TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

This is a Data Analysis Capstone Project which will be dealing with analysing the Automobile Data Set. This data set has three types of entities: (i). the specification of an automobile in terms of various characteristics, (ii). its assigned insurance risk rating, (iii). its normalized losses in use as compared to other automobiles. The data set has about 200 automobiles listed with 26 columns showing characteristics of each car.

In this document I will be sharing how I have analysed this data set using the knowledge I have gained in data cleaning, handling missing data and data visualisations.

DATA CLEANING AND MISSING DATA

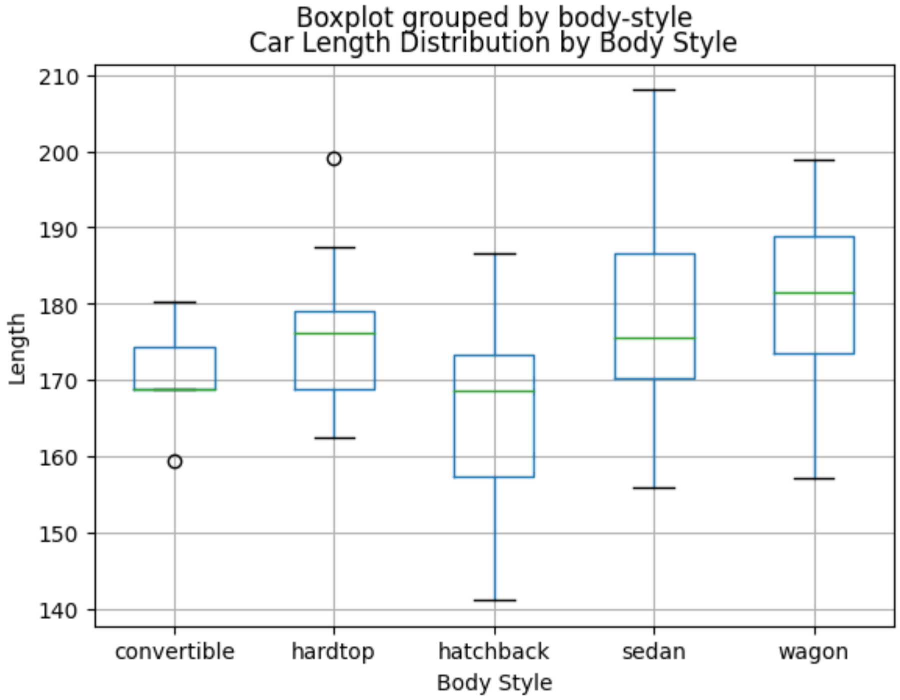
In the data cleaning process I replaced all the question marks with NaN to make it easier for me to take care of this in the following step of dealing with missing data. I followed by filling missing data of the columns normalised-losses, price, horsepower, peak-rpm, bore, stroke with the respective column mean. I also corrected the data types in some columns where is was necessary.

DATA STORIES AND VISUALISATIONS

Below are the three interesting visualisations I generated.

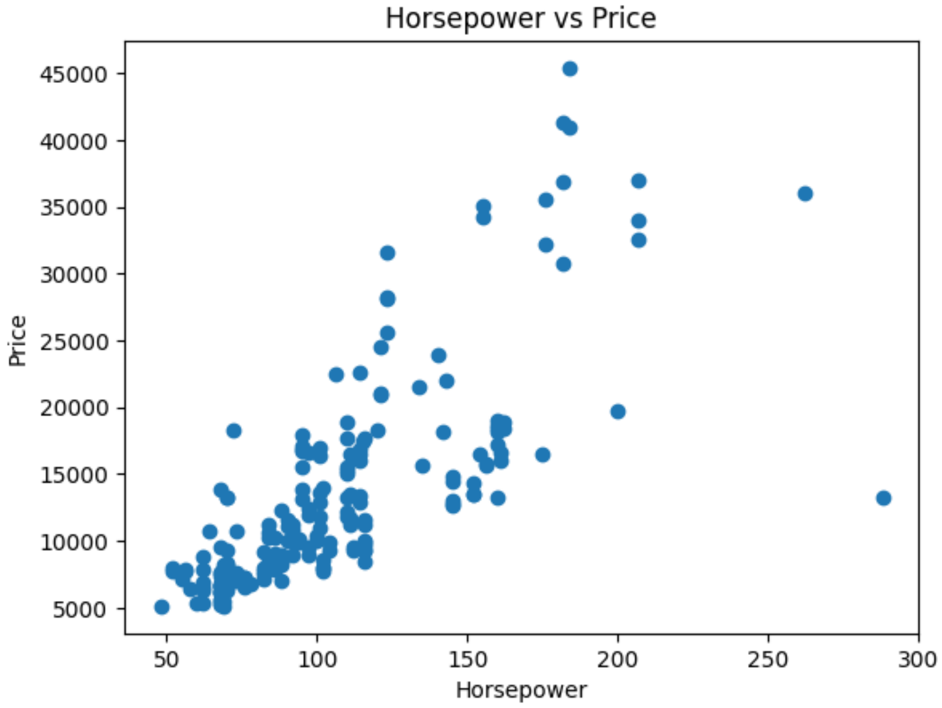
Car Length Distribution by Body Style

This analysis helps in understanding how car lengths differ across various categories of body styles. To do this I used box plot, I found that sedans have the longer length in the list car list and the hatch back cars have the most compact length.



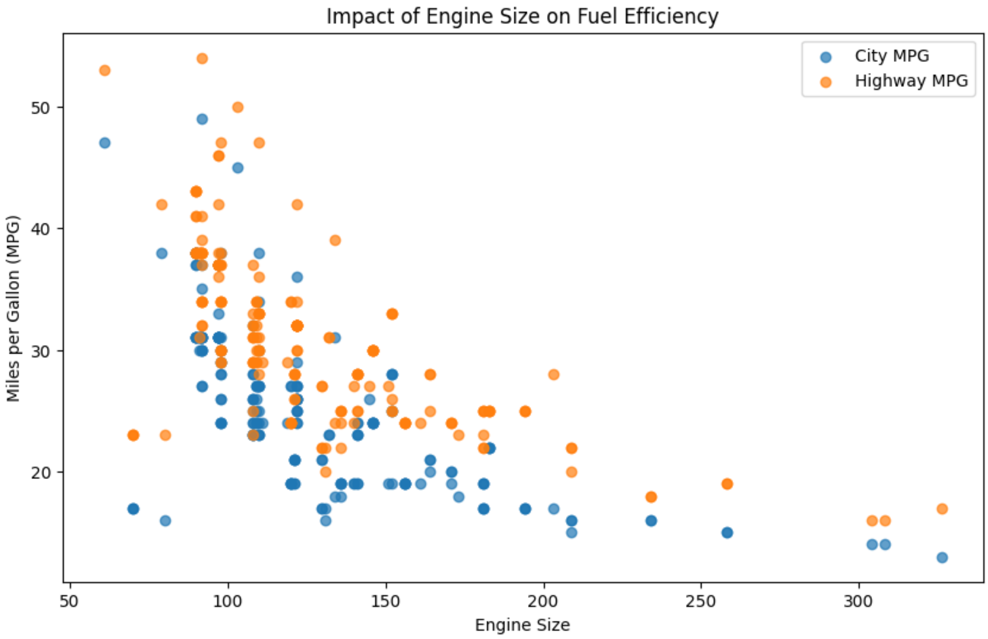
<u>Horsepower vs Price</u>

In this scatter plot I am interested to see how a vehicle's horsepower affects its price. Horsepower is the measure of how quickly the force is produced vehicle's engine. There is a kind of linear relationship here, as the horsepower increases so does the price of the vehicle. This means that cars with high horsepower are most likely to cost more in the market.



Impact of Engine Size on Fuel Efficiency

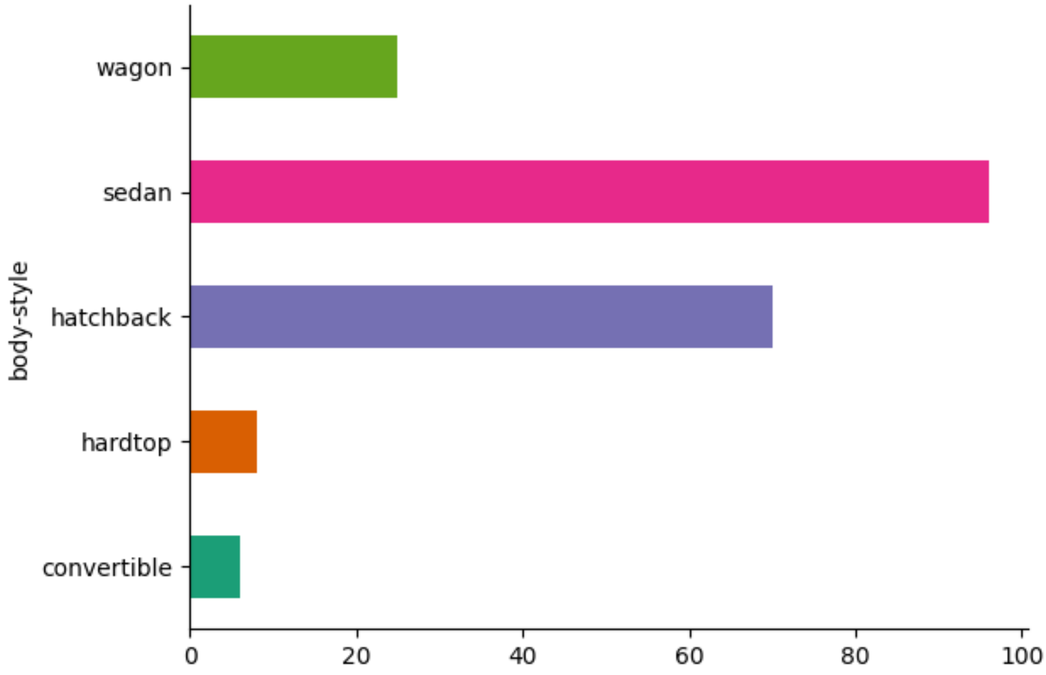
This analysis helps consumers and manufacturers understand how the choice of engine impacts the overall operational cost and environmental footprint of a vehicle. I used a scatter plot to achieve this.
In the plot below it is evident that driving in the city uses more fuel than driving on highways, in other words the bigger the engine size the lower the fuel efficiency due to increased fuel consumption.



Plot of number of cars per body-style

Analysing the number of cars per body style helps in understanding the popularity of different body styles among consumers and can provide insights for manufacturers and marketers in catering to customer preferences.
I this visualisation I first grouped the dataset by body style, counted the number of cars in each body style category and then plotted a bar graph.

From the plot below we see that sedans are the most purchased cars, the reason for that would be because they are bigger and can accommodate families. We can also see that convertibles are the least purchased cars and that is solely because the can only accommodate two people meaning they are not practical for bigger families.



THIS REPORT WAS WRITTEN BY : NONOPHA GEGE