SI618 Data Manipulation and Analysis

Project 2 Report

Exploratory Data Analysis of Traffic Accidents in France

Name: Xinhao Liao
Date: 8 December 2019

Contents

1 Introduction

1.1 Motivation

With the mass production and popularization of automobiles, road accidents have become a big issue in the past decades.

The details of road accidents have long been recorded and stored by the police in many places. Nowadays with the help of data science, we can study the accumulated data, and try to answer question like:

- What's the trend of the change of number of accidents every year and how is the trend affected by the month-of-the-year effects?
- How is the fatal rate affected by some variables like lighting conditions, atmosphere conditions, and collision types?
- For a traffic accident, how does its collision type related with other variables like lighting conditions and atmosphere conditions?

2 Data Source

2.1 Description

The dataset used is available in [1] which constains information about road accidents happening in France from 2005 to 2016. Three of the *csv* files provided, *caracteristics.csv*, *places.csv*, and *users.csv*, will be involved in this project. They respectively provide the basic information, the road conditions, and the victim information of the accidents.

An obvious limitation of this dataset is that most of the variables like lighting conditions, atmosphere conditions, and collision types are actually categorical data, with some integers from 0 to 9 indicating different conditions. This makes some analysis like clustering and principal component analysis meaningless or limited. Other data like year, month, date are integers representing the time. The dataset records 839985 accidents in France happening from 2005 to 2016.

3 Question 1: Accident number trending and month-of-the-year effect

3.1 Method

3.1.1 Data preparation and manipulation

Data from *caracteristics.csv* describing the year and month of accidents are used. Since there are only separate columns about year and month of accidnets, we need to combine them together to form a new column.

3.1.2 Data missing

All rows of data with missing information are omitted.

3.1.3 Challenges

The biggest challenge was to combine the separated year and month columns together to form a new column in *R*. To further label the x axis of plots in some date format, I have to make the new column in the format of some date in that month rather than of month so that it can be transferred to *Date* class of *R*.

3.1.4 Result and analysis

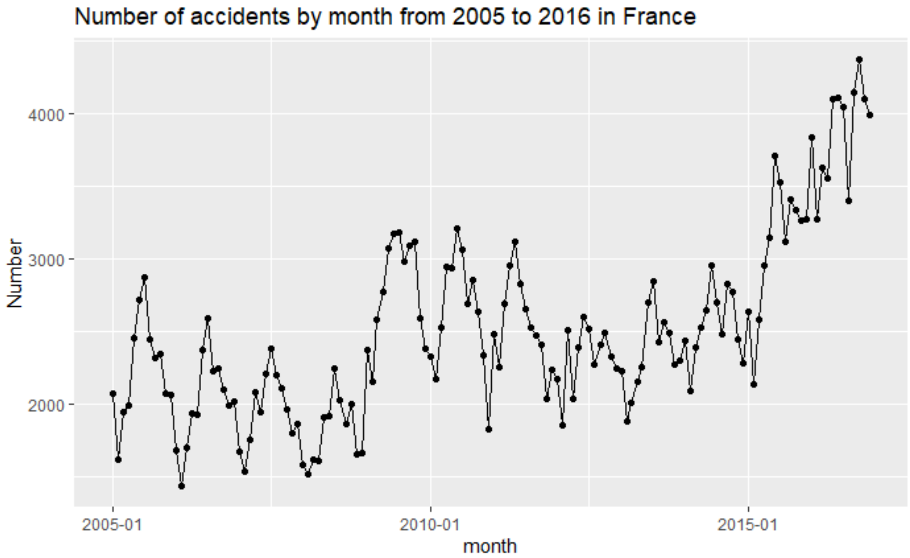

Figure 1: Number of accidents in France from 2005 to 2015 by month.

After processing the data, we need to further group data by month and then calculate the count of accidents in every month from 2005 to 2016. Then, we can draw a graph showing how number of accidents in France changes from 2005 to 2015 by month as shown in Fig. 1 above. However, there's month-of-the-year effects involved in the changes which needs to be eliminated.

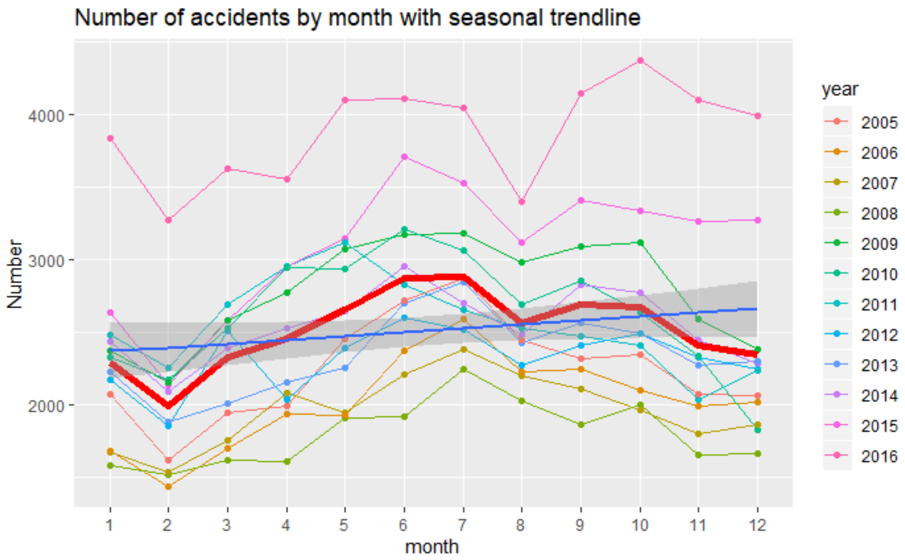

Figure 2: Number of accidents in France from 2005 to 2015 (with the red line as the average value).

Fig. 2 above shows the number of accidents by month from 2005 to 2016 in France, with the red line showing the average values by month. As we can see, the average values greatly fluctuates in a year, which clearly shows the month-of-the-year effect. As we can see, on average there are greatest number of accidents in July in France.

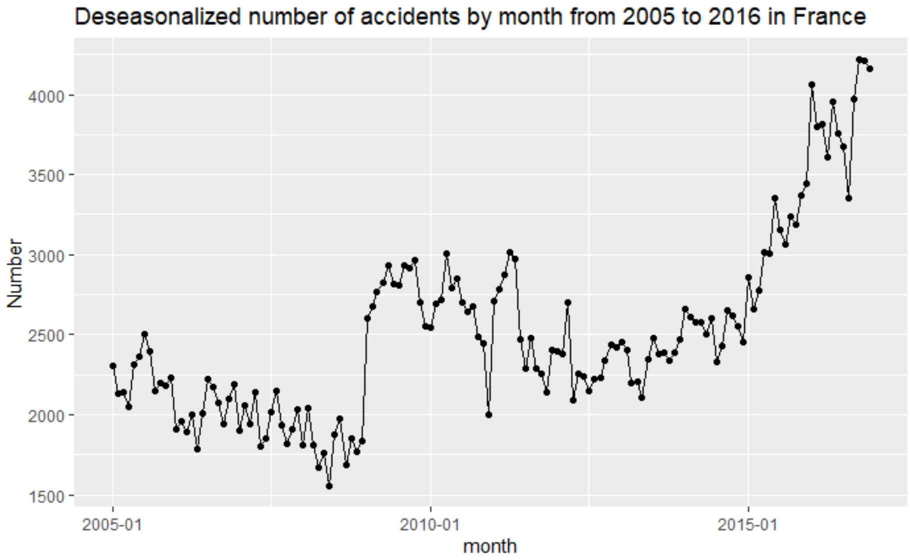

Figure 3: Deseasonalized number of accidents in France from 2005 to 2015 (with the red line as the average value).

To eliminate the month-of-the-year effect, mean variations are calculated and subtracted out from the original counts to make the mean line flat. After that, we can obtain the final result showing in Fig. 3 above.

As we can see, the number of accidents in France grew greatly from 2005 to 2016. The total number almost doubled. Although there was a slight declining trend from 2005 to 2008, the number climbed rapidly after that.

4 Question 2: Variables affecting fatal rates

4.1 Method

4.1.1 Data preparation and manipulation

Data from *caracteristics.csv* describing the lighting condition, the atmosphere condition, and the collision types of accidents, as well as data from *users.csv* describing the severity of accidents are used. Data from the two datasets are joined together based on the accident id number.

4.1.2 Data missing

All rows of data with missing information are omitted.

4.1.3 Challenges

There's no obvious challenge during the analysis. Everything goes smoothly as expected.

4.1.4 Result and analysis

The fatal rates for every condition of every variable is calculated respectively. Graphs showing how the fatal rates varies in different conditions are given below.

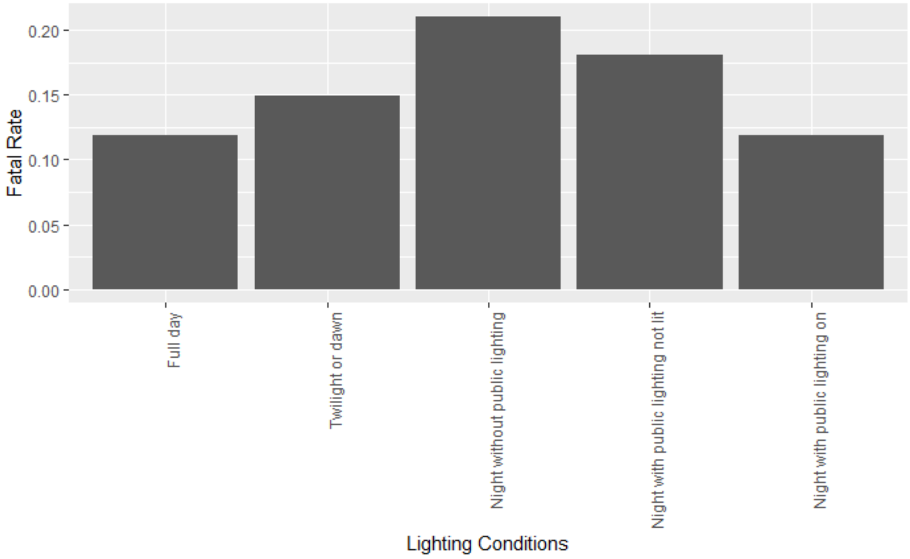

Figure 4: Fatal rates in different lighting conditions.

The plot above shows that the highest fatal rate occurs when it is night without pubic lighting. And both full day and night with public lighting lit gives low fatal rates.

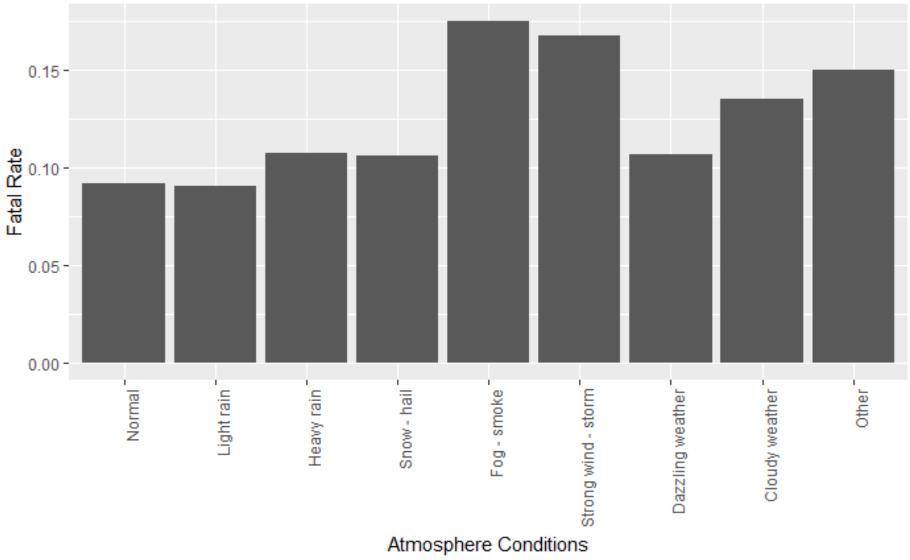

Figure 5: Fatal rates in different atmosphere conditions.

The plot above shows that the highest fatal rate occurs when there is fog or smoke. And both normal condition and light rain condition gives low fatal rates.

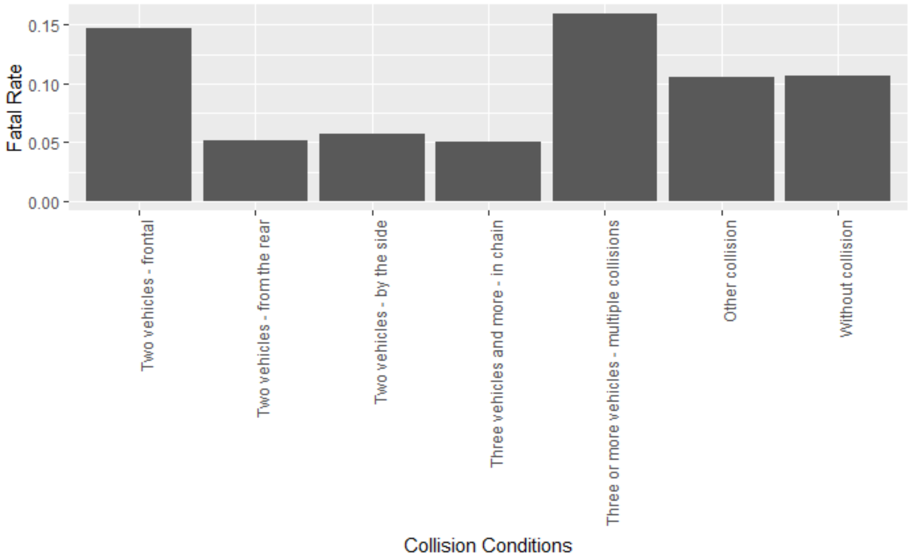

Figure 6: Fatal rates in different collision conditions.

The plot above shows that the two most dangerous collision types are respectively multiple collisions of three or more vehicles and the two vehicles frontal collisions. Collisions of three or more vehicles in chain and of two vehicles from the rear are relatively safer than other conditions.

5 Question 3: Collision types related to other variables

5.1 Method

5.1.1 Data preparation and manipulation

Data from *caracteristics.csv* describing the lighting condition, the atmosphere condition, and the collision types of accidents are extracted used.

5.1.2 Data missing

All rows of data with missing information are omitted.

5.1.3 Challenges

There's no obvious challenge during the analysis.

5.1.4 Result and analysis

Since we are studying relationships between categorical variables, we are not going to check some linear relationship as is done for continuous data. Rather, we can generate 2-d contingency table, and then draw graphs showing the composition of all the cases of different collision type cases.

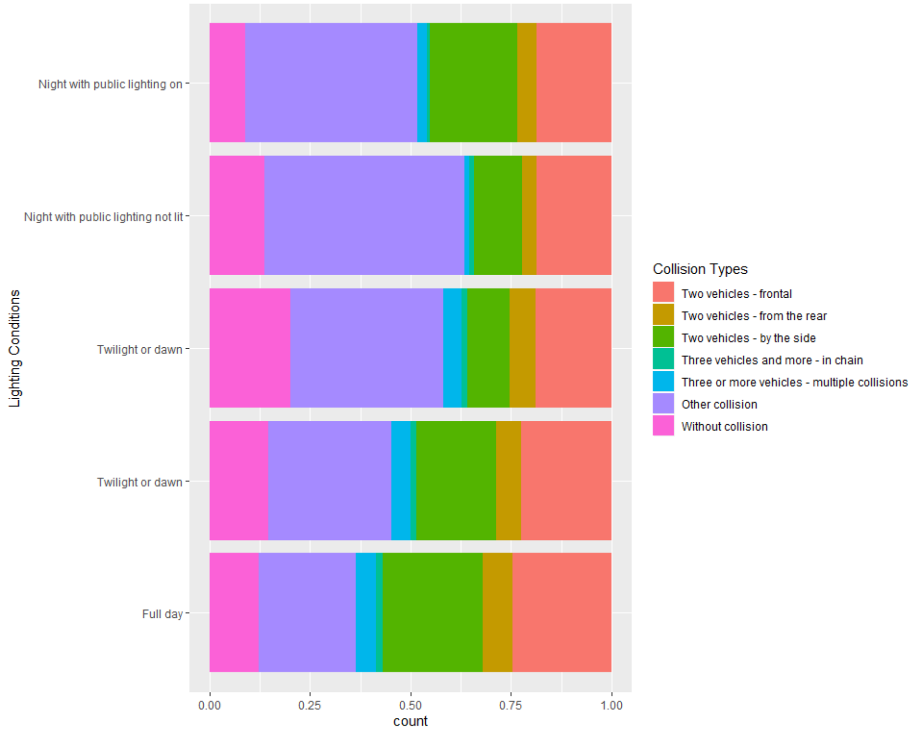

Figure 7: Relationship between collision types and lighting conditions.

The plot above shows that the relationship between collision types and lighting conditions. As we can see, all types of two-vehicle collision and the three-vehicle multiple collisions are most likely to happen in full day. The without-collision accidents mostly occurs at night without public lighting. Other uncovered collision types occurs mostly at night with public lighting unlit.

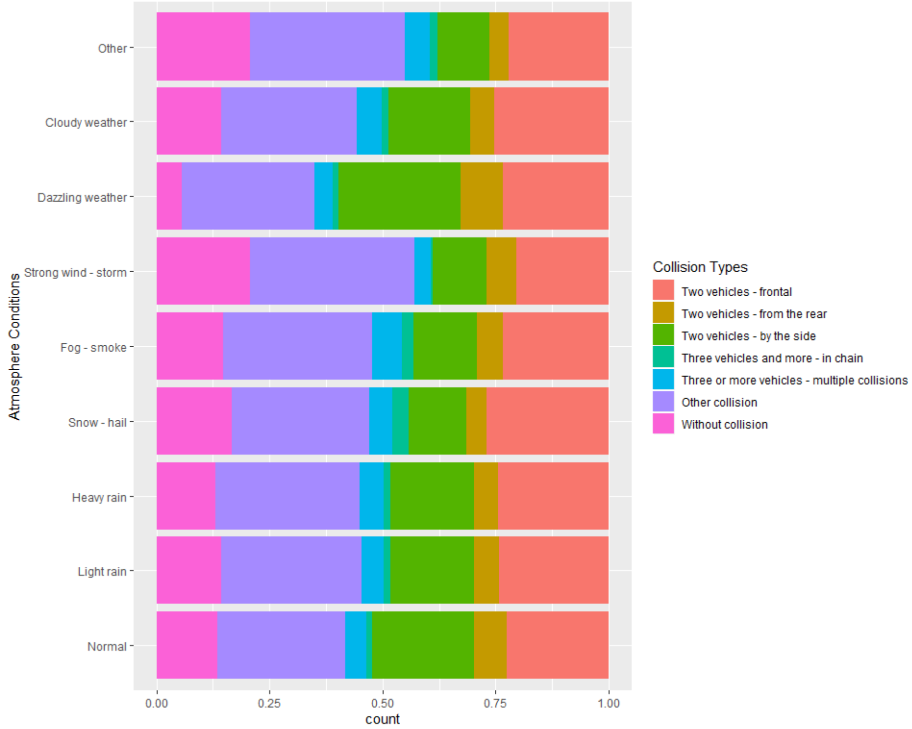

Figure 8: Relationship between collision types and lighting conditions.

The plot above shows that the relationship between collision types and atmosphere conditions. As we can see, two-vehicle collisions by the side or from the rear occur most in the dazzling weather; two-vehicle frontal collisions and three-or-more-vehicle multiple collisions occur most when there's snow or hail.

References

[1] "Accidents in France from 2005 to 2016", *Kaggle*. `https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016`