RELATIONSHIP BETWEEN CAR FEATURES AND ACCIDENTS

1. Introduction

1.1. Background

It has been studied that there is a high relation between the velocity of the car and severe accident cause. It might also seem that the size of the car and the color could have some impact, such as red cars are often bought from people who like to drive fast and so have more possibilities to have a car accident with worst circumstances than others that might buy other cars with another colors.

1.2. Problem

Data can assure or confirm if these arguments are true or not, and if it really has an impact in severe car accidents visual characteristics, such as color or the size of the car. This project aims to predict if there are more traits apart from the driver's ones that can indirectly be inductors of car accident.

1.3. Interest

Public organizations like traffic ones would be very interested in seeing these correlations, as well as car productors, in order to know to which car model should they focus in developing new security structures and programs.

2.Data acquisition and cleaning

There are a lot of public and private companies that develop tables to resume traffic information for different purposes. In this project tables have been created taking different information from different sites, in order to develop data that might be accurate for this investigation and do not have biases. For instance, there is the website of Spanish DGT as a public organization that develops traffic statistics with the objective of achieving a more secure traffic.

2.2. Data cleaning

As in this project a new data information has been created, it has been decided to take the characteristics of 'color', 'size', 'velocity' , 'power', 'value' and 'victims' as traits to be analyzed from the year 2020 and created a data set of approximately 60 rows.

Regarding that the data comes from different sources, not all of the tables contained the values that were investigated in this project. For that reason, the missing values have been filled with approximated data taking into account the rest of the information, even though accuracy might risk.

The application for creating the table was Microsoft Excel and for each column a value was set. Each of the columns contain approximately sixty results. All the features contain very different values, so there is no problem regarding confusion. In resume, the table is composed of six values: 'color of the car', size of the car', 'victims', 'horsepower', 'price' and 'velocity'. Each value has forty rows and the information contained is a mix of integers or strings.

color	size (meters)	victims	horsepower	velocity	price
blue	4	4	111.0	27	13495.0
red	3	3	111.0	27	16500.0
yellow	3	1	154.0	26	16500.0
orange	4	4	102.0	30	13950.0
pink	3	2	62.0	39	5348.0
green	5	0	62.0	38	6338.0
brown	6	0	62.0	38	6488.0
blue	3	0	62.0	37	6918.0
red	4	1	62.0	32	7898.0

Figure 1. Table of the dataset used

Finally, all the redundant information or similar one has been discarded. Cases of cars that had equal velocity, color or size, as well as if it had victims have been joined to one set, as it was not given any reliable added information to the project.

3.Exploratory Data Analysis

3.1. Calculation of the target

All the information that was needed regarding the objective of the project was set in the table created in the previous step. Target was not needed to be calculated but to be found. For example, the aim of the study was to find if there was a correlation between a feature and the severity of the accident, in this case victims. Relations were needed, but no other calculation or other new columns.

3.2. Relationship between velocity and car price

First it was adequate to find the relationship between the price of the cars and its velocity. For that it was used a polynomial regression schema, giving the following result:

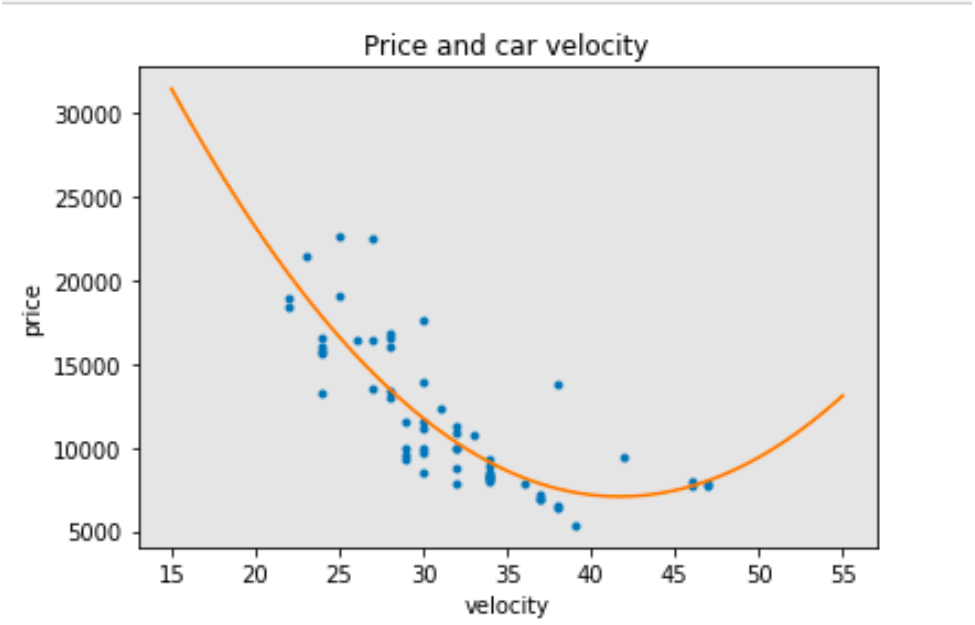

Figure 2.Polynomial regression result

In this example it is shown that most of the cars are between 15.000 and 7.000 €, while they mostly go between 30 and 40 miles per gallon.

3.3. Relationship between victims and velocity

In this example we wanted to know the relation between the number of victims in an accident and the velocity in which the cars were going (counting in miles per gallon). Using a boxplot, it has been observed that the most critical accidents were with cars that were going 24 mpg, between 29 and 30 mpg and 37 mpg, whereas the less were with a velocity of 25 mpg and 22 mpg.

This example assures that there is not a narrow relationship between velocity and the critical an accident can be, because even though there is existence of multiple victim accidents, there are accident where the car was going fast and there are not many victims, whereas other in which the car was not going faster than 24 mpg it implied several victims.

Taking into account this argument, it can be possible to predict that one of the causes of critical accidents is if the car was going slow or fast, and regarding that normally cars go slow within the cities and towns, if the car goes in populated places or not.

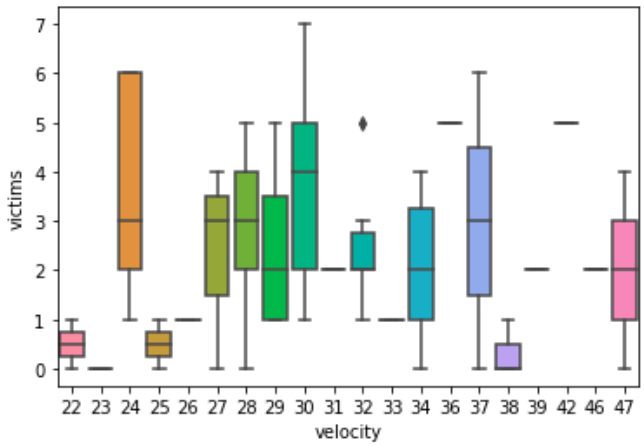

Figure 3. Boxplot between victims and velocity

3.4. Other relationships

Other relationships that were taken into account were the relationship between the size of the car, color and the number of victims. Although people might not think about this characteristic, it can possible be an indirect cause of becoming an accident more or less critical or severe.

First, it was necessary to group the three values: 'color', 'size (meters)' and 'victims'. After grouping it has been used this command:

df_group_one = df_group_one.groupby

Giving the result it has shown a table with the three characteristics and an 'id' for each represented object. With this data a pivot table was also formed and the missing values (NaN) filled with a '0'.

	color	victims	size (meters)
0	black	0	3.666667
1	black	1	3.000000
2	black	5	3.000000
3	black	6	5.000000
4	blue	0	3.000000
5	blue	1	4.000000
6	blue	4	3.666667
7	brown	0	6.000000
8	brown	1	3.000000
9	brown	2	4.000000
10	dark blue	3	3.000000
11	dark blue	5	3.000000
12	dark purple	2	2.000000
13	green	0	5.000000
14	green	1	5.000000
15	green	2	2.000000

Figure 4. Table with values

After regarding the table it was adequate to recreate the data within a heatmap.

	size (meters)															
lor	black	blue	brown	dark blue	dark purple	green	light blue	light green	orange	pink	pintk	purple	red	violette	white	yello
:tims																
0	3.666667	3.000000	6.0	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN	4.0	4.000000	NaN	2.0	5
1	3.000000	4.000000	3.0	NaN	NaN	5.0	NaN	NaN	NaN	5.0	NaN	NaN	3.750000	3.0	4.0	3
2	NaN	NaN	4.0	NaN	2.0	2.0	NaN	NaN	4.5	3.0	4.0	NaN	3.333333	NaN	NaN	3
3	NaN	NaN	NaN	3.0	NaN	NaN	3.0	5.0	4.0	NaN	NaN	NaN	3.000000	3.0	4.0	Na
4	NaN	3.666667	NaN	NaN	NaN	NaN	NaN	NaN	4.0	NaN	NaN	3.0	NaN	4.0	3.0	Na
5	3.000000	NaN	NaN	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.0	NaN	NaN	5.0	Na
6	5.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3.5	NaN	NaN	NaN	NaN	NaN	2.0	Na
7	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na

Figure 5. Pivot table with missing values (NaN)

The heatmap gives data about the big correlation of the different colors and the accident cause. The problem in this case is that all cars have a color, and most of them have very similar ones, which does not give reliable data about indirect causes that cause critical accidents.

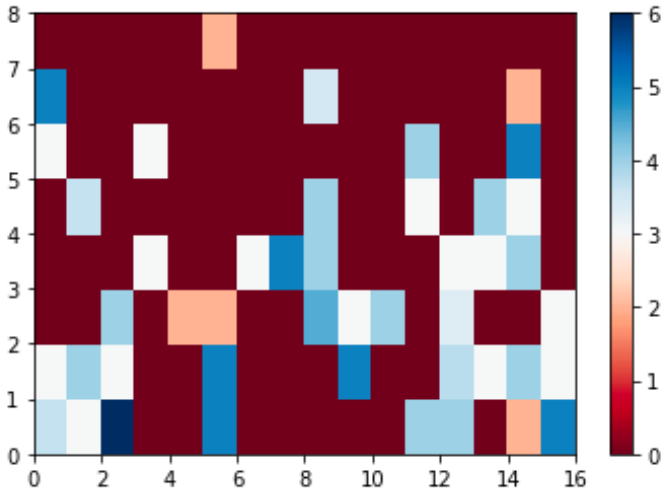

Figure 6. Heat map

4. Predictive Modelling

First it is important to explain that even though it was thought that some characteristics about the car were indirect causes of severe accidents, regression models have shown that is not possible. Only velocity and victim causation can explain somehow the number of victims in one or another accident.

4.1. Predictions

As shown in the previous chapters it has been seen that the best characteristic to analyze a relationship between victims and one indirect feature is the velocity one. An indirect characteristic that becomes directly related to a severe accident cause.

Regarding the standard algorithms a number of 9 test samples were given, whereas another 50 samples were from the training set. The resulting score is 1.0, which might seem that the he predicted values had a narrower range than the previous analyzed ones resulting that the prediction errors were larger as the actual values deviated further from zero. In this case the result was given to 1. This is not desirable, as errors are high.

4.2. Solution to problems

The problem was that it was very common sets with very little victims, whereas severe ones were rare. This made that the model analyzed first the accident with few victims rather the ones with more than three victims.

The solution was to give more importance to the sets that referred to severe accidents not only in the training set but also in the test one. With this solution, all models had similar range and distribution.

4.3. Other problems

Some errors with the data did not permit the analysis of cross validation score and prediction models, as shown in the next figure. This is due to error data or an error of formatting it, which implies not getting the expected results.

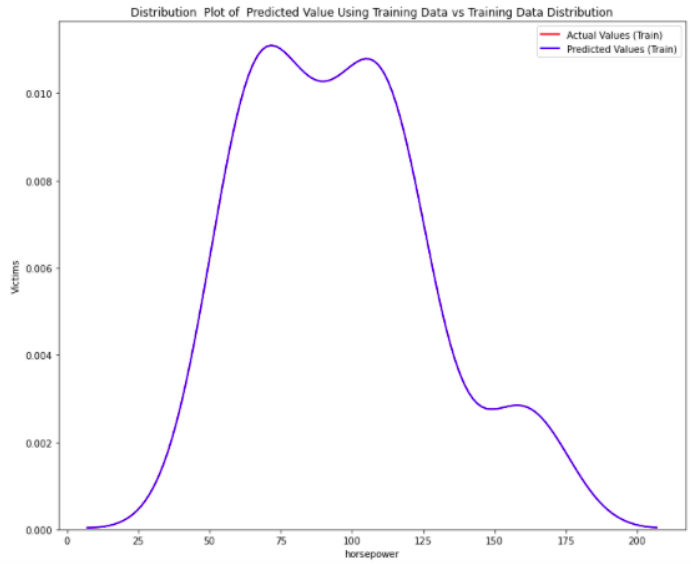

Figure 7. Prediction problems

5. Conclusion

In this project it was able to analyze that even though there are some features that people normally assign to an accident cause, in reality there are not. Regarding the schemas that

have been prepared in this project, the only cause in severe accidents is velocity, and moreover the horsepower of a car, whereas the color and the size of it does not really affect. An unexpected data has been found, and it is that less velocity produces more victims, which may be caused to the place where they occur, for instance in towns or cities. Accidents with high velocity impacts also create some severe accidents, but fewer.

For this analysis there has been produced linear regression models and also polynomial relationships between features, combining them to have a more exact idea of how all characteristics may join together to explain what the indirect cause of a severe accident is.

6. Future analysis

From this analysis it can be expected that even though these features are not indirectly implied in a cause of an accident, it might seem that not all the fault may be from the driver. It is true that the person who drives the car is highly responsible of the cause of the accident, but there might be other indicators to take into account that would create a more severe accident and are not controlled by the driver itself, for example the size of the car or other similar features.