

**TASK**

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

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

**Introduction**

Summary of the data set

The dataset comprised of information pertaining to various components which make up a vehicle. Such as, ‘symboling', 'normalized-losses', 'make', 'fuel-type', 'aspiration', 'num-of-doors', ‘price’ and so on.

Therefore, the goal here was to find out, what are the characteristics that have the most impact on the car price.

**DATA CLEANING**

* **The first step was to find all missing data in the various columns and either drop certain rows or convert the missing values to the mean of the column.**
* **For the missing data in normalized-losses, peak-rpm, horsepower, bore and stroke. The missing values were set to the mean.**
* **For the missing data in the num-of-doors column the value was set to four as typically sedan vehicles have 4 doors.**
* **Lastly, for the missing data in the price column, those rows were dropped. Since the price of the cars is what we are trying to predict in the analysis.**
* **After cleaning the data, the various data points were converted to their relevant d-types needed for the analysis.**
* **For further analysis and to improve the accuracy of the predictive models, prices were binned into 3 categories, ‘low’, ‘medium’ and ‘high’.**
* **The fuel-type categorical variable was also converted into a quantitative variable.**

MISSING DATA

* The dataset had a number of missing values in various columns, this included missing data in the normalized-losses column, num-of-doors column, price column, peak-rpm column, and horsepower column.
* Depending on the column, the missing values were either converted to the mean or the row was dropped.

DATA STORIES AND VISUALIZATIONS

* Given that the question at hand was - What are the characteristics that have the most impact on the car price?
* The EDA (Exploratory Data Analysis) was geared towards - summarizing the main characteristics of the data, gain a better understanding of the dataset, uncover relationships between different variables, and extract important variables for the problem we are trying to solve.
* Chart, bar chart

  Description automatically generatedThe first graph simply analyses on the top 10 vehicles by make.
* Followed after this, is a heatmap to find the dependent variables. As one of the best ways to find the correlation between the features can be done using heat maps. For example: As shown above the Price has a strong correlation with Engine-size of 87%. This is very important because the more the relationship between the variables the more accurate the model will be.

Calendar

Description automatically generated

* Next, is a Boxplot - to view the distribution of the drive-wheel data in relation to price. It's very evident that the Real wheel drive cars are most expensive and front wheel is least expensive cars. Four-wheel drive cars are little higher than the front wheel drive cars. There are fewer four-wheel drive cars in our dataset so this picture might not be very accurate.

Chart, box and whisker chart

Description automatically generated

* Next, are two scatterplots, the first to analyse if engine size can predict the price of a car and second to analyse if horsepower can predict the price of a car?
* From the first scatterplot - The conclusion which can be drawn is that we see that as the Engine-size goes up, the Price of the car also goes up. Positive correlation between “engine-size” and “price”.

Chart, scatter chart

Description automatically generated

* From the second scatterplot - The conclusion which can be drawn is that we see that as the Horsepower goes up, the Price of the car also goes up.

Chart, scatter chart

Description automatically generated

* Chart

  Description automatically generatedChart, bar chart

  Description automatically generatedNext, is an analysis of on average, which type of drive wheel is most valuable. From the data, it seems rear-wheel drive vehicles are, on average, the most expensive.
* Body-style was also included in the analysis. From the heatmap is seems that rear-wheel drive hardtops were the most expensive.
* An Analysis of Variance (ANOVA) was also included.
* The Analysis of Variance (ANOVA) is a statistical method used to test whether there are significant differences between the means of two or more groups. ANOVA returns two parameters:
* F-test score: ANOVA assumes the means of all groups are the same, calculates how much the actual means deviate from the assumption, and reports it as the F-test score. A larger score means there is a larger difference between the means.
* P-value: P-value tells how statistically significant is our calculated score value.
* The ANOVA was performed on the mean prices of two different manufactures.

Graphical user interface, text, application

Description automatically generated

* An analysis was also done, to view on average, which type of fuel-type is most expensive. The conclusion being diesel.

Chart, bar chart

Description automatically generated

* Chart, bar chart, histogram

  Description automatically generatedAn analysis was also done, to view the average prices of the various makes. The conclusion being that on average, Jaguars were the most expensive.
* Lastly, an analysis was also done on the normalized losses on vehicle body-styles and number of doors. Findings: As we understand the normalized loss which is the average loss payment per insured vehicle is calculated with many features of the cars which includes body style and no. of doors. Normalized losses are distributed across different body style, but the two door cars has a greater number of losses than the four door cars.

Chart, bar chart

Description automatically generated

* Ultimately, the goal of the EDA was to view - What are the characteristics that have the most impact on the car price?
* The conclusion which can be drawn from the analysis is that engine-size, horsepower, drive-wheels, body-style, and manufacturer greatly influences the price of a car.

**THIS REPORT WAS WRITTEN BY: DEVAN BOOY**

