Study of several factors for cars accidents in France from 2005 to 2019

Claire Robin

March 2021, Data Mining Project of Data Mining class,
MLDM master Degree, first year

1 Problem Understanding

Every year, road accidents cause numerous injuries, the premature death of many people, as well as important material damages. Therefore, important means are developed for the prevention and design of safer roads.

The majority of available statistics focus on user aggravating factors (alcohol, drugs, fatigue, speed). On the one hand because they are major factors for accidents, on the other hand because they are much used for prevention. The influence of parameters external to the users seems much less studied, or the analyses are not freely available.

One of the particularity of this problem is that we have only positive data, since we have no "non-accident" samples, so we cannot compare the proportion of a variable between a group that has had an accident and one that has not. Actually, we want to find unexpected factors accidents involved in the severity of accidents, so the classification is not a pertinent solution.

2 Data Understanding

The data used are available on the data.gouv.fr website : www.data.gouv.fr/fr/datasets/accidents-de-la-route/. To constitute my dataset, I used a first concatenation of the car accident dataset of the years from 2005 to 2016 available on Kaggle : https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016. I update the dataset with the year 2017, 2018 and 2019 available on the government website. The dataset is made of 3 files, a user file, a place file and a characteristics file.
All the information about the user (date of birth, gender, driver or passengers, severity of injury) are in the first file, it contain 2 142 195 rows and 12 variables. In the place file, we find all the information about the accident place (type of road, condition of the road, maneuver in progress during the accident, it contain 956 . and in the last file, we find information about localisation of the accident, date, weather condition, it contain 958 469 rows and 14 variables. Each accident is identified in the three files by a unique ID number.

All the information are numbers, but except of gps coordinates, all the variables are natural number (as "sexe" which is coded by 1 for the men, 2 for the women).

3 Data Preparation

I first completed each file with data from 2017, 2018, and 2019. Then I deleted some irrelevant columns: in caracteristics, I deleted gps because it did not bring any new information (it indicated if GPS coordinates had been provided) as well as adr (address) because I did not want to use them.

Then, according to the tasks I made, I transformed some variables in integer to be able to visualize each value separately. I also removed some of the columns when it was mainly Na value. For the map, I removed all lines where latitude and longitude were not provided as well as all lines containing outliers (lat , long = 0, 0)

4 Modeling

4.1 Spacial studies

To begin with, a map of the car accident at Saint-Etienne was created 1. To objective was to better understand the distribution of the accident and visualize them to become familiar with the dataset. We can see that some years are missing, this is because the GPS coordinates was not systematic during

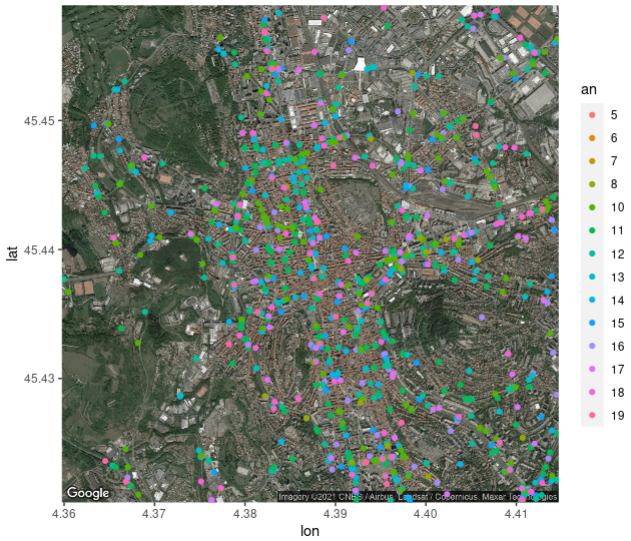

Figure 1: Car accident at Saint-Etienne from 2005 to 2019

the first years. We can see that some roads concentrate the majority of the accidents on the map.

4.2 time studies

Then, we would like to make a temporal study. We already know that the number of accident decrease with the years,the holidays like Easter are the most deadly in France and the number of accidents is highly correlated to traffic and therefore to rush hour. So it's not really interesting. But we would like to see if the impact of a particular measure, a new law or a new regulation, can be observed directly on the proportion of accidents. Since the number of accidents is not constant between the month, we have chosen an axis on the months rather than on the years. No significant variation is visible. We have therefore concentrated our observations only on accidents outside the agglomeration. The objective was to see the impact of the law on the reduction of the speed to 80 km/h on secondary roads. It can be seen from the figure 2 that this measure does not seem to have had an influence on the reduction of accidents.

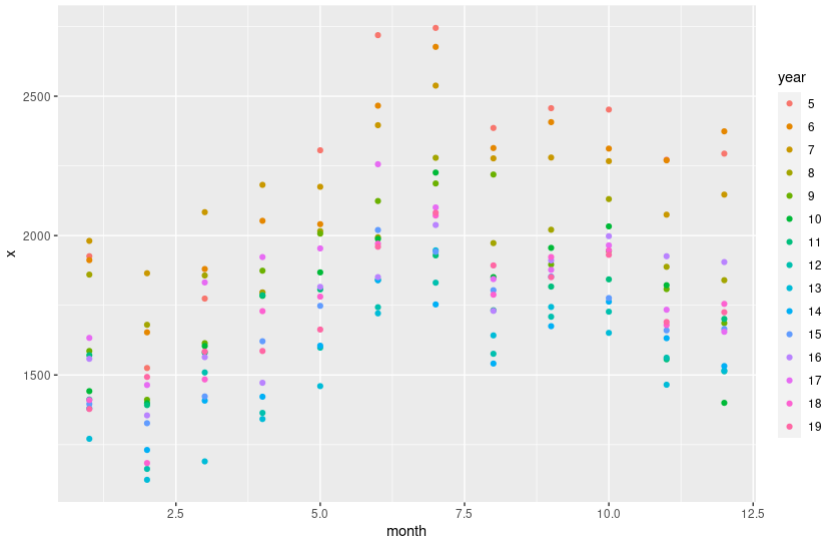

Figure 2: accidents during the months

4.3 studies of some variables

Before to use more complex methods as decision tree, we did a manual study of the relationship between some variables. We can see that young people are the population with the most accidents see 3.

The question of the relationship between accidents and the gender of the driver is a debate that

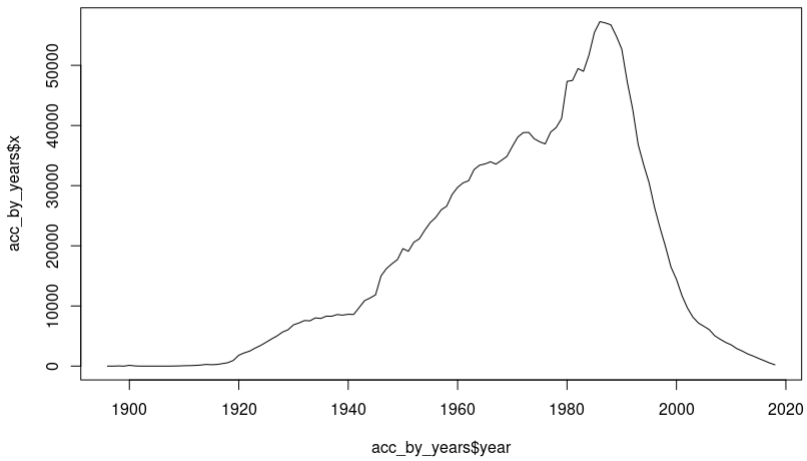

Figure 3: accidents according to the age of the driver

regularly comes up during meals and discussions. We can close this debate by saying that women have 2 times less accidents than men (1 439 318 vs. 702 877), no matter how serious the accident, so they are in general safer drivers. See 4

A last relationship that we would like to study is the relationship between the severity of the

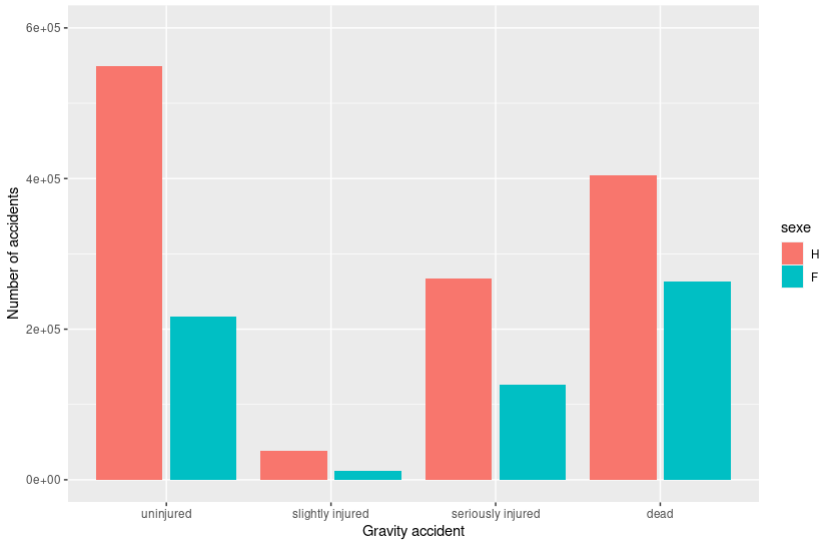

Figure 4: Gravity of the accident taking into account the sexe of the driver

accident and the time. The idea behind it is that we know that the number of accidents follows the traffic, however, we also know that fatigue and alcohol are aggravating factors for accidents. So we want to observe if, at the times when fatigue and alcohol consumption are the most important, this is reflected in the number of serious accidents. On the figure 5 we can see that the number of accidents is clearly lower during the night, however we observe that one accident out of 2 is fatal. I was surprised by the distribution of the data between accidents with no injuries, minor injuries and deaths. So I checked in depth if I had not made a mistake and this is the correct distribution.

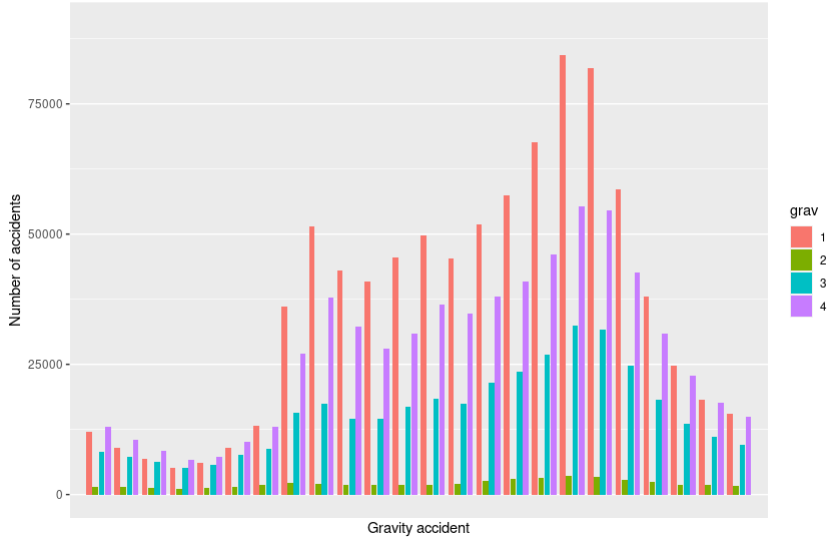

Figure 5: Gravity taking into account the time

4.4 Research of the determining parameters for the severity of accidents

To determine the parameters that most determine the severity of the accident, we made a decision tree. The parameters used to determine the severity of the accident are those that influence the most this variable 6. Not surprisingly, the first parameter is safety, the higher the number, the less safety elements are present or used (no belt, belt, airbag etc.). Age and gender are also rather expected from the previous plots. Finally, the decision tree also takes into account whether the user is a driver (= 1) or a passenger (> 1). However we are in a Simpson paradox, since drivers are clearly more represented than passengers, the driver's place is necessarily strongly correlated to the severity of the accident. We can see on the figure 7 that the first place, the driver place, is over-represented in the most serious accidents because it is over-represented in the dataset.

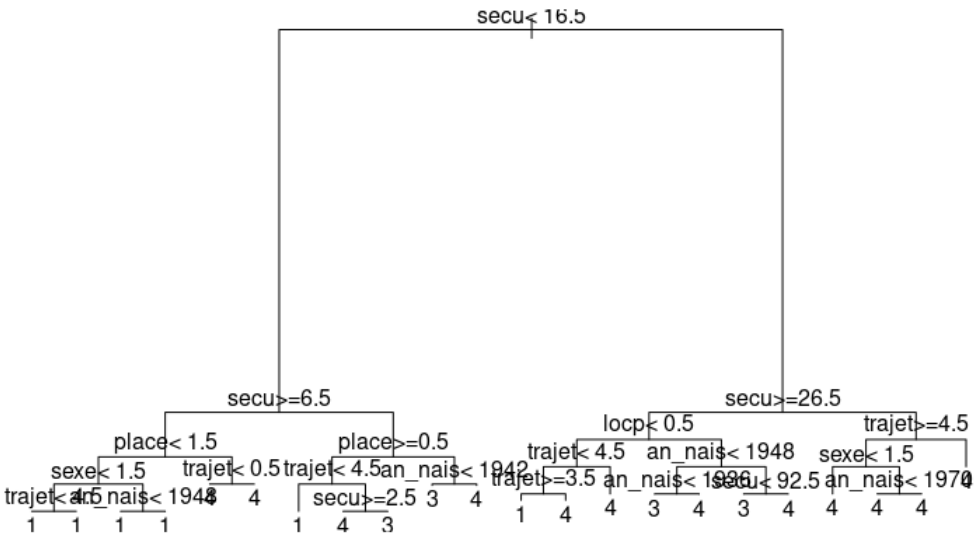

Figure 6: Decision tree on the user file variable

Finally, we tried to study the impact of several variables such as weather, luminosity, road type and intersections, pedestrian location. All of the variables that were promising for new information turned out to be under-provided. This was the case for pedestrian location, pedestrian action, maneuvering. Others were essentially composed of a single value such as luminosity or atmosphere. Finally, others did not show any particular variation (agglomeration, outside agglomeration).

5 Evaluation

We have seen that certain variables such as age, sex, time of day and safety are determinant for the severity of the accident. Many variables are correlated (the place in the car and the severity of the accident). In addition, many promising variables, such as weather, light, intersection, etc., did not prove interesting in the first experiments.

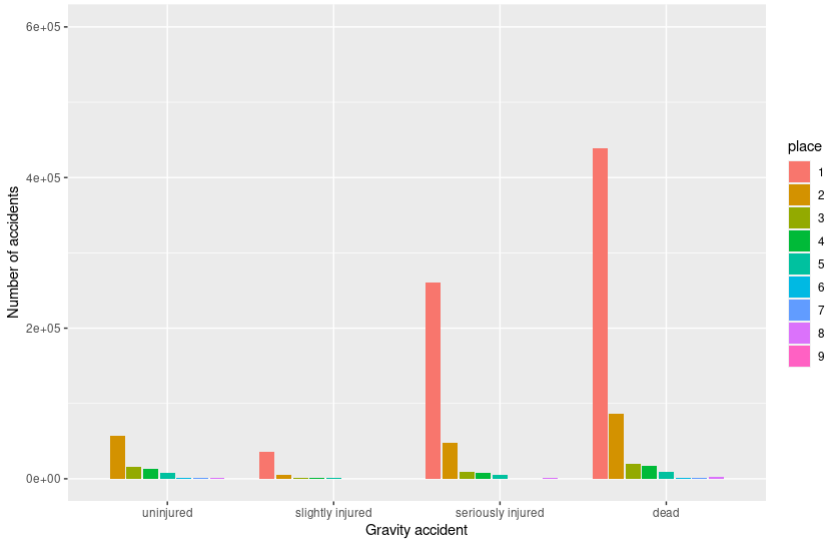

Figure 7: Gravity taking into account the place