Road Accidents in France

-

Sensitization, Prevention, Monitoring

I. The project

We have decided to collaborate on this project using a Github repository, which can be accessed here : https://github.com/maelfabien/DataViz

a. Data Description

In this data visualization project, we will analyze the road traffic accidents from the French national database between 2005 and 2017. The dataset can be downloaded from the following link : https://www.data.gouv.fr/en/datasets/base-de-donnees-accidents-corporels-de-la-circulation/

The data are collected by the police each time a traffic accident occurs in France. The data are then aggregated by the "Observatoire national interministériel de la sécurité routière" (ONISR). The data takes the form of 56 CSV files describing, 4 for each of the 14 years of history :

- The vehicles implied
- The users : passengers/pedestrians
- A description of the location
- Characteristics of the accident

The are 9 attributes in the vehicles dataset (vehicle category, type of obstacle hurt...), 12 attributes for the passenger / user dataset (including age, sex, activity at the time of the accident...), 18 attributes for the location of the accident (road type, luminosity, road width...), and 16 attributes for the characteristics of the accident (date, time, GPS coordinates...). The 56 files represent 254 Mo overall.

There are a little more than 1'500'000 entries in the vehicle dataset, each entry being a single vehicle implied in an accident, 2'000'000 entries for the users dataset, each entry being a single person implied in an accident, and 900'000 entries for the location and characteristics, in which each entry represents a single accident.

Most of the features are ordinal. However, two features are nominal :

- The number of the vehicle, an identifier in the users/passengers dataset that links a passenger to a given car

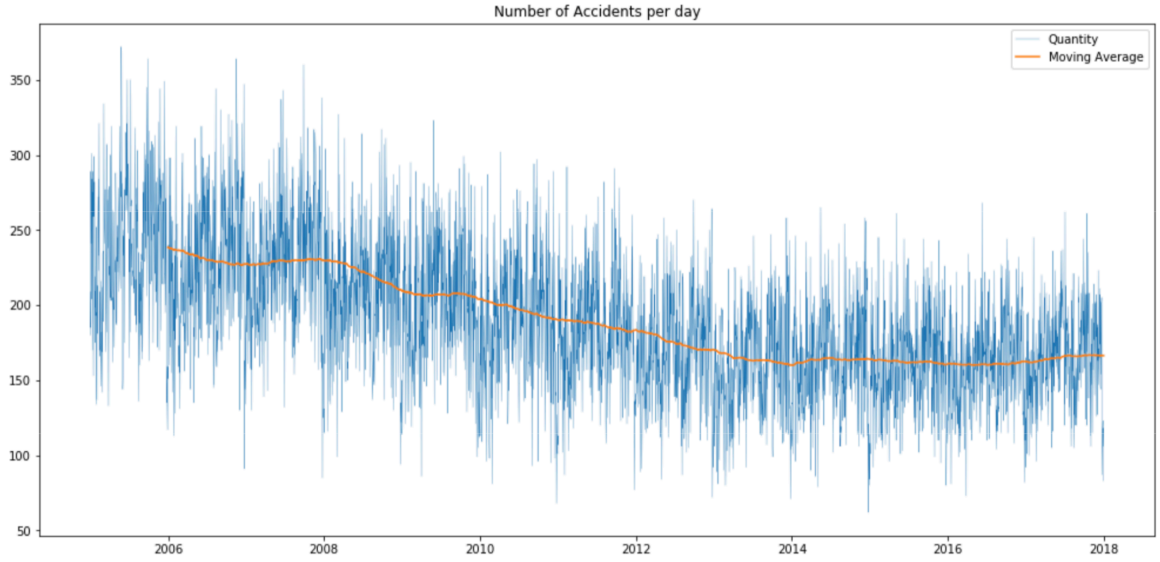
- The address of the accident

Since for our analysis, we did not expect the address to be a relevant feature, we removed this feature.

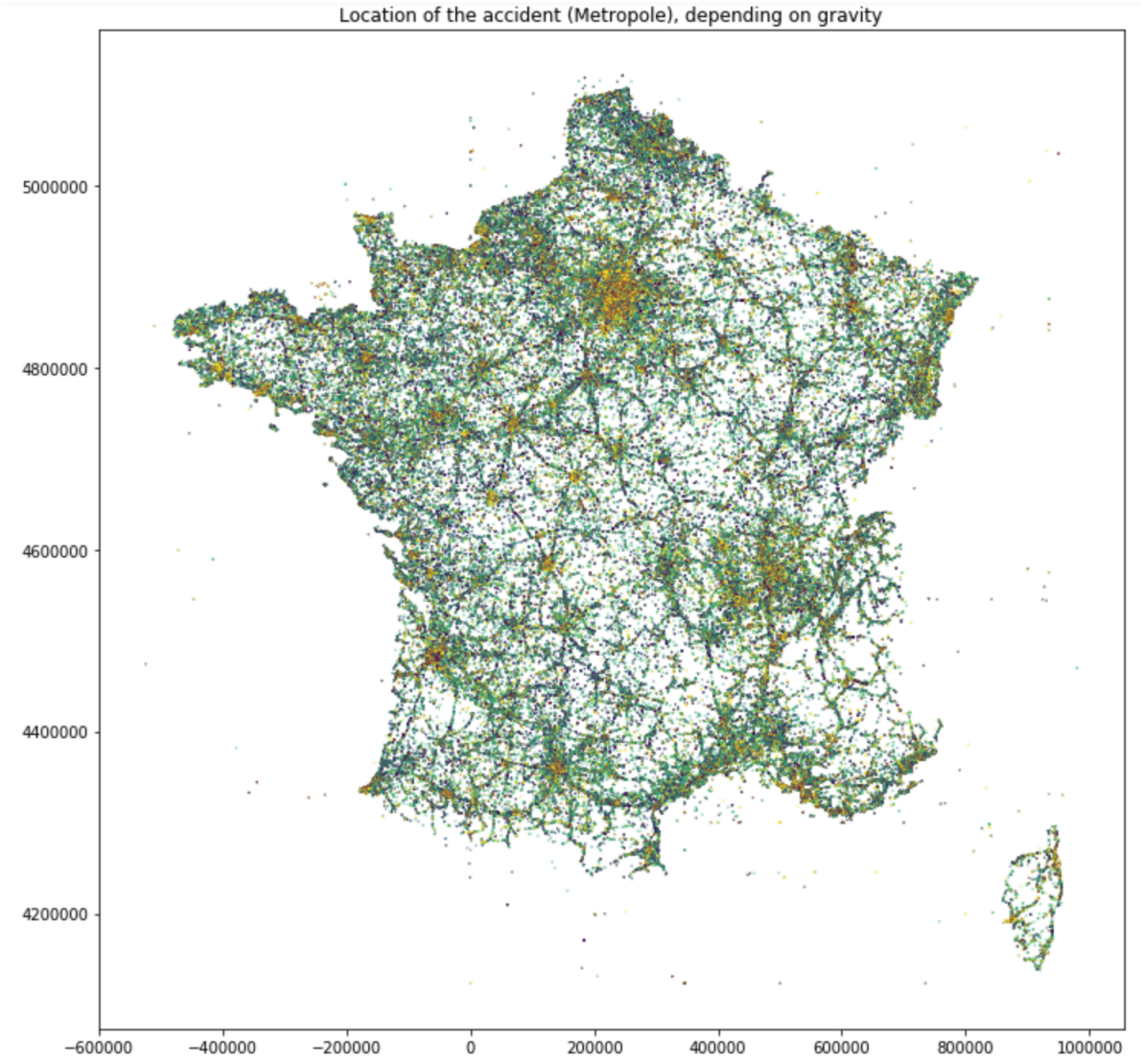
b. Users

Our analysis could bring value to several road authorities in France :
- Communes, that are in charge of communal roads
- Departments, that are in charge of departmental roads
- The state, in charge of the national roads

Actors of the road safety are in general used to comparing data and trends, and to manipulate data from this specific dataset. They are used to seeing time series representing the evolution of the number of deaths or the number of accidents over a certain period of time for performance monitoring purposes, although we can expect users from the state to have additional skills compared to communes for example.



They are also probably used to working with maps at different scales, and visually identifying clusters and dangerous roads based on several filters.



c. Tasks

We except our end users to use our tool to :

- **Sensitize** the youngsters and road criminals on the danger of the roads depending on the transportation mode they are using.
- **Prevent** road dangers through a tool that would advise the best road characteristics for a road rehabilitation, and visually illustrate the outcome of the algorithm
- **Monitor** the dangers of the roads (at difference scales, including Communes, Departments and Nation) by allowing the user to select geographic zones and filters (weather conditions, hour of the day…) to gain additional insights

Since the sensitization part is an external task that implies a presentation of the information to pupils and high school students, we need to adapt this specific design to make it as clear as possible, use relevant colors, and avoid information overload.

We believe that those 3 tasks correspond to the main objectives of all the road authorities in France, and DataViz could help bring additional value in these cases.

II. Sensitization

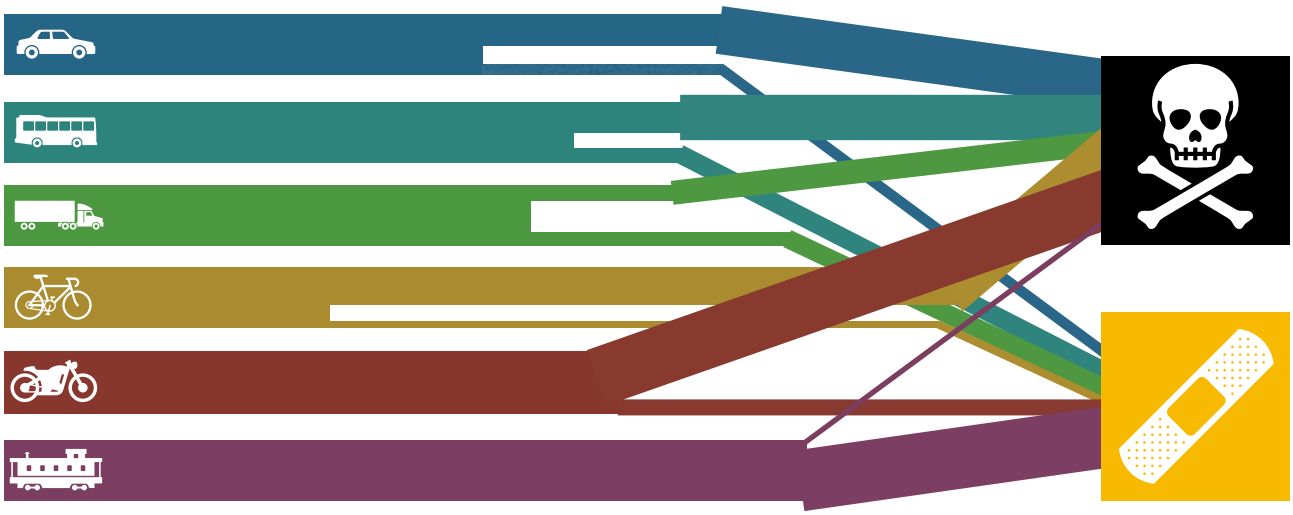
a. Motivation

Local police and communes are working together on sensitization of the youngster on the road dangers. One might believe that dangers for these kids arise once they have their first motorcycle or car. However, if they are involved in an accident, death rates among bike riders and pedestrians are incredibly high compared to other transportation means.

We selected a Sankey diagram (flow chart) to illustrate the dangers of the road, not only for car users, but also for pedestrians, bike riders or motorcyclists. The aim of this design is to convey a simple, easily understandable yet visually efficient message.

b. Design sketch

Our initial sketch of the design looked like this :



We intend to support a filter on the type of roads (communal, departmental, national), and to observe the difference of survival rate of the different categories of road users depending on the speed on the road. This would illustrate how the speed of a car impacts the survival rate of the other road users.

The design helps the reader understand the distribution of the volume of road users accident, depending on the kind of road being used, as well as the effect of speed.

This graph is quite similar to the one of Napoleon's 1812 march. The design initially represented the number of soldiers alive through the different steps of the march. We wanted to build something similar, while displaying the outcome of the accident. Since we have a large number of features, we need to make a trade-off on the complexity of the information displayed, especially if the design is to be used for sensitization among youngsters.

c. Data processing and challenges

One of the challenges is that each database (vehicle, passengers/pedestrians, location, characteristic) does not share a common key to join them. Indeed, the vehicle database distinguishes every vehicle implied in a same crash, whereas the characteristics of an accident only reports one observation for this specific case. On the other hand, the pedestrians and bike riders only appear in the passenger database. Overall, some data processing was needed in order to build a complete table.

Then, one of the other challenges was to find the right tradeoff between the completeness of the content and the visual message our end users could extract from it. Since the tool has been thought as a sensitization tool for young populations, we need to control the number of input classes, middle layers and output classes.

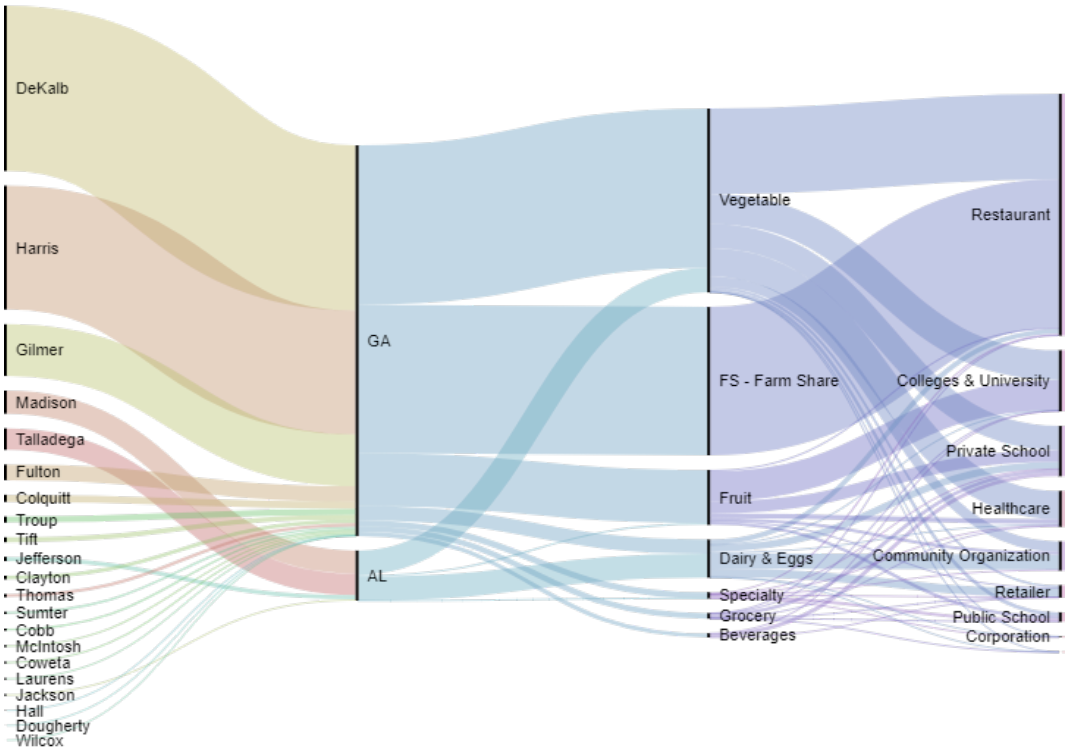
d. Technologies and outcome

We have used Plotly as the main framework for this visualization. The interactivity offered by the JavaScript behind Plotly allow us to display :

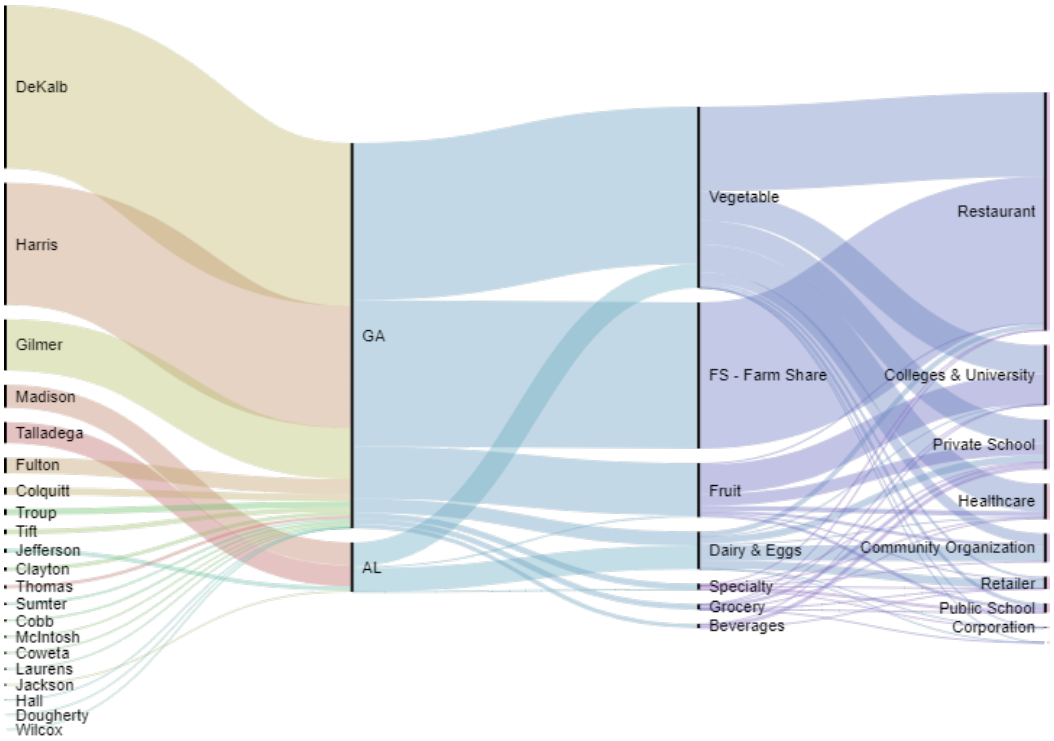
- The input classes : Car, Truck, Motorcycle, Bike, Pedestrian
- The middle layer : Wearing the security equipment or not
- The output layer : Death, harmed, unharmed

This also allows us to highlight the importance of wearing all the necessary security equipments. When the user clicks on a branch of the graph, this highlights the values of this branch specifically, and displays in a tooltip additional information and exact numbers about this branch.

We then exported this graph to D3.js in order to integrate it in a website. The final outcome is the following :



When we select a given flow, the visualization changes and the user sees this information :



e. Limits of the design

Although we do believe that the design manages to display the right message, there are several limits that we would like to highlight :

- The complexity of the graph has been highly reduced in order to make it understandable for the end users
- The interaction is limited to the tooltip on the hover, and could be improved by showing additional middle layers when the user clicks on some buttons or activates some options
- We could develop another reading perspective on the same design by adding several buttons that allow to increasingly augment the complexity of the graph
- We could also use the notion of time to show how this graph has changed over time

III. Prevention

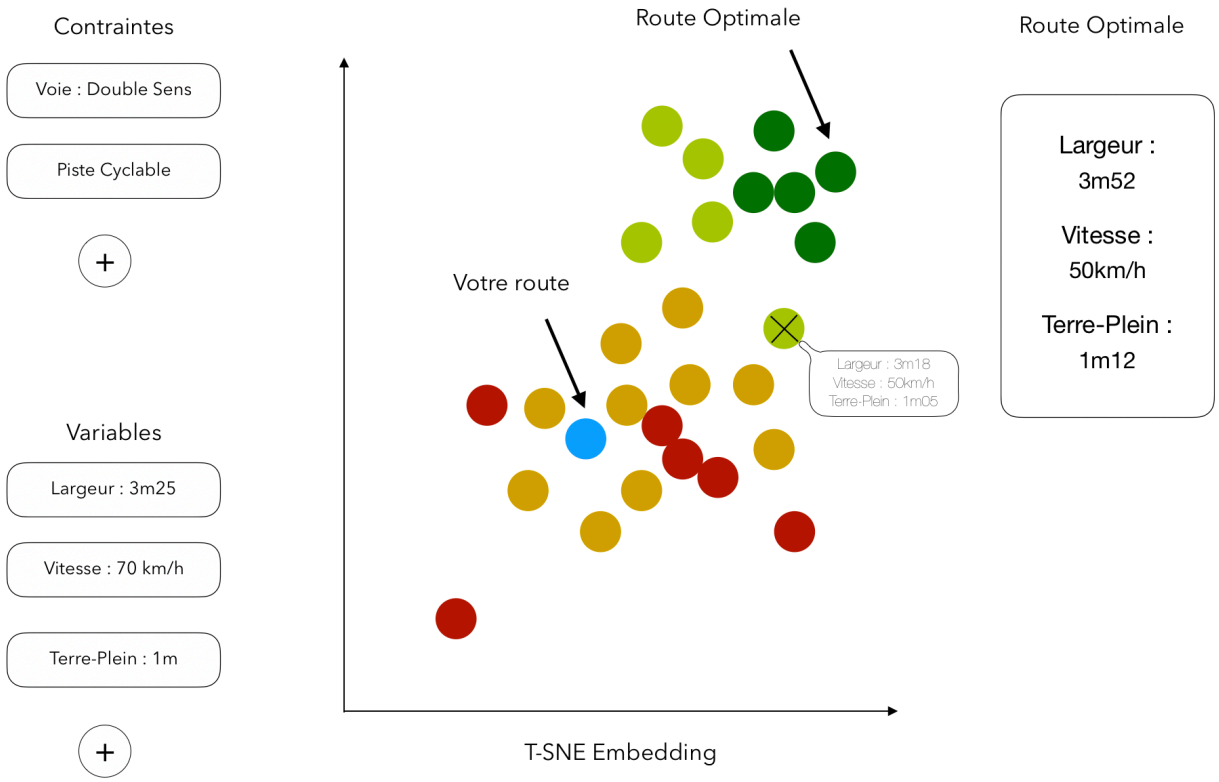
a. Motivation

There are critical decisions taken by road authorities when it comes road rehabilitation or construction. It is indeed necessary to choose the width of the road, the turn angle, the speed of the road, whether the road should have a median strip or not… All these decisions can have a large impact on the future accident profile of this road. Although road authorities in France are used to exploit the datasets we are presenting, we believe that a design that allows them to observe the accident profile of similar roads could be a great decision support tool.

Our aim is also to develop a Machine Learning model able to predict the accident profile of a road given its characteristics. Overall, if our design is successful, it should help prevent some accidents linked to profiles of roads. The main point of embedding is to unlock insights on high dimensional problems.

b. Design sketch

Our initial sketch of the design looked like this :



The user is able to set some constraints, i.e environment linked constants that cannot be changed. On the other hand, he sets some variables (width of a road, speed, median strip width…). Based on these constraints and variables, we can compute a T-SNE embedding of the roads that share the same constraints. This way, the user is able to see where the profile of the manually specified road stands compared to other roads that share the same constraints.

Then, to explore and help his decision, we provide a tooltip that allows the user to display the information related to the road on the graph. The tooltip shows information about the characteristics of each road.

If the user observes that his road stands within a red region, i.e a large number of accident recorded since 2005, it should typically lead the user to challenge his initial hypothesis. In such case, our visualization tool would come as an exploration tool to see if there are some better characteristics that can be chosen among roads that share similar characteristics.

Finally, we provide the user a prediction of the accident profile of his road. We created a scale, that ranges from 1 to 5 and describes the cumulated amount of accidents. We chose to implement a Random Forest Classifier that is trained on the whole dataset. This additional information should simply relate a no-go scenario in case the accident profile is among the worst class for example. The accuracy achieved by this algorithm, although it's not the main focus here, is a little higher than 52% in this 5-class framework.

c. Data processing and challenges

For this design, we are exclusively focusing on the location dataset.

One of the main challenges here is the quality of the data collected. It is not always the main priority for police forces to write down the exact width of the road or the exact width of the median strip… A lot of data cleaning is required on this side, and it reduces the input data.

Then, as each entry represents a single accident, we need to group the roads that share the same characteristics, and compute the overall amount of accidents for each type of road. Once this step is done, we have a little less than 12'000 entries.

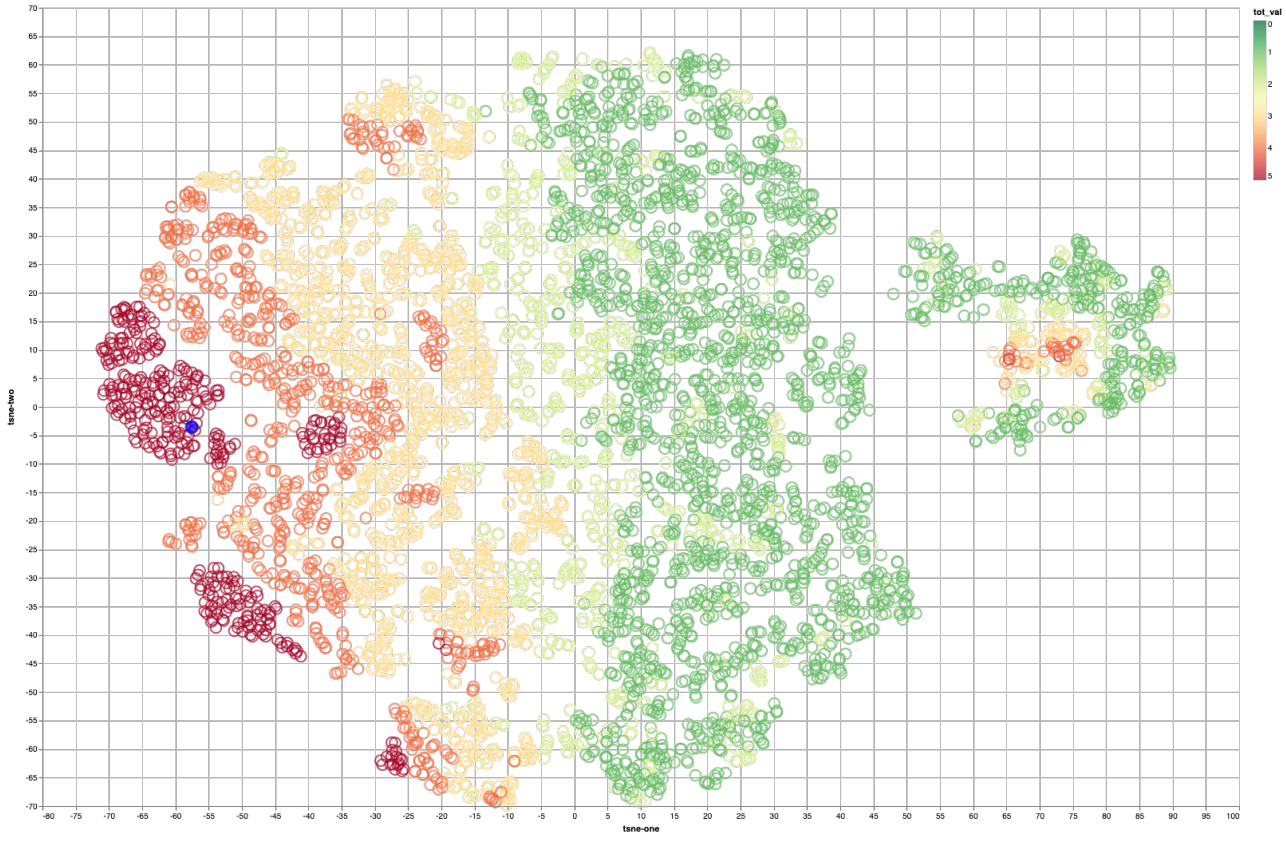
d. Technologies and outcome

We have used Altair as the main support for this graph. Altair offers a simple way to display tooltips, and offers a web support to export a graph in HTML format.

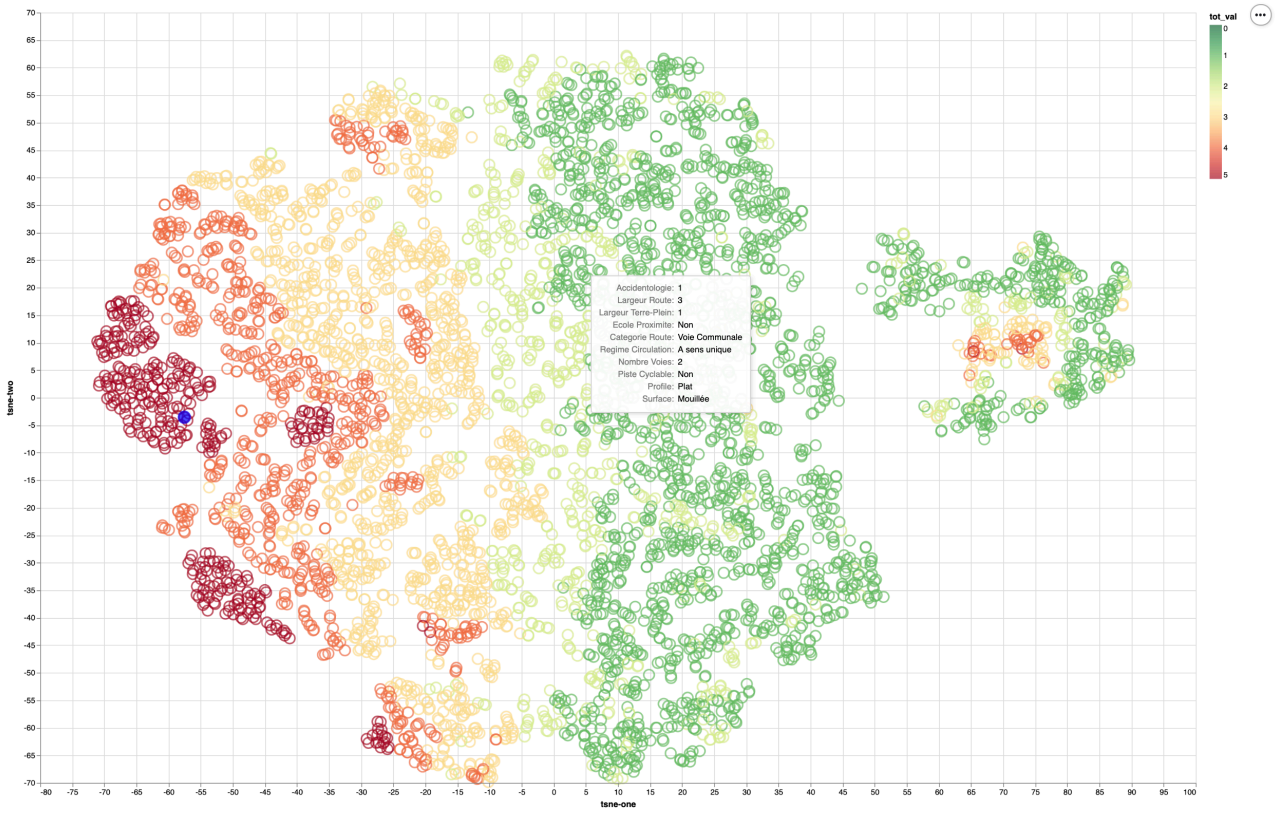
We decided to add interactivity at several levels :

- Through the tooltip
- By allowing the user to select the kind of road he's interested in.

Indeed, we created three buttons (Communal, Departmental, State) so that the different users we are targeting can all filter the most relevant information quickly. We believe that this removes noise from other road types, and avoids to select filters manually. For example, here is the T-SNE embedding of the communal roads in France.



The tooltip looks like this :



The blue dot corresponds to the user's input. We observe in this case that it is among the worse class. If the user wants to optimize the accident profile of the road, he might be interested in exploring the points at the green limit or at the orange limit.

e. Limits of the design

This design, due to its custom functionalities and quite original side, comes with several limitations :

- The main bias is that we have to think in terms of absolute number of accidents, since each record relates a single accident. However, there must be some dominant kind of roads in France, which means that it would be more relevant to think in terms of accident rate than absolute number of accidents
- There could be a cursor to show the evolution over time of this graph
- The interactivity could be improved to allow, for example, the user to add new points on the map
- The algorithm could output a single best road that is the closest (in terms of distance on the T-SNE map) to the input road characteristics
- The T-SNE embedding cannot be used to make a fit-transform and cannot compute the embedding of additional points. Therefore, exploring other dimension reduction techniques

such as PCA could help bring interactivity on this level. Moreover, the T-SNE embedding does not output the same graph at each run

- The data cleaning can be improved, by adding some outlier detection for example

IV. Monitoring

a. Motivation

Monitoring the number of accidents on roads is not an easy task. Communes and departments might be aware of roads in which there are empirically more accidents. However, it might become really hard to deal with the large amount of variables observed at each accident : weather conditions, hour of the day, number of passengers… Moreover, it is really hard to visualize the accident profile of roads at different scales (a single commune, or the whole state), compared to other places (how does a commune perform compared to another), over time (year by year) given filters.

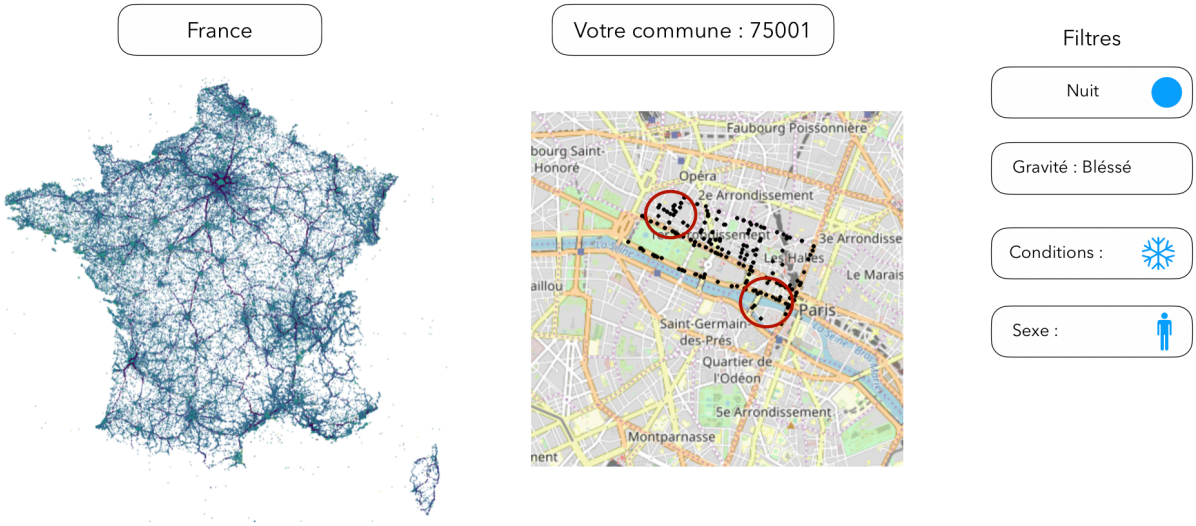
For this reason, we wanted to develop a tool that looks really simple at first sight, i.e two maps and a set of filters. This tool is however really powerful, since it covers all the needs described above. The user can set filters and observe the interaction between them. For example, the end user will be able to monitor the accidents that occurred during the night, implied male drivers, with snowy conditions, at the level of his commune or of the department for example.

The filters can be : night vs.day, death vs. unharmed, weather condition, year, sex of the driver… The main idea behind such a complete tool is to allow authorities to gain insights on a large dimensional problem, with cumulative filters : for example, deaths on snowy days, by night, since 2005. Clusters can then be visually defined, and actions can be taken from this perspective.

A concrete way to apply the insights gained from this design would be for a department to take a look at the design on certain weather conditions, in certain luminance conditions, identify clusters, and decide to take special actions from this observation. It could mean additional security measures, reduced speed, radars…

b. Design sketch

Our initial sketch of the design looked like this :



The main feature is to visualize two maps at the same time, and the location of the accidents on the two maps according to filters. The interesting point it to be able to compare different scales and how they react to filters :

- Commune vs. Commune
- Commune vs. department
- Commune vs. Region
- Commune vs. state
- Department vs. region
- Department vs. state
- Region vs.state

Instead of computing several graphs, all the information is loaded on a single graph. The filters are cumulative, which means that the users can explore the interactions between all filters.

Since we have a lot of features, it was tempting to display them in additional dimensions instead of filters. We could play on the size of dots, on their color, on the marker, on the angle of the marker… In the end, the user would have surely been lost, which is the reason why we chose filters.

c. Data processing and challenges

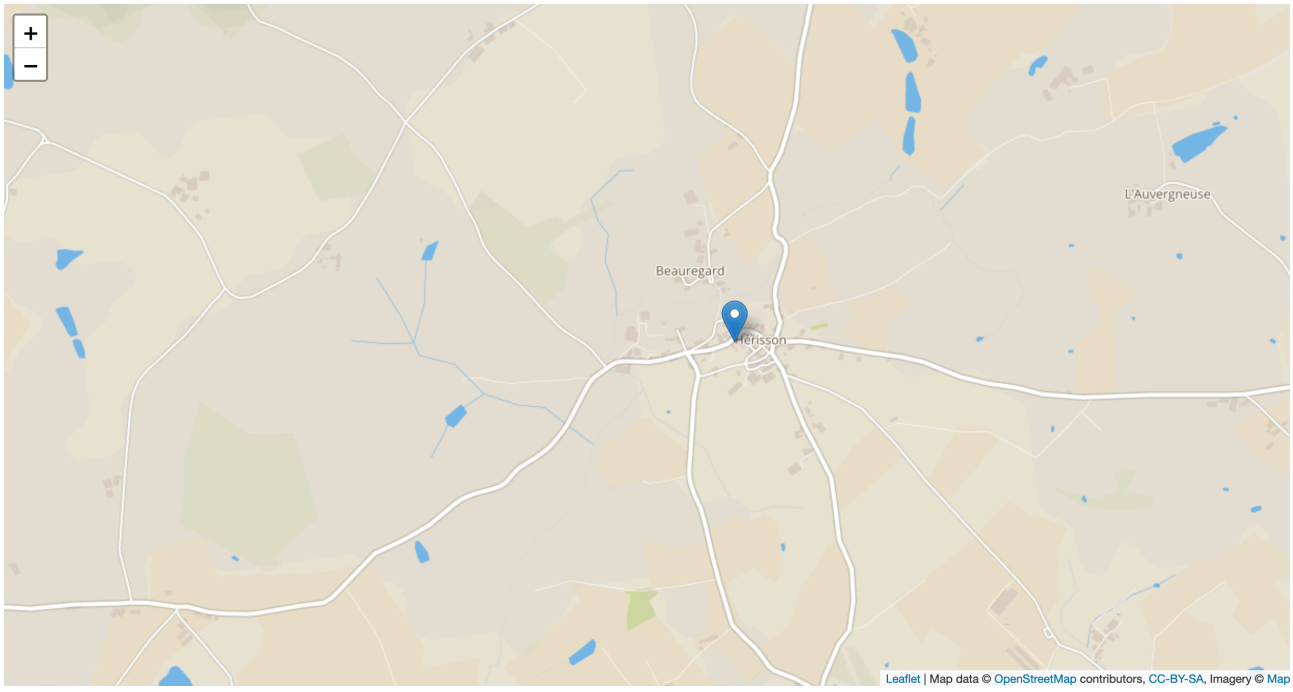
For this visualization, we focus on the characteristics dataset. Many entries in the characteristics dataset do not contain GPS coordinates. These observations had to be removed.

Then, some outliers were located in Switzerland for example or in neighboring countries. We also removed those points. We also chose to only focus on the metropolitan area, and exclude the DOM-TOM of our analysis.

We also expect the user to use the tooltip provided to get the micro-macro view and explore the details of a single accident if needed.

d. Technologies and outcome

We use Folium map plotting tool as our main framework. Folium is relatively easy to configure in Python, and has some good interactive features. We included a tooltip on the hover with details about the accident. The filters can be defined with buttons on the side of the map.



Folium also supports export to the HTML format.

e. Limits of the design

The design is interesting and we believe can bring value to authorities in charge of these questions. However, it suffers from some limitations :

- The data source is updated once a year only, which means that the data of year 2019 are not present at all
- The interactivity remains limited due to the computation time of the different views
- We could expand the filters list

V. Conclusion

We presented in this report the source of data we chose, the users and tasks our design will target, each design individually, along with its advantages and defaults.

The major challenge was to process the data properly, manage the diversity of the features, the outliers, the key to join the tables…

Once we identified the right designs to match the tasks we identified around the users and the datasets, the implementation remained challenging, due either to the concept of the design (i.e the T-SNE tool), the trade-off to make between the interpretability and the completeness (i.e the Sankey diagram) or the amount of data to display and the interactivity requirements (i.e the interactive map).

We really enjoyed working on these topics and improving the concepts of our different designs session after session.