**Car Sales Analysis**

**Introduction**

The aim of the project is to analyze the Car Sales Dataset containing the data of different cars models, manufacturers and other car feature with respect to their sales and pricing. Our task is to analyze the data and to answer the following questions:

* Which Manufacturer has the highest and lowest average sales volume?
* How are the prices of cars distributed?
* Is there a correlation in the numerical variables of the dataset?
* Which is the most popular car brands?
* What is the top 3 fuel-efficient cars with an engine size above 2.5 litres?

The essence of this analysis is to show if there has been a substantial growth in the production of cars as well as their pricing and efficiency throughout the years. The analysis would be done using python to creates charts and perform statistical analysis in an easy-to-understand method.

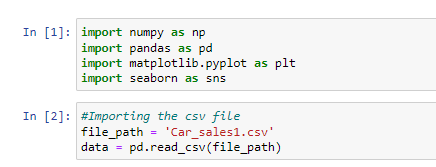
**Data Sourcing**

The data is gotten from a google drive link uploaded to a WhatsApp group as the fourth assessment in the PSP Data Analytics class facilitated by Mr. Joseph Elijah. The folder contained the car sales csv file and the assessment question file.

**Data Preparation**

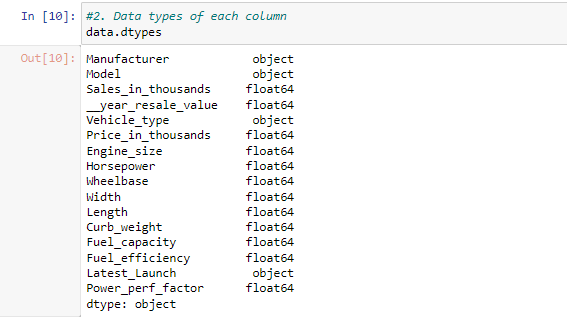
Before analyzing, we’ll first prepare the python environment by importing necessary python libraries. Numpy is imported to work with number/integers, Pandas; to work on Series or Dataframe and also to manipulate data while Matplotlib and Seaborn are used to visualize data.

Next, we read in the data in csv format as shown in the image below



**Data Exploration**

Let’s now do an exploration of our data.

* First, we check for the number of rows and columns in the data. There are 157 rows and 16 columns in the dataset
* Next, we check for the datatypes of each column in the car sales data using the “*dtypes”* function. 

We can see that all columns with string values are classified as “objects” and columns with numbers or decimals are “*float”* data types, except for the “Launch\_date” column classified as objects instead of ***datetime objects.***

* We check for missing values in the dataset using **“*.isna()”*** function. We count them per columns using the SUM function.

**Data Cleaning**

Furthermore, we clean the data properly before starting the analysis

**Handling Missing Values**

After checking for missing values, we now handle the missing values in the following ways

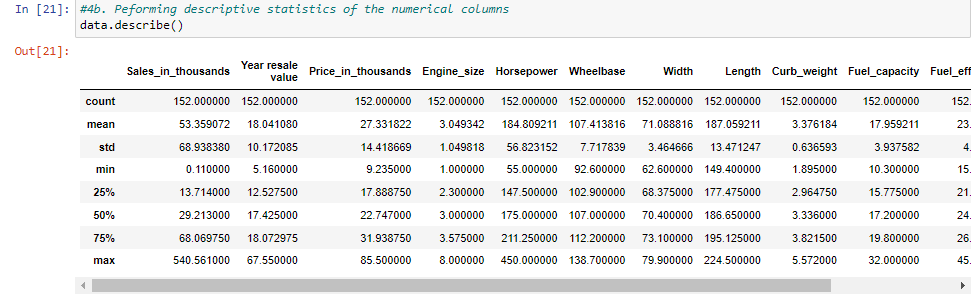
* There are about 36 null/missing values in the **year\_resale\_value** columns. We handle them by first finding the average of the whole columns and then replace the null values with them
* Next, we drop every other rows with missing values.

**Renaming Columns**

We now clean the name of the **\_\_year\_resale\_value** column by removing the leading underscore

We check for duplicates in the dataset, but there are no duplicates.

**Descriptive Statistics**

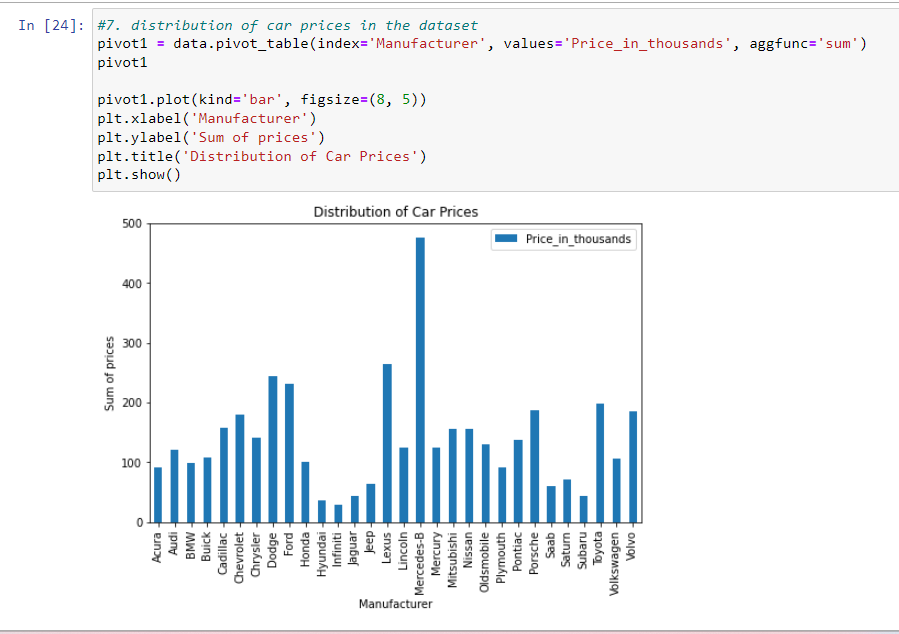
Here, we will do a descriptive statistic of the numerical columns to find the mean, median, mode and others. We use the describe() function with specifying all, because we need just the numerical columns. 

**Exploratory Data Analysis**

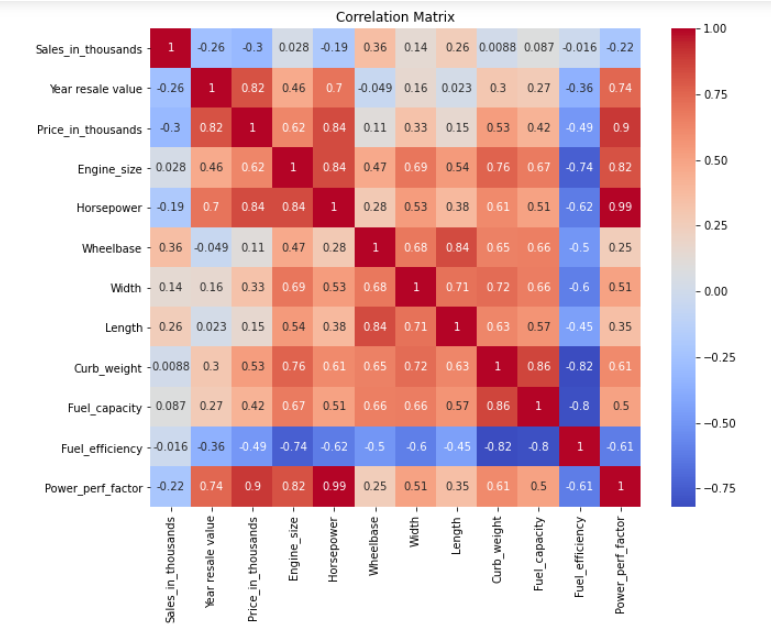
1. **Which Manufacturer has the highest and lowest average sales volume?**

We check for the highest and lowest by grouping the Manufacturer with the mean of Sales in thousands and then sorting them in ascending or descending order respectively. Instead of sorting, we can also use **idmax()** and **idxmin()** to find the highest and lowest average sales volume respectively

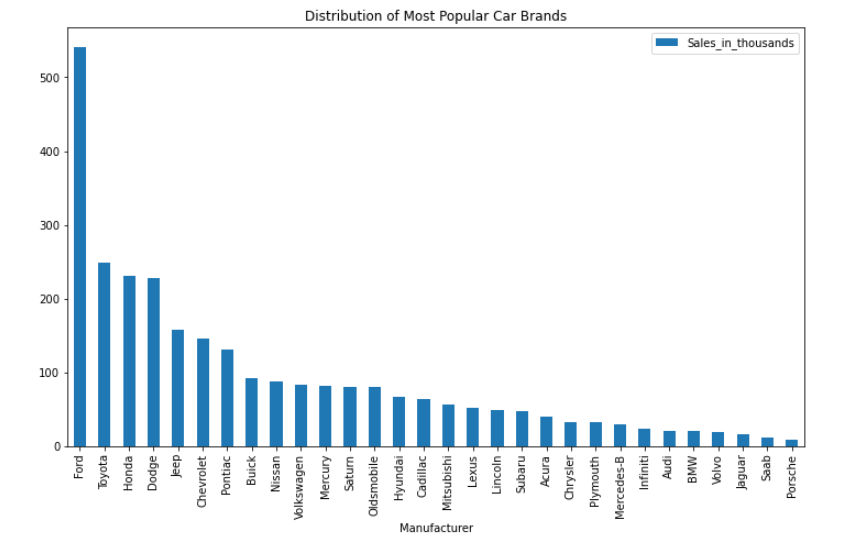
1. **Distribution of car prices in the dataset**

Here, we group and plot the Manufacturer with the sum of the Price in thousand 

1. **Correlation matrix of numerical variables in the dataset**

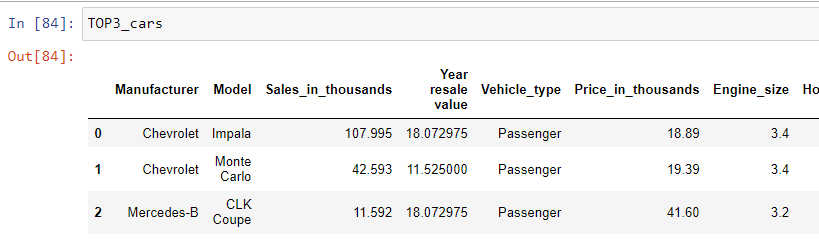


1. **Which is the most popular car brands?**



By grouping the Manufacturer together with the maximum sales in thousand, we can see that Ford, particularly the F-Series model is the most popular car.

1. **What is the top 3 fuel-efficient cars with an engine size above 2.5 litres?**

****

From the image above, the Chevrolet Impala, Chevrolet Monte Carlo and Mercedes CLK Coupe model are the 3 top fuel efficiency cars with an engine size above 2.5 litres.

**Conclusion/Recommendation**

Based on our analysis, we’ve been able to provide valuable and data driven answers to our business questions. Recommendation is that a perfect analysis can be done on a wide/large dataset. Visualizations can then be done on powerbi and other advanced analytics tools.