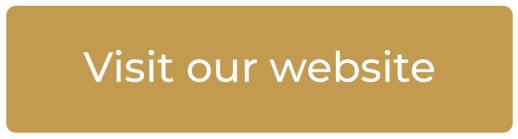


**TASK**

**Exploratory Data Analysis on the Automobile Data Set**

[](http://www.hyperiondev.com/portal/)

**Introduction**

Summary of the data set

This data set consists of three types of entities:

1. Vehicle’s specification of an auto in terms of various characteristics
2. Vehicle’s assigned insurance risk rating,
3. Vehicle’s normalized losses in use as compared to other cars.

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

-Cleaned the num-of-door column by removing the records which consist of invalid value ‘?’.

-Cleaned the bore column by finding the number of invalid values. And the output displayed that there were invalid valued in the data.

-Cleaning the horsepower column by going through the column itself and replaced it with a normalized mean where there were invalid values.

-Cleaned the peak rpm column by replacing it with a normalized mean where there were missing values.

-Cleaned the price column by replacing it with a normalized mean where there were invalid values.

-Cleaned the normalized losses column by replacing the invalid values with a normalized mean.

-Cleaned the stroke column by replacing the non-numeric values to null and convert the datatype.

-Cleaned the data and checked for unnecessary columns and dropped them.

MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT

-Calculated the total number of missing values and it was displayed that there were 8 missing values.

-Checked for the values that are missing in the num-of-doors column ‘?’ and replaced them with a ‘0’.

-Calculated the sum of fields that contained replacements and it was displayed that the stroke column contained 4 and the bore column contained 4 too.

-Looked for the number of records having ‘?’ value for normalized losses and the output was 0.

-Dropped all the columns that contained missing values.

-Replaced the ‘?’ record and called the function to find empty columns.

-Replaced the non-numeric value to null.

-Replaced the ‘?’ record with numeric values.

-Called all valued which were not numeric and the output summed up to 227 of non-numeric values.

-I then set the missing values to mean of price and converted the datatype to integers

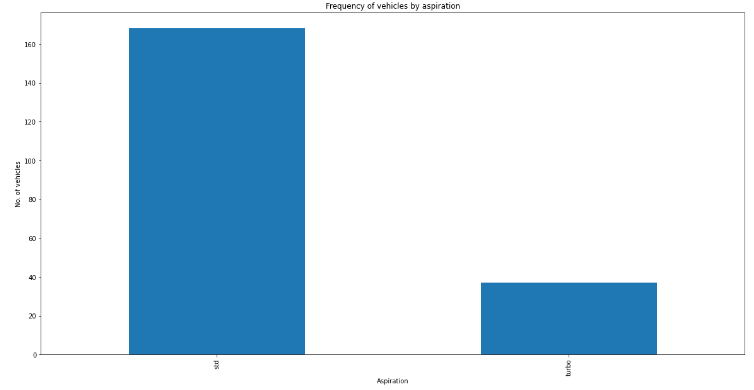
DATA STORIES AND VISUALIZATIONS

# THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALIZATIONS OF THE DATA

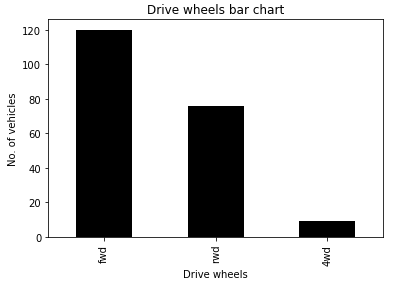
First: Plotted a scatter plot based on the number of doors of the cars and the prices of the vehicles, in order to determine the relationship between the two.and it has been determined that the two door has the most highest prices and the two door together with the four door have no relationship between each other.



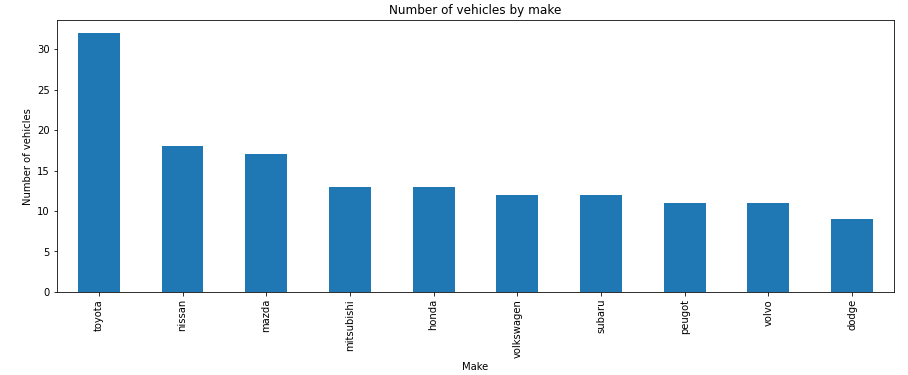
Second: Plotted a bar graph for the frequency of vehicles by aspiration. It was then determined by the visualization that there were higher amount of vehicles using the std aspiration than the turbo aspiration.



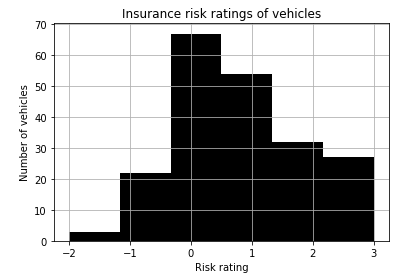
Third: Plotted a bar graph for the drive wheels of the vehicles and it then was determined by the visualization that the fwd wheel was used by most vehicles than the other type of wheels, followed by the rwd wheel type and then the 4wd wheel type that had a used by a least amount of vehicles.



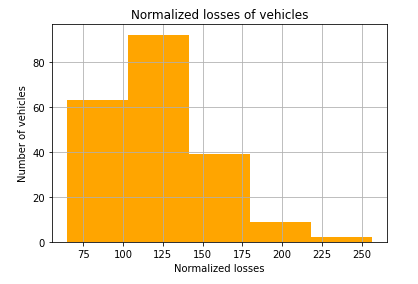
Forth: Plotted a top ten records bar graph that will display the type of car make that are mostly in demand/used, the visualization then displayed the Toyota make as the most frequently used make of all and the Dodge make the least used when it comes to the Top ten make.



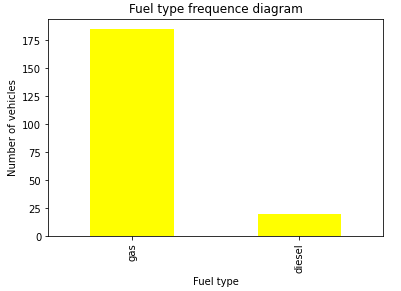
Fifth: Plotted an insurance risk rating histogram displaying the insurance risk rating of vehicles, it has been determined by the visualization that most vehicles are rated are at risk.



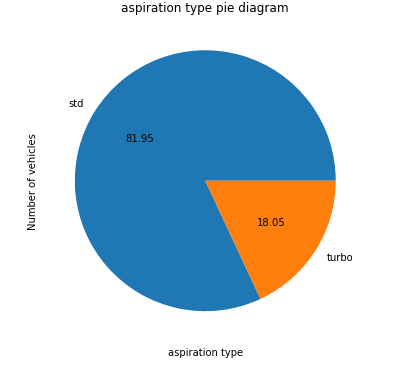
Sixth: Plotted a histogram of Normalized losses on insurance on vehicles. The visualization has displayed that there is a high relative average loss payment per insured vehicle.



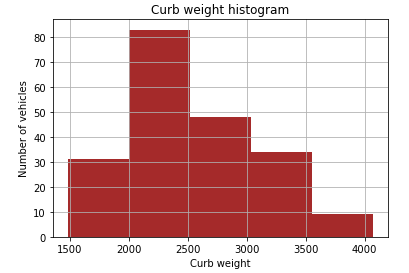
Seventh: Plotted a bar graph regarding the fuel vs no. Of vehicles graph to find out the mostly used fuel type by vehicles between the diesel type and the gas type. It was then determined by the graph that the gas type is mostly used than diesel.

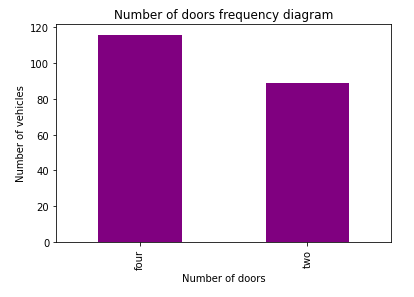


Eighth: Plotted a Pie Chart to view the type of aspiration used in a much visible suitable graph for the information as there are also percentages to make it more clear and readable. It has been determined that the turbo is used by 18,05% of vehicles and std is used by 81,95% of the cars which happens to be the highest.



Nine: Plotted a bar Curb weight histogram to determine how much do most cars weigh and it has been determined that most cars are between 2000 and 2500 of curb weight.



Tenth: Plotted a Number of Door bar graph to determine which number of doors is mostly used by cars, and it has been determined by the graph that four doors is the most frequently used than Two doors.  


# ENSURE THIS DOCUMENT IS NEAT AND CAN BE ADDED IN YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY : Nombuso Dlamini**

