***Case Study-Exploratory Data Analysis***

***Submitted by: T. Sujaya***

***INTRODUCTION***

In Statistics Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. EDA essentially helps us in knowing our data better .It’s a good way of quality control for mistakes in data entry .It also helps us in finding outliers that influence our output. If we are performing analysis that requires certain distributional assumptions to be fulfilled, EDA will help us to find whether we are fulfilling those assumptions or not. It also helps us in finding missing data.

Discrete data has a finite or countably infinite set of values, which may or may not be represented as integers. The attributes hair colour, smoker, medical test and drink size each have a finite number of values and so are discrete. Continuous data represents measurements, there possible values cannot be counted and can only be described using intervals on the real number line. It can be uncountably infinite. Nominal data represents characteristics such as a person’s gender, marital status. It can take on numerical values (such as 1 indicating male and 2 indicating female), but those numbers don’t have mathematical meaning .Parameters in statistics is an important component of any statistical analysis. In simple words, a parameter is any numerical quantity that characterizes a given population or some aspect of it. Note that all the distributions depend on some parameters like λ, μ or other .These parameters determine some characteristics of the shape of the pdf/pmf of a random variable. It can be location, spread, skewness etc. When certain parameters are considered fixed and known we get a family of distributions. In probability and statistics, an exponential family is a set of probability distributions of a certain form. The exponential families include many of the most common distributions, including normal, exponential, gamma, chi-squared, beta and many others.

***METHODOLOGY***

***Exploratory Data Analysis for individual variables***

One of the most important aspects of describing a distribution is the central value around which the observations are distributed. Any arithmetical measure which is intended to represent the centre or central value of a set of observations is known as measures of central tendency. They summarize a sample or population by a single typical value. The two most commonly used measures of central tendency for numerical data are the mean and median. Mean is defined as the average of all the observations. Mean can be influenced profoundly by outliers. The trimmed mean is obtained after removing 5% of extreme observations from both ends. Averages do not give us a complete picture of the distribution. Median is a score found at the exact middle of the set of values.It does not get affected by extreme observations. Mode is the most frequently occurring value in the set of scores. In addition to it we should have a measure of variation of observations. The scatteredness or variation of observations from their average is called dispersion. They are different measures of dispersion like the range, standard deviation, interquartile range and mean deviation. Interquartile range is a better measure than the range as it makes use of 50% of the data. Skewness and kurtosis are used to describe the nature of the distribution. Skewness means lack of symmetry .A distribution is said to be symmetrical when the values are uniformly distributed around the mean. A measure of peakedness or convexity of a curve is known as kurtosis. Standard error is the standard deviation of sample means.

Bar chart is used to visually compare the data from different groups. The histogram is a commonly used display .The range of observed values is subdivided into equal intervals and then the cases in each interval are obtained .A display closely related to the histogram is the stem and leaf plot. Stem and Leaf plots are a method for showing the frequency with which certain classes of values occur. The STEM is the left hand column which contains the tens digits. The LEAVES are the lists in the right hand column, showing all the ones digits for each of the tens, twenties and thirties. The horizontal leaves in the stem and leaf plot correspond to the vertical bars in the histogram and the leaves have lengths that equal the numbers in frequency tables. The advantage of using stem and leaf over a histogram is that we get to see the entire data in a tabular form along with their respective frequencies. A display that further summarizes the information about the distribution of values is the box plot. It displays summary statistics for the distribution. It plots the median, 25th percentile, 75th percentile and values that are deviating from the rest.50% of the cases lie within the box. The length of the box corresponds to the interquartile range which is the difference between 1st and 3rd quartiles. It identifies extreme values which are more than 3 box lengths from the upper or lower edge of the box. The values which are more than 1.5 box lengths are known as outliers. The median lies within the box. Length of the box depicts variability of observations. If median is closer to the bottom of the box than the top, the data are positively skewed otherwise it is negatively skewed.

Missing value imputation is a method by which we substitute plausible values for the missing scores. Normality can be tested by plotting a normal plot .In a normal probability plot each observed value is paired with its expected value from the normal distribution. In a situation of normality it is expected that points will fall on a straight line.

Besides these visual displays, the statistical tests are Shapiro-Wilk and the Lilliefors. The Lilliefors test is based on the modification of Kolmogrov-Smirnov test for the situation when means and variances are not known but are estimated from the data. Kolmogrov-Smirnov is a non parametric test and Shapiro-Wilk is a parametric test .Kolmogrov Smirnov is typically used for larger data sets. Shapiro-Wilk is used for smaller data sets (i.e. with less than 50 observations).In statistics the p-value is a function of the observed sample results(a statistic) that is used for testing a statistical hypothesis. Before the test is performed, a threshold value is chosen, called the significance level of the test, usually 5% or 1% and denoted as α.

***Decision Rule*** –If p-value>0.05, we may accept the null hypothesis at 5% level of significance (here α=5).Otherwise we may reject our null hypothesis.

***Exploratory Data Analysis for Multiple variables***

In terms of correlation graphical plots are called scatter plots. Scatter plots can show you visually the strength of the relationship and direction of the relationship between the variables and whether the outliers exist. Instead of plotting several variables on the same axis (which can be difficult to interpret with several variable pairs), it is possible to plot a matrix of scatter plots. This type of plot allows you to see the relationship between all combinations of many different pairs of variables. Correlation is the measure of strength and direction of the relationship between the variables. The direction of the relationship can be either positive or negative. The strength of the relationship is measured from -1 to 1.The farther the value is from 0, the stronger the relationship. Partial correlation is a type of correlation that allows us to look at the relationship between two variables when the effects of a third variable are held constant.

Consider the mtcars data set available in R. We have exported it to spss using function (write.csv (mtcars,”mtcars.sav”)).The data was extracted from the 1974 Motor Trend US Magazine and comprises gasoline mileage in miles per gallon and ten aspects of automobile design and performance for 32 automobiles(1973-74 models).It's a data frame with 32observations on 11 variables. The source of data is Henderson and Velleman(1981),building multiple regression models interactively.Biometrics,37,391-441.

We have the variables and their data types as follows:

***Table 1.1: Description of Variables in Our Dataset***

|  |  |  |
| --- | --- | --- |
| S.No. | Variable Name | Type |
| 1 | Miles /(US) gallon (mpg) | Continuous |
| 2 | Number of cylinders (cyl) | Discrete |
| 3 | Displacement (disp) | Continuous |
| 4 | Gross horsepower (hp) | Continuous |
| 5 | Rear axle ratio (drat) | Continuous |
| 6 | Weight (wt) | Continuous |
| 7 | (1/4)th mile time(qsec) | Continuous |
| 8 | V/S(vs)(0=Vengine1=straightengine) | Discrete |
| 9 | Transmission (0=automatic,1=manual)(am) | Discrete |
| 10 | Number of forward gears(gear) | Discrete |
| 11 | Number of carburetors(carb) | Discrete |

***We start EDA with mpg***.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Table 1.2 DESCRIPTIVES(mpg)*** | | | Statistic | Std. Error |
| Miles per gallon | Mean | | 20.09 | 1.065 |
| 95% Confidence Interval for Mean | Lower Bound | 17.92 |  |
| Upper Bound | 22.26 |  |
| 5% Trimmed Mean | | 19.89 |  |
| Median | | 19.20 |  |
| Variance | | 36.324 |  |
| Std. Deviation | | 6.027 |  |
| Minimum | | 10 |  |
| Maximum | | 34 |  |
| Range | | 24 |  |
| Interquartile Range | | 8 |  |
| Skewness | | .672 | .414 |
| Kurtosis | | -.022 | .809 |

1) First of all we see the mean and the median to check if they are approximately equal. Since there is a difference of nearly one unit it is not normally distributed i.e. might be skewed (As in a normal distribution mean, median and mode are equal).

2).Here mean>median and coefficient of skewness is greater than zero .We can infer that the distribution is positively skewed. Moreover the coefficient is insignificant as its absolute value is not greater than twice its standard error (0.672<0.828).

3) Ideal value for kurtosis is 0. The ratio of kurtosis to standard error is -0.0271 .Hence we conclude that data is normal. A negative value for kurtosis indicates that the observations cluster less and have shorter tails (becoming like those of a box shaped uniform distribution).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Table 1.3:Tests of Normality*** | | | | | | |
|  | ***Kolmogorov-Smirnova*** | | | Shapiro-Wilk | | |
| Statistic | Df | Sig. | Statistic | Df | Sig. |
| Miles per gallon | .126 | 32 | .200\* | .948 | 32 | .123 |
| \*. This is a lower bound of the true significance. a. Lilliefors Significance Correction | | | | | | |

We set the following hypothesis:

Ho-The observed distribution fits the normal distribution.

H1-The observed distribution does not fit the normal distribution.

So from the table above we infer that the p value is 0.200(which is >0.05).Hence we accept the null hypothesis at 5% level of significance and conclude that data fits the normal distribution.

|  |
| --- |
| ***A stem and leaf plot shows all data values .From the table on the left we infer from the first row that the observations10, 10, 13, 14, 14 appear in the data. The digit with 1 in the tens place occurs 5 times. Every row can be interpreted similarly.*** |

***Table1.4: Stem-and-Leaf Plot***

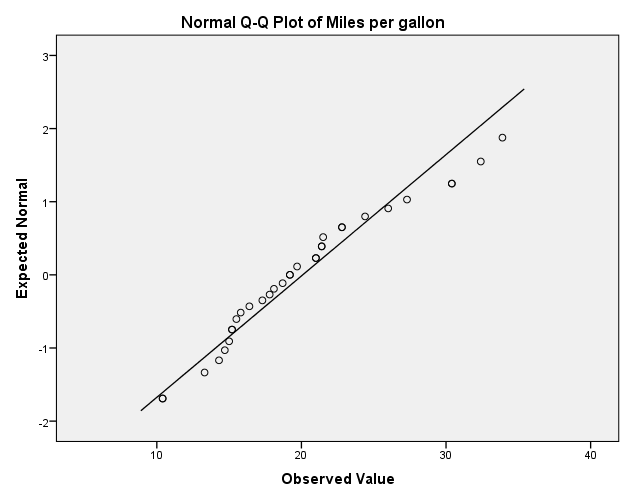
|  |  |
| --- | --- |
| ***Frequency*** | ***Stem and Leaf*** |
| 5.00 | 1 . 00344 |
| 13.00 | 1 . 5555567788999 |
| 8.00 | 2 . 11111224 |
| 2.00 | 2 . 67 |
| 4.00 | 3 . 0023 |
| Stemwidth: | 10 |
| Eachleaf: | 1 case(s) |

***Figure 1.1: Histogram***



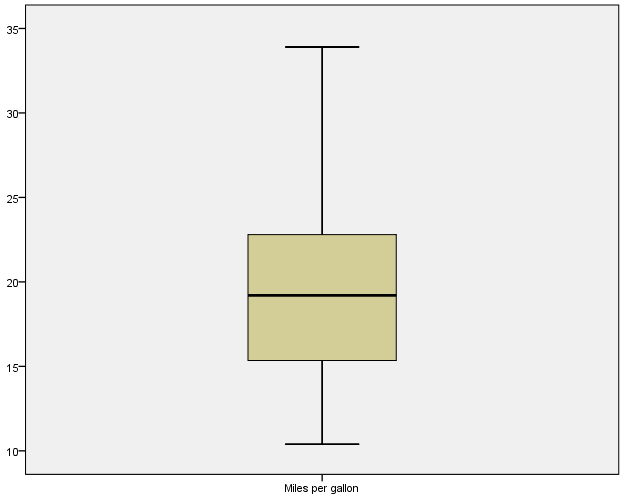
The gap between 27.5 and 30 reveals that there are no frequencies. The distribution could be bimodal. This might suggest that the sample is not homogeneous but possibly its elements came from two different populations, each more or less normally distributed.

***Figure 1.2: Normal Q-Q plot***



As the points are more or less close to the straight line except at around 30.So from above we infer that the data is fairly normally distributed except at around unit 30.

***Figure 1.3******Box plots***



As no points are lying above and below the upper and lower edge of the lines (near 10 and 35) we infer that there are no outliers. As the median (=20) is in the centre of the box, the values are not skewed.

***EDA with (no of cylinders)***

| ***Table 1.5 : Descriptive Statistics*** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | N | Minimum | Maximum | Mean | | Std. Deviation | Skewness | |
|  | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Statistic | Std. Error |
| cyl | 32 | 4 | 8 | 6.19 | .316 | 1.786 | -.192 | .414 |

1) Smaller value of standard deviation implies that the data is less spread out.

2) As coefficient of skewness <0,it implies that it is negatively skewed.Moreover the coefficient of skewness is insignificant as its absolute value is not greater than twice its standard error (0.192<0.828).

***Figure 1.4: Bar chart***



The bar chart shows that the mtcars data has been split into three classes depending on the number of cylinders.

***Figure 1.5: Box plot***

******

As no points are lying above or below the edges of the box, we infer that there are no outliers.

***EDA with displacement***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Table 1.6:Descriptives*** | | | Statistic | Std. Error |
| Displacement | Mean | | 230.72 | 21.909 |
| 95% Confidence Interval for Mean | Lower Bound | 186.04 |  |
| Upper Bound | 275.41 |  |
| 5% Trimmed Mean | | 226.34 |  |
| Median | | 196.30 |  |
| Variance | | 15360.800 |  |
| Std. Deviation | | 123.939 |  |
| Interquartile Range | | 222 |  |
| Skewness | | .420 | .414 |
| Kurtosis | | -1.068 | .809 |

1) Here mean>median and coefficient of skewness is greater than zero .We can infer that the distribution is positively skewed. Moreover the coefficient is insignificant as its absolute value is not greater than twice its standard error (0.420<0.828).

2) Ideal value for kurtosis is 0.The ratio of kurtosis to standard error is -1.3201 which is not less than -2.Hence we conclude that data is normal. A negative value for kurtosis indicates that the observations cluster less and have shorter tails (becoming like those of a box shaped uniform distribution)

3) As standard deviation is large, it implies the observations are more spread out.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Table 1.7:Tests of Normality*** | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | Df | Sig. | Statistic | Df | Sig. |
| Displacement | .195 | 32 | .003 | .920 | 32 | .021 |
| -Lilliefors Significance Correction  We set the following hypothesis :  Ho-The observed distribution fits the normal distribution.  H1-The observed distribution does not fit the normal distribution.  So from the table above we infer that the p value is 0.003(which is <0.05).Hence we reject the null hypothesis at 5% level of significance and conclude that data does not fit the normal distribution. | | | | | | |

***Figure 1.6: Histogram***



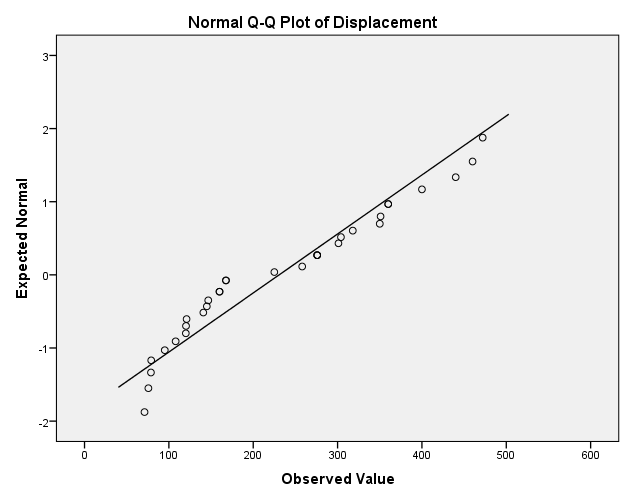
So from the above figure we see that the data is not normally distributed. The distribution could be bimodal. This might suggest that the sample is not homogeneous but possibly its elements came from two different populations.

***Table 1.8: Stem-and-Leaf Plot***

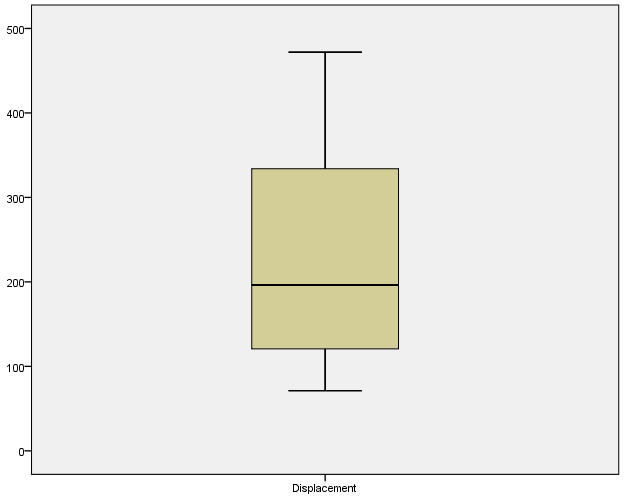
|  |  |
| --- | --- |
| Frequency | Stem & Leaf |
| 5.00 | 0 . 77779 |
| 7.00 | 1 . 0222444 |
| 4.00 | 1 . 6666 |
| 1.00 | 2 . 2 |
| 4.00 | 2 . 5777 |
| 3.00 | 3 . 001 |
| 4.00 | 3. 5566 |
| 2.00  2.00 | 4. 04  4.67 |
| Stmwidth: | 100 |
| Each leaf: 1 case(s) | |

A stem and leaf plot shows all data values .From the table on the left we infer from the sixth row that the observations 30, 30, 31 appear in the data. The digit with 3 in the tens place occurs 3 times. Every row can be interpreted similarly.

***Figure 1.7: Normal Q-Q Plot***

  
As the points are deviating from the straight line, hence it’s not normally distributed.

***Figure 1.8: Box Plot***



We can see from the box plot that sample median is 200.The width of the box is around 200 units which is the interquartile range. There are no outliers.

***EDA with Gross horsepower***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Table 1.9:Descriptives*** | | | Statistic | Std. Error |
| Gross horsepower | Mean | | 146.69 | 12.120 |
| 95% Confidence Interval for Mean | Lower Bound | 121.97 |  |
| Upper Bound | 171.41 |  |
| 5% Trimmed Mean | | 142.76 |  |
| Median | | 123.00 |  |
| Variance | | 4700.867 |  |
| Std. Deviation | | 68.563 |  |
| Minimum | | 52 |  |
| Maximum | | 335 |  |
| Range | | 283 |  |
| Interquartile Range | | 85 |  |
| Skewness | | .799 | .414 |
| Kurtosis | | .275 | .809 |

1) First of all we see the mean and the median to check if they are approximately equal. As we observe the difference we infer that it is not normally distributed i.e. might be skewed (As in a normal distribution mean, median and mode are equal).

2) The trimmed mean is 142.76 and is less than the mean (146.69).A considerable amount of difference indicates presence of outliers.

3).Here mean>median and coefficient of skewness is greater than zero .We can infer that the distribution is positively skewed. Moreover the coefficient of skewness is insignificant as its absolute value is not greater than twice its standard error (0.799<0.828).

4) Ideal value for kurtosis is 0. The ratio of kurtosis to standard error is 0.3399.Hence we conclude that data is normal. A positive value for kurtosis indicates that the observations cluster more and have longer tails (than those in normal distribution).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Table 1.10:Tests of Normality*** | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | Df | Sig. | Statistic | df | Sig. |
| Gross horsepower | .166 | 32 | .024 | .933 | 32 | .049 |
| a. Lilliefors Significance Correction | | | | | | |

We set the following hypothesis:

Ho-The observed distribution fits the normal distribution.

H1-The observed distribution does not fit the normal distribution.

So from the table above we infer that the p value is 0.024(which is <0.05).Hence we reject the null hypothesis at 5% level of significance and conclude that data does not fit the normal distribution.

***Figure 1.9: Histogram***



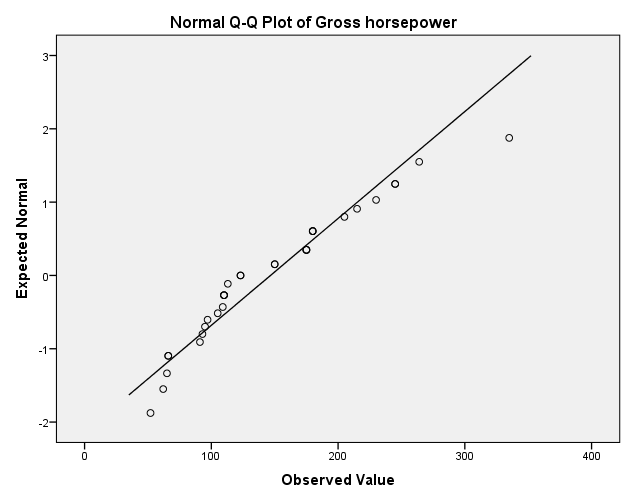
When a histogram has a bell shape, it suggests that data may have come from a normal distribution.Since we do not have a bell curve so we conclude that data is not normal.

***Table1.11: Stem-and-Leaf Plot***

|  |  |
| --- | --- |
| Frequency | Stem & Leaf |
| 9.00 | 0 . 566669999 |
| 8.00 | 1 . 00111122 |
| 8.00 | 1 . 55777888 |
| 5.00 | 2 . 01344 |
| 1.00 | 2 . 6 |
| 1.00 | Extremes (>=335) |
| Stemwidth: | 100 |
| Each leaf: | 1 case(s) |

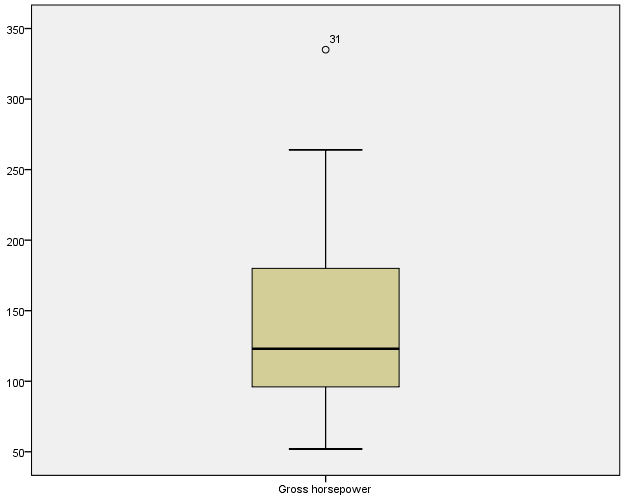
|  |
| --- |
| ***A stem and leaf plot shows all data values .From the table on the left we infer from the fourth row that the observations 20, 21, 23, 24, 24 appear in the data. The digit with2 in the tens place occurs 5 times. Every row can be interpreted similarly.*** |

***Figure 1.10: Normal Q-Q plot***



Since points do not lie on the straight line, the data is not normal.

***Figure 1.11: Box Plot***



As the box plot is not symmetric with the median line it suggests data has not come from a normal distribution .It also indicates the presence of outliers.

So we will perform EDA again after removing the outlier and using missing value imputation.

The missing value has been calculated by the method of median of nearby points.

| ***EDA with gross horsepower*** | | | | |
| --- | --- | --- | --- | --- |
| ***Table1.12*** | ***Descriptives*** | ***Horsepower*** | Statistic | Std. Error |
| MEAN(hp,1) | Mean | | 140.656 | 10.4882 |
| 95% Confidence Interval for Mean | Lower Bound | 119.265 |  |
| Upper Bound | 162.047 |  |
| 5% Trimmed Mean | | 138.917 |  |
| Median | | 123.000 |  |
| Variance | | 3520.104 |  |
| Std. Deviation | | 59.3305 |  |
| Minimum | | 52.0 |  |
| Maximum | | 264.0 |  |
| Range | | 212.0 |  |
| Interquartile Range | | 84.5 |  |
| Skewness | | .460 | .414 |
| Kurtosis | | -.749 | .809 |

After removing the outlier we observe the following:

1)The mean of the variable has considerably reduced while the median remains unchanged.As mean and median are not equal it implies that the distribution may be skewed.Coefficient of skewness is positive implying positive skewness.

2)A negative value of kurtosis indicates shorter tails.

3)We see a drastic reduction in the variance implying less spread of data.

| ***Table 1.13: Tests of Normality*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | Df | Sig. | Statistic | df | Sig. |
| MEAN(hp,1) | .148 | 32 | .072 | .944 | 32 | .096 |
| a. Lilliefors Significance Correction | | | | | | |

We set the following hypothesis:

Ho-The observed distribution fits the normal distribution.

H1-The observed distribution does not fit the normal distribution.

So from the table above we infer that the p value is 0.072 (which is >0.05).Hence we accept the null hypothesis at 5% level of significance and conclude that data does not fit the normal distribution.

***Figure 1.12: Histogram***



Clearly from the graph we infer that the distribution appears to be normally distributed.

***Figure 1.13: Normal Q-Q Plot***



We infer that after removing the outliers the data is normally distributed.

***Figure 1.14: Box plot***



From the figure we infer that there are no outliers.As the central line lies near the bottom of the box it implies data is slightly negatively skewed.

***EDA with Rear axle ratio***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Table 1.14:Descriptives*** | | | Statistic | Std. Error |
| Rear axle ratio | Mean | | 3.597 | .0945 |
| 95% Confidence Interval for Mean | Lower Bound | 3.404 |  |
| Upper Bound | 3.789 |  |
| 5% Trimmed Mean | | 3.579 |  |
| Median | | 3.695 |  |
| Variance | | .286 |  |
| Std. Deviation | | .5347 |  |
| Minimum | | 2.8 |  |
| Maximum | | 4.9 |  |
| Range | | 2.2 |  |
| Interquartile Range | | .8 |  |
| Skewness | | .293 | .414 |
| Kurtosis | | -.450 | .809 |

1. Since the value of the mean and trimmed mean are approximately equal so it indicates absence of outliers.
2. As the standard deviation is small the observations are less spread out.
3. As kurtosis is negative, it implies tails are lighter and we will assume that it is flatter than what the normal distribution will allow.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Table 1.15:Tests of Normality*** | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | Df | Sig. | Statistic | Df | Sig. |
| Rear axle ratio | .160 | 32 | .037 | .946 | 32 | .110 |
| a. Lilliefors Significance Correction | | | | | | |

We set the following hypothesis:

Ho-The observed distribution fits the normal distribution.

H1-The observed distribution does not fit the normal distribution.

So from the table above we infer that the p value is 0.037(which is <0.05).Hence we may reject the null hypothesis at 5% level of significance and conclude that data does not fit normal distribution.

***Figure 1.15: Histogram***

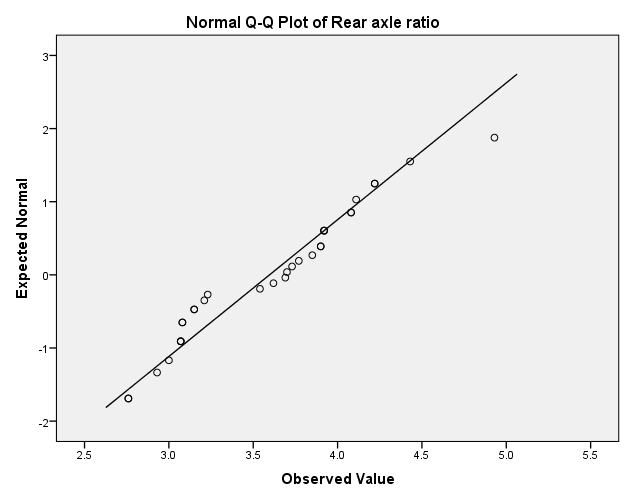


As there are some gaps in the histogram it indicates either presence of outliers before 3.5 and after 4.5 or no frequencies in that interval.

|  |  |
| --- | --- |
| ***Table 1.16: Stem-and-Leaf Plot*** | |
| Frequency | Stem & Leaf |
| 3.00 | 2 . 779 |
| 10.00 | 3 . 0000001122 |
| 12.00 | 3 . 566777899999 |
| 6.00 | 4 . 001224 |
| 1.00 | 4 . 9 |
| Stem width: | 1. 0 |
| Each leaf: | 1 case(s) |

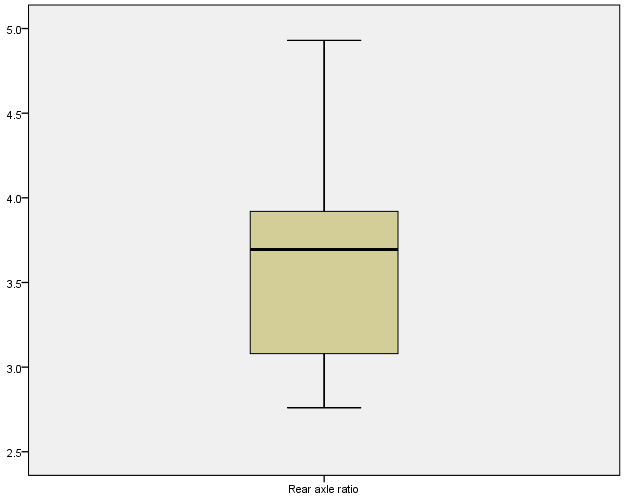
***Figure 1.16: Normal Q-Q Plot***

|  |
| --- |
| ***A stem and leaf plot shows all data values .From the table on the left we infer from the first row that the observations 27, 27, 29 appear in the data. The digit with2 in the tens place occurs 3 times. Every row can be interpreted similarly.*** |



As some points in the data are going astray it implies that data is not normally distributed.

***Figure 1.17: Box Plot***



There are no outliers.It is positively skewed as central line is near the top of the box.

***EDA with weight***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Table 1.17:Descriptives*** | | | Statistic | Std. Error |
| Weight | Mean | | 3.2173 | .17297 |
| 95% Confidence Interval for Mean | Lower Bound | 2.8645 |  |
| Upper Bound | 3.5700 |  |
| 5% Trimmed Mean | | 3.1889 |  |
| Median | | 3.3250 |  |
| Variance | | .957 |  |
| Std. Deviation | | .97846 |  |
| Minimum | | 1.51 |  |
| Maximum | | 5.42 |  |
| Range | | 3.91 |  |
| Interquartile Range | | 1.19 |  |
| Skewness | | .466 | .414 |
| Kurtosis | | .417 | .809 |

Since the value of the mean and trimmed mean are approximately equal so it indicates absence of outliers. As the standard deviation is small the observations are less spread out .As kurtosis is positive, it implies tails are heavier and we will assume that it is peaker than what the normal distribution will allow.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Table 1.18:Tests of Normality*** | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | Df | Sig. | Statistic | df | Sig. |
| Weight | .136 | 32 | .142 | .943 | 32 | .093 |
| a. Lilliefors Significance Correction | | | | | | |

We set the following hypothesis:

Ho-The observed distribution fits the normal distribution.

H1-The observed distribution does not fit the normal distribution.

So from the table above we infer that the p value is 0.142(which is >0.05).Hence we may accept the null hypothesis at 5% level of significance and conclude that data fits the normal distribution.

***Figure 1.18: Histogram***

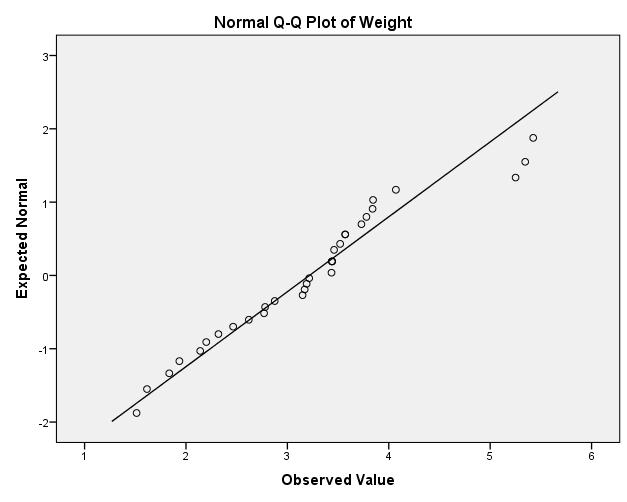


If we consider data till 4 we will see a bell shaped curve indicating that the distribution of the data is normal.

|  |  |
| --- | --- |
| ***Table 1.19: Stem-and-Leaf Plot*** | |
| Frequency | Stem & Leaf |
| 4.00 | 1 . 5689 |
| 4.00 | 2 . 1234 |
| 4.00 | 2 . 6778 |
| 9.00 | 3 . 111244444 |
| 7.00 | 3 . 5557788 |
| 1.00 | 4 . 0 |
| .00 | 4 . |
| 1.00 | 5 . 2 |
| 2.00 | Extremes (>=5.3) |
| Stemwidth | 1.00 |
| Each leaf | 1 case(s) |

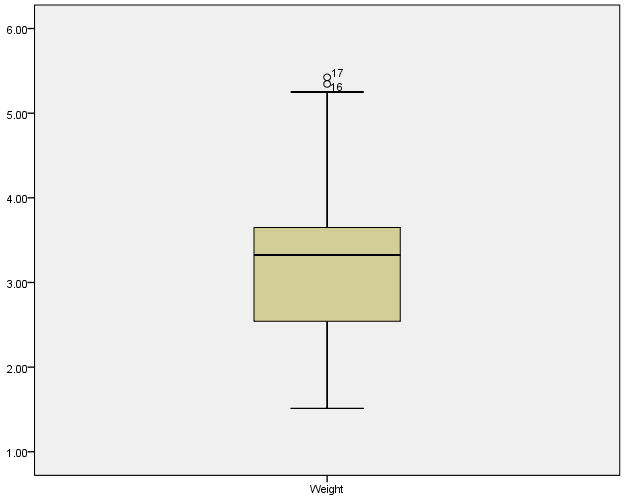
|  |
| --- |
| ***A stem and leaf plot shows all data values .From the table on the left we infer from the first row that the observations 15,16,18,19 appear in the data. The digit with 1 in the tens place occurs 4 times. Every row can be interpreted similarly.*** |

***Figure 1.19: Normal Q-Q Plot***



Except the outliers all the points are more or less lying on the straight line which indicates that the distribution of the data is normal.

***Figure 1.20: Box plot***



The median is around 3.5 and the dots above the horizontal line around 5.2 indicate presence of outliers. As the median is closer to the top we conclude that the data is skewed.

After removing outliers we will use missing value imputation and then conduct EDA.

| ***EDA with weight*** | | | | |
| --- | --- | --- | --- | --- |
| ***Table 1.20:*** | ***Descriptives*** | (Weight) | Statistic | Std. Error |
| SMEAN(wt) | Mean | | 2.99769 | .121945 |
| 95% Confidence Interval for Mean | Lower Bound | 2.74898 |  |
| Upper Bound | 3.24640 |  |
| 5% Trimmed Mean | | 3.02316 |  |
| Median | | 3.16000 |  |
| Variance | | .476 |  |
| Std. Deviation | | .689825 |  |
| Minimum | | 1.513 |  |
| Maximum | | 4.070 |  |
| Range | | 2.557 |  |
| Interquartile Range | | 1.001 |  |
| Skewness | | -.636 | .414 |
| Kurtosis | | -.483 | .809 |

1) Skewness has changed from 0.466 to -0.636 after removing the outlier. It implies negative skewness and the coefficient is insignificant.

2) Negative kurtosis indicates that the observations cluster less and have shorter tails.

| ***Table 1.21:Tests of Normality*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | df | Sig. | Statistic | df | Sig. |
| SMEAN(wt) | .143 | 32 | .094 | .942 | 32 | .084 |
| a. Lilliefors Significance Correction | | | | | | |

We set the following hypothesis:

Ho-The observed distribution fits the normal distribution.

H1-The observed distribution does not fit the normal distribution.

So from the table above we infer that the p value is 0.094(which is >0.05).Hence we may accept the null hypothesis at 5% level of significance and conclude that data fits the normal distribution.

***Figure 1.21: Histogram***



It appears to be normally distributed.

***Figure 1.22: Normal Q-Q Plot***



As the points lie along the straight line, we infer that the data is normally distributed.

***Figure 1.23: Box plot***



Figure indicates absence of outliers.As the central line is nearer to the top of the box it implies data is slightly positively skewed.

***EDA with (1/4)th mile time***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***(1/4)th mile time Table 1.22:Descriptives*** | | | Statistic | Std. Error |
|  | Mean | | 17.8488 | .31589 |
| 95% Confidence Interval for Mean | Lower Bound | 17.2045 |  |
| Upper Bound | 18.4930 |  |
| 5% Trimmed Mean | | 17.8079 |  |
| Median | | 17.7100 |  |
| Variance | | 3.193 |  |
| Std. Deviation | | 1.78694 |  |
| Minimum | | 14.50 |  |
| Maximum | | 22.90 |  |
| Range | | 8.40 |  |
| Interquartile Range | | 2.02 |  |
| Skewness | | .406 | .414 |
| Kurtosis | | .865 | .809 |

As the standard deviation is small the observations are less spread out .As kurtosis is positive, it implies tails are heavier and we will assume that it is more peaker than what the normal distribution will allow.Skewness is positive implying positive skewness.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Table 1.23:Tests of Normality*** | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | Df | Sig. | Statistic | df | Sig. |
| 1/4 th mile time | .073 | 32 | .200\* | .973 | 32 | .594 |
| \*This is a lower bound of the true significance. a.Lilliefors Significance Correction | | | | | | |

We set the following hypothesis:

Ho-The observed distribution fits the normal distribution.

H1-The observed distribution does not fit the normal distribution.

So from the table above we infer that the p value is 0.200(which is >0.05).Hence we may accept the null hypothesis at 5% level of significance and conclude that data fits the normal distribution.

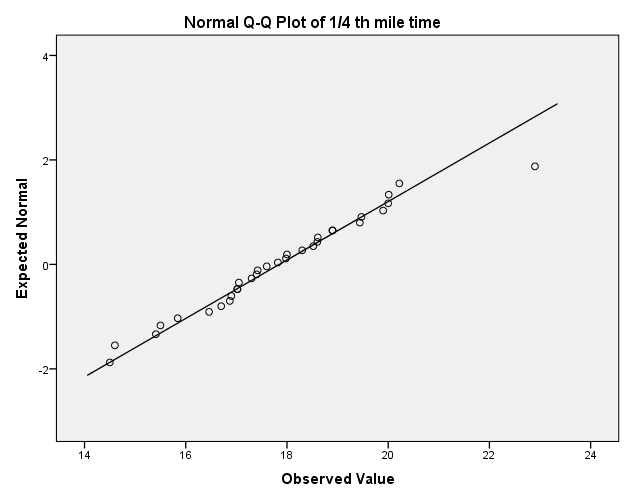
***Figure 1.24: Histogram***



The gap after 20 indicates presence of outlier or that there is no frequency in that interval. Ignoring that, the shape of the curve is bell shaped. So the distribution of the data is normal.

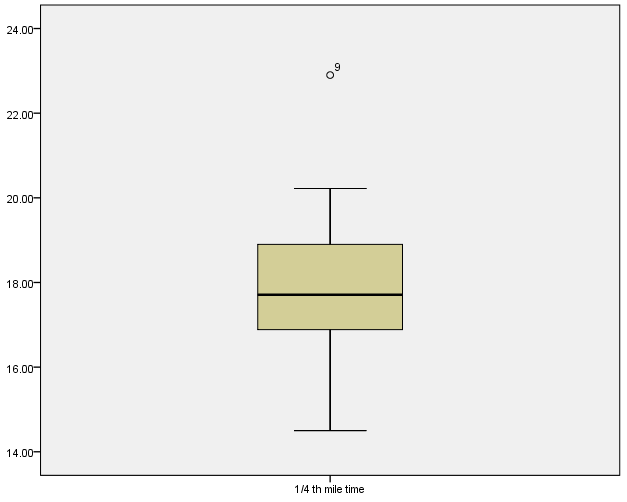
|  |  |
| --- | --- |
| ***Table 1.24:Stem-and-Leaf Plot*** | |
| Frequency | Stem & Leaf |
| 2.00 | 14 . 56 |
| 3.00 | 15 . 458 |
| 4.00 | 16 . 4789 |
| 9.00 | 17 . 000344689 |
| 7.00 | 18 . 0356699 |
| 3.00 | 19 . 449 |
| 3.00 | 20 . 002 |
| 1.00 | Extremes (>=22.9) |
| Stemwidth | 1.00 |
| Each leaf | 1 case(s) |

***Figure 1.25: Normal Q-Q Plot***



Almost all points are lying on or around the straight line so the distribution of the data is normal.

***Figure 1.26: Box plot***



There is an outlier around 23.We delete the outlier and treat it as a missing value. Then we compute the missing value using series mean. Finally now we perform the EDA.

***EDA with (1/4)th mile time***

.

| **Table 1.25** | **:Descriptives** | (qsec) | Statistic | Std. Error |
| --- | --- | --- | --- | --- |
| SMEAN(qsec) | Mean | | 17.6858 | .27062 |
| 95% Confidence Interval for Mean | Lower Bound | 17.1339 |  |
| Upper Bound | 18.2377 |  |
| 5% Trimmed Mean | | 17.7243 |  |
| Median | | 17.6429 |  |
| Variance | | 2.344 |  |
| Std. Deviation | | 1.53086 |  |
| Minimum | | 14.50 |  |
| Maximum | | 20.22 |  |
| Range | | 5.72 |  |
| Interquartile Range | | 1.95 |  |
| Skewness | | -.248 | .414 |
| Kurtosis | | -.400 | .809 |

1) As the standard deviation is small the observations are less spread out.

2) Coefficient of skewness has changed from 0.406 to -0.248 implying negative skewness. Coefficient of skewness is insignificant.

3) As kurtosis is negative, it indicates that the observations cluster less and have shorter tails.

| **Table 1.26: Tests of Normality** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | df | Sig. | Statistic | df | Sig. |
| SMEAN(qsec) | .078 | 32 | .200\* | .971 | 32 | .514 |
| a. Lilliefors Significance Correction | | | | | | |
| \*. This is a lower bound of the true significance. | | | | | | |

We set the following hypothesis:

Ho-The observed distribution fits the normal distribution.

H1-The observed distribution does not fit the normal distribution.

So from the table above we infer that the p value is 0.200(which is >0.05).Hence we may accept the null hypothesis at 5% level of significance and conclude that data fits the normal distribution.

***Figure 1.27: Histogram***



As the figure shows bell shaped curve ,it implies data follows normal distribution.

***Figure 1.28:Normal Q-Q Plot***

  
It indicates that the data is normally distributed.

***Figure 1.29: Box Plot***



It indicates absence of outliers.Data is almost symmetrical.

***EDA with V/S***

| ***Table 1.27: Statistics (v/s)*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 0 | 18 | 56.3 | 56.3 | 56.3 |
| 1 | 14 | 43.8 | 43.8 | 100.0 |
| Total | 32 | 100.0 | 100.0 |  |

The left hand column shows the valid categories and total number of participants. The frequency column gives the number of cars who have the following values. The percent column is the percentage of each value, including missing values. Cumulative percent is the percentage of subjects in a category plus the categories listed above it.56.3% are v engines and 43.8 are straight engines.

***Figure 1.30: Bar chart***



The bar chart presents a bar for each level (0=v engine,1=straight engine) of the nominal variable. So we have a frequency of 18 for v engines and 14 for straight engines.

| ***EDA with am***  ***Table 1.28:Statistics(am)*** | | | |
| --- | --- | --- | --- |
|  |  | Frequency | Percent | | Valid Percent | Cumulative Percent |
| Valid | 0 | 19 | 59.4 | | 59.4 | 59.4 |
| 1 | 13 | 40.6 | | 40.6 | 100.0 |
| Total | 32 | 100.0 | | 100.0 |  |

59.4% of all cars are having automatic transmission and 40.6% have a manual transmission.

***Figure 1.31: Bar chart***



Automatic transmission has a frequency of 19 and manual transmission has a frequency of 13.

| ***EDA with gear***   | ***Table1.29:DescriptiveStatistics(gear)*** | | --- | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | N | Minimum | Maximum | Mean | Std. Deviation | Skewness | |
|  | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error |
| Gear | 32 | 3 | 5 | 3.69 | .738 | .582 | .414 |

1) Smaller value of standard deviation implies that the data is less spread out.

2) As coefficient of skewness >0, it implies that it is positively skewed. Moreover the coefficient of skewness is insignificant as its absolute value is not greater than twice its standard error (0.582<0.828).

***Figure 1.32: Bar chart***

The bar chart shows that the mtcars data has been split into three classes depending on the number of gears.

***Figure 1.33: Box plot***

| It indicates absence of outliers. |
| --- |

***EDA with carb***

| ***Table 1.30:Descriptive Statistics(carb)*** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | N | Minimum | *Maximum* | Mean | Std. Deviation | Skewness | |
|  | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error |
| Carb | 32 | 1 | 8 | 2.81 | 1.615 | 1.157 | .414 |

As coefficient of skewness >0, it implies that it is positively skewed. Moreover the coefficient of skewness is significant as its absolute value is greater than twice its standard error (1.157>0.828).

***Figure 1.34: Box plot***



| It indicates presence of outliers.  We perform EDA again after removing the outlier.  ***EDA with carb***   | ***Table 1.31:Descriptive Statistics(carb)*** | | | | | | | | | --- | --- | --- | --- | --- | --- | --- | --- | |  | N | Minimum | Maximum | Mean | Std. Deviation | Skewness | | |  | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | | SMEAN(carb) | 32 | 1.0 | 6.0 | 2.645 | 1.3087 | .444 | .414 |   As coefficient of skewness >0, it implies that it is positively skewed. Moreover the coefficient of skewness is insignificant as its absolute value is not greater than twice its standard error (0.444<0.828).  ***Figure 1.35: Bar chart*** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

The bar chart shows that the mtcars data has been split into six classes depending on the number of carburetors.

***Figure 1.36: Box plot***

It indicates absence of outliers.

***Figure1.37: Scatterplot matrix***



Mpg and disp show high negative correlation pattern. Mpg and drat show moderately positive correlation pattern. Mpg and hp show high negative correlation pattern. Mpg and wt show high negative correlation pattern. Mpg and qsec show low positive correlation pattern. Disp and drat show high negative correlation pattern. Hp and disp show high positive correlation pattern.Disp and wt show moderately positive correlation pattern. Qsec and disp show low negative correlation pattern. Drat and hp show moderately negative correlation pattern. Drat and wt show moderately negative correlation pattern. Drat and qsec show low positive correlation pattern. Hp and wt show moderately positive correlation pattern. Hp and qsec show moderately negative correlation pattern. Wt and qsec show low negative correlation.

***Correlations-We set the following Hypothesis for the correlation part.***

|  |
| --- |
| ***Ho: r=0 ie. Correlation is not significant***  ***H1: Correlation is significant*** |

| ***Table No.1.32*** | ***Correlations*** | mpg | disp | drat | MEAN(hp,1) | SMEAN(wt) | SMEAN(qsec) |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Mpg | Pearson Correlation | 1 | -.848\*\* | .681\*\* | -.808\*\* | -.781\*\* | ***.439\**** |
| Sig. (2-tailed) |  | .000 | .000 | .000 | .000 | ***.012*** |
| Disp | Pearson Correlation | -.848\*\* | 1 | -.710\*\* | .855\*\* | .659\*\* | ***-.427\**** |
| Sig. (2-tailed) | .000 |  | .000 | .000 | .000 | ***.015*** |
| Drat | Pearson Correlation | .681\*\* | -.710\*\* | 1 | -.507\*\* | -.677\*\* | .040 |
| Sig. (2-tailed) | .000 | .000 |  | .003 | .000 | .828 |
| MEAN(hp,1) | Pearson Correlation | -.808\*\* | .855\*\* | -.507\*\* | 1 | .604\*\* | -.648\*\* |
| Sig. (2-tailed) | .000 | .000 | .003 |  | .000 | .000 |
| SMEAN(wt) | Pearson Correlation | -.781\*\* | .659\*\* | -.677\*\* | .604\*\* | 1 | -.290 |
| Sig. (2-tailed) | .000 | .000 | .000 | .000 |  | .108 |
| SMEAN(qsec) | Pearson Correlation | .439\* | -.427\* | .040 | -.648\*\* | -.290 | 1 |
| Sig. (2-tailed) | .012 | .015 | .828 | .000 | .108 |  |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | | | | |
| \*. Correlation is significant at the 0.05 level (2-tailed). | | | | | | | |

From the above table we infer that for the underlined values we reject the null hypothesis at 5% level of significance and conclude that correlation is significant.

***Table No1.33: Correlations***

| Mpg | Cyl | disp | Drat | vs | am | gear | MEAN(hp,1) | SMEAN(wt) | SMEAN(qsec) | SMEAN(carb) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1.000 | -.911\*\* | -.909\*\* | .651\*\* | .707\*\* | .562\*\* | .543\*\* | -.881\*\* | -.698\*\* | ***.439\**** | -.635\*\* |
| . | .000 | .000 | .000 | .000 | .001 | .001 | .000 | .000 | ***.012*** | .000 |
| -.911\*\* | 1.000 | .928\*\* | -.679\*\* | -.814\*\* | -.522\*\* | -.564\*\* | .898\*\* | .762\*\* | -.535\*\* | .559\*\* |
| .000 | . | .000 | .000 | .000 | .002 | .001 | .000 | .000 | .002 | .001 |
| -.909\*\* | .928\*\* | 1.000 | -.684\*\* | -.724\*\* | -.624\*\* | -.594\*\* | .865\*\* | .706\*\* | ***-.444\**** | .537\*\* |
| .000 | .000 | . | .000 | .000 | .000 | .000 | .000 | .000 | ***.011*** | .002 |
| .651\*\* | -.679\*\* | -.684\*\* | 1.000 | ***.447\**** | .687\*\* | .745\*\* | -.539\*\* | -.633\*\* | .058 | -.122 |
| .000 | .000 | .000 | . | ***.010*** | .000 | .000 | .001 | .000 | .754 | .505 |
| .707\*\* | -.814\*\* | -.724\*\* | .447\* | 1.000 | .168 | .283 | -.752\*\* | -.464\*\* | .771\*\* | -.620\*\* |
| .000 | .000 | .000 | .010 | . | .357 | .117 | .000 | .007 | .000 | .000 |
| .562\*\* | -.522\*\* | -.624\*\* | .687\*\* | .168 | 1.000 | .808\*\* | ***-.445\**** | -.697\*\* | -.162 | -.136 |
| .001 | .002 | .000 | .000 | .357 | . | .000 | ***.011*** | .000 | .376 | .458 |
| .543\*\* | -.564\*\* | -.594\*\* | .745\*\* | .283 | .808\*\* | 1.000 | ***-.437\**** | -.559\*\* | -.181 | .028 |
| .001 | .001 | .000 | .000 | .117 | .000 | . | ***.012*** | .001 | .323 | .880 |
| -.881\*\* | .898\*\* | .865\*\* | -.539\*\* | -.752\*\* | -.445\* | -.437\* | 1.000 | .621\*\* | -.598\*\* | .704\*\* |
| .000 | .000 | .000 | .001 | .000 | .011 | .012 | . | .000 | .000 | .000 |
| -.698\*\* | .762\*\* | .706\*\* | -.633\*\* | -.464\*\* | -.697\*\* | -.559\*\* | .621\*\* | 1.000 | -.258 | .312 |
| .000 | .000 | .000 | .000 | .007 | .000 | .001 | .000 | . | .154 | .082 |
| .439\* | -.535\*\* | -.444\* | .058 | .771\*\* | -.162 | -.181 | -.598\*\* | -.258 | 1.000 | -.602\*\* |
| .012 | .002 | .011 | .754 | .000 | .376 | .323 | .000 | .154 | . | .000 |
| -.635\*\* | .559\*\* | .537\*\* | -.122 | -.620\*\* | -.136 | .028 | .704\*\* | .312 | -.602\*\* | 1.000 |
| .000 | .001 | .002 | .505 | .000 | .458 | .880 | .000 | .082 | .000 | . |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | | | | | | | |
| \*. Correlation is significant at the 0.05 level (2-tailed). | | | | | | | | | | |

From the above table we infer that for the underlined values we reject the null hypothesis at 5% level of significance and conclude that correlation is significant.

***Next course of action after EDA***-We will build a linear regression model of mileage on the rest of the attributes of the car and calculate the significance of overall regression, R2 and adjusted R2

Perform multicollinearity diagnostics and removal using correlation matrix and variance inflation factors. We will build models which are free from multicollinearity and calculate the significance of individual variables,significance of overall regression, R2 and adjusted R2.

***Appendix***

| ***Table No.1.34: Partial Correlations*** | | | | |
| --- | --- | --- | --- | --- |
| Control Variables | | | SMEAN(wt) | Disp |
| cyl & vs & am & gear & SMEAN(carb) & drat & MEAN(hp,1) & SMEAN(qsec) & mpg | SMEAN(wt) | Correlation | 1.000 | -.479 |
| Significance (2-tailed) | . | .021 |
| df | 0 | 21 |
| disp | Correlation | -.479 | 1.000 |
| Significance (2-tailed) | .021 | . |
| df | 21 | 0 |

We notice that the partial correlation between weight and displacement is -0.479 which is considerably less than the correlation when none of the other variable effect was controlled for(r=0.659).In fact the correlation coefficient has reduced by 0.18 units. Although this correlation is still statistically significant (its p value is still below 0.05), the relationship has diminished.

***Reference***

Fundamentals of mathematical statistics by S.C .Gupta and V.K. Kapoor

SPSS for intermediate statistics by Nancy L. Leech, Karen C. Barrett, George A. Morgan

Additional notes given by Abhishek K. Umrawal Sir

Google was also referred for some interpretations.