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

1. **INTRODUCTION**

The ‘automobile’ dataset contains data of cars. The dataset contains 205 rows and 26 columns and includes data on make of the car and details about car dimensions, fuel types/consumption, engine characteristics, and price.

This EDA explores the following:

1. Which cars in the dataset are the most powerful at the lowest price?
2. The relationship between various body styles and:
   1. Fuel consumption;
   2. Fuel type;
   3. Weight; and
   4. Horsepower
3. **DATA CLEANING**

There are 205 rows and 26 columns.

The following columns were not used, and so were dropped: 'symboling', 'normalized-losses', 'bore', 'stroke', 'compression-ratio'.

After printing out all the columns and plotting a bar chart of all value counts of each column, the following was noticed:

* In the ‘make’ column, the name ‘peugot’ is incorrect spelling; and
* The ‘price’ and ‘horsepower’ columns are objects; values need to be converted to floats.

1. **MISSING DATA**

At first glance, there appears to be no missing/NaN values. Utilizing the **df.isna().sum()** command, Seaborn heatmap, and Missingno Matrix, no null values were picked up **.** However, upon closer inspection, there are ‘?’ NaN values in the ‘num-of-doors’ (2/205), peak-rpm’ (2/205), ‘stroke’ (4/205), and ‘price’ (4/205) columns. As a result, the ‘?’ values will be replaced with 0, as the percentage of ‘?’ values is small and won’t adversely affect the subsequent calculations and analyses (0,98% and 1,95% respectively).

1. **DATA STORIES AND VISUALIZATIONS**
   1. **WHICH CARS IN THE DATASET ARE THE MOST POWERFUL AT THE LOWEST PRICE?**
      1. **Introduction**

Which cars in the dataset are the most powerful at the lowest price? To figure this out, we will take the following steps:

* Calculate the power-to-weight ratio of all the cars in the dataset; and
* Compare power-to-weight ratio to prices.
  + 1. **Power-to-Weight Ratio**

The power-to-weight ratio (PWR) is a measurement of actual performance of any engine or power source. It is calculated by taking the horsepower and dividing it by the weight:

PWR = BHP / Weight in Tonnes.

Chart, scatter chart

Description automatically generated

* + 1. **Findings**

The scatterplot of PWR/Price shows that there is a large cluster of cars at $10 000 and less, at a PWR of between 35 and 40. A trend of increase in price as the PWR increase can be observed. There are a handful of cars with PWR above 60. Some of these are at $30 000 and upwards, but since we are looking for the highest PWR at the lowest price, there are two cars at under $20 000 whose PWR is above 60:

* ID 75 Mercury
* ID 105 Nissan

The other car with PWR over 60 and price under $20 000 is null, as the price is 0.

* 1. **DATA EXPLORATION OF BODY STYLES**
     1. **Introduction**

**Chart, pie chart

Description automatically generated**

There are five body styles represented in the data:

* Sedans (46,83%);
* Hatchback (34.15%);
* Wagon (12,20%);
* Hardtop (3,90%); and
* Convertible (2.93%).

Sedans and hatchbacks make up over 80% of the body styles, with the remaining +/- 20% comprising of wagon, hardtop, and convertible.

* + 1. **Body Styles, Fuel Consumption, and Fuel Type**

Fuel consumption is divided into two sections:

* Highway MPG (miles per gallon); and
* City MPG.

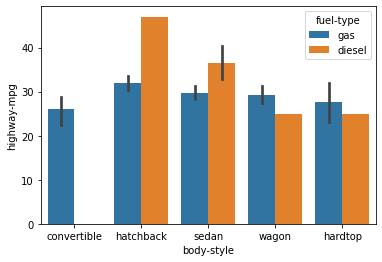
Highway driving is more fuel efficient that city driving, so it is expected that the data here reflects this.

Also, regarding fuel types, there are two types:

* Gas (petrol); and
* Diesel

**(Note: there is no diesel fuel type for the ‘convertible’ body type).**

* + - 1. **Body Styles/Highway Driving**

****

Gas (Petrol): The convertible category has the lowest highway mpg (i.e., it uses the most fuel). The other four body styles have better fuel consumption (i.e. higher values), with ‘hatchback ‘, having the best/highest highway miles per gallon. This is sensible as hatchbacks are the smallest cars, with the smallest engines and the lowest weight.

Diesel: Continuing the trend, ‘hatchback’ has the highest/most fuel-efficient highway mpg of any categories, over 40mpg. ‘Sedan’ mpg is around 35mpg, with ‘wagon’ and ‘hardtop’ having the lowest mpg at around 25mpg (the least fuel-efficient diesels).

While hatchback and sedan highway mpg are higher for diesel than gas-as is expected-, the diesel mpg for the wagon and hardtop body styles are less than their gas mpg. It is surprising that diesel is usually more fuel efficient than gas.

* + - 1. **Body Styles/City Driving**

**Chart, bar chart

Description automatically generated**

Gas (Petrol): City mpg follows similar trends to highway mpg. The ‘convertible’ body style is the thirstiest at around 20 mpg, while ‘hatchback’ the most frugal at about 26 mpg.

Diesel: Again, ‘hatchback’ has the highest city mpg of all the body styles, and again, diesel city mpg is very close to gas city mpg for the ‘wagon’ and ‘hardtop’ body styles.

Analysis: With regards to the ‘hardtop’ body style, the anomaly in the data could be due to it making up such a small percentage of the sample (3.90%), and so with such a small number of data points, any anomalies can be safely ignored (this is also applicable to the ‘convertible’ body style which makes up only 2.93% of the data). However, the ‘wagon’ category needs further investigation, as it has sufficient data points to warrant looking for the factors that could make diesel cars less fuel efficient than gas/petrol cars. This might include greater weight and greater horsepower. And so, this will be explored below.

* + 1. **Weight & Fuel Consumption**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

General Trend: The hypothesis is that there is an inverse relationship between weight and fuel consumption, i.e., an increase in weight results in lower mpg. The scatterplots (‘*Weight & Fuel Consumption*) showing the distribution of weight/highway-mpg and weight/city-mpg indeed show the hypothesis to be true. The downward trend applies along the mean as well as those cars with higher mpg values (better fuel consumption) than the mean. The scatterplots show this clearly.

* + 1. **Body Style & Weight**

**Chart, bar chart

Description automatically generated**

Gas (Petrol): The ‘hatchback‘ body style is the lightest of all the body styles, with ‘convertible’ being the heaviest at around 2700.

Diesel: Again, and as expected, the smallest body style ‘hatchback’ is the lightest, with ‘wagon’ and ‘hardtop’ being the heaviest-by a considerable margin-at almost 3500.

While diesel hatchbacks are slightly lighter than the gas hatchbacks, and while diesel sedans are slightly heavier than gas sedans, the diesel wagon and hardtop body styles are significantly heavier than the gas equivalents. Given the fact that an increase in weight results in poorer fuel consumption (i.e., lower highway mpg and city mpg values), the heavier weight of the diesel versions of the wagon and hardtop body styles may be contributing to making the diesel variants less fuel efficient than the gas variants.

* + 1. **Body Styles/Horsepower**

Chart, bar chart

Description automatically generated

General Trend: As with the relationship between horsepower and fuel consumption, above, the hypothesis is that there is an inverse relationship between horsepower and fuel consumption, i.e., an increase in horsepower results in lower mpg. Once again, the scatterplots showing the distribution of weight/highway-mpg and weight/city-mpg indeed show the hypothesis to be true.

Gas (Petrol): The ‘wagon’ body style generates the least amount of horsepower among the gas cars, followed by ‘hatchback’ and ‘sedan’.

Diesel: The ‘hatchback’ category has the weakest/lowest horsepower of all the body styles, followed by sedan, wagon, and hardtop. ‘Hatchback’ and ‘sedan’ diesel horsepower is considerably lower than their respective gas equivalent. This makes sense as car models have gas and diesel engine variants, with diesel engines having less horsepower than the gas engines. But surprisingly, the ‘wagon’ body style’s diesel horsepower is more than its gas horsepower. It can be explained that since the more horsepower a car has, the less fuel efficient it is, the fact that the ‘wagon’s’ diesel horsepower output is greater than its gas equivalent may be yet another factor in the ‘wagon’ category’s diesel engines being less fuel efficient than its gas ones.

* 1. **CAVEATS**

The automobile dataset is small, with only 205 entries. The accuracy and veracity of the data is further called into question with regards to the segmentation of the data into the different body styles. This is evident by the large error bars on the bars in the various bar graphs.

For example, regarding the two fuel types (gas and diesel), gas comprises of 185 out of the 205 total cars in the dataset (90.24%), while diesel only makes up 20/205 cars (9.76%). This means that data analysis tat incorporates fuel type is skewed. Another example, the ‘wagon’ body style totals 25 vehicles. Only 3 of the 25 cars are diesel (12%), so the most likely explanation for diesel wagons being less fuel efficient than gas/petrol wagons is that the diesel wagons in the dataset are not representative and have skewed the analysis & results.

**Shape, rectangle

Description automatically generated**