Task 1: Data Acquisition & Preparation

In the first step, it is importing data set into a python environment using python package pandas. The next step is fundamental insight from data by checking its data types before doing analysis. After the next step is data cleaning with various methods, and the final step is data preparation for analysis.

1.1 Importing Data and Libraries

The libraries are used, such as pandas, seaborn, and matplotlib for data loading and analysis. The automobile data set loaded into Jupiter notebook for cleaning, analysis, and basic insight of data. The panda's library is used to load three data set into three panda's data frames using the pandas read_csv method. The imported data is checked by head and tail methods to verify with source data set and it is same [1].

1.1 Importing Libraries and Data

```
import pandas as pd
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
import seaborn as sn

df1 = pd.read_csv("data1.csv")
df2 = pd.read_csv("data2.csv")
df3 = pd.read_csv("data3.csv")
```

1.2 Merging Data

To merge three data frames of automobile car data set into one with 27 attributes. First, we used the method merge on data frame 1 & 2 on column name ID and then append with this merge data frame with data frame 3. As a result, we get the merge automobile car data frames into one single data frame, which contains all data of cars [1].

1.2 Merging Data Frames

```
#Merge dataframe 1&2
merged_inner = pd.merge(left=df1, right=df2, how='inner', left_on='id', right_on='id')

#Create final data set of cars which contains all three files data with 27 attributes
cars = merged_inner.append(df3, ignore_index = True)
cars
```

1.3 Data Cleaning

To clean the automobile data set, we used various methods.

1.3.1 Checking Missing Values

The missing values are checking using panda method isnull().sum() or isnull().values.any(). In this data set, we found missing values, NAN, on two columns normalised-losses and price. If we will replace NAN value to 0 then the car which contain 0 in normalised-losses and price is meaning less and the record does not contain more than 50% missing value so we cannot drop that record too. To handle the missing values for normalised losses, price columns is to calculated mean value on these columns of the entire data set and replace Nan value to mean value by fillna () method used [1] [7] [8].

1.3.2 Checking Duplicate Values

The duplicate values are checking using the pandas' method duplicate (). In this data set, we found two rows are duplicate and all the attribute value identical. To remove the duplicate values will use drop_duplicates () methods [1] [8].

1.3.3 Checking Impossible Values

To check the impossible values on this data set using histogram on all numeric columns. The histograms give a clear idea about values which are unexpected values which are not the in-between range of attribute. As a result, there are no impossible values in this data set [1] [6].

1.3.4 Checking Typo Errors

While analysing columns num-of-doors and fuel-system typo errors found in these columns corresponding spelling mistake, wrong input value in the string. In the num-of-doors column found wrong input value, 4L replace with four and 2L replace with two. Also, in column fuel system occurred spelling mistake error and replaced ‘Mpfi’ to ‘mpfi’ [1] [8].

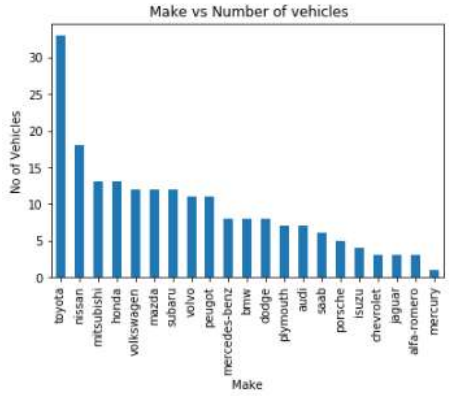
1.3.5 Removing White Space

To check white space in the car dataset, the white space found in column fuel type. In fuel type column values are like gas, diesel, and ‘Diesel ‘ with space. It can be fixed by replace value ‘Diesel ‘ to diesel. Other white space in data set on all string value base columns can be fixed by using the str.strip() method to remove the front and end of the word white space[1] [11].

Task 2: Data Exploration

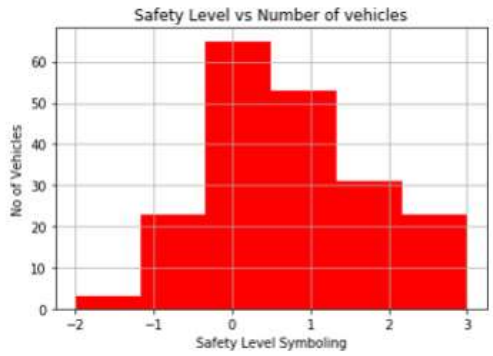
2.1 Explore and Identify Important Columns

2.1.1 Highest No of vehicle Maker



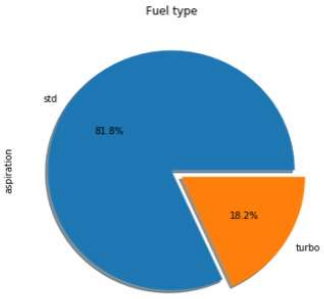
The customer wants to know which car company is produced more car models at an affordable price with the best performance. In this case, there are mainly ten different types of car manufacturing companies, but it is often essential to know who has the most significant number of car manufacture. To do this, the bar chart is one of the trivial solutions which lets us know the total number of cars manufactured by a different company. Toyota is the highest car maker with more than 40%, and Nissan is the second-highest car maker [10] [9].

2.1.2 Safety Level in No of vehicles



The car manufacturer and customer give the safety level is the highest priority. To achieve this, we are selecting a symbolling column from this car dataset and create the histogram for different safety levels. As per shown in the histogram, safety level -3 has no values that imply it is challenging to manufacture fully protected and perfect car. The most frequency value in this chart is safety level 0, that means all car manufactures are not compromising on car safety. In this dataset, there are few cars with safety level 3, and it is indicating that there is no safety in that level 3 cars [10] [9].

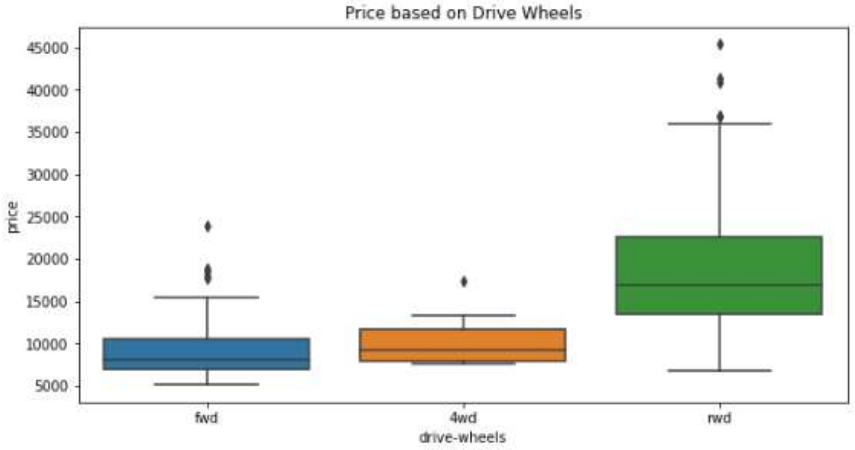
2.1.3 Fuel Type in Vehicle



The customer is selecting the car based on their specific needs of fuel type and engine performance. As shown in this pie chart most of customer and manufactures preferred the standard engine which is more than 80%. The turbo engine fuel type car is less because it's more expensive than standard engine cars [10].

2.2 Data Descriptive Statistics

2.2.1 Price – Drive Wheels



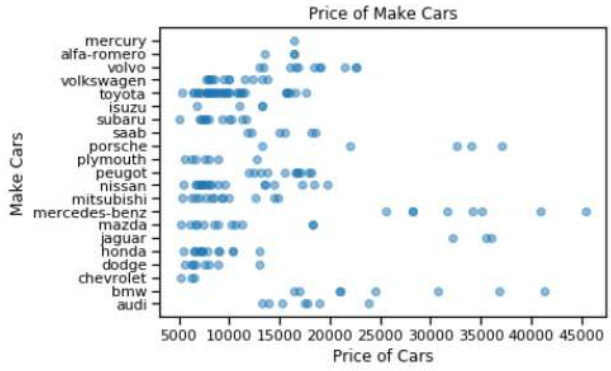
The plausible hypothesis to be determined about the drive wheels of the car and the price of the car. As shown in the box plot, it is clear that the real-wheel-drive cars are the most expensive, and the front wheel is the least expensive car. Four-wheel drive cars are a little cheaper than real-wheel-drive cars, but it is expensive than the front-wheel-drive cars. The number of records of four-wheel drive cars in our data set is very less so this box plot is not very accurate [10] [5].

```
cars[cars['drive-wheels'] == '4wd']
```

	id	symboling	normalised-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower
1	26247	2	164.000000	audi	gas	std	four	sedan	4wd	front	...	136	mpfi	3.19	3.40	8.0	115
99	10215	2	83.000000	subaru	gas	std	two	hatchback	4wd	front	...	108	2bbl	3.62	2.64	8.7	73
115	35959	0	102.000000	subaru	gas	std	four	sedan	4wd	front	...	108	2bbl	3.62	2.64	9.0	82
131	37724	0	81.000000	toyota	gas	std	four	wagon	4wd	front	...	92	2bbl	3.05	3.03	9.0	62
145	34893	0	85.000000	subaru	gas	turbo	four	wagon	4wd	front	...	108	mpfi	3.62	2.64	7.7	111
161	19387	0	85.000000	subaru	gas	std	four	wagon	4wd	front	...	108	2bbl	3.62	2.64	9.0	82
187	32794	0	102.000000	subaru	gas	turbo	four	sedan	4wd	front	...	108	mpfi	3.62	2.64	7.7	111
190	16229	0	120.745342	audi	gas	turbo	two	hatchback	4wd	front	...	131	mpfi	3.13	3.40	7.0	160
191	34127	0	91.000000	toyota	gas	std	four	wagon	4wd	front	...	92	2bbl	3.05	3.03	9.0	62

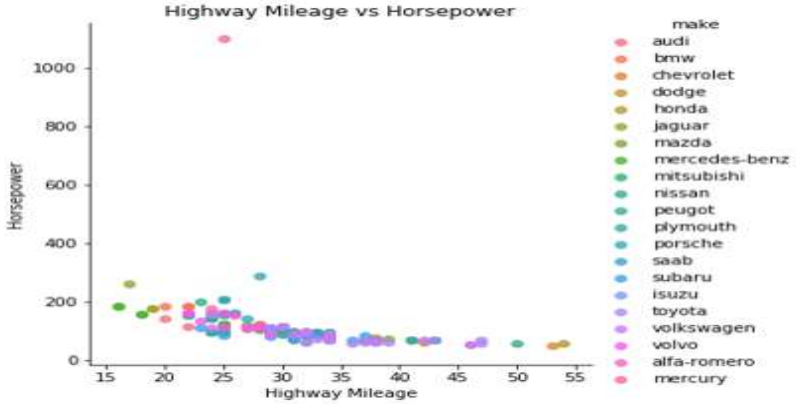
9 rows × 27 columns

2.2.2 Price – Make cars



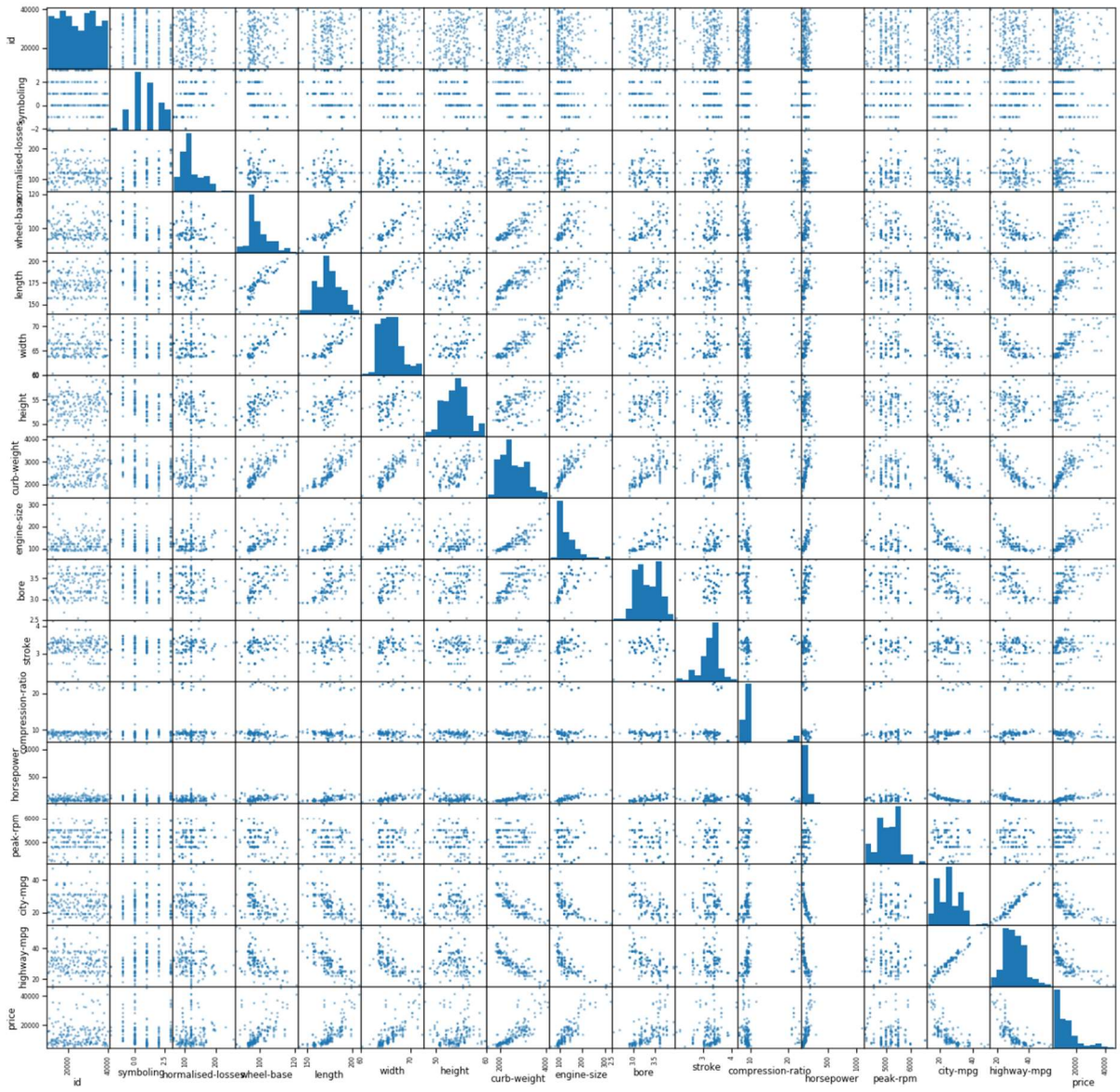
The plausible hypothesis to be determined about the make of the car based on the price of the car. According to the scatter plot, the Mercedes Benz brand car price is highest, then the other brand cars. The second highest brand by price is BMW. The lowest price of car brand is Chevrolet. The range of cars produced by manufacturers is 5200 to 45500. The scatter plots are giving clear picture of formed data in pattern [10] [9].

2.2.3 Highway Mileage – Horsepower



The plausible hypothesis to be determined about highway mileage vs horsepower via maker of the car. As shown in the scatter plot, it is clear that high horsepower cars have low highway mileage and vice versa. Moreover, a high horsepower car is costly. The high horsepower car is Audi with low highway mileage and high mileage car is dodge with less horsepower. The scatter plots are giving clear picture of formed data in pattern [4].

2.3 Scatter Matrix



The given dataset all numeric values scatter matrix is plotted as shown in the above figure with size 17. The scatter matrix diagonal axes contain the bar charts of the 17 numeric values, and others are scatter plots [10].

References

1. Yu Sun 2020, 'Lecture 04 Data Wrangling, Visualization and Management', COS60008 Introduction to Data Science learning materials on, Swinburne University of Technology, viewed 10 April 2020.
2. Yu Sun 2020, 'Lecture 05 Data Wrangling, Visualization and Management II', COS60008 Introduction to Data Science learning materials on, Swinburne University of Technology, viewed 10 April 2020.
3. https:// www.rapidinsight.com/7 data cleanup terms explained visually
4. https://seaborn.pydata.org/generated/seaborn.scatterplot.html
5. https://medium.com/@sriramselvank/let-us-do-data-analysis-with-python-db2cb6eca43f
6. https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b
7. https://www.tutorialspoint.com/python_data_science/python_data_cleansing.htm
8. https://towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d
9. https://towardsdatascience.com/5-minute-guide-to-plotting-with-pandas-e8c0f40a1df4

10. https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html
11. https://www.geeksforgeeks.org/python-pandas-series-str-strip-lstrip-and-rstrip/