Project Report for Data Analytics: SS2018

Saikat Roy, Vijayesh Kumar Das and Pooja Bhatia
Institute for Informatics
University of Bonn
Summer Semester 2018

***Abstract*—We investigated the *France 2005-2016 Road Accidents Dataset* as part of the final project of the Data Analytics lecture in SS 2018 at the University of Bonn. The following report is a summarization of findings on the data using various analytical techniques.**

I. Introduction

This report is a collection of data visualization and analytics techniques applied on the *2005-2016 France Road Accidents Dataset* from Kaggle. The following report contains a short desription of the dataset and why it was chosen, the motivation behind the analysis, the methodology and results from the analysis followed by a conclusion and possible improvements.

II. Motivation

The improving conditions of roads and traffic management in a country ideally should provide an improvement in traffic conditions and gradual reduction in the number of traffic accidents witnessed by a country. The presence of data for road accidents over a large temporal duration such as a decade allows us to validate this rather general hypothesis empirically. This is considered to be the main motivation for our analysis of the road accidents dataset used in this work.

Accidents are associated with a number of causative factors many of which are co-occurring and the incidents of which also vary with time. A secondary goal of this project was to perform a basic analysis of the variation of these factors (such as road conditions, age of users involved, weather conditions, illumination) associated with road accidents and if possible track their variation over time.

III. Dataset Description

The *Accidents in France from 2005 to 2016* Dataset is available in Kaggle at the following URL as of the time of writing this report. The main dataset has the *caracteristics, holidays, places, users, vehicles* datasets as comma-separated value (csv) files. We primarily work with the *caracteristics*, *places* and briefly with the *users* datasets.

IV. Methodology

The dataset in question is suitable for a variety of visualizations to analyze the distribution of various factors in the context of road accidents. We analyze the distribution of various causative factors using visualizations such as bar graphs and line charts and try to talk about common trends in the same. In addition of investigating the distribution of accidents in monthly, yearly and hourly, we also look into the usual suspects such as road surface condition, atmospheric conditions, lighting, road type and additionally also check the distribution of people by age involved in accidents.

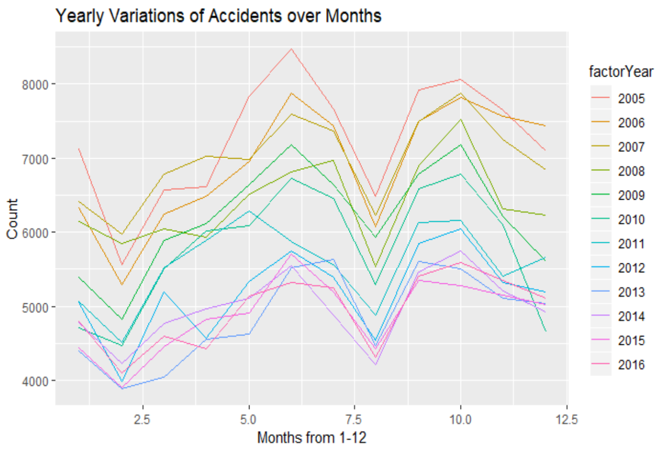

Fig. 1. Yearly Variations of Accidents over Months (1-12)

Assumption: *It is assumed as part of our analysis methodology that the logging of accidents over the course of the year is complete and geolocalized to a constant region in France.*

V. Results and Analysis

For ease of understanding, the results and associated figures have been organized according to order in the accompanying *R Markdown PDF*.

A. Yearly Variations of Accidents over Months (1-12)

It is clearly seen in the plot in Figure 1 that for yearly variation over months, there is decrease in the number of accidents as from the period of 2005 to 2016. This clearly reinforces our central assumption that gradual improvement in roads and improvement in traffic management does result in reduced number of road accidents. This plot also demonstrates a uniquely that accidents over the years, although reducing, do follow a more or less consistent pattern, with consistent peaks and dips each year.

B. Monthly Variations of Accidents over Years (2005-2016)

The plot in Figure 2 reinforces with finality that accidents have decreased in time over the last decade due to improving road conditions. A clear negative trend is observed for the line plot for every month for each of the 11 years in analysis.

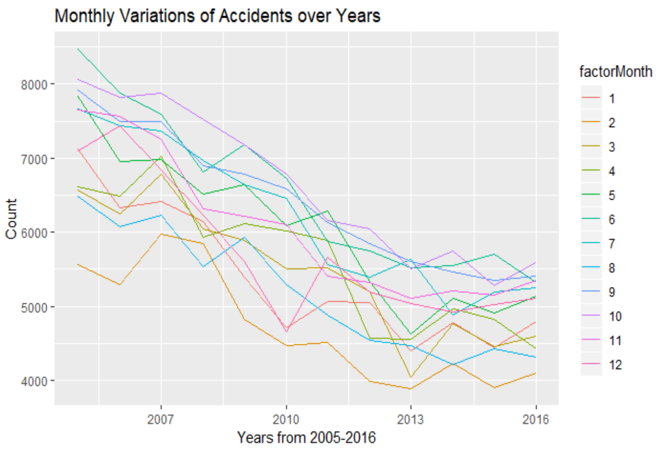

Fig. 2. Monthly Variations of Accidents over Years (2005-2016

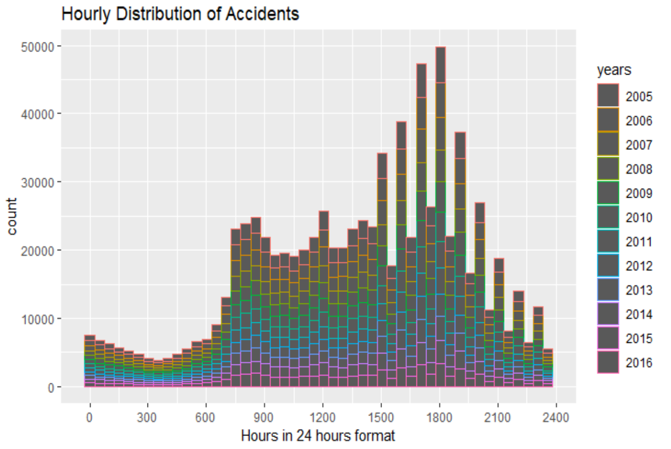

Fig. 3. Hourly Distribution of Accidents

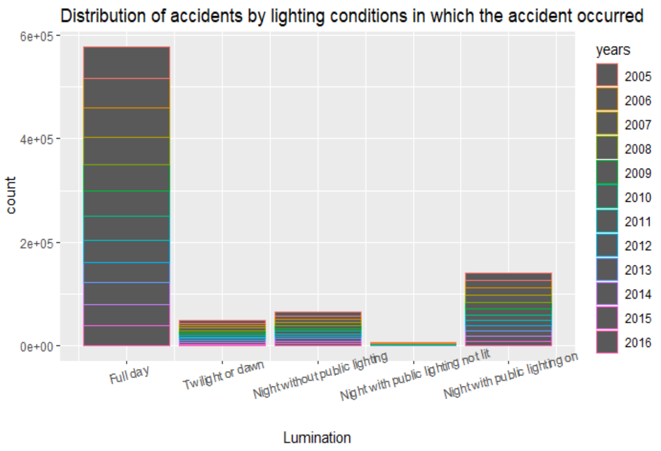

Fig. 4. Distribution of accidents by lighting conditions in which the accident occurred

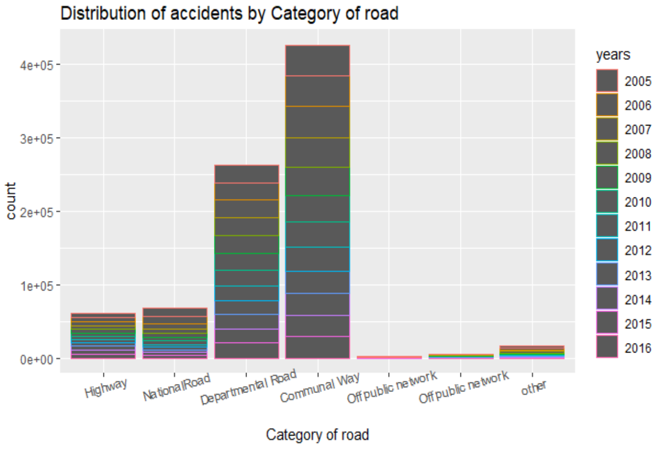

Fig. 5. Distribution of accidents by Category of Road

C. Hourly Distribution of Accidents

The hourly distribution of accidents in Figure 3, shows a clear peak of accident count at around 1700-1800 hrs. This is coincident with the fact that that specific timing coincides with the period during which people in any region usually finish their workday and are probably on the road (many in private vehicles) on the way to their homes. This results in an higher amount of traffic in the roads, which coincides with a higher accident count. Also, it is interesting to note that this trend has not changed much over more than a decade.

D. Distribution of accidents by lighting conditions in which the accident occurred

While one may believe that accidents usually happen in bad lighting conditions, in Figure 4, a significant number of accidents are seen to happen in what is classified in the dataset as *Full Day* conditions. This reinforces the idea that empirically accidents happen when the drivers have the *least* reason to be cautious. Alternatively, this can also be because more vehicles are on the road during the day, thus increasing the probability of accidents. Night with public lighting, while sizable and an obvious condition when accidents can happen, comes a distant second.

E. Distribution of accidents by Category of Road

In Figure 5, most accidents in France are seen to happen on what is defined as a *Communal Road* which are among the smaller roads maintained by communities. This is understood that such roads in general are narrower with higher incidents of accidents. The second is Departmental Roads which are slightly larger. There is a clear pattern with decreasing number of accidents with the size of the road as national roads and highways have significantly less number of such events.

F. Distribution of accidents by Atmospheric Conditions

Figure 6, shows that most accidents happen in conditions of clear weather reinforcing our road lighting analysis that*most accidents happen when there is least need to be careful.* Understandably the second major atmospheric state during accidents is seen to be light rain. Interestingly cloudy also is seen to be another state where accidents occur — a possible explanation can be that obvious indications of oncoming

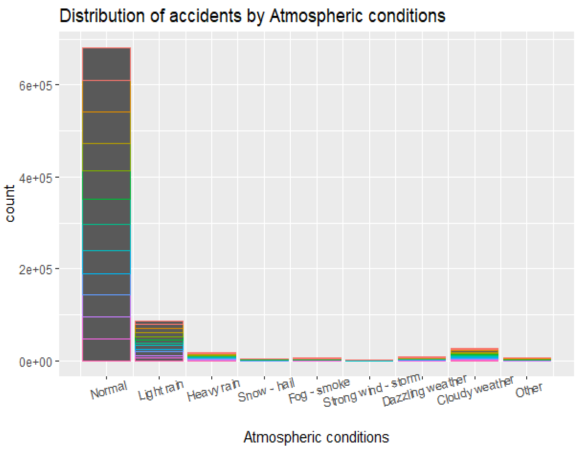

Fig. 6. Distribution of accidents by Atmospheric Conditions

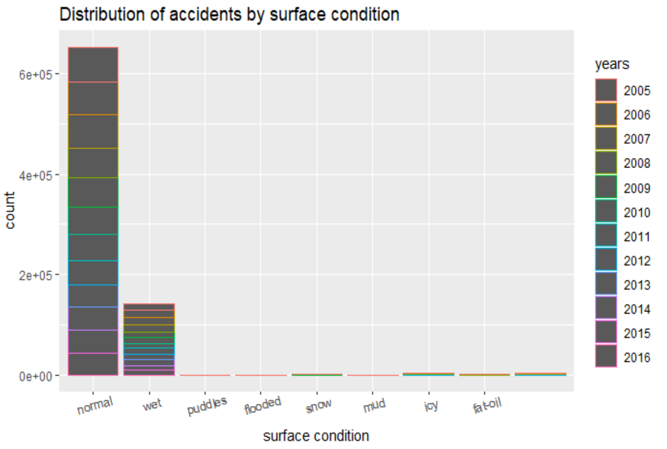

Fig. 7. Distribution of accidents by surface condition of Road during accident

downpour might encourage people to drive rashly in an effort to *not get caught in the rain*, thus leading to accidents.

G. Distribution of accidents by surface condition of Road during accident

Again in Figure 7, the surface condition of roads is seen to coincide with Figure 6 and the associated analysis of atmospheric conditions that accidents usually happen in clear conditions while the second most prevalent is during rain and associated wet road surfaces.

H. Distribution of Accidents according to the age of persons involved

Figure 8 shows the distribution of ages of people involved in accidents in the 11 years under analysis. This shows a clear peak around 25-30 years of age. This may not, however, lead to the obvious conclusion of *young people drive rashly*. It can also indicate that a higher number of younger people may be on the roads (that is, possessing vehicles and driving) and this may lead to more of them being involved in accidents.

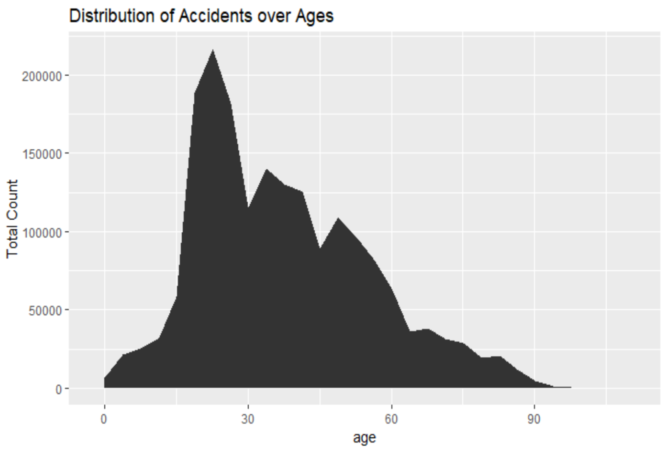

Fig. 8. Distribution of Accidents according to the age of persons involved

VI. Possible Improvements

- The analysis of various factors co-occurring during accidents is an issue that could be explored.
- The age distribution of people in accidents over the years can be explored to see if the age of those involved have changed over the years - thus helping target ads and other materials helping educate on road accidents towards a particular people of a particular age.

VII. Conclusion

In conclusion it is seen that our premise of the decreasing incidents of accidents over the years does hold based on the analysis of the data of French accidents from 2005-2016. Also, the analysis of various factors of the environment during accidents as well as the age of those involved reveal insightful information which can be used for targeted preventative measures.