Haladó adatelemzési módszerek - Accidents in France

Júlia Jankó, Péter Mészáros, Anna Gergály

March 2024

Contents

1 Introduction

The dataset we selected for our project details the accidents that have happened on the roads of France from 2005 to 2016. The dataset contains in-depth information about the vehicles, people and road involved in the accident as well as the weather and other driving conditions. Our goal with this project is to gain insight into what factors play a significant role in roadside accidents and their severity; aiming to be able to predict accident chances and consequences.

We downloaded the dataset from Kaggle where it has been already aggregated for the years 2005-2016, however the data is originally from an official french government website, where the data stored by year. The data is organized into 5 .csv files, one of which contains holiday dates, the other four containing data pertaining to the accidents: characteristics, places, users, vehicles.

2 Description of files

2.1 Places

Contains data about the place the accident took place, the road quality and other characteristics. There are a lot categorical data describing road type, condition, accident situation, road curvature,

traffic type, terrain and infrastructure. These are fairly well filled out and can be used for the creation of dummy columns.

There are a few unknown fields in this file, dealing with a PR value and the french road numbering system, which we do not have further information about.

2.2 Characteristics

Contains the characteristics of the accident. A lot of categorical columns can be found in this .csv file. The categorical data is represented in numerical coding. For most of them I would introduce dummy columns except if it has something to do with space/time/continuity that can be expressed using numbers. It has a few time columns like: Year, Month, Day. For these we could add an additional column combining these three numbers. With that we could make some continuity in the data. Express that 2012.12.31. and 2013.01.01. are very close to each other.

There are a few problematic columns here. Mostly the spacial data is incorrect or too specific. Address is very specific, it has a street name and very few data have the same address. The long-lat columns have about 230 000 rows filled out seemingly correctly out of the 900 000.

2.3 Vehicle

Contains the properties of the vehicles. Flow direction and Occupants columns have a very high 92%-99% null ratio. Flow direction's only gives information about the numbering of the postal address numbers. Occupants is only filled out when public transportation is involved in the accident.

2.4 Users

Contains information about the users (people involved in the accident). Birth year, user type (e.g. driver, passenger, pedestrian), severity are essential information and are almost never missing. Severity could serve as a main goal of the data analysis, extremities (death and no injury) are the most common. Birth may be useful when considering the driver. There is additional information about pedestrians, their location, action and whether they were accompanied at the time of the accident. Safety equipment is also noted and when used together with the severity of the accident, could reveal useful insights. The type of trip is missing in 29% of cases. Seat in the vehicle is also described, but between 2004 and 2008, it was not recorded.

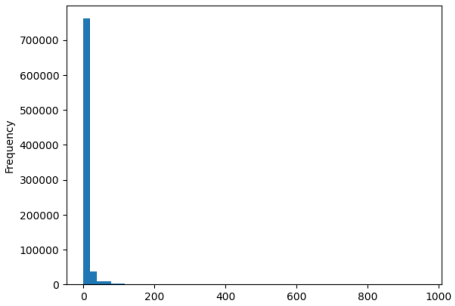
3 Data preparation and visualisation

Because many variables are categorical, we could not use correlation to measure their pairwise association, instead, we used Cramer's V for this purpose.

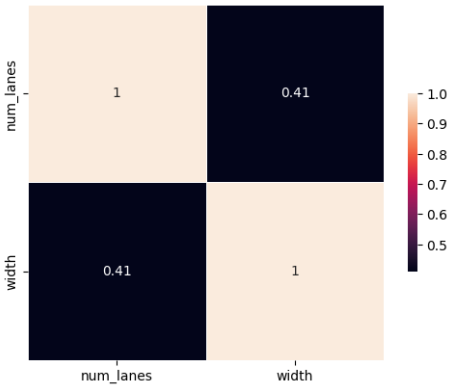
3.1 Places

We have removed quite a few columns while preparing this file. The road numbering system is very opaque and seems unusable for both that reason and because the available data is fairly sparse: v1 and v2 are missing for more than half of rows and road number is 0 for more than half of them.

We also removed the columns relating to a so-called 'pr value' (pr and pr distance) as the site contained no information regarding the meaning of this field and a large percentage of the data was missing. There were several columns concerning the width of the road (number of lanes, width, central width) of which the latter two were removed.

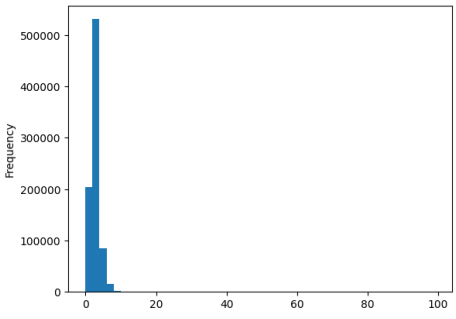


(a) The distribution of central width

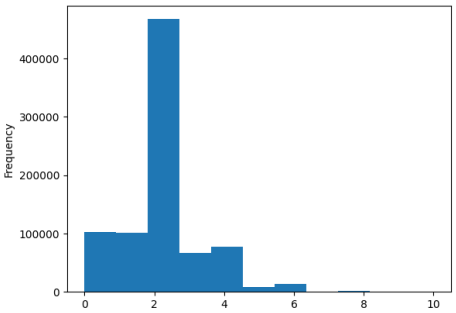


(b) Correlation of width and number of lanes

Figure 1: Width and central width



(a) The distribution of lane numbers before data cleansing



(b) ...and after

Figure 2: Data cleansing on number of lanes

Central width had an overwhelming majority of 0 values1a, while width contained very outlandish values ranging from -81 to 999 meters and had a fairly high correlation with number of lanes1b. The number of lanes was mostly alright, but contained way too large values: up to 99, so these were removed.1

Missing values in each column have been filled with either the other/not applicable value (0 in most cases) or with the most common/most likely value for the field (2 lanes for traffic, normal road conditions).

The values in this file are almost exclusively categorical and so the used correlation metric was Cramer's V (Figure 3).

3.2 Characteristics

The lighting column values were renumbered from 1 to 5 in a way that now 1 means there was more lighting and 5 means no lighting instead of random numbering of the categories in this. The address column contained a lot of missing values and also the values we had were not really useful. The most common street address was "AUTOROUTE A1" with 2816 rows which was very little out of the 839.985 rows so we removed it.

Longitude, latitude and gps columns were removed because department and municipality columns provided enough information to visualise the data and not too detailed. Also these columns contained a lot of null values.

Other than these the other columns contained very few null values and they were filled with the most common values.

There were quite a few correlations in the characteristics file (Figure 4). The most obvious is the holiday column and the date. Holiday is an added column based on the holiday.csv file. This column has a 1 if it's a holiday and a 0 if it's not so it is understandable why it correlates with the time. Another strong correlation is the Time-Lighting columns. While the time contains the time of day, lighting contains if it's a full day, twilight or dawn, night with or without public lighting. With this information it makes sense that these correlate so much. Localisation contains if the accident was in built in or out of agglomeration so it is also not surprising that it correlates with the value of lighting, because public lighting is more likely in built-in areas.

3.3 Vehicle

This data contained very few missing values. For fix obstacles and mobile obstacles columns, the nulls were filled with the majority value (0) also signaling that if it is null there were probably no obstacle. For the shock and maneuver columns they were also filled with the most common value. There were only a few hundred missing values compared to the 1.4 million row data so it shouldn't affect the training.

Next was the flow direction column. It contains information about the ascending or descending of the postal address number however there were 1.3 million missing data so the column got removed.

When investigating the correlation of the vehicle's properties, there weren't any strong correlations (Figure 5). There is a mild correlation between fix obstacle and mobile obstacle, probably because there was either a mobile or a fix obstacle during the accident. The other correlation is between maneuver and shock. This makes sense since the maneuver could highly influence where the car was damaged.

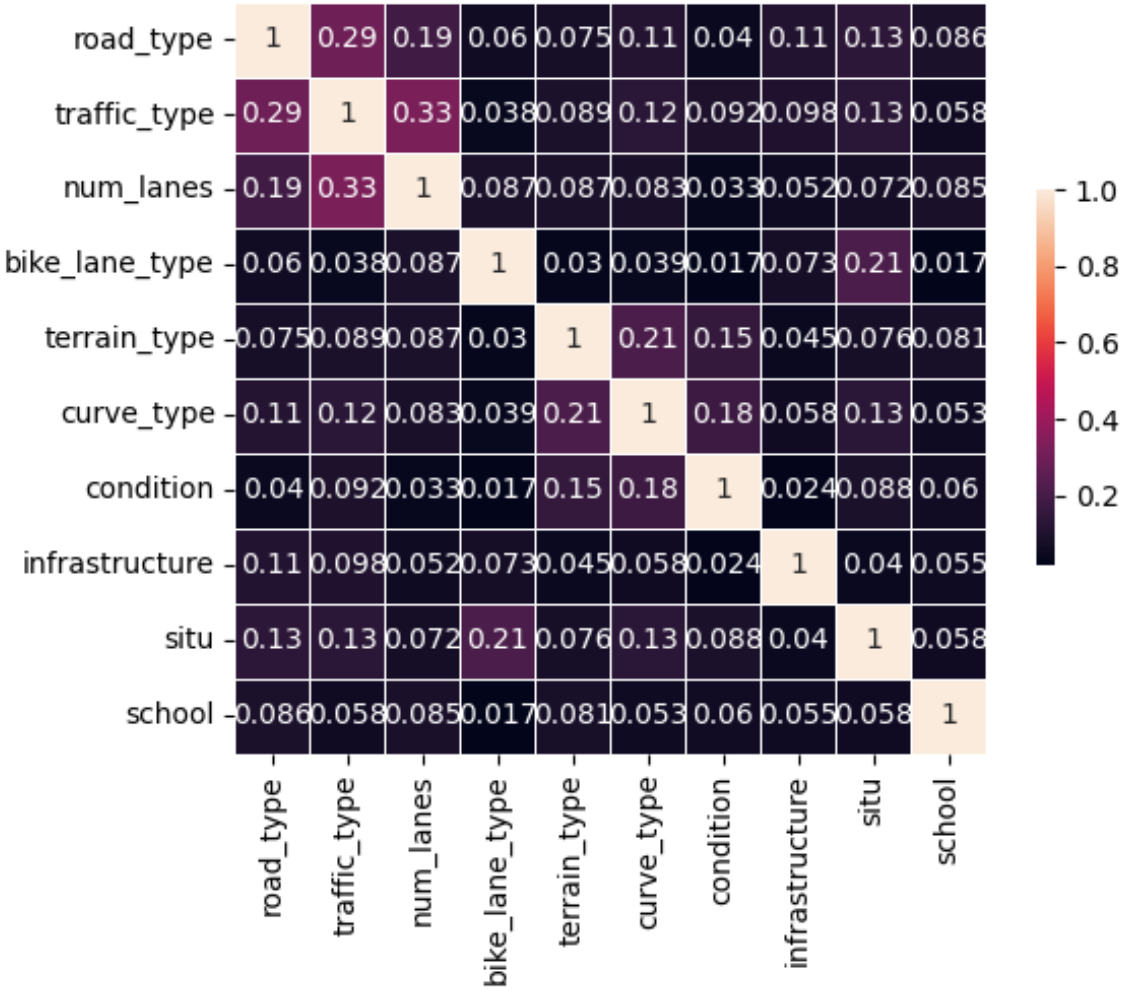


Figure 3: Correlation via Cramer's V in the fields of the places file

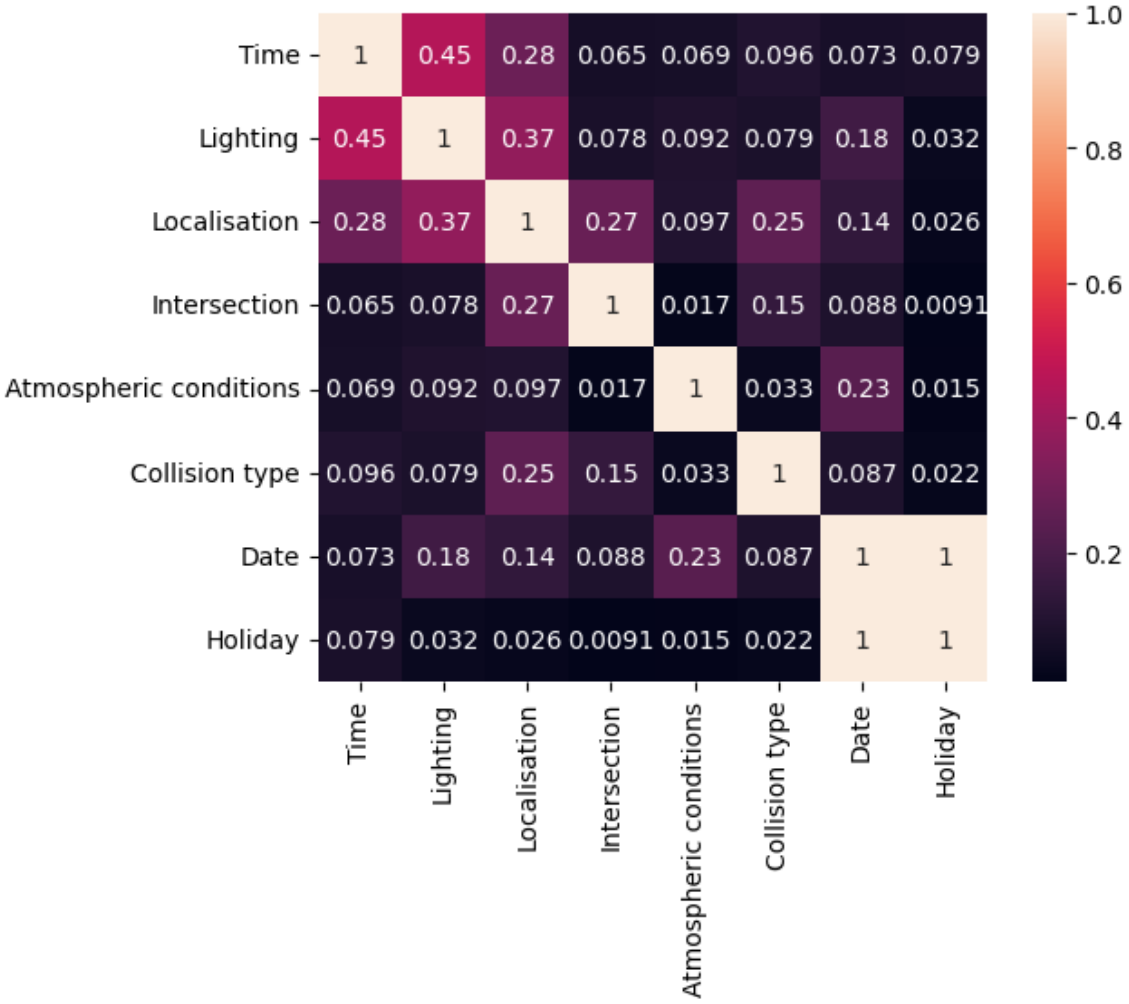


Figure 4: Correlation via Cramer's V in the fields of the characteristics file

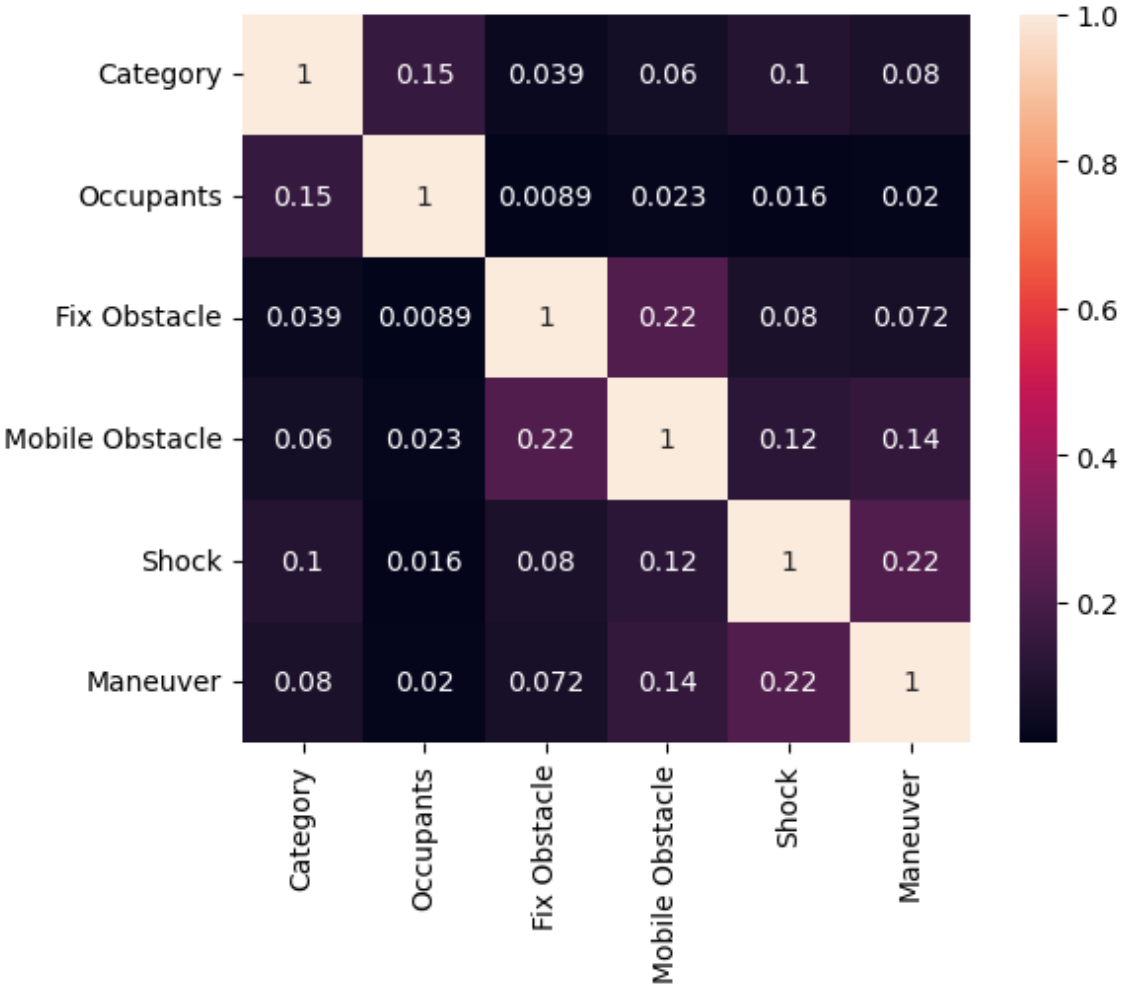
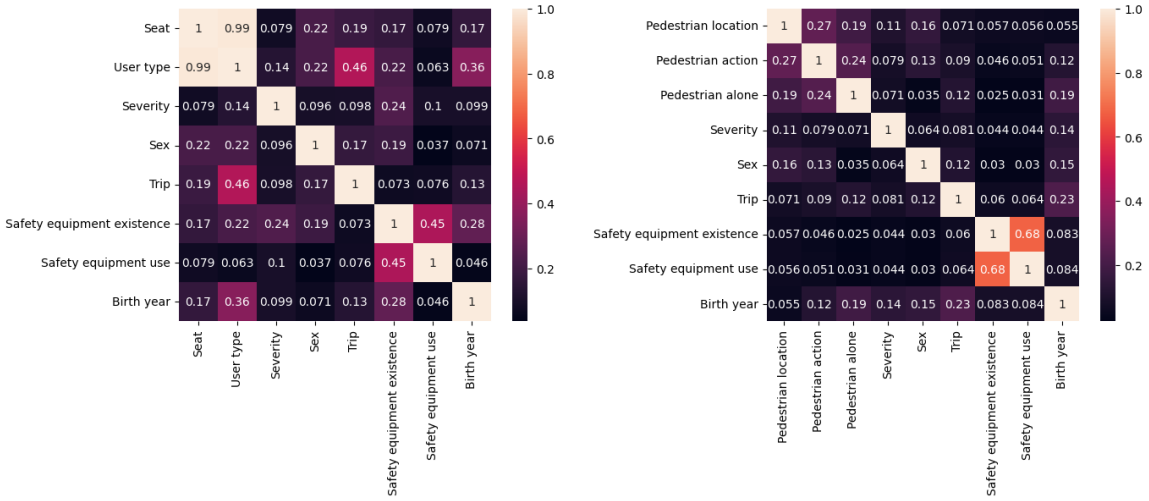


Figure 5: Correlation via Cramer's V in the fields of the vehicles file



(a) Not pedestrians (i.e. drivers and passengers)

(b) Pedestrians

Figure 6: Association of the user columns depending on user type

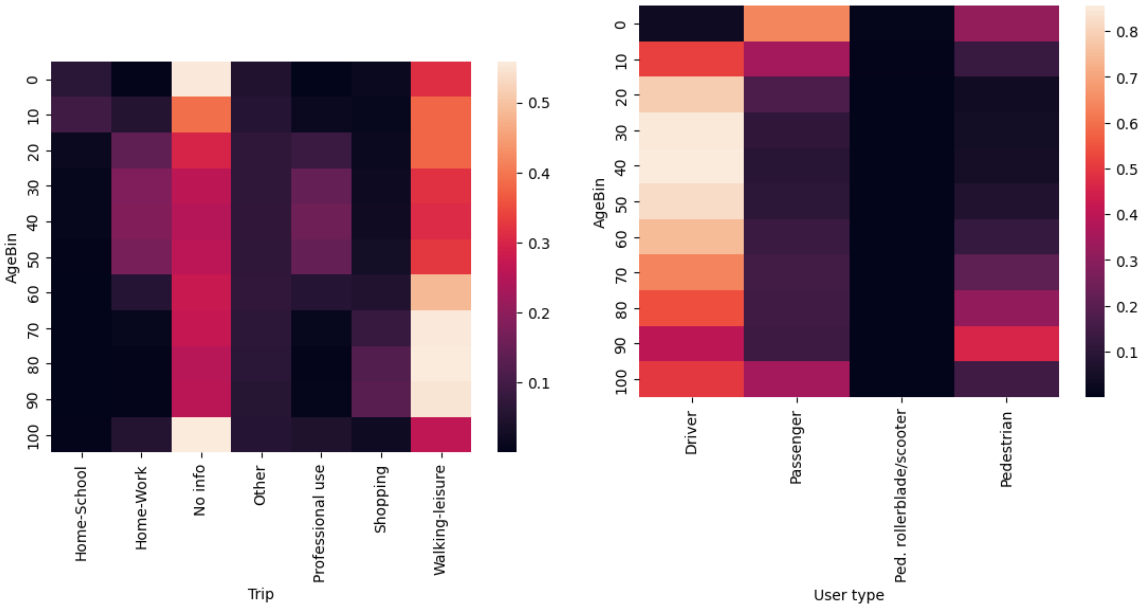
3.4 Users

In this case, not many data was missing, most columns were almost filled out. Seat was missing in four years (between 2005 and 2008), in that case it was filled with 0, to indicate, that it is unknown. User type, severity and sex were not missing. Trip was not given in just a couple of cases (about 300). Severity variable was reordered to become a scale which reflects the severity of the accident (1 meaning no harm, 4 meaning death). Safety equipment was not given in about 2% of cases, it was filled with the unknown category. Safety equipment consists of two digits, the first one referring to the existence of a safety equipment (and type of safety equipment), the second one referring to whether it was used or not (or could not be determined). Pedestrian variables had seldom (0.08%) missing values.

The association of the variables can be seen on Figure 6. Separating pedestrians and non-pedestrians (i.e. drivers and passengers) is necessary, because pedestrians have additional variables (location, action and whether they were alone) and some variables are not applicable to them (e.g. seat in the vehicle). Another reason for making the distinction, is that an accident can have different outcomes whether one is in a vehicle (and therefore 'protected' in a sense) or not.

The seat and user type have a high association because in a sense seat further refine the type of user, more precisely the seat of the passenger. In more than 99.9% of cases the driver was in the front left seat (which is expected, because people drive on the right in France). High association is also between safety equipment existence and use, but that could be due to the missing values (i.e. both variables being 0).

Figure 7 shows the distribution of trip and user types in age intervals with a 10 year step. Younger and very old people tend to be passengers 30-50 year old people are mostly drivers. Most of the people had a trip of type "Walking-leisure", but also in many cases the data is not known.



(a) Trip

(b) User type

Figure 7: Trip and user type category popularities in age intervals. AgeBin refers to age intervals by 10 years, e.g. 0-10, 10-20, etc.

3.5 Merging the dataset

To be able to effectively train a model, we concluded that we should merge our data into one big dataset. In order to do this we can't enumerate each vehicle's properties in the accident's row but rather make as many rows as many vehicles took part in the accident. Because each vehicle contains at least one person the vehicle rows all had to be expanded.

Now we have 1.876.005 rows. Each row contains the characteristics of the accident they took part in, the properties of the place and vehicle they were in and the person's attributes.

3.6 Visualization on map

We have made two visualizations on a map, using the department codes of the injuries, see Figure **??**. (Departments refer to geographical regions/administrative units). Some regions around Paris had no data, but that could be due to department changes along the years, it needs more investigation. The average severity of injuries vary a bit by department, areas with lesser accidents had higher average severity.

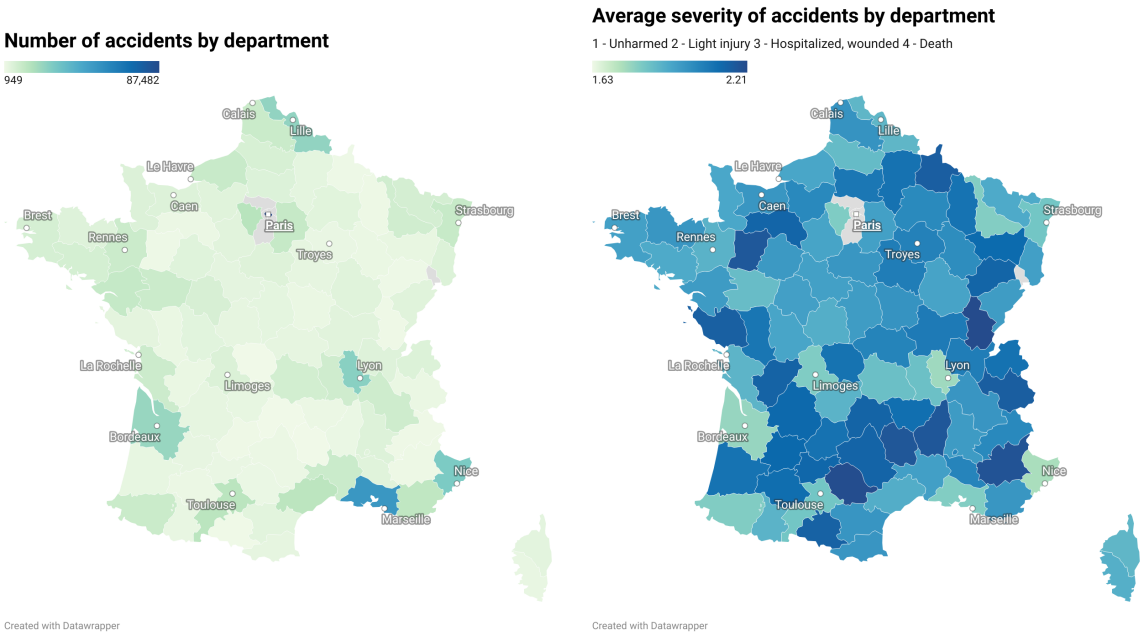


Figure 8: Visualizations by departments.