PROJECT REPORT

CAR ACCIDENTS IN BARCELONA - 2017

Submitted By:

Sherin Joseph

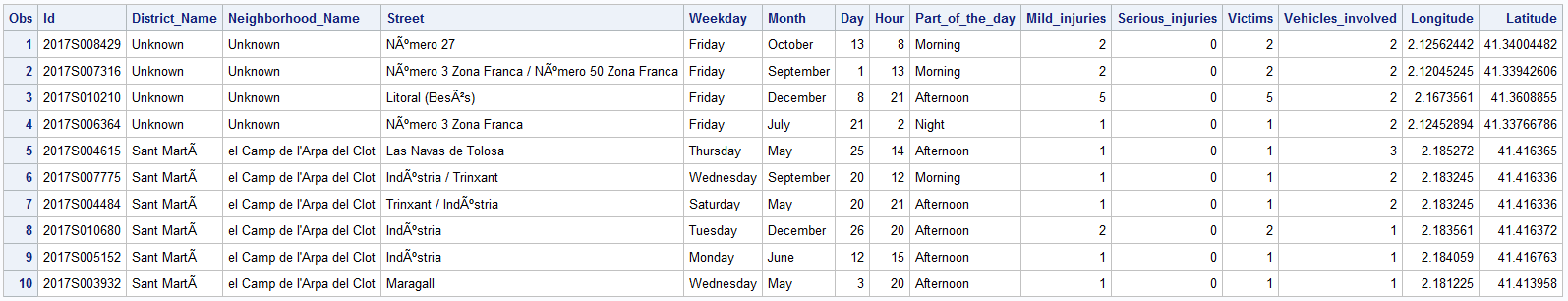
23 / Nov / 2021

# INTRODUCTION

The project emphasis on the data acquisition, data cleaning and analysis of car accidents that occurred in the state of Barcelona in the year 2017. The data is obtained in the form of csv file from the state department for detailed study and analysis of the car accidents occurred and acquire valuable insights from the data. After the analysis recommendations are provided to improve the traffic system and take necessary actions to reduce the accidents in the coming years.

# BACKGROUND

The observational data shows that there was a total of 10,339 accidents that took place in Barcelona in the year 2017. The key attributes in the dataset are District Name, Neighborhood Name, Weekday, Month, Hour of day, Injuries and vehicles involved in the accident. The sample of dataset is shown below.



The data consists of both numerical and categorical data as shown below.

Table

Description automatically generated

We have chosen “**Victims**” as the target variable and the rest as the predictor variables.

# OBJECTIVES

The objectives of the analysis are the following:

* Which district and neighborhood accounted for the maximum number of accident victims in 2017.
* Are victims affected different across different districts.
* What time of the day has the maximum victims reported and why so.
* Are victims affected significantly different on weekdays than weekends and why so.
* Which month of the year has highest and least victims and how are they significantly different from other months.

# METHODOLOGY

We have used Base SAS 9.4 software for the analysis of the data set. Statistical operations such as Chi-square test and Multinomial Logistic Regression are done to draw valuable insights from the data provided. The relevant variables such as “District Name”, “Neighborhood Name”, “Weekday”, “Month”, “Hour of day”, “Mild Injuries”, “Major Injuries” and “vehicles involved” are chosen as the X variables and the “Victims” is taken as the target variable. We have found that there are no missing values in the data and hence all observations are considered for analysis. The variable “Street” is excluded since it does not play significant role in predicting the victims. The variable “Hour” is further split into time frames “12AM-6AM”, “6AM-12PM”, “12PM-6PM” and “6PM-12AM”. Hence part of day variable which has same data is excluded. The variables “Latitude” and “Longitude” are excluded from modelling because they carry geological information and the same can be obtained from “District Name” and “Neighborhood Name”. There are no missing values in the dataset and hence all observations are taken into consideration for the analysis purpose. However, there are some “Unknown” values in the District Name and Neighborhood field and they are replaced by the mode of their corresponding columns. The final outlook of the analysis is as shown below:Diagram

Description automatically generated

# EDA

The univariate analysis of the month reveals that months are almost equally distributed but however, November has the highest accident rate of almost 9.59% and August has the least with 6.31% of accidents reported in the year 2017. Along with that the part of the month such as “start”, “Mid” and “End” of the month account almost same number of accidents and hence those does not have significant importance in the analysis. We could see that “Friday” has maximum number of accidents making almost 17% and “Sunday” reported the least with 8.66%. The time of the data has an important role in the accidents because the afternoon time especially between 12PM and 6PM reported to have the highest accident rate compared to other times.

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

The geographic data analysis shows that the District “Eixample” and the neighborhood “la Dreta de l’ Eixample” has reported the highest numer of accidents during the year. The areas that had least counts were “Gracia” district and the neighborhood “Torre BarA”.

Chart, bar chart

Description automatically generated Chart, table

Description automatically generated

On taking in to consideration about the injuries reported, major count of the mild inuries reported were less than or equal to 4 and the serious injuries were 0.

Chart, bar chart

Description automatically generated Chart, histogram

Description automatically generated

The victims less than or equal to 1 were extremely higher than others and the accident involving vehicles less than or equal to 3 were reported way higher than other groups.

Chart, bar chart

Description automatically generated Chart

Description automatically generated

On analysis of the target variable “Victims” across the study variable District, “Eixample” had the hishest victims in all categories such as less than or equal to 1, 2 - 3, 4 - 5 and greater than or equal to 6.

Chart, bar chart

Description automatically generated

When the victims are studied corresponding to the variables such as mild inuries and serious injuries, we could see that as these parameters incerases, the number of victims also increases. This is demonstarted by the graphs below.

Chart

Description automatically generated Chart, bar chart

Description automatically generated

Same is the case with the variable vehicles involved because, as the number of vehicles involved in the accidents are more, the chances of having more victims are more. This is explained as shown in the stacked bar char below.

Chart, bar chart

Description automatically generated

As mentioned before, the weekdays are subjected to accidents and hence creating more victims. We could see that across the the categories of victims, Friday stands top in the list and Sunday falls in the last position creating the least number of victims.

Chart, bar chart

Description automatically generated

Since the accidents are more during the afternoons, victims also increase during that time period. Early moring time from 12AM to 6AM has the least victims because of the least traffic during that time period. The comparison of the timings and the victims reported is as below:

Chart, bar chart

Description automatically generated

A geographic representation in the form of scatter plot that plots the victims across the latitude and longitude tells us the areas which has reported highest victims and this information is useful for creating new policies expecially for the high accident prone areas.

Chart, scatter chart

Description automatically generated Map

Description automatically generated

**Chi-Square Summary**

For finding out the predictor variables that have significant effect on the target variable “victims”, we have performed chi-square test since the target variable is categorical. A summary of the chi-square test tells us that the variables such as “District Name”, “Neighborhood Name”, “Weekday”, “Month”, “Hour”, “Mild Injuries”, “Major Injuries” and “vehicles involved” have a significant impact on the victims. The summary chart is furnished below for reference.

Table

Description automatically generated

# MODELLING

The target varibale beeing categorical, we have done Logistic Regression model to evaluate the odds of different study variables that has impact on Victims. The reference group for the victims is “>=6”. In the case of the predictors, the reference group that were chosen are District Name (“Eixample”), Neighborhood Name (“la Dreta de l’Eixample”), Weekday (“Sunday”), Month (“August”), Hour (“12AM-6AM”), Mild Injuries (“<=4”), Serious Injuries (“2”) and vehicles involved (“>=6”). The code for the logistic regression is shown below.

Graphical user interface, text, application, email

Description automatically generated

The log of odds ratio of the predictor variables are obtained as the result of the logistic regression and the major results are as follows.

Graphical user interface, application

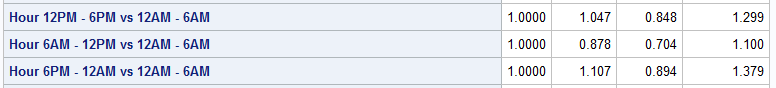
Description automatically generated with medium confidence

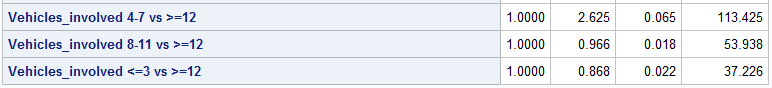
Table

Description automatically generated

A picture containing table

Description automatically generated





# CONCLUSION

To conclude, the project gave the following insights:

* The odd of District “Eixample” and neighborhood “la Dreta de l’Eixample” is significantly different from others.
* The odd of time 12PM-6PM to have victims >=6 is 1.047 times the odd of period 12AM-6AM.
* The odd of time 6PM-12AM to have victims >=6 is 1.107 times the odd of period 12AM-6AM.
* Even though the accidents are lower on Sundays, the odd of Sunday to have victims >=6 is greater than the odd of all other days.
* The month of November reported to have maximum accidents and victims reported and is significantly different from others.
* When the reported victims are >=6, the odd of vehicles group 4-7 is higher compared to the rest of the groups.

# RECOMMENDATIONS

* There should be more traffic signals and speed breakers installed in the “Eixample” district and neighborhood “la Dreta de l’Eixample” because of their increased accident rates.
* The speed limit during Afternoon time should be reduced to less than 50km/hr. because of more vehicles on road and more accidents happening during that time.
* More Traffic police should be deployed during the weekdays compared to weekends, especially at the time when school/college disperses and end of office hours.
* On Sundays, more ambulance and medical crew should be deployed in hospitals because in case of accidents, the chances of casualties over 6 are higher.
* There should be advanced education policies implemented to the drivers in the high accident reported areas so that the awareness will reduce the accidents in the future.