# Exploratory Data Analysis

## Univariate Analysis

### Price

A graph with red lines

Description automatically generated

count 1435.000000

mean 10720.915679

std 3608.732978

min 4350.000000

25% 8450.000000

50% 9900.000000

75% 11950.000000

max 32500.000000

Figure 4‑1 Boxplot of Price and summary statistics of price

A graph of a graph

Description automatically generated with medium confidence

Figure 4‑2 Histogram of price

Price is the response variable of the study and it represents the car price of used cars. According to the histogram, it can be seen that price is unimodal positively skewed. Range of the car prices are spread between 4k and 32k and 9900 is the median value. According to the Boxplot it can be clearly seen that there are many outliers and among them 3 are highly deviated outliers. But there is not enough evidence to remove those outliers. And also it can be seen that 50% of the prices are between 8k and 12k.

### Age

A graph of a growing graph

Description automatically generated

count 1435.000000

mean 56.027178

std 18.532943

min 1.000000

25% 44.000000

50% 61.000000

75% 70.000000

max 80.000000

Figure 4‑3 Histogram of Age and Summary Statistics of Age

When considering the price of a used car, age of the car is very important and in this dataset age has been measured in months. According to the histogram of age, it can be seen that it is negatively skewed and unimodal. It follows a distribution from 1 months to 80 months with 61 as the median age. There is no any outliers

### KM

A green rectangular object with a cross

Description automatically generated with medium confidence

count 1435.000000

mean 68571.782578

std 37491.094553

min 1.000000

25% 43000.000000

50% 63451.000000

75% 87041.500000

max 243000.000000

Figure 4‑4 Boxplot of KM and Summery Statistics of KM

When buying a used car it is important to know the miles of the car that has run. Here it has been measured in Kilometers and boxplot according to the collected data as above. According to histogram it has a positively skewed distribution and it is unimodal. Distribution of KM varies in a huge range from 1 to 24k with 63k as the median value. There are few outliers that can be easily seen by the boxplot and 50% of data fall in between 43K to 87k. Even though this is a used car dataset, 1km driven cars don’t show as outliers according to the dataset. When investigating those, it showed that their age is considerably high for more cases. So keep every such observation as it is and do the analysis.

A green graph with white text

Description automatically generated

Figure 4‑5 Histogram of KM

### Fuel Type

A red circle with a triangle and a number of percentages

Description automatically generated

Figure 4‑6 Pie Chart of Fuel Type

Fuel Type is another main feature that is considered when buying both used cars and new cars because preference for the fuel type is dependent on many factors such as availability of the fuel type, cost for the fuels and ease to reach. In this data set cars have 3 fuel types and they are Petrol, Diesel and CNG which is compressed natural gas. In this dataset 88.1% of cars fuel type is Petrol and 10.7% of cars fuel type is Diesel. Only 1.18% of cars use CNG as fuel. So most of the cars used Petrol as the fuel type in this dataset and very few cars used CNG.

### HP (Horse Power)

A graph with a purple rectangle

Description automatically generated

count 1435.000000

mean 101.491986

std 14.981408

min 69.000000

25% 90.000000

50% 110.000000

75% 110.000000

max 192.000000

Figure 4‑7 Boxplot of Horse Power and Summary Statistics of Horse Power

Horse Power is another main feature that is considered when buying a car. It is the power that the engine produces. Here almost every cars’ horsepower is less than 116 and there is only one outlier which has 192 horsepower and it is a possible amount of horsepower. Thus there is not enough evidence to remove it from the dataset. 50% of cars’ horsepower is between 90 and 110 and median horse power is also 110.

### Transmission Type

A green circle with a orange triangle

Description automatically generated

Figure 4‑8 Pie Chart of Transmission Type

Based on the transmission type also customers preference for cars is varying. Even though there are four car transmission types named Automatic (AT), Manual(MT), Automated Manual (AM) and Continuously Variable(CVT) here have considered only whether a car is automatic or not. Pie chart of transmission type is as above and according to it, 94.4% of cars are not automatic and only 5.57% of cars are automatic. That means almost every car in this dataset is not automatic.

### Weight

A graph with a purple line

Description automatically generated with medium confidence

count 1435.000000

mean 1072.311498

std 52.246637

min 1000.000000

25% 1040.000000

50% 1070.000000

75% 1085.000000

max 1615.000000

Figure 4‑9 Boxplot of Weight and Summary statistics of the weight

Weight of the car is another feature that is considered when buying a car and here weight is measured in kilograms. According to the Histogram, it can be seen that there is a positively skewed distribution and it is unimodal. And also there are few outliers and they can be clearly identified by the corresponding boxplot. Among the outliers it can be clearly seen that 3 highly deviated outliers and because of them distribution is highly skewed. But since those weight of cars present, there is not enough evidence to remove them from the dataset. 50% of the car’s weight is between 1040kg and 1085kg. The median weight is 1070kg.

A graph with a bar graph

Description automatically generated with medium confidence

Figure 4‑10 Histogram of weight

### Metallic Color

A yellow and green pie chart

Description automatically generated

Figure 4‑11 Pie Chart of Color Type

Color of the car also can be a reason for the customers to choose a car. In here have considered whether the car has a metallic color or not. According to the pie chart, it can be seen that 67.5% of cars have metallic color and 32.5% of cars have no metallic color. Thus in the dataset, there are more metallic color cars than non-metallic color cars.

### CC (Engine Capacity)

A diagram with red lines

Description automatically generated

count 1435.000000

mean 1566.735192

std 186.767122

min 1300.000000

25% 1400.000000

50% 1600.000000

75% 1600.000000

max 2000.000000

Figure 4‑12 Boxplot of CC and Summary Statistics of CC

Engine capacity of the car is another important feature that is considered when buying a car and it is the measurement of the total volume of the cylinders in the engine. According to the boxplot, it can be seen that engine capacity of almost every car of this study is below 1900 and there are only three outliers which have engine capacity around 2000. Median engine capacity is 1600 and 50% of cars’ engine capacity is between 1400 and 1600.

### Doors

A blue rectangular object with a black pole

Description automatically generated with medium confidence

count 1435.000000

mean 4.032753

std 0.952667

min 2.000000

25% 3.000000

50% 4.000000

75% 5.000000

max 5.000000

Figure 4‑13 Boxplot of Doors and Summary statistics of doors

Number of doors also can be a reason to consider when buying a car. So in this study it has been considered too and the boxplot of it is as above. According to that, the median number of vehicles have 4 doors and 75% of vehicles have the number of doors in 3, 4 or 5.

## Bivariate Analysis

### Association between Price and Age of cars

A graph with a red line

Description automatically generated

Figure 4‑14 Scatterplot of Age vs. Price

When considering the relationship between Price and the Age of cars, it can be seen that there is a strong negative relationship according to the corresponding scatterplot. So it can be concluded that the price of a used car will depend on the age of the car in a negative manner. Thus when the age of the car is increasing the price of the used cars will decrease.

### Association between Price and KM

A graph showing a red line and green dots

Description automatically generated

Figure 4‑15 Scatterplot of KM vs. Price

To consider the relationship between Price and kilometers that the car has been driven, the corresponding scatterplot has been drawn. According to the graph it can be seen that there is a strong negative relationship. Therefore it can be said that used car prices will decrease with the number of kilometers that it has driven. So the price of a used car will depend on the number of kilometers in a negative manner.

### Association between Price and Fuel Type

A graph of different colored squares

Description automatically generated with medium confidence

Figure 4‑16 Boxplot of Fuel Type vs. Price

Considering the relationship between the fuel type and the price is also important. So according to the boxplot of Price vs. Fuel Type, it can be seen that the median price of Petrol cars is higher than other two types of fuel and the median price of Diesel and CNG cars is approximately equal. Price of diesel Cars varies in a broad range and the price of CNG cars is varying in a small range compared to the other two types. Both diesel and petrol fuel types’ prices have outliers and diesel cars have highly deviated outliers.

### Association between Price and Horsepower

To consider the relationship between Price and the Horsepower of the car, the corresponding scatterplot was drawn as below. According to that, it can be seen that there is a moderately strong positive relationship between price and horsepower. So when increasing the horsepower price of used cars is also increasing. Thus Horse power has a positive impact on the car price.

A graph with a red line

Description automatically generated

Figure 4‑17 Scatterplot of Horse Power vs. Price

### Association between Price and the Transmission Type

A graph of different colored boxes

Description automatically generated with medium confidence

Figure 4‑18 Boxplot of Transmission Type vs. Price

When considering the relationship between Price and the transmission type, it can be seen that the price of non-automatic cars is varying in a larger range than automatic cars according to the corresponding boxplot. But the median price of both types of cars is approximately equal. When looking at the maximum priced car and the lowest priced car in the dataset both belong to non-automatic transmission type.

### Association between Price and Metallic color

A graph of different colored squares

Description automatically generated with medium confidence

Figure 4‑19 Boxplot of Metallic Color vs. Price

When considering the boxplot of price and metallic color, it can be seen that the median price of both types of cars is approximately equal and the interquartile price range of having metallic color cars is larger than the price of cars which do have not metallic colors. Both types of cars have outliers. Since both median prices are approximately equal, there might not be a strong impact on the price if the metallic color of the car is present or not.

### Association between Price and Engine Capacity

A graph with red lines

Description automatically generated

Figure 4‑20 Scatterplot of Engine Capacity vs. Price

Considering the relationship of price and the engine capacity is also important. According to the scatterplot between Price and engine capacity, it can be seen that there is a very weak positive relationship between price and the engine capacity. So it can be said that the engine capacity of cars does not highly increase the price of used cars according to this dataset.

### Association between Price and Doors

A graph with a line and a red line

Description automatically generated

Figure 4‑21 Scatterplot of Doors vs. Price

When considering the scatterplot between price and number of doors, it can be seen that there is a weak positive relationship. Therefore it can be said that the numbers of doors doesn’t have a huge impact on the price of used cars.

A graph of a chart

Description automatically generated with medium confidence

Figure 4‑22 Side by side boxplot of Doors vs. Price

Relationship of the number of doors with the price can be clearly seen by a side by side boxplot. According to that, it can be seen that the highest median price taken by cars which have 5 doors and its price also vary in a huge range. Cars which have only 2 doors have the least median price.

### Association between Price and the weight of the car

A graph with a red line

Description automatically generated

Figure 4‑23 Scatterplot of Weight vs. Price

To consider the relationship between price and the weight of the car corresponding scatterplot has been drawn. According to that, it can be seen that there is a strong Positive relationship. So it can be say that weight of the car has huge impact on the price of the car.

### Correlation Matrix

A screenshot of a computer

Description automatically generated

Figure 4‑24 Correlation Matrix (Heat Map)

By considering the correlation matrix, relationships among quantitative variables can be clearly seen. Pearson correlations among them have been calculated. According to this also it can be observed that variables Age has a strong relationship and KM and weight has moderately strong relationships with the Price. Doors and CC have weak positive relationships with Price and there is a moderately strong relationship between CC and Weight according to the correlation matrix.

Correlations between Price and categorical variables “Automatic” and “MetColor” were calculated by using the Point- biserial Correlation and there were respectively 0.0339 and 0.1076 correlations. Thus there are weak positive relationships between Price and the Automatic and, Price and the metColor.