IBM Machine Learning Professional Certificate

- (Exploratory Data Analysis) -

Choe C.S

October 2021

Main Objective

Objective

- This report aims to perform Exploratory Data Analysis on Vehicle Price dataset.

Data Set

- The data set used for this analysis is Vehicle Price from Kaggle.com.

Steps Involved

1) Perform Data Wrangling – Handling missing data, drop unwanted columns and data formatting to convert feature column unit for better analysis.

2) Feature Engineering – Perform feature engineering to overcome outliers in data set.

3) EDA – To study, analyze and gain better understanding of each feature variables and its affects on the target variable.

4) Hypothesis Testing – Generate statistical hypothesis to determine and validate desired significance level of data set.

Import Data

The data set was import from Kaggle.com.

- The table at the bottom shows the top 5 rows for the data set

	car_ID	symboling	CarName	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	wheelbase	...	enginesize	fuelsystem	boreratio
0	1	3	alfa-romero giulia	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47
1	2	3	alfa-romero stelvio	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47
2	3	1	alfa-romero Quadrifoglio	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68
3	4	2	audi 100 ls	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19
4	5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19

Data Wrangling

- There are a total of 204 rows with 26 column of data .
- There is a total of 8 “float64”, 8 “int64” and 10 “object” columns.
- Data formatting will be required on ‘**citympg** and **highwaympg‘** column for data unit conversion.
- Column **'car_ID','symboling’** and **'CarName’** will be drop.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   car_ID            205 non-null    int64
 1   symboling         205 non-null    int64
 2   CarName           205 non-null    object
 3   fueltype          205 non-null    object
 4   aspiration        205 non-null    object
 5   doornumber        205 non-null    object
 6   carbody           205 non-null    object
 7   drivewheel        205 non-null    object
 8   enginelocation    205 non-null    object
 9   wheelbase         205 non-null    float64
 10  carlength         205 non-null    float64
 11  carwidth          205 non-null    float64
 12  carheight         205 non-null    float64
 13  curbweight        205 non-null    int64
 14  enginetype        205 non-null    object
 15  cylindernumber    205 non-null    object
 16  enginesize        205 non-null    int64
 17  fuelsystem        205 non-null    object
 18  boreratio         205 non-null    float64
 19  stroke            205 non-null    float64
 20  compressionratio  205 non-null    float64
 21  horsepower        205 non-null    int64
 22  peakrpm           205 non-null    int64
 23  citympg           205 non-null    int64
 24  highwaympg        205 non-null    int64
 25  price             205 non-null    float64
dtypes: float64(8), int64(8), object(10)
memory usage: 41.8+ KB
```

Data Wrangling – Drop Columns

- The column **'car_ID','symboling'** and **'CarName'** is drop from the data frame.

	car_ID	symboling	CarName	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	wheelbase	...	enginesize	fuelsystem	boreratio
0	1	3	alfa-romero giulia	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47
1	2	3	alfa-romero stelvio	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47
2	3	1	alfa-romero Quadrifoglio	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68
3	4	2	audi 100 ls	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19
4	5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19

Before

Drop unwanted column

	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	wheelbase	carlength	carwidth	carheight	...	enginesize	fuelsystem	boreratio
0	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	...	130	mpfi	3.47
1	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	...	130	mpfi	3.47
2	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	...	152	mpfi	2.68
3	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	...	109	mpfi	3.19
4	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	...	136	mpfi	3.19

After

Data Preprocessing – Handling Missing Data

- From the figure on the right, there a no missing data present in the data set.
- All columns shown are of value ‘0’ representing a total of zero missing data.
- Therefore, the process for handling missing data is not required.

```
fueltype          0
aspiration        0
doornumber        0
carbody           0
drivewheel        0
enginelocation    0
wheelbase         0
carlength         0
carwidth          0
carheight         0
curbweight        0
enginetype        0
cylindernumber    0
enginesize        0
fuelsystem        0
boreratio         0
stroke            0
compressionratio  0
horsepower        0
peakrpm           0
citympg           0
highwaympg        0
price             0
dtype: int64
```

Data Wrangling – Data Formatting

- The columns, '**citympg** and **highwaympg** are of mpg unit.
- It is preferred to convert the unit to km per litter (km/l) for better analysis of data.
- Therefore, data formatting for unit conversion is applied.
- The figure on the right shows the unit conversion of both columns.

	citympg	highwaympg
0	21	27
1	21	27
2	19	26
3	24	30
4	18	22
...	...	...
200	23	28
201	19	25
202	18	23
203	26	27
204	19	25

Before

	city_km_l	highway_km_l
0	8.928024	11.478888
1	8.928024	11.478888
2	8.077736	11.053744
3	10.203456	12.754320
4	7.652592	9.353168
...	...	...
200	9.778312	11.904032
201	8.077736	10.628600
202	7.652592	9.778312
203	11.053744	11.478888
204	8.077736	10.628600

After

Feature Engineering – Log Transformation

- As some numerical variables are slightly skewed to the right, it is important to perform transformation on the columns that are rightly skewed to overcome outliers.
- In the data set columns, **'compressionratio', 'enginesize', horsepower', 'wheelbase'** and **'carwidth'** are of skewness > 0.75.
- Therefore, **log-transformation** from **numpy** library was applied to normalize the skewness of these columns.

Before

```
compressionratio    2.610862
enginesize          1.947655
horsepower          1.405310
wheelbase           1.050214
carwidth            0.904003
dtype: float64
```


After

```
compressionratio    2.349716
price               1.777678
wheelbase           0.883387
enginesize          0.857828
carwidth            0.813993
```

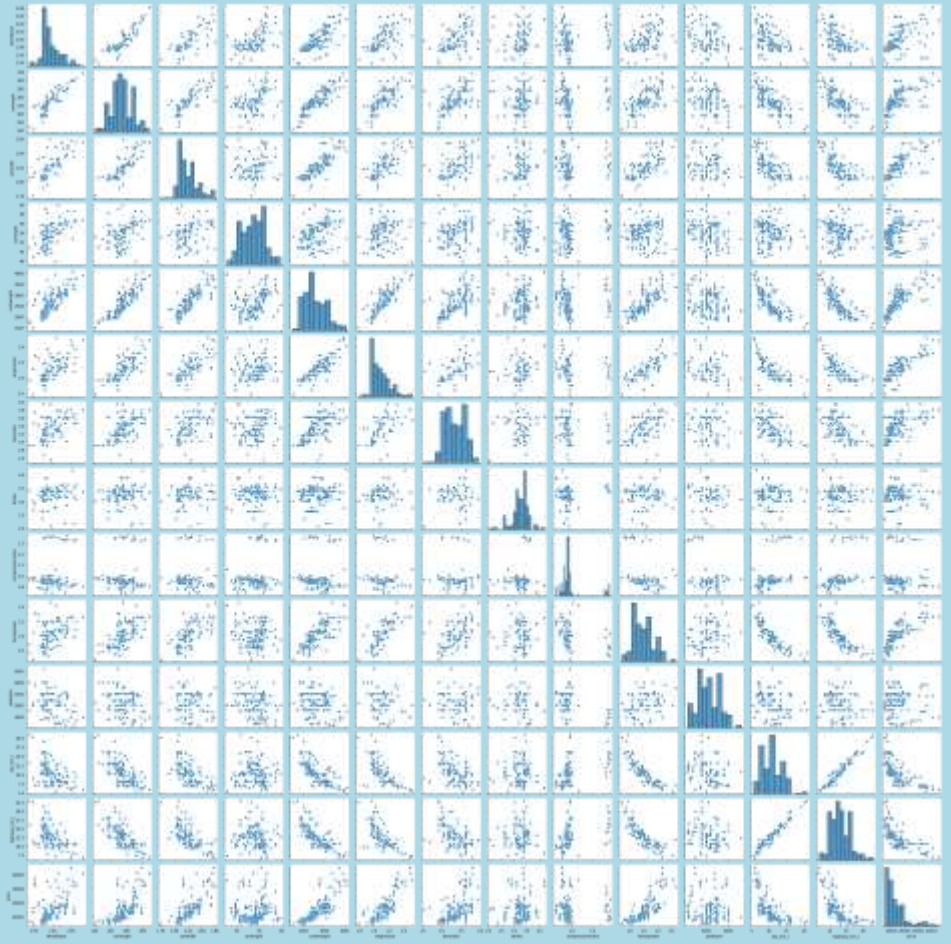
EDA – Descriptive Statistics

- The table below shows descriptive analysis of all the numerical variables of data set.

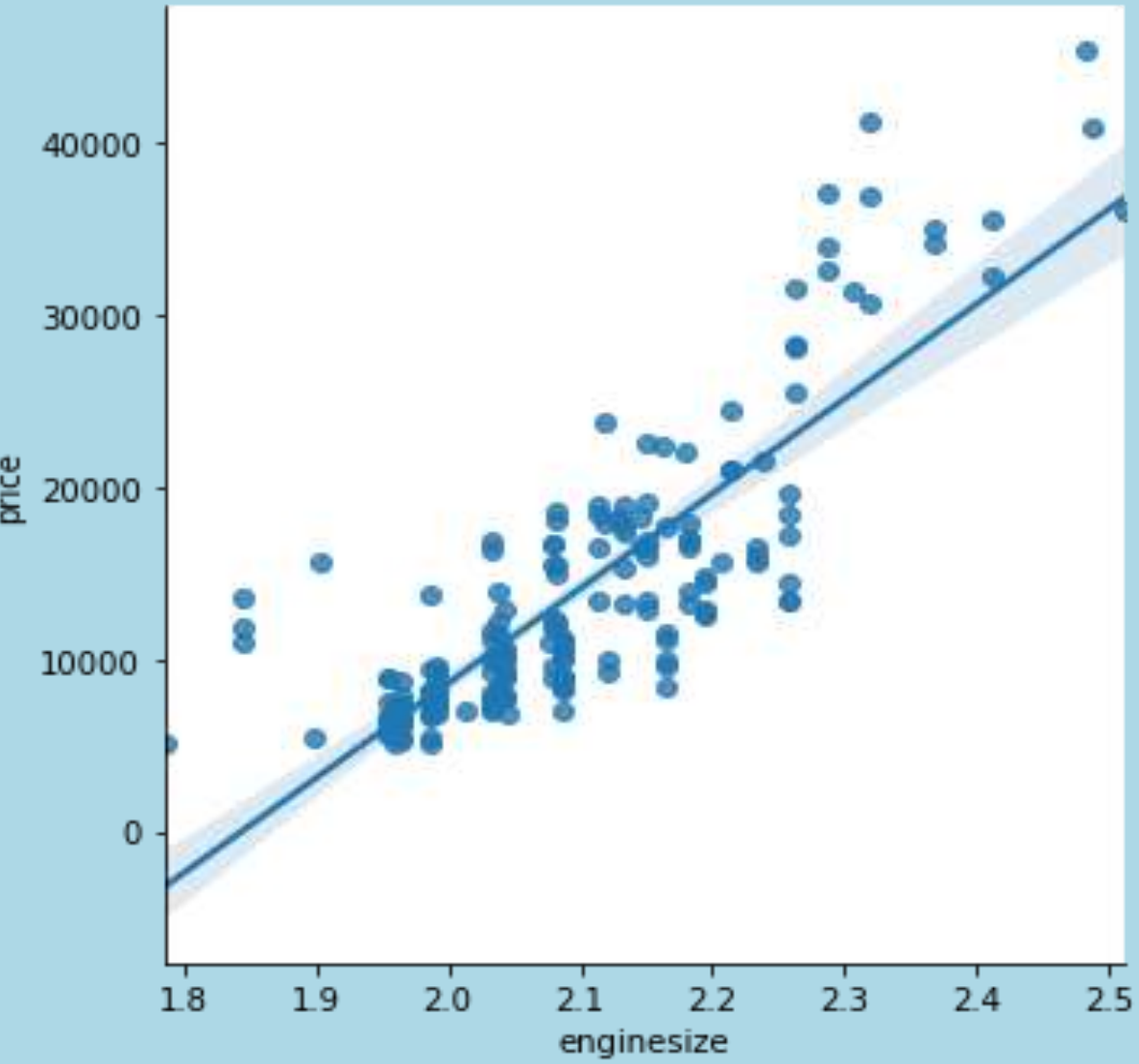
	wheelbase	carlength	carwidth	carheight	curbweight	enginesize	boreratio	stroke	compressionratio	horsepower	peakrpm	city_km
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.00000
mean	1.993791	174.049268	1.818712	53.724878	2555.565854	2.084694	3.329756	3.255415	0.984715	1.990500	5125.121951	10.72192
std	0.025781	12.337289	0.013950	2.443522	520.680204	0.122891	0.270844	0.313597	0.122573	0.149786	476.985643	2.78135
min	1.937518	141.100000	1.780317	47.800000	1488.000000	1.785330	2.540000	2.070000	0.845098	1.681241	4150.000000	5.52687
25%	1.975432	166.300000	1.806858	52.000000	2145.000000	1.986772	3.150000	3.110000	0.934498	1.845098	4800.000000	8.07773
50%	1.986772	173.200000	1.816241	54.100000	2414.000000	2.079181	3.310000	3.290000	0.954243	1.977724	5200.000000	10.20345
75%	2.010300	183.100000	1.825426	55.500000	2935.000000	2.149219	3.580000	3.410000	0.973128	2.064458	5500.000000	12.75432
max	2.082426	208.100000	1.859138	59.800000	4066.000000	2.513218	3.940000	4.170000	1.361728	2.459392	6600.000000	20.83205

EDA – (Pairplot)

- The plot on the right shows the overall statistical relationship between each variables.
- As the plot is filled with large amount of data, it is rather difficult to analyze the variables' relationship.
- Therefore, the next few slides will break the variables into 2 groups, numerical variables and categorical variables to perform visual exploration.



EDA – Visual exploration

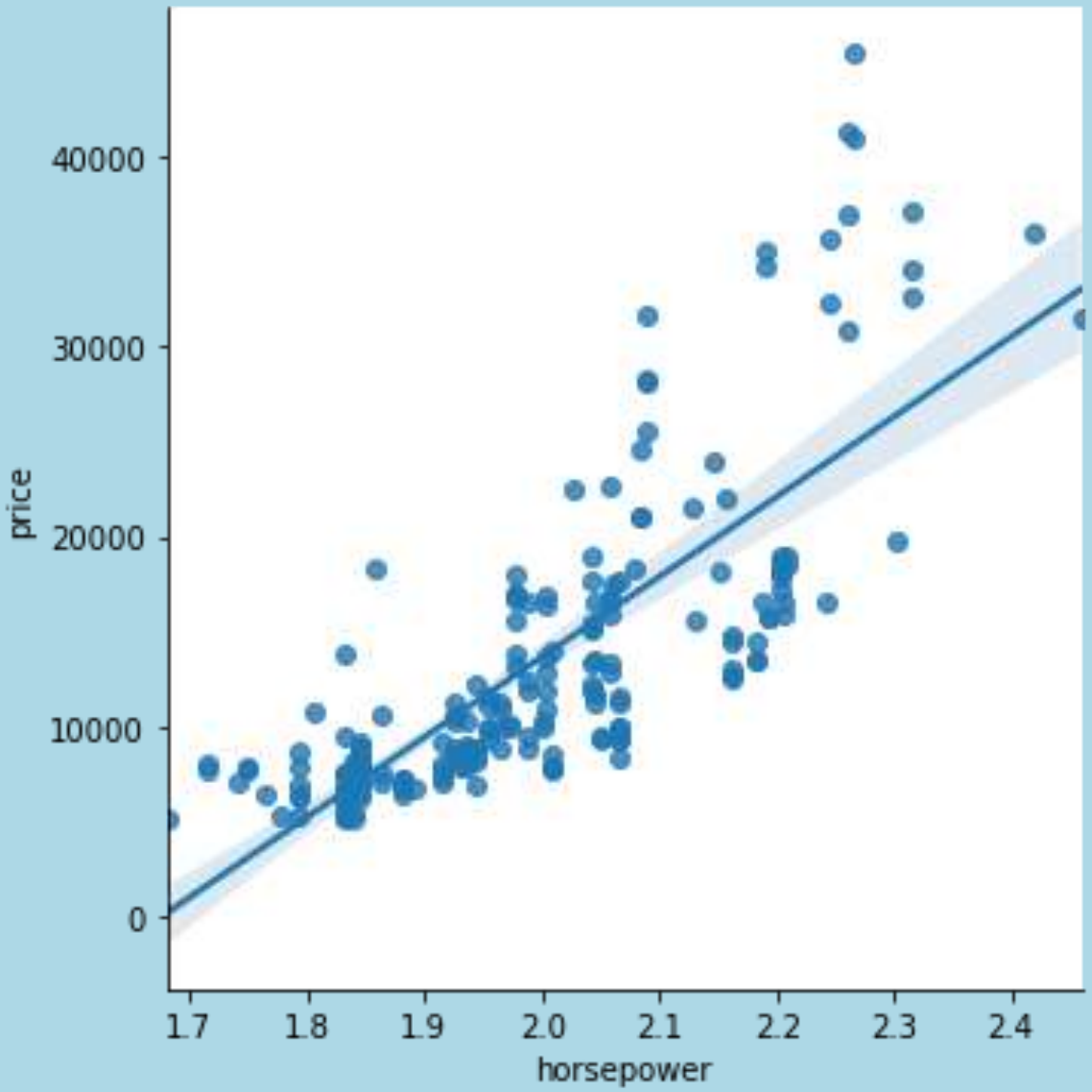


- The scatter line plot on the right shows the relationship between **Engine Size** and **Price** of vehicle.
- From the plot there seems to be a positive linear relationship between both variables. With increase of **Engine Size** the **Price** increases as well.
- The points are highly saturated and close to the line between Engine Size range of value 2.0 to 2.3 with price range of roughly 5000 to $25000.
- The Pearson Coefficient is 0.84566 and p-value is 2.91e-57. This suggest that the correlation is highly positive and **Engine Size** plays an **important role** in determining the **price** of vehicle.

```
Pearson Coefficient is: 0.8456606586416033
```

```
P-value is: 2.913491424941682e-57
```

EDA – Visual exploration (Numerical data)

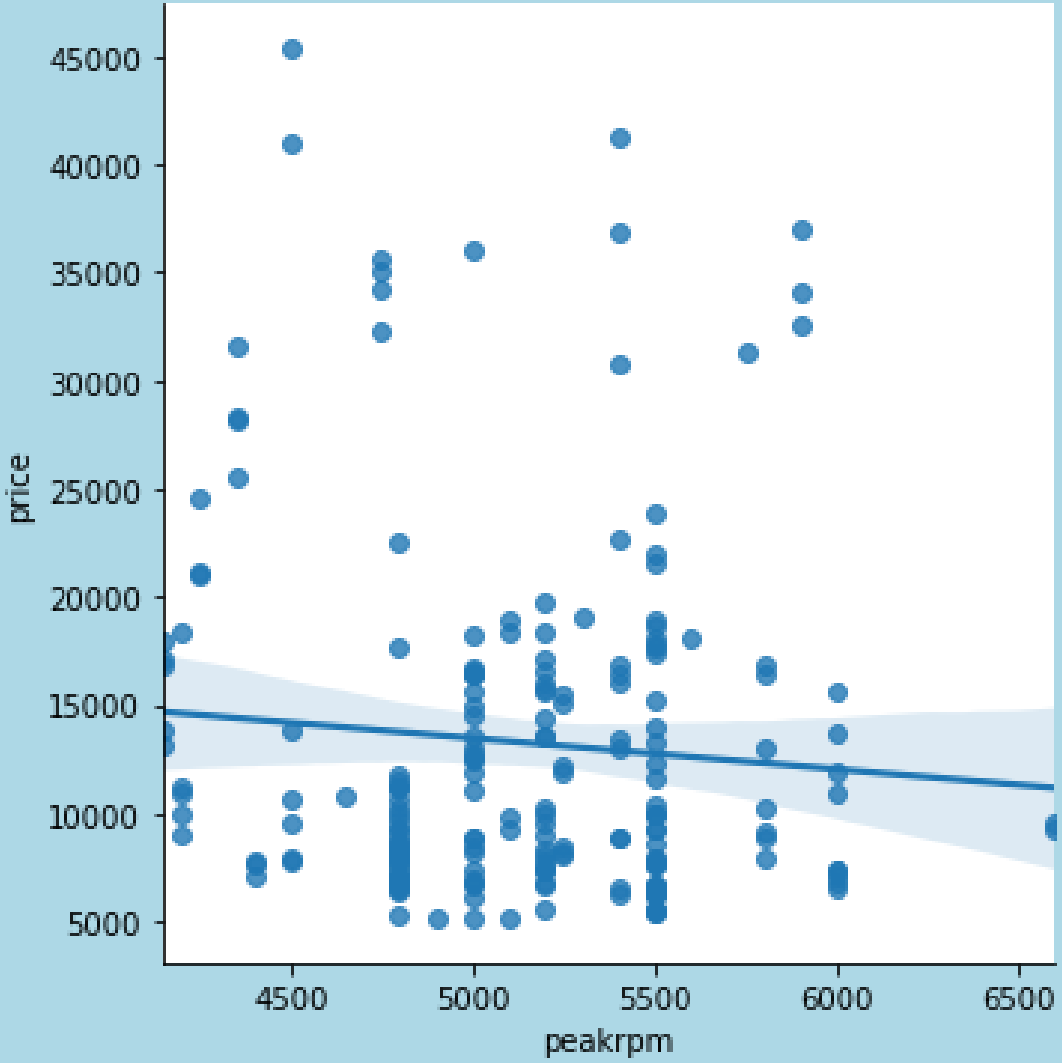


- The scatter line plot on the right shows the relationship between **Horse Power** and **Price** of vehicle.
- From the plot there seems to be a positive linear relationship between both variables. With increase of **Horse Power** the **Price** increases as well.
- The points are highly saturated and close to the line between Engine Size range of value 1.8 to 2.2 with price range of roughly 10000 to $20000.
- The Pearson Coefficient is 0.788 and p-value is 1.053e-44. This suggest that the correlation is highly positive and **Horse Power** plays an **important role** in determining the **Price** of vehicle.

```
Pearson Coefficient is: 0.7883579131818567
```

```
P-value is: 1.0528954530448985e-44
```

EDA – Visual exploration (Numerical data)

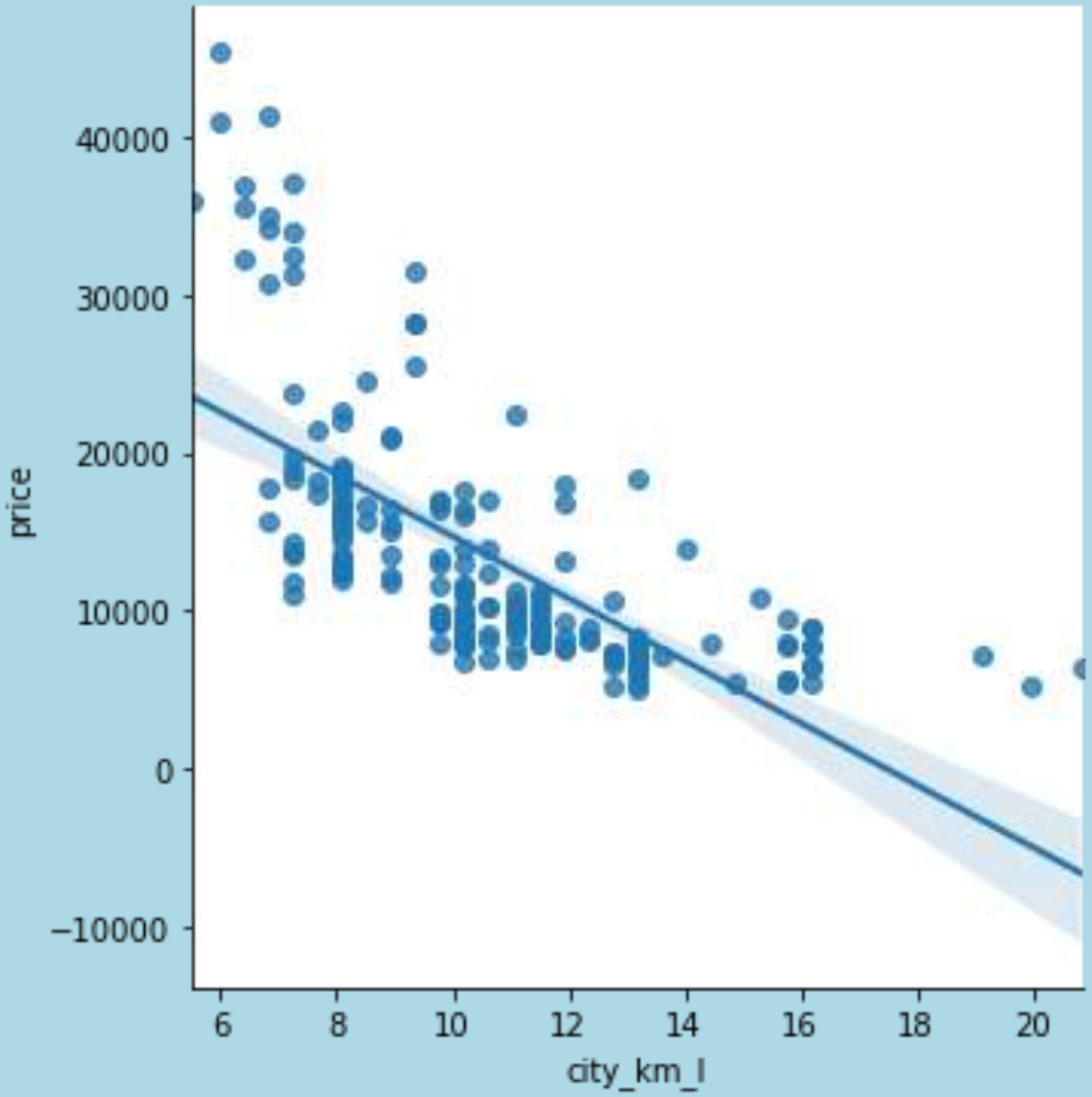


- The scatter line plot on the right shows the relationship between **Peak Rpm** and **Price** of vehicle.
- From the plot there seems to not have a linear relationship between both variables. With increase of **Peak Rpm** the **Price** does increases nor decrease.
- The Pearson coefficient is -0.08526 and p-value is 0.224. This suggest that the weak correlation and **Peak Rpm** will not play an **important role** in determining the **Price** of vehicle.

```
Pearson Coefficient is: -0.08526715027785689
```

```
P-value is: 0.22414123444667824
```

EDA – Visual exploration (Numerical data)

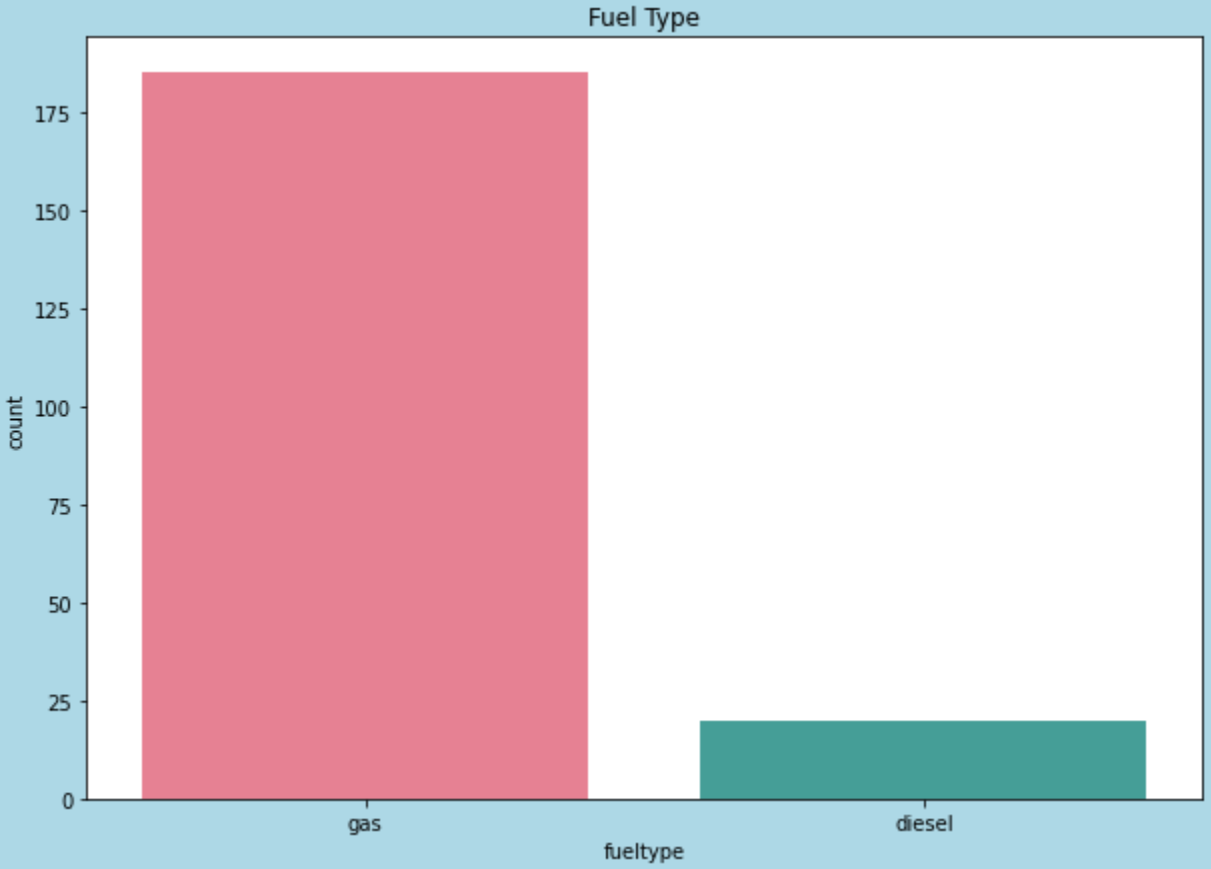


- The scatter line plot on the right shows the relationship between **fuel** consumption for **City in km/l** and **Price** of vehicle.
- From the plot there seems to be a negative linear relationship between both variables. With increase of **Horse Power** the **Price** decreases.
- The points are highly saturated and close to the line between Engine Size range of value 5 to 16 with price range of roughly $20000 to $5000.
- The Pearson Coefficient is -0.686 and p-value is 7.98 e- 30. This suggest that the correlation is highly negative and **City in km/l** plays an **important role** in determining the **Price** of vehicle.

```
Pearson Coefficient is: -0.6857513360270397
```

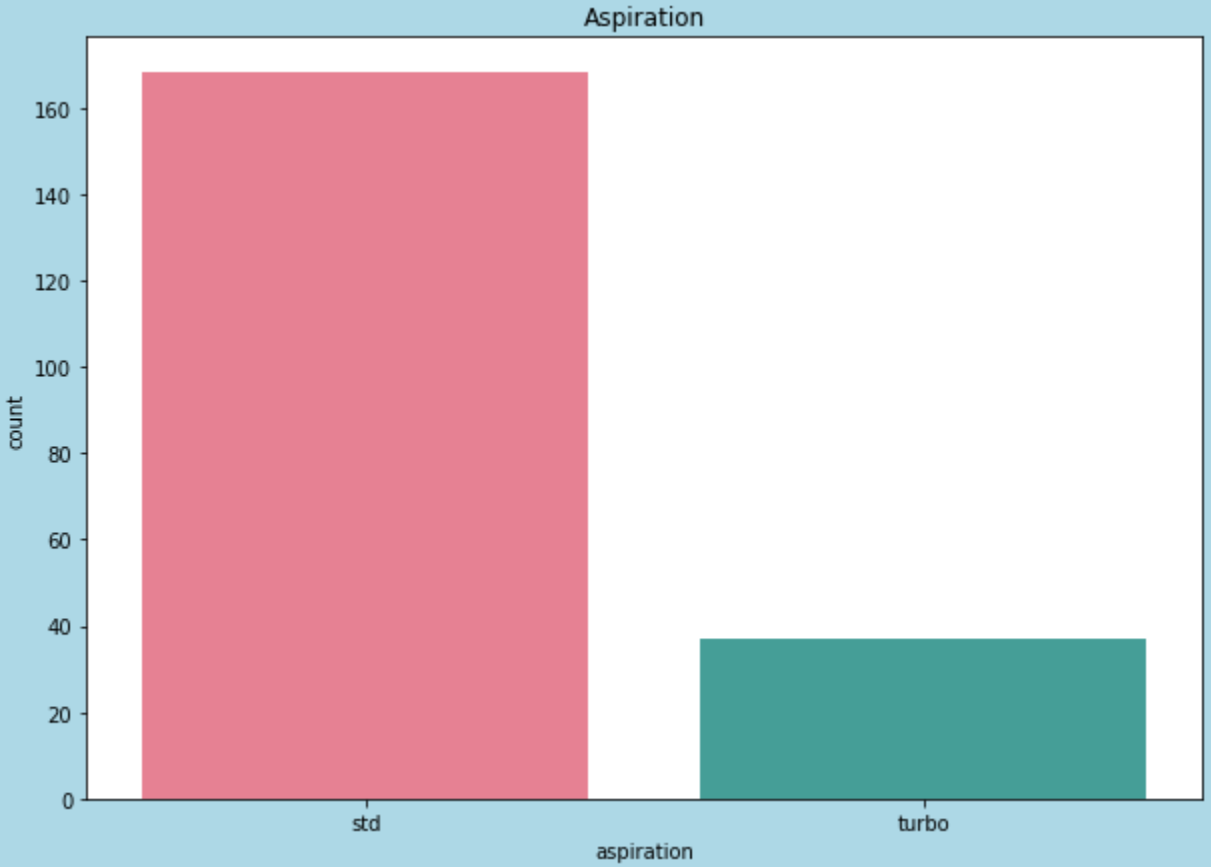
```
P-value is: 7.978684249663976e-30
```

EDA – Visual exploration (Categorical data)



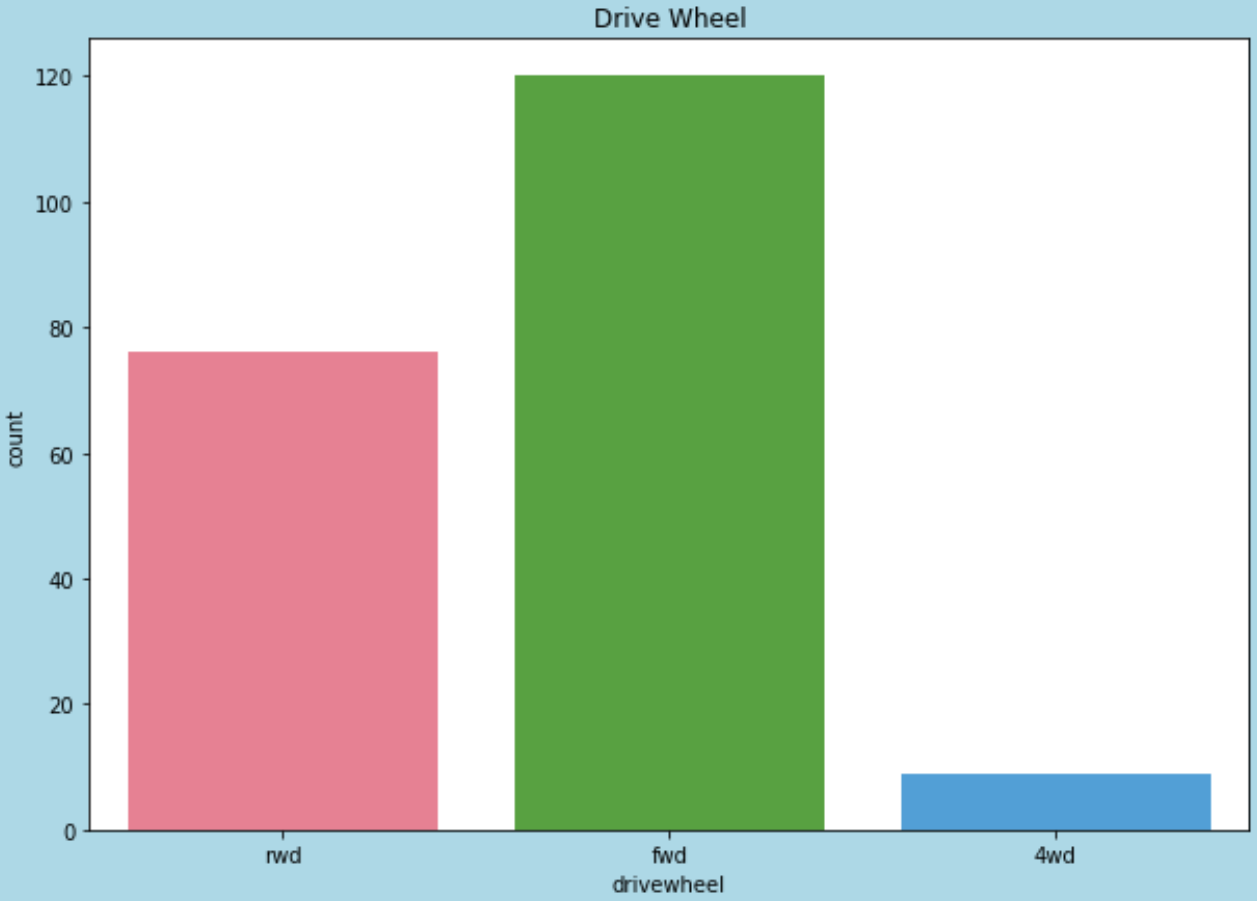
- The bar plot shows the total '**Fuel Type**' in the data set.
- There is a total of 185 '**Gasoline**' fuel type vehicles.
- There is a total of 20 '**Diesel**' fuel type vehicles

EDA – Visual exploration (Categorical data)



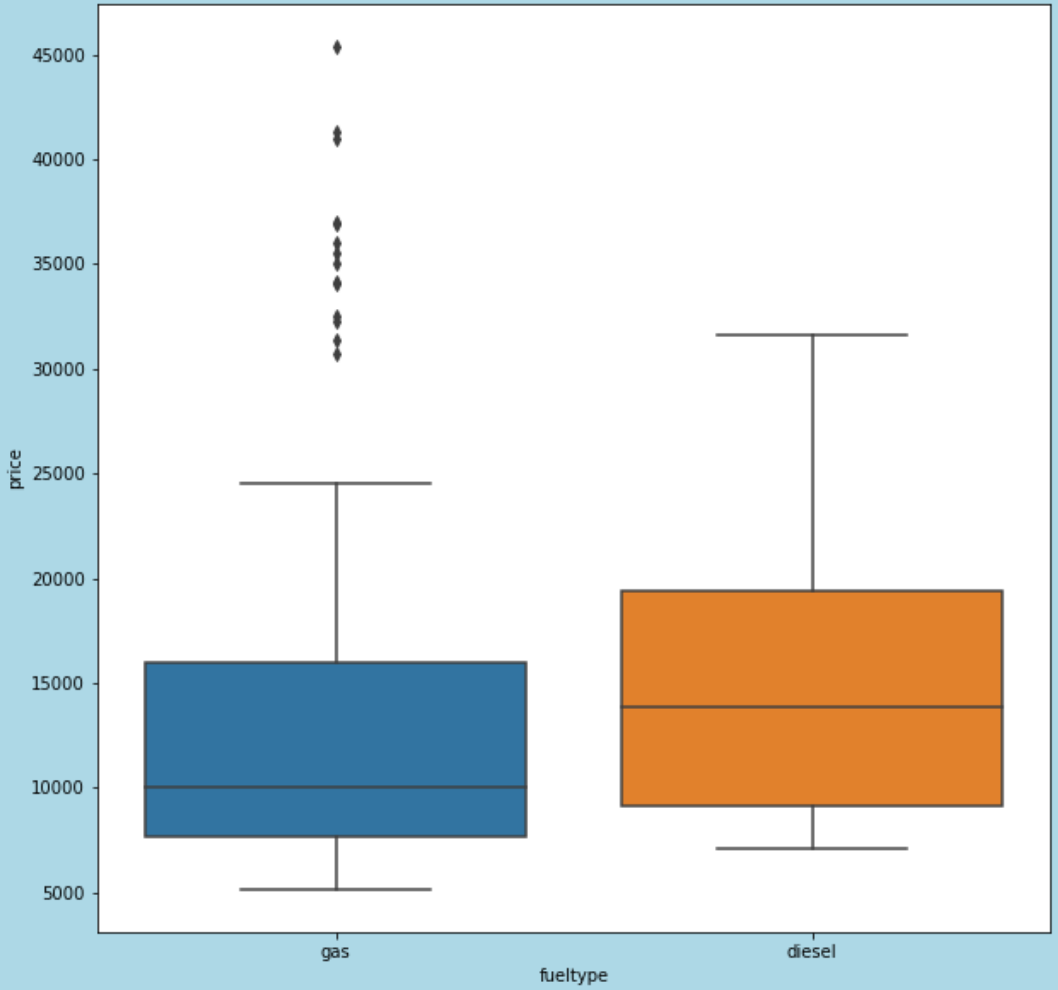
- The bar plot shows the total '**Aspiration Type**' in the data set.
- There is a total of 168 '**Standard**' aspiration type vehicles.
- There is a total of 37 '**Turbo**' aspiration type vehicles.

EDA – Visual exploration (Categorical data)



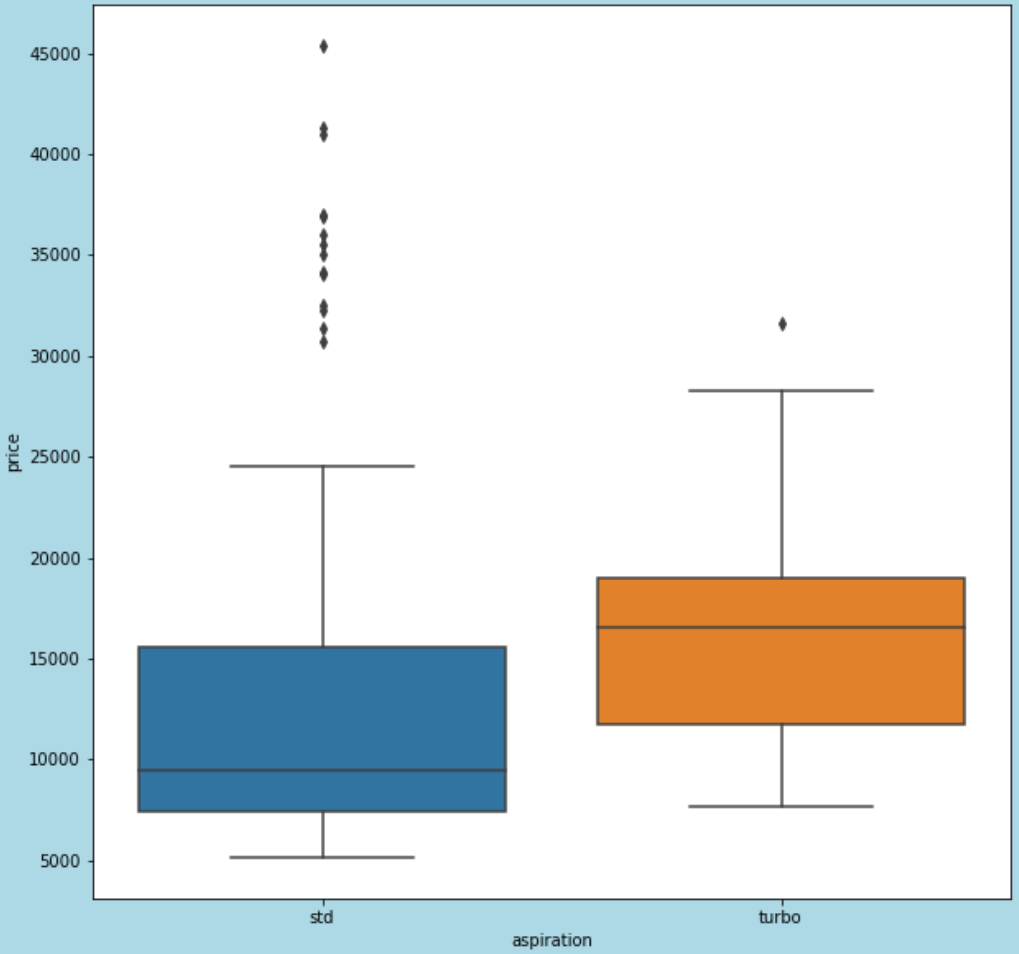
- The bar plot shows the total '**Drive Wheel Type**' in the data set.
- There is a total of more than 120 '**Forward Wheel Drive**' type vehicles.
- There is a total of roughly 76 '**Rear Wheel Drive**' type vehicles.
- There is a total of roughly 9 '**4 Wheel Drive**' type vehicles

EDA – Visual exploration (Categorical data)



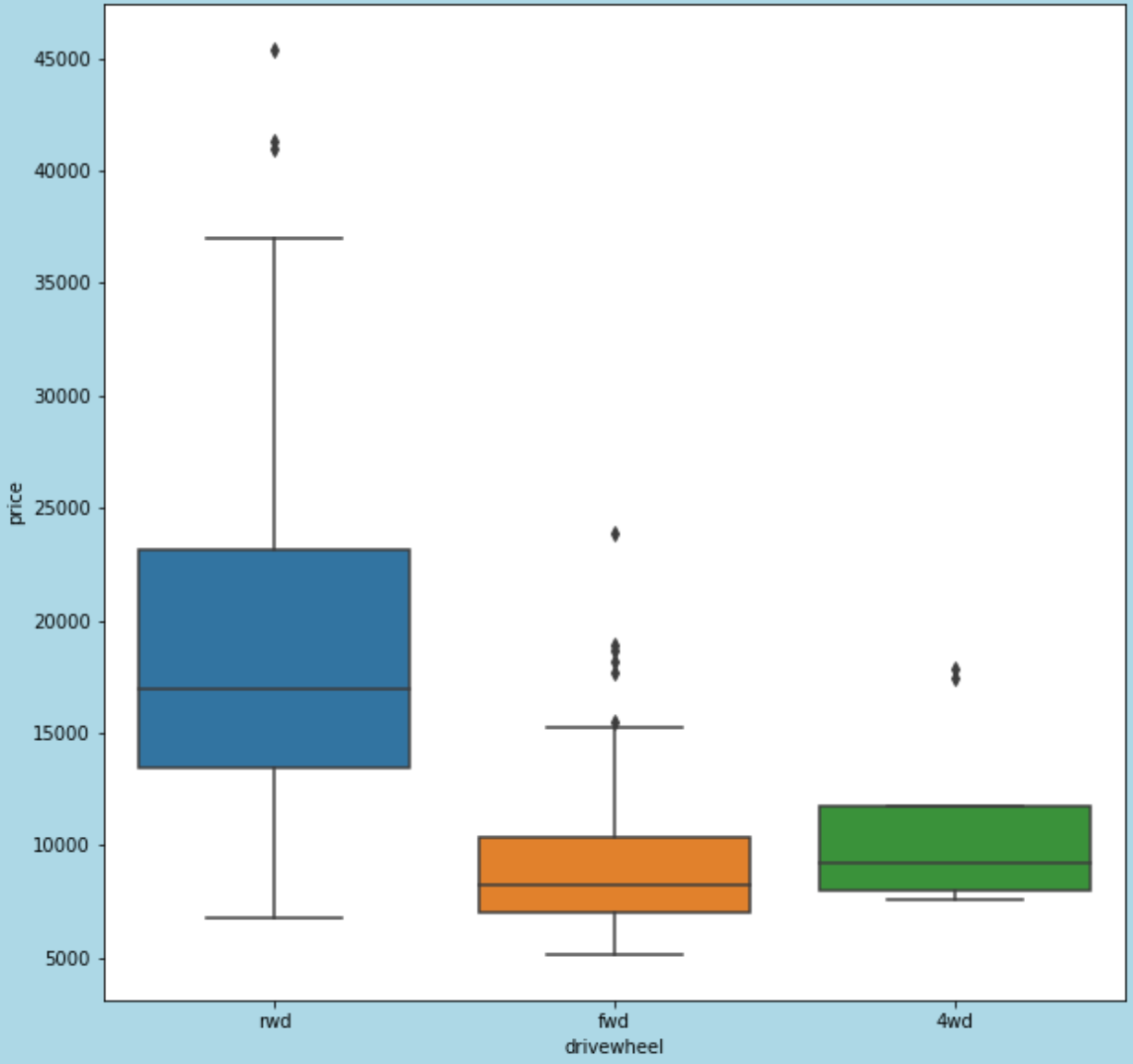
- The box plot shows '**Price** 'of '**Fuel Type**' in the data set.
- The minimum price for 'Gasoline Fuel' is roughly $5000 and $7000 for 'Diesel Fuel'.
- The average price for 'Gasoline Fuel' is roughly $10000 and $150000 for 'Diesel Fuel'.
- The maximum price for 'Gasoline Fuel' is roughly $25000 and $32500 for 'Diesel Fuel'.
- There are many outliers for 'Gasoline Fuel' but no outliers for 'Diesel' typefuel.

EDA – Visual exploration (Categorical data)



- The box plot shows '**Price** 'of '**Fuel Type**' in the data set.
- The minimum price for 'Gasoline Fuel' is roughly $5000 and $7000 for 'Diesel Fuel'.
- The average price for 'Gasoline Fuel' is roughly $10000 and $150000 for 'Diesel Fuel'.
- The maximum price for 'Gasoline Fuel' is roughly $25000 and $32500 for 'Diesel Fuel'.
- There is more outliers for 'Standard' aspiration compared to 'Turbo' aspiration.

EDA – Visual exploration (Categorical data)



- The box plot shows '**Price** 'of '**Drive Wheel Type**' in the data set.
- The minimum price for 'Rear Wheel Drive is roughly $6000, 'Forward Drive' is $5000 and $7000 for '4 Wheel Drive'.
- The average price for 'Rear Wheel Drive is roughly $16500, 'Forward Drive' is $7500 and $8500 for '4 Wheel Drive'.
- The maximum price for 'Rear Wheel Drive is roughly $37000, 'Forward Drive' is $15500 and $12000 for '4 Wheel Drive'.
- All 3 Drive Wheels have low amount of outliers.

Hypothesis Testing

Hypothesis I - Chi Square Test

Null Hypothesis (H0): There is no significant association between Categorise 'Fuel Type' and 'Aspiration'.

Alternative Hypothesis (Ha): There is evidence of significant association between Categorise 'Fuel Type' and 'Aspiration', both categorise are not independant.

Hypothesis II - T-test

Null Hypothesis (H0): There is no significant difference between 'Engine Size' and 'Horse Power'.

Alternative Hypothesis (Ha): There is significant difference between 'Engine Size' and 'Horse Power'.

Hypothesis III - T-test

Null Hypothesis (H0): There is no significant difference between 'Stroke' and 'Bore Ratio'.

Alternative Hypothesis (Ha): There is significant difference between 'Stroke' and 'Bore Ratio'.

Result – (Hypothesis III)

Hypothesis III - T-test

Null Hypothesis (H0): There is no significant difference between 'Stroke' and 'Bore Ratio'.

Alternative Hypothesis (Ha): There is significant difference between 'Stroke' and 'Bore Ratio'.

```
Ttest_indResult(statistic=-2.568760698249295, pvalue=0.010560971300206863)
```

- The p-value is 0.010561 which is smaller than 0.05.
- The Null Hypothesis (H0) is rejected and the Alternative Hypothesis is accepted (Ha).
- As a result there is significant difference between mean of the variables.

Improvements

- From the plots, we can observed that there were still outliers present. Future improvements can be done on reducing the outliers of certain column to ensure better results in analysis.

- Furthermore, Feature Scaling can be applied to data set to scale the features to normalize the range of independent variables.

Appendix

Link to Code

https://github.com/cs-robot-collab/IBM-Machine-Learning-Professional-Certificate/blob/master/Exploratory%20Data%20Analysis%20for%20Machine%20Learning.ipynb