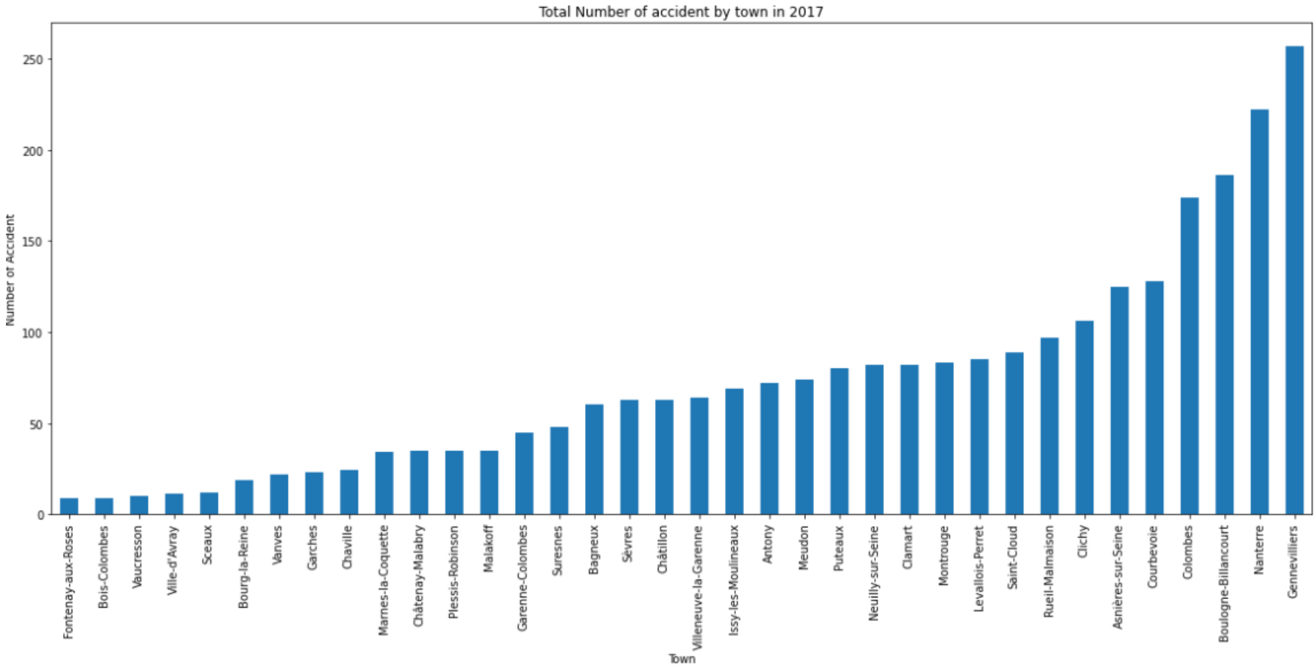


Figure 3 Number of accident by town in 2017

We can see that the ranking is quite the same if we only analyses the last year.

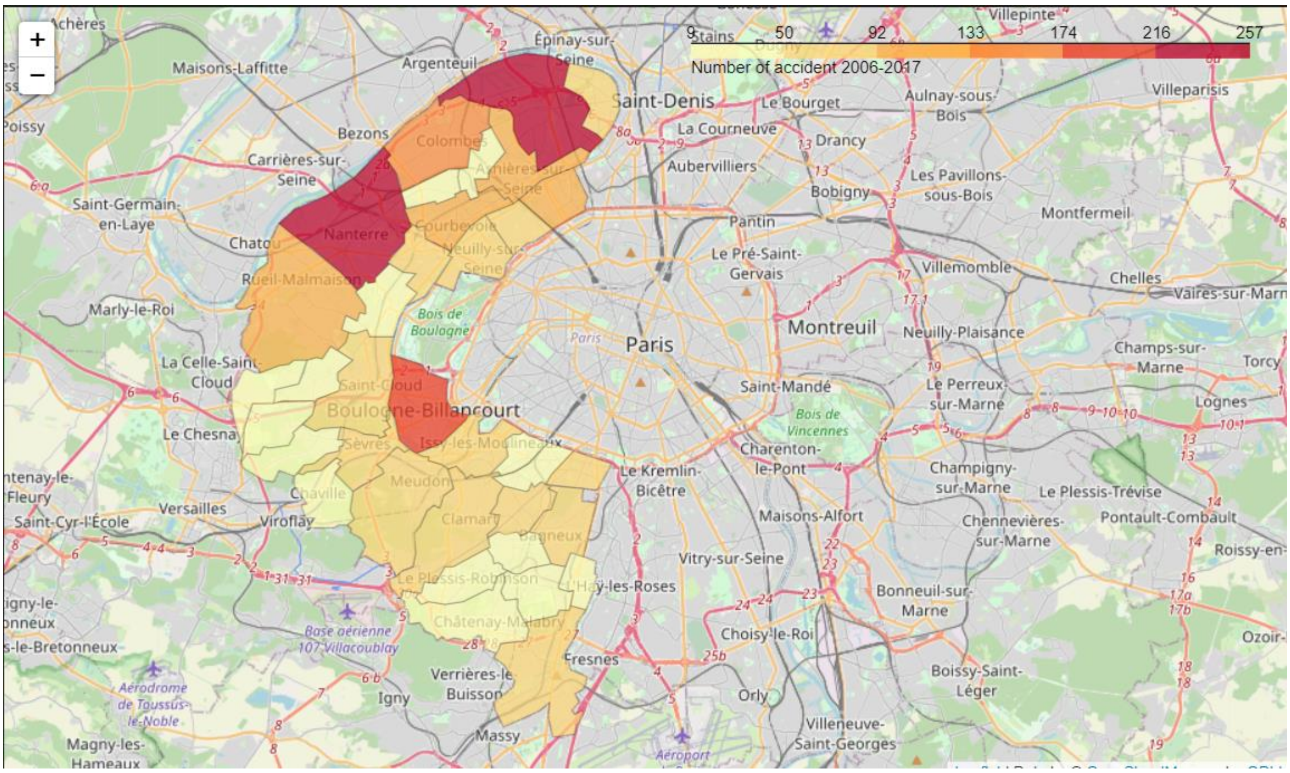


Figure 4 Accident by town in 2017

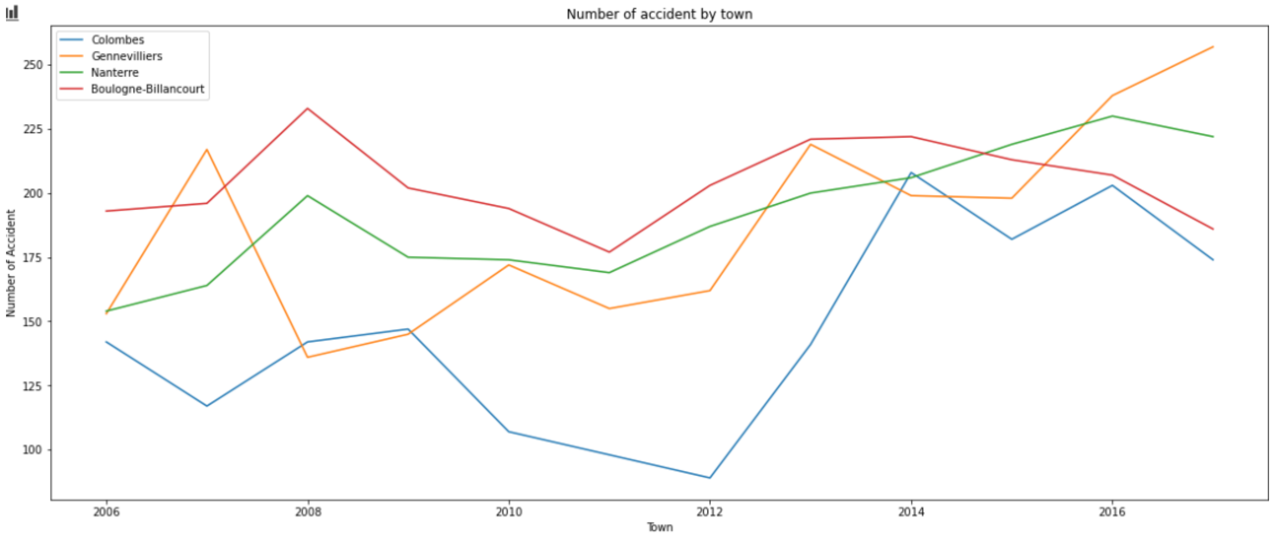


Figure 5 : Evolution of the number of accidents for the top 4 towns

We can observe 2011/2012 present à decrease period, but from 2012 to 2017 there a raise trend for all these cities.

Exploration of the top 3 town with foursquare

Nanterre

```
Hotel                   6
Plaza                   5
Japanese Restaurant     4
Supermarket             3
Park                    3
```

Gennevilliers

```
Hotel                      7
Supermarket                7
Furniture / Home Store     3
Japanese Restaurant        3
Sporting Goods Shop        2
```

Boulogne-Billancourt

```
French Restaurant      20
Tennis Court           10
Italian Restaurant      7
Bakery                  5

Bistro                  4
```

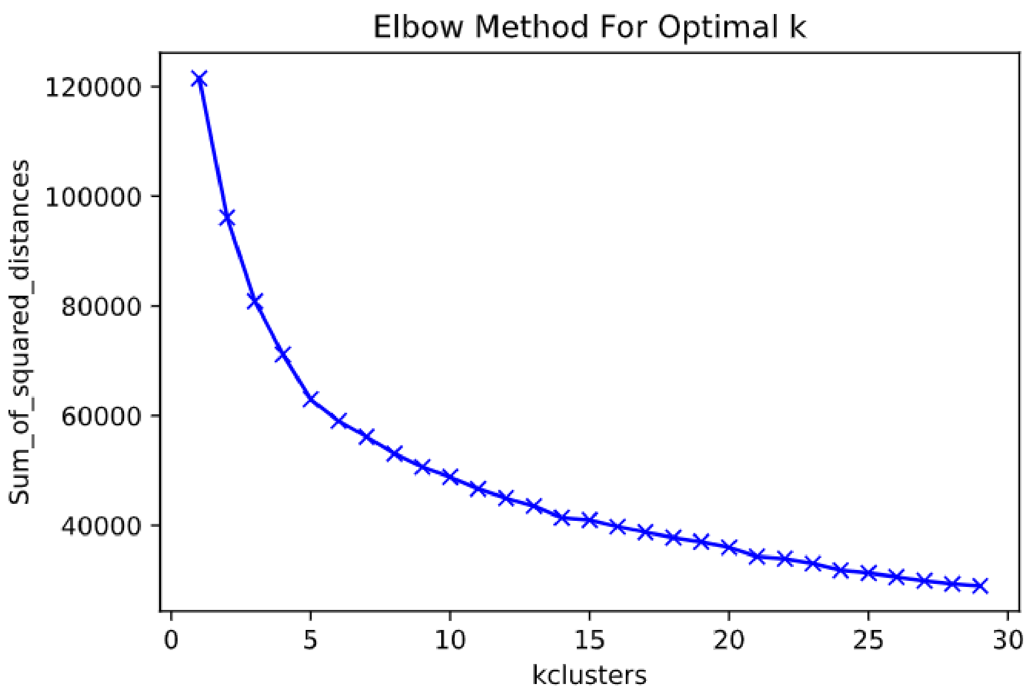
Nanterre and Gennevilliers has quite the same venues profile, with hotel in first place and Plaza/Supermarket in second place. Both these cities are business centers, this explains why there is so much hotel. Nanterre and Gennevillers are also town with a lot a large set of building.

Boulogne has another type of profile. The presence of French/Italian Restaurant that are generally expensive restaurant and Tennis court show that Boulogne is a rich city with high income resident.

Clustering with K_Means

First Analysis

I search the best number of clusters with the elbow method. I chose 6 clusters.



N° Cluster	K-Means Results	Comment
0	(7435, 23) NB_USAGER 2.036046 NB_VEH 1.010491 NB_PIE 0.955346 NB_VEL 0.021654 NB_CYC 0.068460 NB_MOT 0.132213 NB_VL 0.737458 NB_PL 0.017888 NB_TC 0.032818	Accident involving pedestrian, occurring most of the time during the Day.

	DAY 0.357902 NIGHT 0.642098 LETHAL 0.014391 LIGHT 0.802959 SERIOUS 0.182650 CHAIN_COLLI 0.000000 FRONT_COLLI 0.000000 MULTIPLE_COLLI 0.000000 OTHER_COLLI 0.918628 REAR_COLLI 0.000000 SIDE_COLLI 0.000000 WITHOUT_COLLI 0.000134 DRUNK_DRUG 0.002690 CLUSTER_LABEL 0.000000	
1	(5192, 23) NB_USAGER 2.116718 NB_VEH 2.015601 NB_PIE 0.000963 NB_VEL 0.015986 NB_CYC 0.021572 NB_MOT 1.029661 NB_VL 0.914099 NB_PL 0.024461 NB_TC 0.009823 DAY 0.000385 NIGHT 0.999615 LETHAL 0.007512 LIGHT 0.809707 SERIOUS 0.182781 CHAIN_COLLI 0.003082 FRONT_COLLI 0.049499 MULTIPLE_COLLI 0.012712 OTHER_COLLI 0.001541 REAR_COLLI 0.196456 SIDE_COLLI 0.655431 WITHOUT_COLLI 0.000193 DRUNK_DRUG 0.001541 CLUSTER_LABEL 1.000000	Accident between two vehicles, involving a motorcycle, occurs during the night and which has a “light” severity most of the time. Generally, it’s a side collision.
2	(5832, 23) NB_USAGER 2.190158 NB_VEH 1.994856 NB_PIE 0.004973 NB_VEL 0.164781 NB_CYC 0.419753 NB_MOT 0.000000 NB_VL 1.324931 NB_PL 0.053841 NB_TC 0.031550 DAY 0.000000 NIGHT 1.000000 LETHAL 0.002915 LIGHT 0.897119 SERIOUS 0.099966	Accident between two cars, occurs during the night with a light severity.

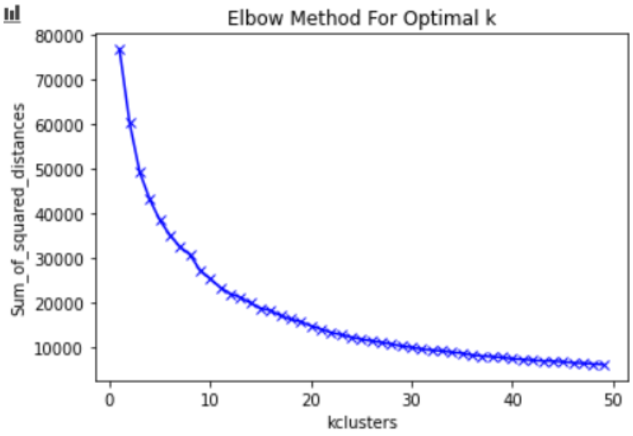
	CHAIN_COLLI 0.001372 FRONT_COLLI 0.057785 MULTIPLE_COLLI 0.003086 OTHER_COLLI 0.005316 REAR_COLLI 0.266461 SIDE_COLLI 0.576646 WITHOUT_COLLI 0.007373 DRUNK_DRUG 0.003944 CLUSTER_LABEL 2.000000	
	(4182, 23) NB_USAGER 1.140124 NB_VEH 1.001674 NB_PIE 0.002869 NB_VEL 0.035629 NB_CYC 0.205404 NB_MOT 0.375179 NB_VL 0.371593 NB_PL 0.011239 NB_TC 0.002630 DAY 0.442850 NIGHT 0.557150 LETHAL 0.018173 LIGHT 0.792922 SERIOUS 0.188905 CHAIN_COLLI 0.000000 FRONT_COLLI 0.000478 MULTIPLE_COLLI 0.000000 OTHER_COLLI 0.003109 REAR_COLLI 0.000717 SIDE_COLLI 0.000000 WITHOUT_COLLI 0.941655 DRUNK_DRUG 0.016499 CLUSTER_LABEL 3.000000	Accident involving only one vehicle.
	--------cluster 4-------- (6609, 23) NB_USAGER 2.141776 NB_VEH 1.998033 NB_PIE 0.003783 NB_VEL 0.085641 NB_CYC 0.173400 NB_MOT 0.521259 NB_VL 1.134967 NB_PL 0.057043 NB_TC 0.025722 DAY 1.000000 NIGHT 0.000000 LETHAL 0.007414 LIGHT 0.861401 SERIOUS 0.131185 CHAIN_COLLI 0.001967 FRONT_COLLI 0.049327 MULTIPLE_COLLI 0.009230 OTHER_COLLI 0.004237	Accident between a car and another type of vehicles occurring during the day with a light severity.

	REAR_COLLI 0.250416 SIDE_COLLI 0.590558 WITHOUT_COLLI 0.011348 DRUNK_DRUG 0.004085 CLUSTER_LABEL 4.000000	
	--------cluster 5-------- (2474, 23) NB_USAGER 3.947049 NB_VEH 2.991512 NB_PIE 0.014956 NB_VEL 0.010914 NB_CYC 0.040016 NB_MOT 0.256669 NB_VL 2.597817 NB_PL 0.058205 NB_TC 0.027890 DAY 0.381164 NIGHT 0.618836 LETHAL 0.010105 LIGHT 0.841148 SERIOUS 0.148747 CHAIN_COLLI 0.395311 FRONT_COLLI 0.014147 MULTIPLE_COLLI 0.329426 OTHER_COLLI 0.010914 REAR_COLLI 0.067502 SIDE_COLLI 0.092158 WITHOUT_COLLI 0.007680 DRUNK_DRUG 0.006871 CLUSTER_LABEL 5.000000 dtype: float64	Accident involving 3 vehicles (cars most of the time) and 3 or 4 person.

Second K-Means Analysis

I removed the type of collision to analyses the types of accident only with the day/night, types of vehicle and severity information.

Again, I use the elbow method and I select to search 12 clusters.



N° Clsuter	K-Means Results	Comment
0	--------cluster 0-------- (3364, 15) NB_VEH 1.869798 NB_PIE 0.002675 NB_VEL 0.020809 NB_CYC 0.024078 NB_MOT 1.003864 NB_VL 0.755648 NB_PL 0.042806 NB_TC 0.022592 DAY 1.000000 NIGHT 0.000000 LETHAL 0.016350 LIGHT 0.983056 SERIOUS 0.000595 DRUNK_DRUG 0.003567 CLUSTER_LABEL 0.000000 dtype: float64	Accident involving at least one motorcycle occurring during the day with a light severity
1	--------cluster 1-------- (3873, 15) NB_VEH 1.885102 NB_PIE 0.002324 NB_VEL 0.227989 NB_CYC 0.719339 NB_MOT 0.022980 NB_VL 0.791376 NB_PL 0.076168 NB_TC 0.047250 DAY 0.007230 NIGHT 0.992770 LETHAL 0.003357 LIGHT 0.994836 SERIOUS 0.001807 DRUNK_DRUG 0.002840 CLUSTER_LABEL 1.000000	Night accident with a light severity that involve a scooter and a car.
2	--------cluster 2-------- (4032, 15) NB_VEH 1.002976 NB_PIE 0.805060	Night accident that involve a car and a pedestrian.

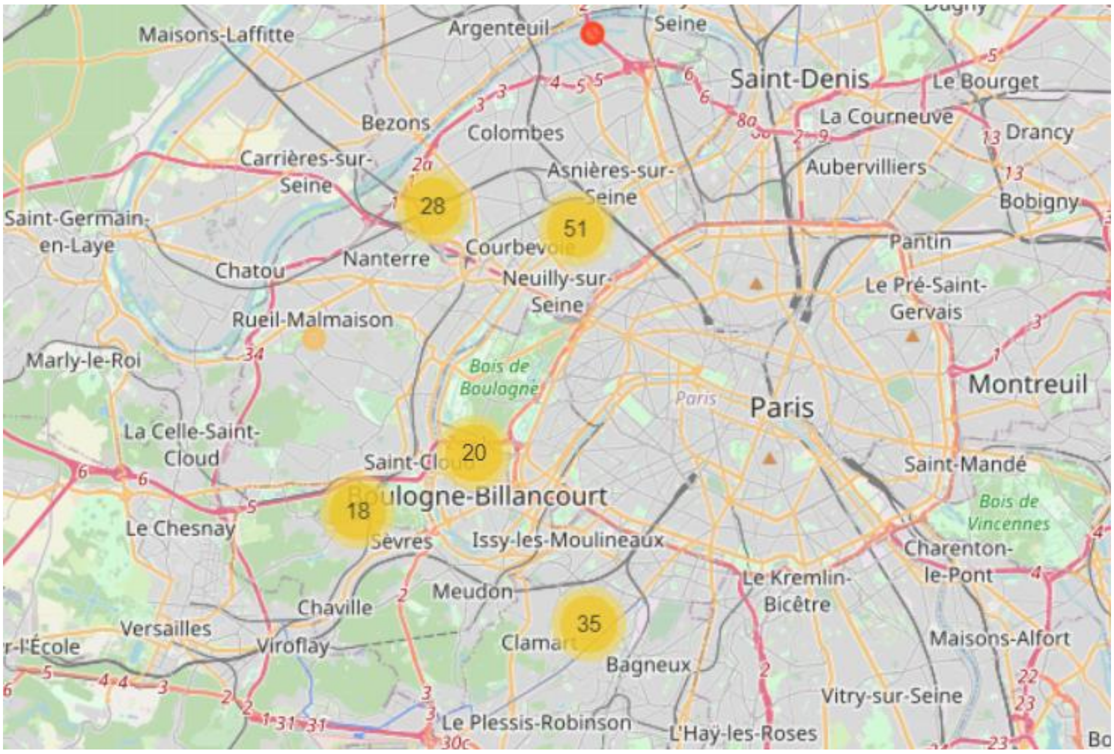
	NB_VEL 0.021081 NB_CYC 0.065476 NB_MOT 0.000248 NB_VL 0.881448 NB_PL 0.012401 NB_TC 0.022321 DAY 0.000000 NIGHT 1.000000 LETHAL 0.010169 LIGHT 0.989831 SERIOUS 0.000000 DRUNK_DRUG 0.003224 CLUSTER_LABEL 2.000000	
3	--------cluster 3-------- (4410, 15) NB_VEH 2.073016 NB_PIE 0.004082 NB_VEL 0.016100 NB_CYC 0.002041 NB_MOT 1.031293 NB_VL 0.997959 NB_PL 0.017914 NB_TC 0.007710 DAY 0.000000 NIGHT 1.000000 LETHAL 0.008163 LIGHT 0.991837 SERIOUS 0.000000 DRUNK_DRUG 0.001587 CLUSTER_LABEL 3.000000	Night accident that involve a car and motorcycle.
4	--------cluster 4-------- (3163, 15) NB_VEH 2.074297 NB_PIE 0.010117 NB_VEL 0.139108 NB_CYC 0.295289 NB_MOT 0.045210 NB_VL 1.496680 NB_PL 0.068606 NB_TC 0.029402 DAY 1.000000 NIGHT 0.000000 LETHAL 0.006007 LIGHT 0.991780 SERIOUS 0.002213 DRUNK_DRUG 0.005375 CLUSTER_LABEL 4.000000	Day accident with a light severity that involve mostly cars.
5	--------cluster 5-------- (1267, 15) NB_VEH 3.395422 NB_PIE 0.006314 NB_VEL 0.001579 NB_CYC 0.011839 NB_MOT 0.074980 NB_VL 3.269140	Accident that involve more than 2 cars.

	NB_PL 0.029992 NB_TC 0.007893 DAY 0.379637 NIGHT 0.620363 LETHAL 0.005525 LIGHT 0.857932 SERIOUS 0.136543 DRUNK_DRUG 0.007893 CLUSTER_LABEL 5.000000	
6	--------cluster 6-------- (1503, 15) NB_VEH 1.001331 NB_PIE 0.872921 NB_VEL 0.022621 NB_CYC 0.095143 NB_MOT 0.115103 NB_VL 0.692615 NB_PL 0.029940 NB_TC 0.045908 DAY 0.292748 NIGHT 0.707252 LETHAL 0.013972 LIGHT 0.000000 SERIOUS 0.986028 DRUNK_DRUG 0.007319 CLUSTER_LABEL 6.000000	Accident that involve a car and a pedestrian with a serious severity.
7	--------cluster 7-------- (2295, 15) NB_VEH 2.044444 NB_PIE 0.015251 NB_VEL 0.003922 NB_CYC 0.015251 NB_MOT 0.000000 NB_VL 2.000000 NB_PL 0.017865 NB_TC 0.007407 DAY 0.000000 NIGHT 1.000000 LETHAL 0.003922 LIGHT 0.996078 SERIOUS 0.000000 DRUNK_DRUG 0.006536 CLUSTER_LABEL 7.000000	Night accident involving 2 cars with a light severity.
8	--------cluster 8-------- (1577, 15) NB_VEH 1.008244 NB_PIE 0.299937 NB_VEL 0.046925 NB_CYC 0.000000 NB_MOT 0.930247 NB_VL 0.000000 NB_PL 0.014585 NB_TC 0.016487 DAY 0.000000 NIGHT 1.000000	Night accident involving only one vehicle, a motorcycle.

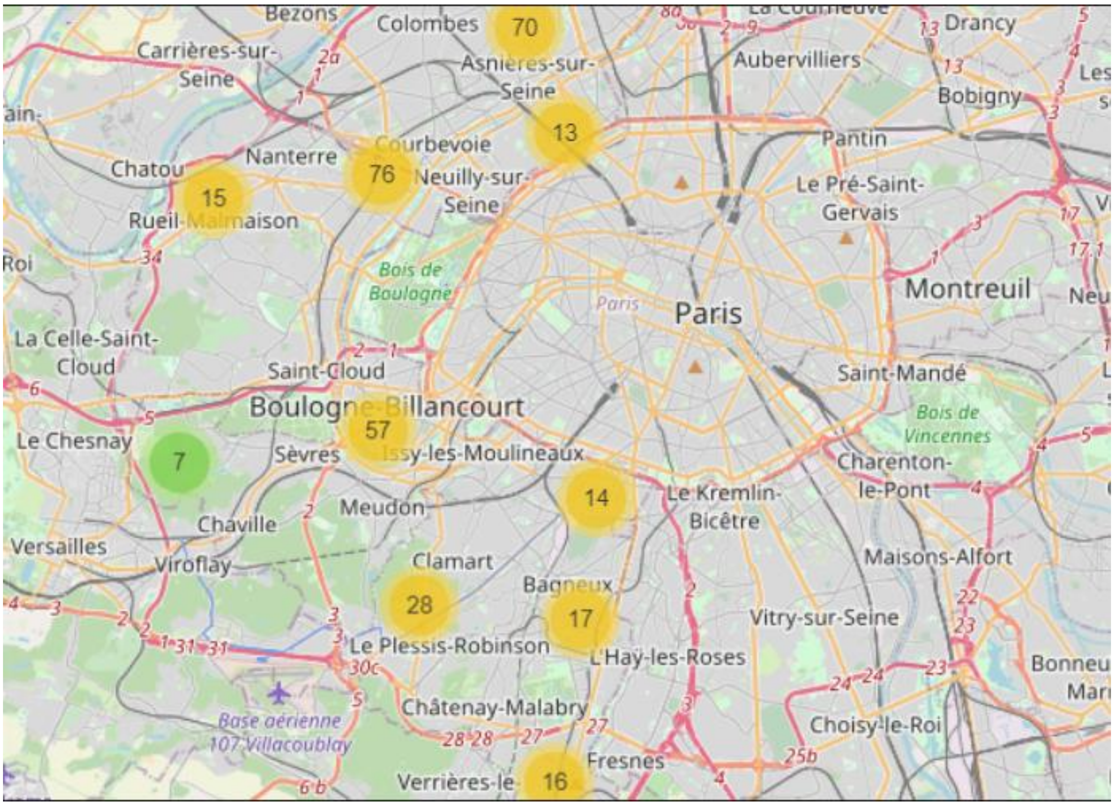
	LETHAL 0.029803 LIGHT 0.820545 SERIOUS 0.149651 DRUNK_DRUG 0.001268 CLUSTER_LABEL 8.000000	
9	--------cluster 9-------- (3225, 15) NB_VEH 1.001240 NB_PIE 0.631628 NB_VEL 0.030388 NB_CYC 0.122481 NB_MOT 0.079690 NB_VL 0.728992 NB_PL 0.017984 NB_TC 0.021705 DAY 1.000000 NIGHT 0.000000 LETHAL 0.017984 LIGHT 0.982016 SERIOUS 0.000000 DRUNK_DRUG 0.011783 CLUSTER_LABEL 9.000000	Day accident that involve only one a vehicle, mostly a car, and pedestrian.
10	--------cluster 10-------- (1649, 15) NB_VEH 2.060643 NB_PIE 0.009703 NB_VEL 0.068526 NB_CYC 0.186173 NB_MOT 0.630685 NB_VL 1.098241 NB_PL 0.054579 NB_TC 0.022438 DAY 0.000000 NIGHT 1.000000 LETHAL 0.002426 LIGHT 0.000000 SERIOUS 0.997574 DRUNK_DRUG 0.002426 CLUSTER_LABEL 10.000000	Night accident with a serious severity that involve a car and another type of vehicle, mostly motorcycle.
11	--------cluster 11-------- (1366, 15) NB_VEH 1.722548 NB_PIE 0.011713 NB_VEL 0.049780 NB_CYC 0.149341 NB_MOT 0.576867 NB_VL 0.883602 NB_PL 0.045388 NB_TC 0.017570 DAY 1.000000 NIGHT 0.000000 LETHAL 0.002196 LIGHT 0.000000 SERIOUS 0.997804 DRUNK_DRUG 0.017570	Day accident with a serious severity.

	CLUSTER_LABEL 11.000000	

map of the accident caused by alcohol or drug



Map of lethal accident



Discussion

The results presented above show that the road safety does not improve in the Haut de Seine. All the awareness campaign does not give good results. In the top 3 towns that we have analyze the number of accidents has an increase trend. These three cities are crossed by high traffic roads that leads to Paris and its suburbs. A reflexion about ways to reduce the traffic on these roads should be done by municipalities and government.

The clustering analysis enable us to analyses most common accident profile. This can help to lead a road safety campaign and to realize enhancement on the road infrastructure.

Conclusion

There is a lot more things to do with this data, we can realize clustering analysis with the type of road for example.

We can also analyze only accident involving pedestrian and bicycle in order to promote that kind of clean mobility solution.