Historic Road Accidents in France – A Study

Carlos Natale

Stephen Waller

Ehsan Jafari

Report 1 : Exploration, Data Visualisation and Data Pre-Processing

**Context**

For this project we are to conduct a detailed analysis of the history of road accidents in France, using data provided by the French government, via the ministry of the Interior and Overseas Territories.

The purpose of this is to analyse the data available, and from a technical viewpoint, clean and format the data to try and understand any correlation between data items that would provide insight and statistical significance of the data objects, which *could* lead to improved predictions of the likelihood of road accidents and their seriousness.

This analysis, in the long term could have benefits to both the population of France, and the French government in an economic sense, as the reduction of accidents or their severity would have benefits, due to the socio-economic impact of road accidents – such as road closures, injuries or even death.

* Context of the project's integration into your business.
* From a technical point of view.
* From an economic point of view.
* From a scientific point of view.

**Objectives**

The main objective of this project/report is to have the data analysed, processed, cleaned and combined as required, in order to have a singular dataset that can be used for complex statistical analysis and modelling.

All members of the group have differing levels of expertise in data science, but we are all working towards improving our knowledge as part of the training with DataScientest.

As this is a project/dataset provided by DataScientest I am not aware of this being completed by other members of our cohort,

* What are the main objectives to be achieved? Describe in a few lines.
* For each member of the group, specify the level of expertise around the problem addressed?
* Have you contacted business experts to refine the problem and the underlying models? If yes, detail the contribution of these interactions.
* (Are you aware of a similar project within your company, or in your entourage? What is its progress? How has it helped you in the realization of your project? How does your project contribute to improving it?).

**Understanding and manipulation of data**

**Framework**

The data that is used for this project is sourced form the Ministry of the Interior and Overseas territories , and is freely published by the French government via the following webpage;

[Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2022 - data.gouv.fr](https://www.data.gouv.fr/en/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2022/)

The data on this site contains details of road accidents from 2005 to 2022 – for the purposes of this project, as a team, we have agreed to focus on the data sets from 2019-2022, as 4 years worth of recent data would be able to give a valuable enough insight, and reducing the risk of results being skewed by using more historic data where the severity of an accident may be affected by differing standards in vehicle safety etc.

The datasets themselves consisted of 4 .csv files per year - “usagers”, “vehicules”, “lieux” and “carcteristiques” - users, vehicles, places and characteristics related to an accident.

These data sets are quite large in volume with the 2022 data across all 4 .csv containing 331,759 rows of data for that year alone.

* Which set(s) of data(s) did you use to achieve the objectives of your project?
* Are these data freely available? If not, who owns the data?
* Describe the volume of your dataset?

## Relevance

As a team we have agreed on the “grav” (Severity) field as the target variable for this project as a whole, this field describes the severity of injuries as a result of this accident. Ranging from “unharmed” to “Killed”, this target was chosen as the severity of an accident can be directly linked to a number of factors in the datasets, such as location and speed.

There are limitations to the data, as this data is only present where an accident has been logged by a law enforcement unit (Police etc.) , so any accidents that have occurred without a report being written, or attendance of law enforcement, would not be present on the dataset(s), so a complete picture may not be possible.

Variables that seem the most relevant to the project, and this stage appear to be “Lighting Conditions”, “Weather Conditions”, “Time of Day”, “Road Category” and “Gender” - however further detailed analysis would need to be conducted to confirm the variables most linked to our target.

-

* Which variables seem most relevant to you with regard to your objectives?
* What is the target variable?
* What features of your dataset can you highlight?
* Are you limited by some of your data?

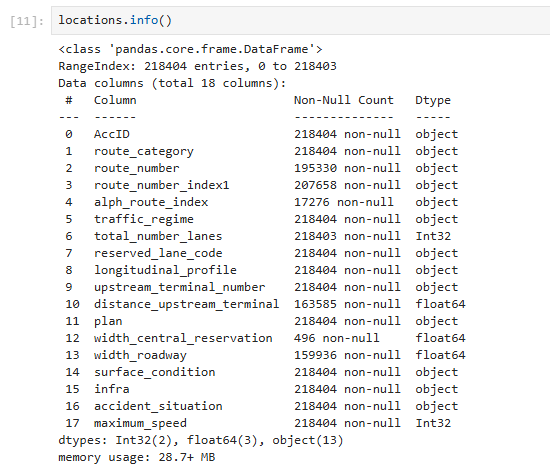
## Pre-processing and feature engineering

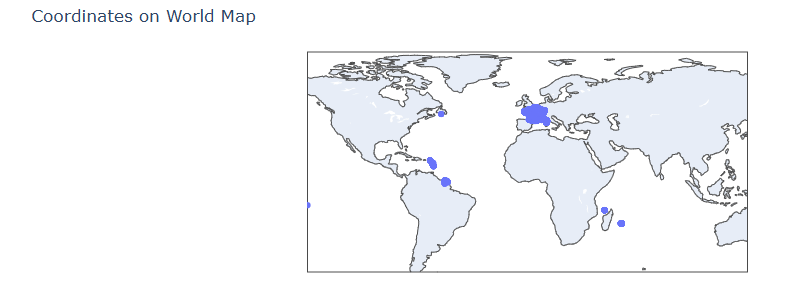
There was a large amount of pre-processing done for this data, due to the fact that there were 16 different .csv files across 4 distinct data sets.

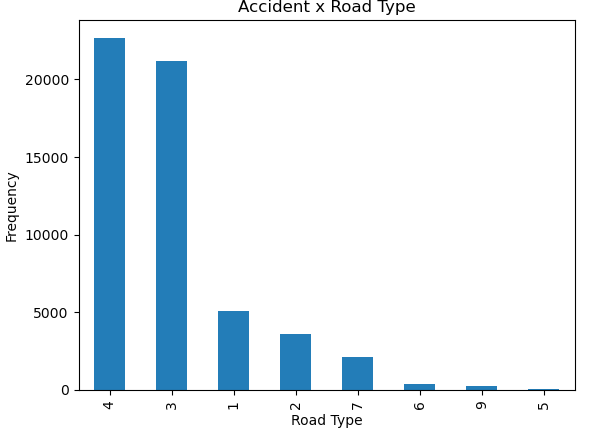
The first part of this process was to combine the differing years for each distinct dataset (i.e. the “lieux” dataset for 2019-2022) – as well as converting the column names into English to enable easy reading by the entire team as English is a more common language than French.

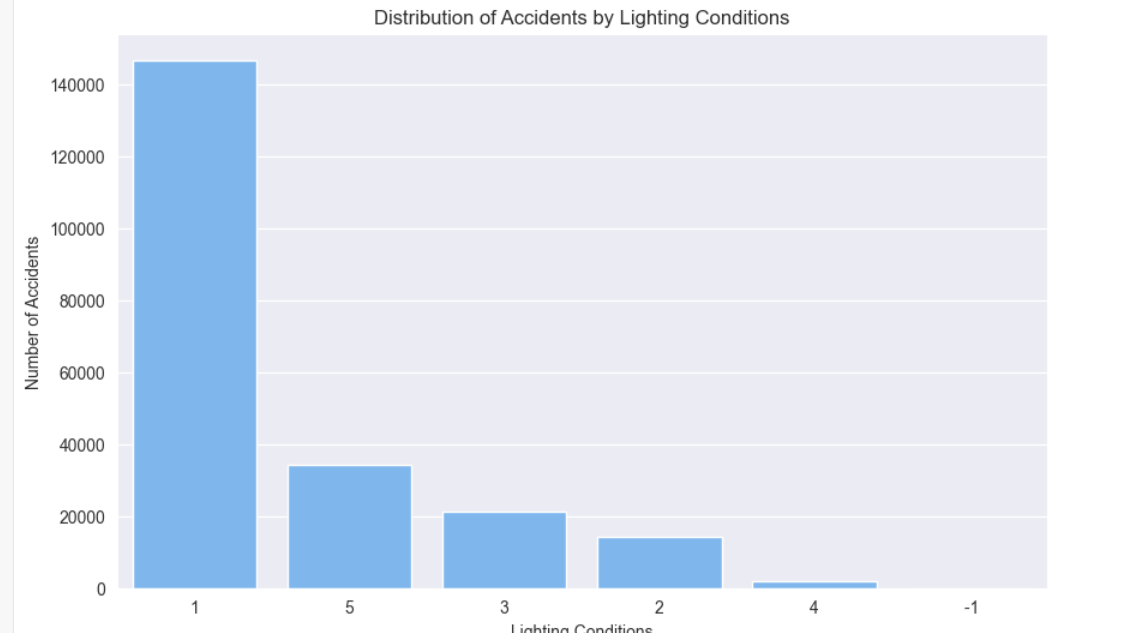
After this was completed, the next stage of pre-processing was to analyse the data, especially looking for Null/NaN values and values that are out of range. This was done via a combination of data analysis (looking at he total number of rows compare to amount of Blank/Null values) as well as various visualisation techniques such as Bar Plots plots to identify outliers in the dataframes.

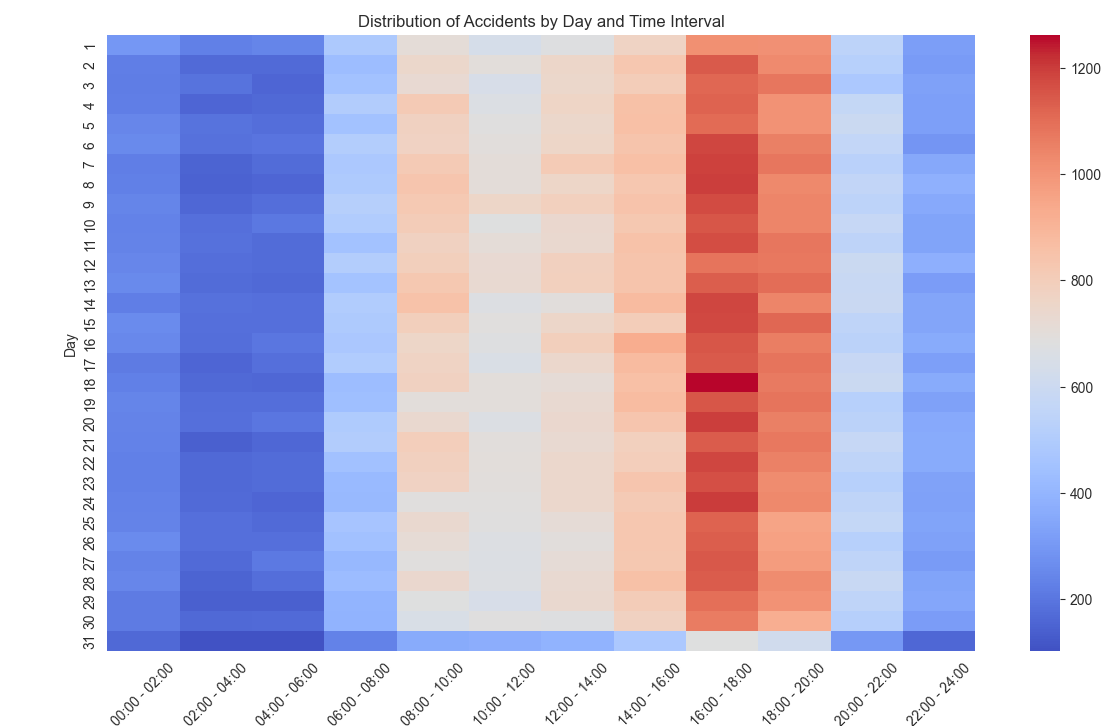
Data cleaning continued using data visualisation methods to determine what data items are key to keep in our final combined dataframe, as well as to give valuable insight onto what data items are less significant, and therefore cam be dropped. Sme examples of these are detailed below.

[[1]](#footnote-1)

[[2]](#footnote-2)

[[3]](#footnote-3)

[[4]](#footnote-4)

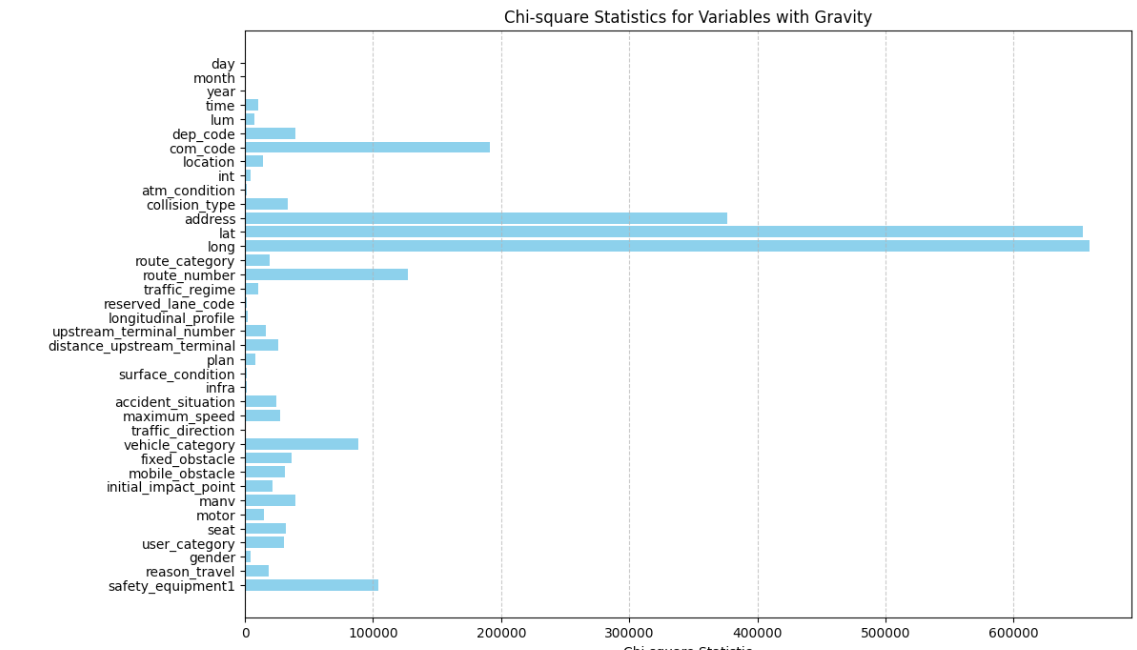
[[5]](#footnote-5)

Other element of dimension reduction have been considered – such as combining the Year/Month/Day fields into an overall date item (therefore reducing 3 columns into 1), as well as proving outliers in some data items that make these rows candidates for removal – such methods would be analysis such as chi-Squared on data items or IQR analysis to identify outliers .

## Visualizations and Statistics

As data processing and visualisation has continued, with a cleansed and combined set of data we are able to produce a set of visualisations based upon this data – in order to see what variables in our complete dataset are relevant to model against and have a predictions against the “grav” (Severity) target variable.

A statistical analysis of the regions of France present in the data set, using the IQR method of identifying outliers, also confirmed that French overseas territories are outliers in the dataset(s) are should be removed .

[[6]](#footnote-6)

Relationships between our target variable of “grav” (Severity) have been identified via various methods of analysis, for example chi-square analysis has been done on the overall dataset against the “grav” field showing that there are key variables that are significant to be used to model against in he future – and example of this would be the “route number” which gives details on specific roads / road types – which have a correlation with the severity of an accident. As part of modelling, this would be a good data item to use, to see how strong the correlation between road types and severity – a hypothesis of this would be more rural roads would have a higher severity of accident.

* Have you identified relationships between different variables? Between explanatory variables? and between your explanatory variables and the target(s)?
* Describe the distribution of these data, distribution, outliers.. (pre/post processing if necessary)
* Present the statistical analyzes used to confirm the information present on the graphs.
* Draw conclusions from the elements noted above allowing them to project themselves into the modeling part

**Assessment methods:**

**Reconstituted professional situation: from a set of company data, the candidate must implement various pre-processing and data augmentation to make them usable through machine learning techniques.**

1. This is showing that the field "width\_central\_reservation" only has 496 Non-Null values out of 218404 - making it an ideal candidate for removal as non-significant [↑](#footnote-ref-1)
2. This is a world map produced plotting the accidents and their location – showing French Overseas territories are included – this could skew the results due to differing reporting methods and/or locals laws [↑](#footnote-ref-2)
3. This bar Plot shows the Accident Frequency x Road Type – as you can see the occurrence of “5” (Off-Grid) is very minor – and therefore a prime candidate for removal as not significant enough to form part of modelling. [↑](#footnote-ref-3)
4. The graph shows the distribution of accidents by lighting conditions. Most accidents occur under light condition 1 (Daylight), with nearly 145,000 incidents. [↑](#footnote-ref-4)
5. This distribution shows a strong link between the frequency of an accident and the time of day it occurred – this data would be a key item to determine the severity (as our target Variable) [↑](#footnote-ref-5)
6. Diagram showing the Chi-square Statistics for various variables against our Target of “Grav”(Severity) – this is showing some key variables that are to be picked and used in our forthcoming model. [↑](#footnote-ref-6)