Case Study – Exploratory Data Analysis

Submitted by: Saikat Ghosh

Due Date: 24th August

Date of Submission: 24th August

Supervisor's Remarks

Late Submission:

Plagiarism:

Completeness:

Quality of Content:

Results and Interpretations:

Additional Remarks:

EXPLORATORY DATA ANALYSIS (EDA)

In statistics**, exploratory data analysis (EDA)** is an approach for analysing data sets to summarize their main characteristics, often with visual methods.

Exploratory data analysis was promoted by **John Tukey** to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.

The particular graphical techniques employed in EDA are quite simple, consisting of various techniques of

- ✓ Plotting the raw data with the help of histograms, bar charts, probability plots to get the frequency distribution
- ✓ Plotting simple statistics such as mean plots, box plots, and main effects plots of the raw data to detect outliers and anomalies
- ✓ Testing the distribution of the data so that validity of the underlying assumptions can be checked

Dataset:

In order to perform EDA, We use "mtcars" dataset from R

Description:

The *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).data was extracted from the 1974

A data frame with 32 observations on 11 variables.

mpg	**Miles/(US) gallon**
cyl	**Number of cylinders**
disp	**Displacement (cu.in.)**
hp	**Gross horsepower**
drat	**Rear axle ratio**
wt	**Weight (lb/1000)**
qsec	**1/4 mile time**
vs	**V/S**
am	**Transmission (0 = automatic, 1 = manual)**
gear	**Number of forward gears**
carb	**Number of carburetors**

Source:

Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411.

1. EDA for Individual Variables

a) For Continuous Variables

In the given data set we have the following continuous variables:

1. mpg – Miles/(US) gallon
2. disp – Displacement
3. hp – Gross Horsepower
4. drat – Rear axle ratio
5. wt – Weight (lb/1000)
6. qsec – ¼ mile time

For the EDA of Continuous Variables, we will use the following measures/tools:

- Descriptive Statistics

Like Mean, Median, Mode etc. to get an insight about the data
• Coeffecient of Skewness > 0: +vely skewed or right skewed,
• Coeffecient of Skewness < 0: -vely skewed or left skewed, and
• Coeffecient of Skewness = 0: symmetric.
The ratio of kurtosis to its standard error can be used as a test of normality (that is, you can reject normality if the ratio is less than -2 or greater than +2).

- Histogram (Overlaid with normal probability curve)

To know about the distribution of data and compare its proximity with the normal distribution

- Q-Q Plot, KS Test and Shapiro Wilks Test
- To test whether the data is Normally distributed or not with hypotheses

H_0 : Sample comes from a normal population

H_1 : Sample does not comes from a normal population

- Box Plot

To know if there are any outliers in the data

- Stem and Leaf Plot

A stem-and-leaf plot is a device for presenting quantitative data in a graphical format, similar to a histogram, to assist in visualizing the shape of a distribution. The stem-and-leaf display is drawn with two columns (usually separated by a vertical line or '.'). The stems are listed to the left of the vertical line. It is important that no numbers are skipped, even if it means that some stems have no leaves. The leaves are listed in increasing order in a row to the right of each stem.

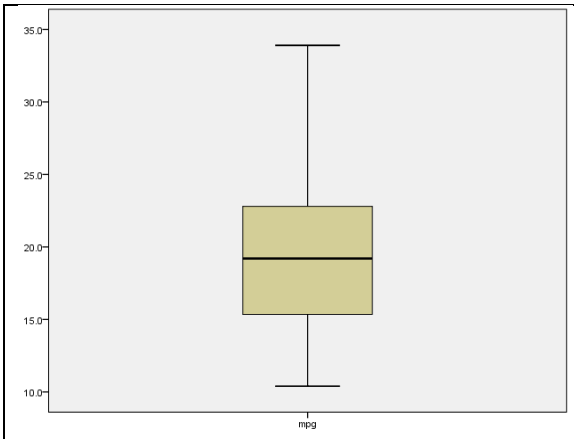
1. mpg – Miles/(US) gallon

Descriptives

			Statistic	Std. Error
mpg	Mean		20.091	1.0654
	95% Confidence Interval for Mean	Lower Bound	17.918	
		Upper Bound	22.264	
	5% Trimmed Mean		19.893	
	Median		19.200	
	Variance		36.324	
	Minimum		6.0269	
			10.4	
	Std. Deviation		33.9	
	Minimum		23.5	
	Interquartile Range		7.5	
	Skewness		.672	.414
	Kurtosis		-.022	.809

- Skewness (.672) > 0, Distribution is positively skewed.
- The ratio of kurtosis to its standard error = -0.022/0.809 = -0.0272 > -2 i,e we accept normality.

Hence the distribution follows normal distribution.



From the box plot we can observe that mpg has no outliers.

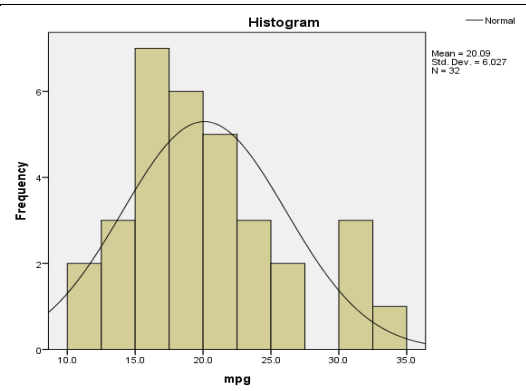
mpg Stem-and-Leaf Plot

Frequency Stem & Leaf

5.00 1 . 00344
13.00 1 . 5555567788999
8.00 2 . 11111224
2.00 2 . 67
4.00 3 . 0023

Stem width: 10.0
Each leaf: 1 case(s)

From it we can observe frequency and also we know about the shape of the Histogram.
We can also guess about the data that it follows normal distribution or not.



From the normal curve on the histogram we can conclude that it almost follow Normal Distribution.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
mpg	.126	32	.200*	.948	32	.123

*. This is a lower bound of the true significance.

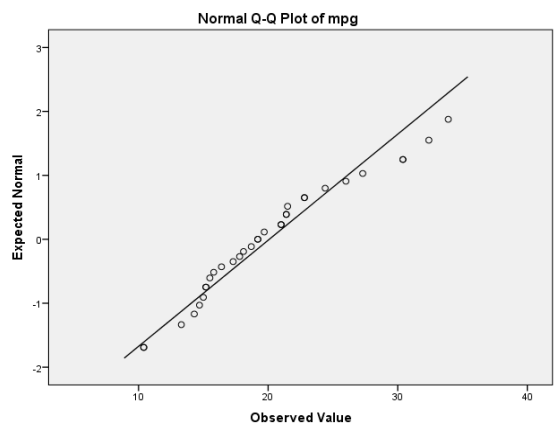
a. Lilliefors Significance Correction

H_0 : The distribution is normal
H_1 : The distribution is not normal

Conclusion:

From **Kolmogorov-Smirnov** Test with Sig. = 0.20 (>0.05), we may conclude that at 5% l.o.s. H_0 is accepted i.e., "mpg" is normally distributed.

From **Shapiro-Wilk** Test with Sig. = 0.123 (>0.05), we may conclude that at 5% l.o.s. H_0 is accepted i.e., "mpg" is normally distributed



Inference: From the QQ Plot, we can observe that "mpg" is almost normally distributed.

2.DISP (Displacement)

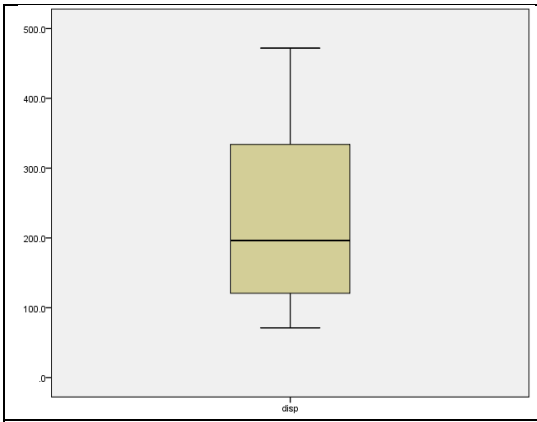
Descriptives

			Statistic	Std. Error
disp	Mean		230.722	21.9095
	95% Confidence Interval for Mean	Lower Bound	186.037	
		Upper Bound	275.407	
	5% Trimmed Mean		226.340	
	Median		196.300	
	Variance		15360.800	
	Std. Deviation		123.9387	
	Minimum		71.1	
	Maximum		472.0	
	Range		400.9	
	Interquartile Range		221.5	
	Skewness		.420	.414
	Kurtosis		-1.068	.809

Skewness (.420) > 0, Distribution is positively skewed.

The ratio of kurtosis to its standard error = -1.068/.809 = -1.3201 > -2 i,e we do not reject normality.

Hence the distribution follows normal distribution.



From Box Plot we can see that Disp has no Outliers

disp Stem-and-Leaf Plot

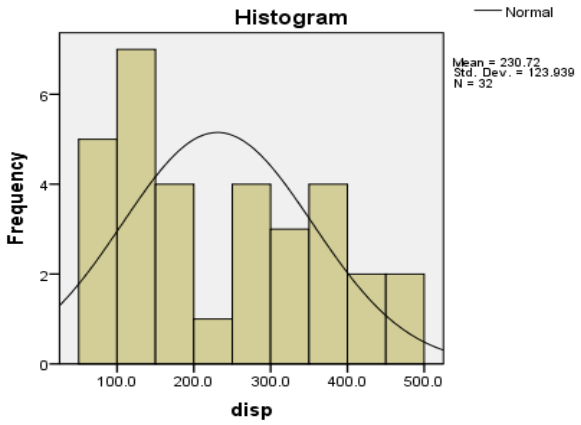
Frequency Stem & Leaf

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 4 . 04
2.00 4 . 67

Stem width: 100.0

From it we can observe frequency and also we know about the shape of the Histogram.
We can also guess about the data that it follows normal distribution or not.

Each leaf: 1 case(s)



Inference: From the normal curve on histogram it can be observed that "disp" is not normally distributed.

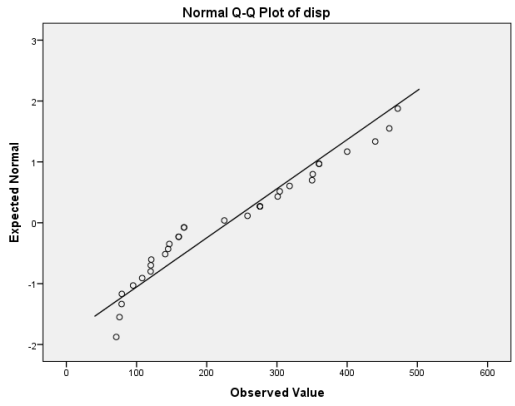
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
disp	.195	32	.003	.920	32	.021

a. Lilliefors Significance Correction

Conclusion

- From **Kolmogorov-Smirnov** Test with Sig. = 0.003 (<0.05), we may conclude that at 5% l.o.s. H_0 is rejected i.e., "disp" is not normally distributed.
- From **Shapiro-Wilk** Test with Sig. = 0.021 (<0.05), we may conclude that at 5% l.o.s. H_0 is rejected i.e., "disp" is not normally distributed.



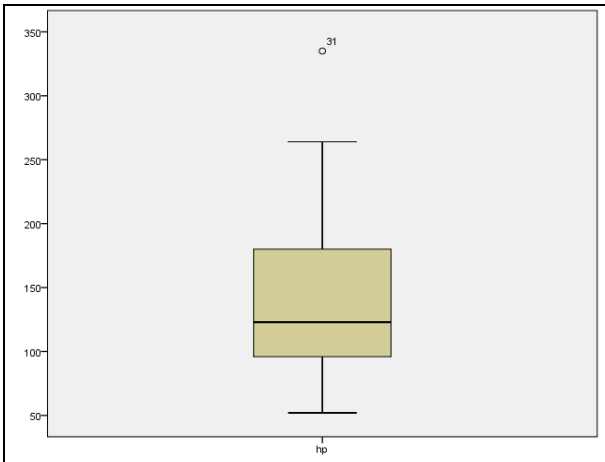
Inference: From the QQ Plot, it can be observed that "disp" is not normally distributed.

3.HP (Gross Horsepower)

Descriptives

			Statistic	Std. Error
hp	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



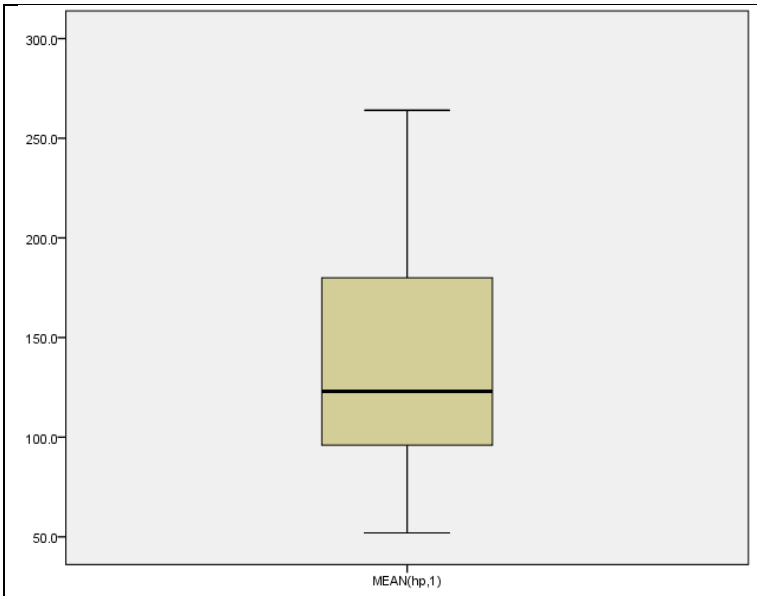
From the Box Plot we can see that 31st observation is outlier.
In order to remove the outliers from the data, we are replacing that observation (which is behaving like an outlier) with **mean of any two nearby points.** Missing value technique helps in replacement of such values.

After removing the outlier

Descriptives

			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

- We see that, **Coefficient of Skewness** i.e(.460) > 0 therefore it is Positively(or right) Skewed
- We see that, **Ratio of Kurtosis** to its Standard Error is (-0.9258) > -2 therefore the Normality is accepted



Now outlier is removed

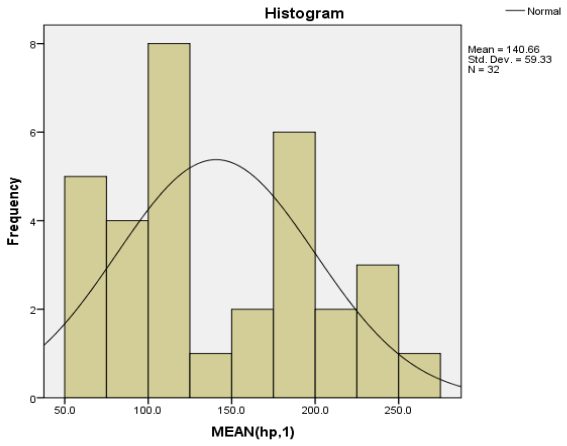
MEAN(hp,1) Stem-and-Leaf Plot

Frequency Stem & Leaf

9.00 0 . 566669999
9.00 1 . 001111224
8.00 1 . 55777888
5.00 2 . 01344
1.00 2 . 6

Stem width: 100.0
Each leaf: 1 case(s)

From it we can observe frequency and also we know about the shape of the Histogram.
We can also guess about the data that it follows normal distribution or not.



Inference: From the normal curve on histogram it can be observed that "hp" is not normally distributed.

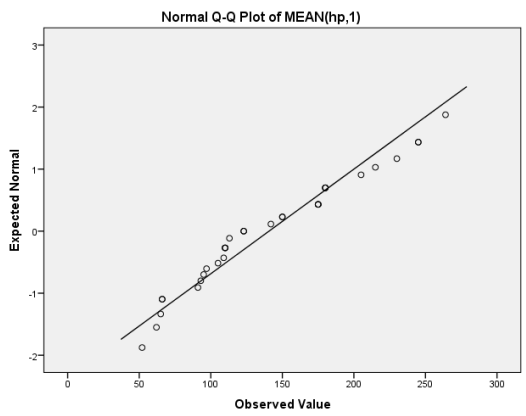
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
MEAN(hp,1)	.148	32	.072	.944	32	.096

a. Lilliefors Significance Correction

Conclusion

- From **Kolmogorov-Smirnov** Test with Sig. = 0.024 (<0.05), we may conclude that at 5% l.o.s. H_0 is rejected i.e., "hp" is not normally distributed.
- From **Shapiro-Wilk** Test with Sig. = 0.049 (<0.05), we may conclude that at 5% l.o.s. H_0 is rejected i.e., "hp" is not normally distributed.



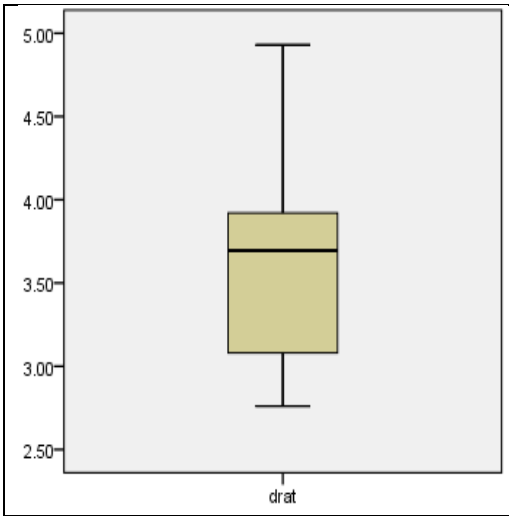
- Inference: From the QQ Plot, it can be observed that "hp" is not normally distributed.

4.DRAT (Rear Axle Ratio)

Descriptives

			Statistic	Std. Error
drat	Mean		3.5966	.09452
	95% Confidence Interval for Mean	Lower Bound	3.4038	
		Upper Bound	3.7893	
	5% Trimmed Mean		3.5794	
	Median		3.6950	
	Variance		.286	
	Std. Deviation		.53468	
	Minimum		2.76	
	Maximum		4.93	
	Range		2.17	
	Interquartile Range		.84	
	Skewness		.293	.414
	Kurtosis		-.450	.809

- We see that, **Coefficient of Skewness** i.e(.293) > 0 therefore it is Positively(or right) Skewed
- We see that, **Ratio of Kurtosis** to its Standard Error is (-0.5562) < -2 therefore the Normality is rejected



Inference: From the Boxplot, it can be observed that "drat" has no outliers.

drat Stem-and-Leaf Plot

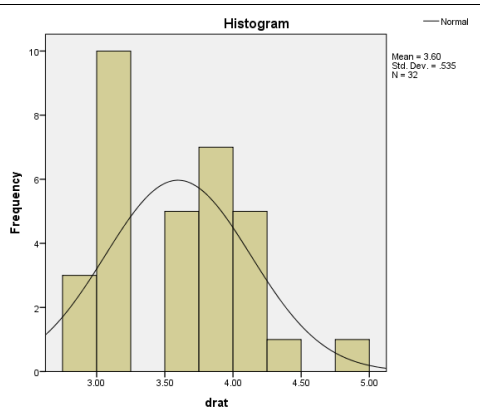
From it we can observe frequency and also we know about the shape of the Histogram.

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: 1.00
Each leaf: 1 case(s)

We can also guess about the data that it follows normal distribution or not.



Inference: From the normal curve on histogram it can be observed that "drat" is not normally distributed.

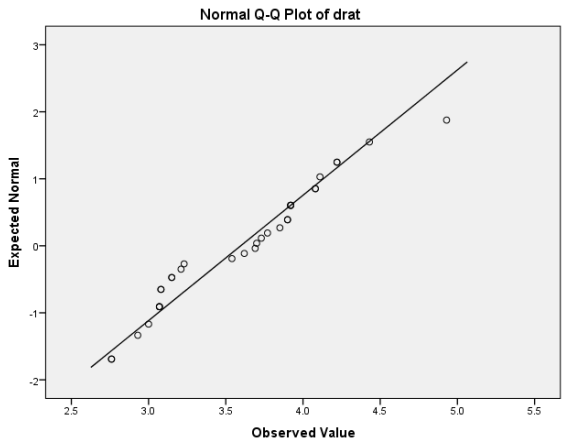
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
drat	.160	32	.037	.946	32	.110

a. Lilliefors Significance Correction

Conclusion

- From **Kolmogorov-Smirnov** Test with Sig. = 0.037 (<0.05), we may conclude that at 5% l.o.s. H_0 is rejected i.e., "drat" is not normally distributed.
- From **Shapiro-Wilk** Test with Sig. = 0.110 (>0.05), we may conclude that at 5% l.o.s. H_0 is accepted i.e., "drat" is normally distributed.

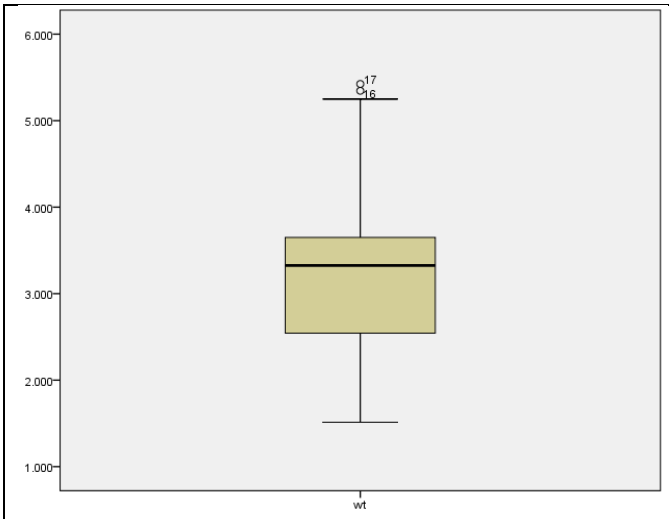


Inference: From the QQ Plot, it can be observed that "drat" is almost normally distributed.

5.WT : Weight (lb/1000)

Descriptives

			Statistic	Std. Error
wt	Mean		3.21725	.172968
	95% Confidence Interval for Mean	Lower Bound	2.86448	
		Upper Bound	3.57002	
	5% Trimmed Mean		3.18885	
	Median		3.32500	
	Variance		.957	
	Std. Deviation		.978457	
	Minimum		1.513	
	Maximum		5.424	
	Range		3.911	
	Interquartile Range		1.186	
	Skewness		.466	.414
	Kurtosis		.417	.809



From Box Plot WE can observe that "drat" has two outliers,that is 16th or 17th observations.

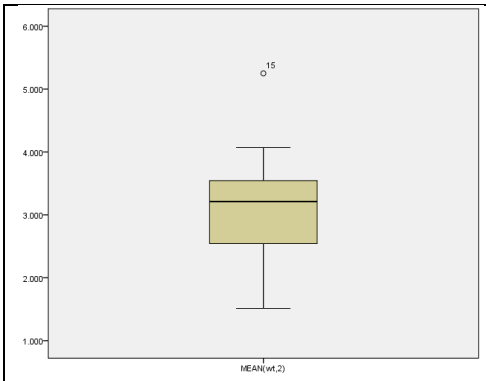
After removing outliers

Descriptives

			Statistic	Std. Error
MEAN(wt,2)	Mean		3.08142	.140891
	95% Confidence Interval for Mean	Lower Bound	2.79407	
		Upper Bound	3.36877	
	5% Trimmed Mean		3.07054	
	Median		3.21125	
	Variance		.635	
	Std. Deviation		.796998	
	Minimum		1.513	
	Maximum		5.250	

- We see that, **Coefficient of Skewness** i.e(.028) > 0 therefore it is (or right) Skewed
- We see that, **Ratio of Kurtosis** to its Standard Error is (0.8603) > -2 therefore the Normality is accepted

Range	3.737	
Interquartile Range	1.054	
Skewness	.028	.414
Kurtosis	-.679	.809



Here after removing the outliers, we plotted "Box Plot" again and we found outliers. We did not remove it because removing this, we will lose information.

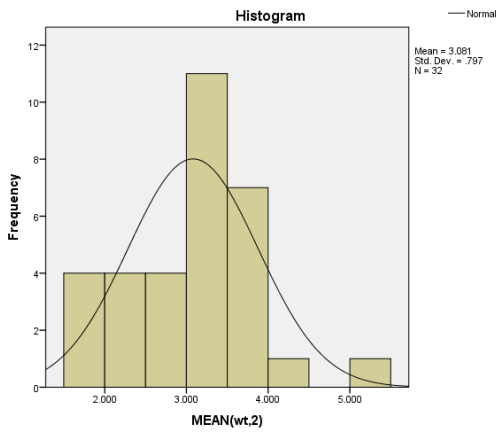
MEAN(wt,2) Stem-and-Leaf Plot

Frequency Stem & Leaf

4.00 1 . 5689
4.00 2 . 1234
4.00 2 . 6778
11.00 3 . 11122244444
7.00 3 . 5557788
1.00 4 . 0
1.00 Extremes (>=5.3)

Stem width: 1.000
Each leaf: 1 case(s)

From it we can observe frequency and also we know about the shape of the Histogram.
We can also guess about the data that it follows normal distribution or not.



Inference: From the normal curve on histogram it can be observed that "wt" is normally distributed.

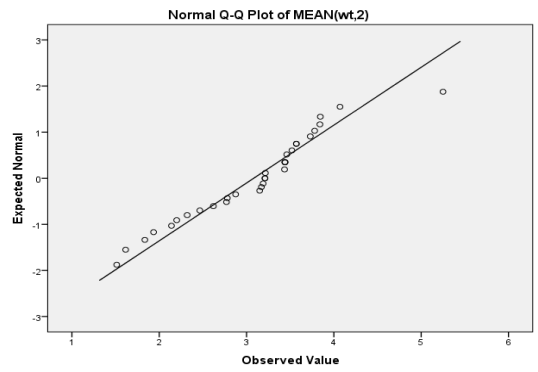
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
MEAN(wt,2)	.159	32	.038	.953	32	.172

a. Lilliefors Significance Correction

Conclusion

- From **Kolmogorov-Smirnov** Test with Sig. = 0.142 (>0.05), we may conclude that at 5% l.o.s. H_0 is accepted i.e., "wt" is normally distributed.
- From **Shapiro-Wilk** Test with Sig. = 0.093 (>0.05), we may conclude that at 5% l.o.s. H_0 is accepted i.e., "wt" is normally distributed.

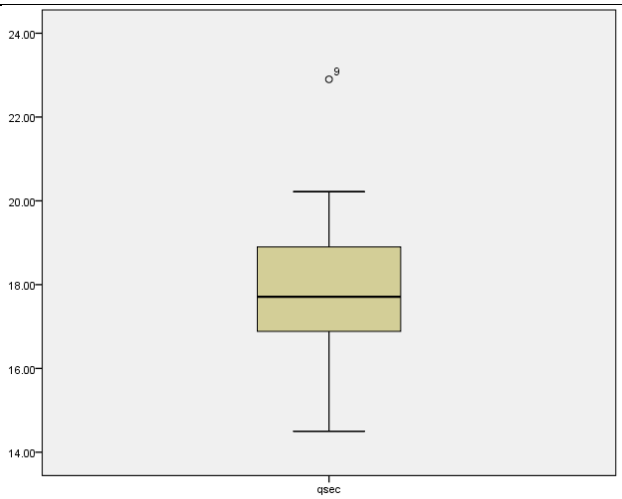


Inference: From the QQ Plot, it can be observed that "wt" is almost normally distributed.

6.QSEC (1/4 Mile Time)

Descriptives

			Statistic	Std. Error
qsec	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



9^{th} observation is outlier

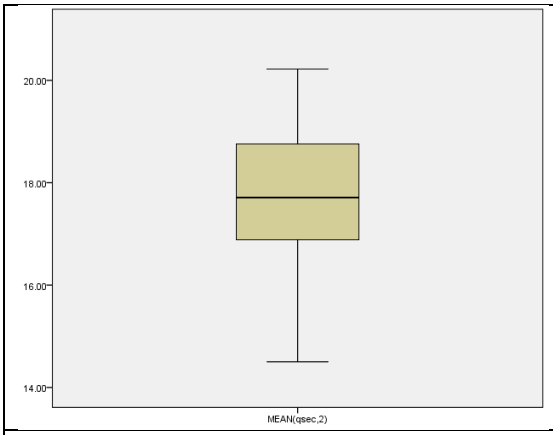
After Removing Outliers

Descriptives

		Statistic	Std. Error
MEAN(qsec,2)	Mean	17.7038	.27122
	95% Confidence Interval for Mean — Lower Bound	17.1506	
	95% Confidence Interval for Mean — Upper Bound	18.2569	
	5% Trimmed Mean	17.7442	
	Median	17.7100	
	Variance	2.354	
	Std. Deviation	1.53423	
	Minimum	14.50	
	Maximum	20.22	
	Range	5.72	
	Interquartile Range	1.95	
	Skewness	-.282	.414
	Kurtosis	-.411	.809

We see that, **Coefficient of Skewness** i.e (-.282) < 0 therefore it is Negatively(or left) Skewed

We see that, **Ratio of Kurtosis** to its Standard Error is (-0.5080) > -2 therefore the Normality is accepted



Now outlier is removed.

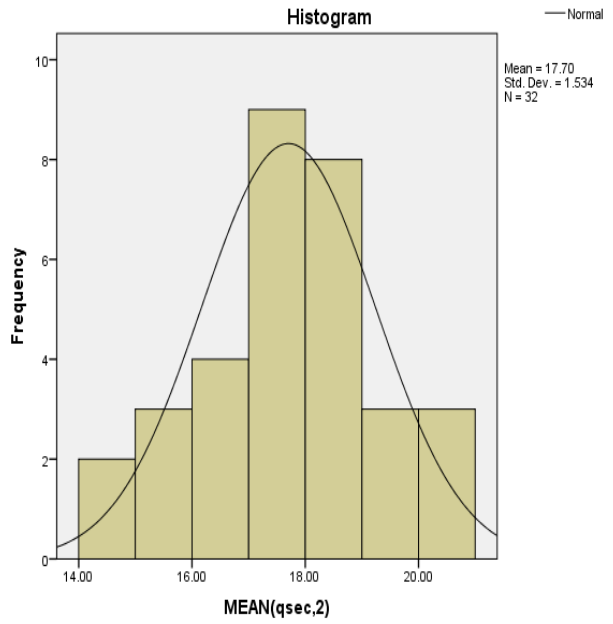
MEAN(qsec,2) Stem-and-Leaf Plot

Frequency Stem & Leaf

2.00 14 . 56
3.00 15 . 458
4.00 16 . 4789
9.00 17 . 000344689
8.00 18 . 02356699
3.00 19 . 449
3.00 20 . 002

Stem width: 1.00
Each leaf: 1 case(s)

From it we can observe frequency and also we know about the shape of the Histogram.
We can also guess about the data that it follows normal distribution or not.



Inference: From the histogram it can be observed that "qsec" is almost normally distributed.

Tests of Normality

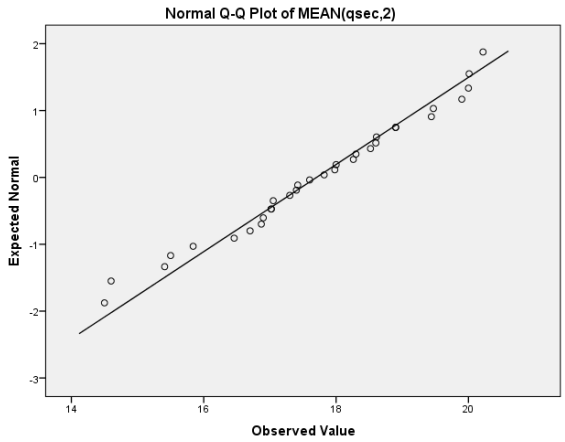
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
MEAN(qsec,2)	.075	32	.200*	.970	32	.509

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Conclusion

- From **Kolmogorov-Smirnov** Test with Sig. = 0.2 (>0.05), we may conclude that at 5% l.o.s. H_0 is accepted i.e., "qsec" is normally distributed.
- From **Shapiro-Wilk** Test with Sig. = 0.594 (>0.05), we may conclude that at 5% l.o.s. H_0 is accepted i.e., "qsec" is normally distributed.



Inference: From the QQ Plot, it can be observed that "qsec" is almost normally distributed.

b) For Discrete / Categorical Variables

In the given data set we have the following discrete variables:

1. cyl – Number of cylinders
2. vs – V/S
3. am – Transmission (0=automatic, 1=manual)
4. gear – Number of forward gears
5. carb – Number of carburetors

For the EDA of Discrete Variables, we will use the following measures/tools:

- Frequency Table

To get the frequency of each data point

- Descriptive Statistics

Like Mean, Median, Mode etc. to get an insight about the data

• Coeffecient of Skewness > 0: +vely skewed or right skewed,
• Coeffecient of Skewness < 0: -vely skewed or left skewed, and
• Coeffecient of Skewness = 0: symmetric.

- Bar plot

To represent the frequency distribution of data

- Stem and Leaf Plot

A stem-and-leaf plot is a device for presenting quantitative data in a graphical format, similar to a histogram, to assist in visualizing the shape of a distribution. The stem-and-leaf display is drawn with two columns (usually separated by a vertical line or '.'). The stems are listed to the left of the vertical line. It is important that no numbers are skipped, even if it means that some stems have no leaves. The leaves are listed in increasing order in a row to the right of each stem.

1. CYL (Number of Cylinders)

cyl

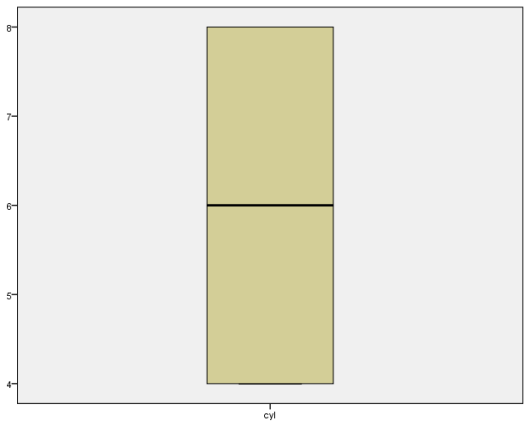
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	4	11	34.4	34.4	34.4
	6	7	21.9	21.9	56.3
	8	14	43.8	43.8	100.0
	Total	32	100.0	100.0	

Descriptives

			Statistic	Std. Error
cyl	Mean		6.19	.316
	95% Confidence Interval for Mean	Lower Bound	5.54	
		Upper Bound	6.83	
	5% Trimmed Mean		6.21	
	Median		6.00	
	Variance		3.190	
	Std. Deviation		1.786	
	Minimum		4	
	Maximum		8	
	Range		4	
	Interquartile Range		4	
	Skewness		-.192	.414
	Kurtosis		-1.763	.809

From the table we can observe various types of central tendency of the data.

Box Plot:

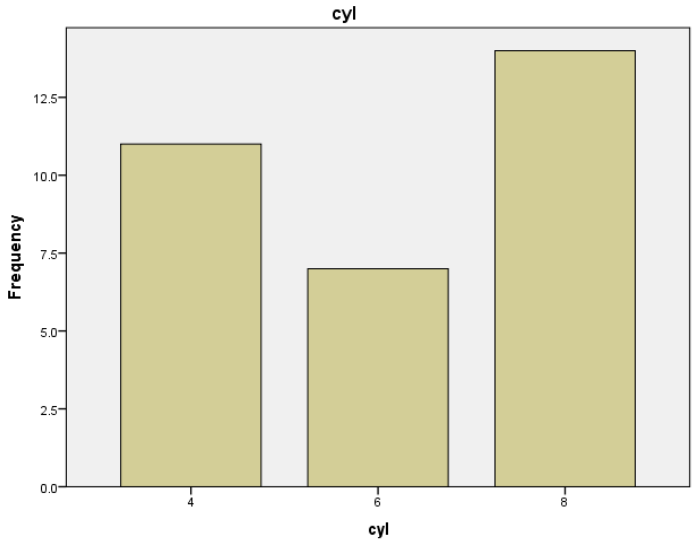


cyl Stem-and-Leaf Plot

Frequency Stem & Leaf

11.00 4 . 00000000000
.00 4 .
.00 5 .
.00 5 .
7.00 6 . 0000000
.00 6 .
.00 7 .
.00 7 .
14.00 8 . 00000000000000

Stem width: 1
Each leaf: 1 case(s)



2.VS (V/S)

vs

		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	

Descriptives

			Statistic	Std. Error
vs	Mean		.44	.089
	95% Confidence Interval for Mean	Lower Bound	.26	
		Upper Bound	.62	
	5% Trimmed Mean		.43	
	Median		.00	
	Variance		.254	
	Std. Deviation		.504	
	Minimum		0	
	Maximum		1	
	Range		1	
	Interquartile Range		1	
	Skewness		.265	.414
	Kurtosis		-2.063	.809

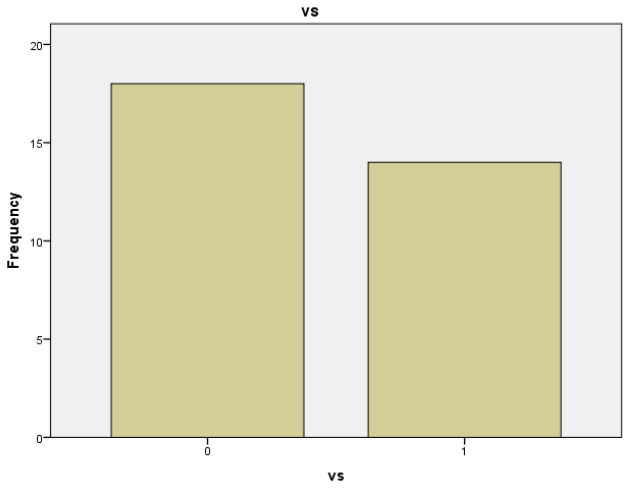
From the table we can observe various types of central tendency of the data

vs Stem-and-Leaf Plot

Frequency Stem & Leaf

18.00 0 . 000000000000000000
.00 0 .
.00 0 .
.00 0 .
.00 0 .
14.00 1 . 00000000000000

Stem width: 1
Each leaf: 1 case(s)



3.AM (Transmission (0=automatic, 1=manual)

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	

Descriptives

			Statistic	Std. Error
am	Mean		.41	.088
	95% Confidence Interval for Mean	Lower Bound	.23	
		Upper Bound	.59	
	5% Trimmed Mean		.40	
	Median		.00	
	Variance		.249	
	Std. Deviation		.499	
	Minimum		0	
	Maximum		1	
	Range		1	
	Interquartile Range		1	
	Skewness		.401	.414

From the table we can observe various types of central tendency of the data

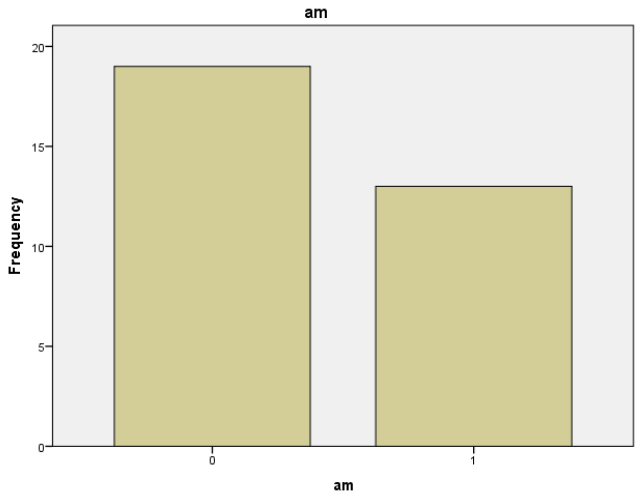
	Kurtosis	-1.967	.809

am Stem-and-Leaf Plot

Frequency Stem & Leaf

19.00 0 . 0000000000000000000
.00 0 .
.00 0 .
.00 0 .
.00 0 .
13.00 1 . 0000000000000

Stem width: 1
Each leaf: 1 case(s)



4.GEAR (Number of forward gears

gear

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	15	46.9	46.9	46.9
	4	12	37.5	37.5	84.4
	5	5	15.6	15.6	100.0
	Total	32	100.0	100.0	

Descriptives

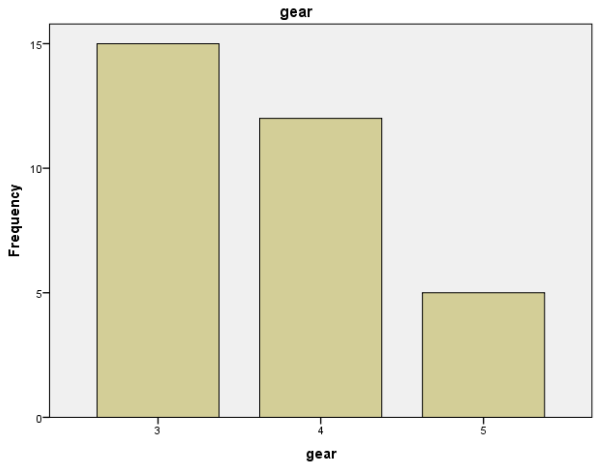
			Statistic	Std. Error
gear	Mean		3.69	.130
	95% Confidence Interval for Mean	Lower Bound	3.42	
		Upper Bound	3.95	
	5% Trimmed Mean		3.65	
	Median		4.00	
	Variance		.544	
	Std. Deviation		.738	
	Minimum		3	
	Maximum		5	
	Range		2	
	Interquartile Range		1	
	Skewness		.582	.414
	Kurtosis		-.895	.809

From the table we can observe various types of central tendency of the data

gear Stem-and-Leaf Plot

Frequency	Stem &	Leaf
15.00	3 .	000000000000000
.00	3 .	
12.00	4 .	000000000000
.00	4 .	
5.00	5 .	00000

Stem width: 1
Each leaf: 1 case(s)



5.CARB (Number of carburettors)

carb

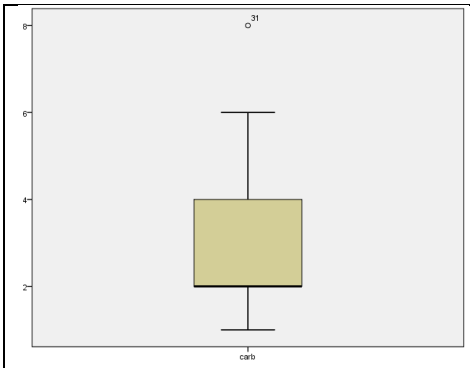
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	21.9	21.9	21.9
	2	10	31.3	31.3	53.1
	3	3	9.4	9.4	62.5
	4	10	31.3	31.3	93.8
	6	1	3.1	3.1	96.9
	8	1	3.1	3.1	100.0
	Total	32	100.0	100.0	

Descriptives

			Statistic	Std. Error
carb	Mean		2.81	.286
	95% Confidence Interval for Mean	Lower Bound	2.23	
		Upper Bound	3.39	
	5% Trimmed Mean		2.67	
	Median		2.00	
	Variance		2.609	
	Std. Deviation		1.615	
	Minimum		1	

From the table we can observe various types of central tendency of the data

Maximum	8	
Range	7	
Interquartile Range	2	
Skewness	1.157	.414
Kurtosis	2.020	.809



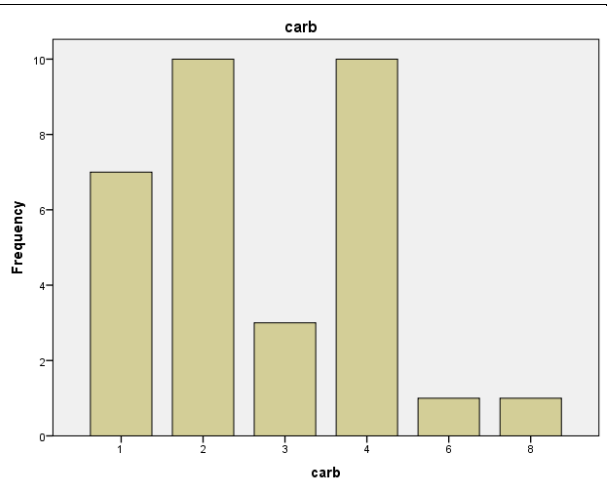
We can see 31st observation is outlier.

carb Stem-and-Leaf Plot

Frequency Stem & Leaf

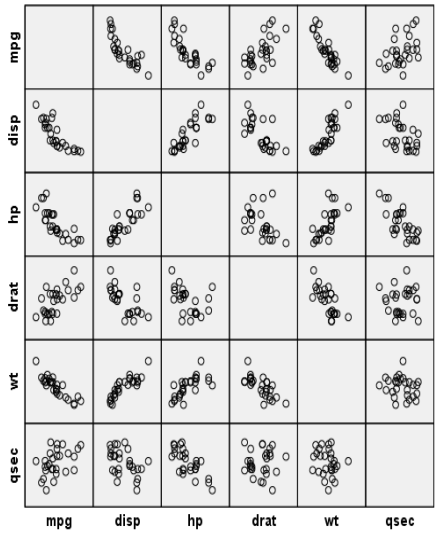
7.00 1 . 0000000
10.00 2 . 0000000000
3.00 3 . 000
10.00 4 . 0000000000
.00 5 .
1.00 6 . 0
1.00 Extremes (>=8)

Stem width: 1
Each leaf: 1 case(s)



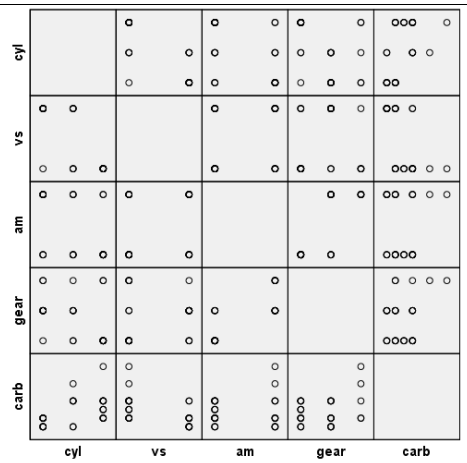
PAIR-WISE SCATTER PLOT

For Continuous Variables



mpg		High negative correlation	High negative correlation	Positive Correlation	High negative correlation	Low positive Correlation
disp	High negative correlation		High positive correlation	Negative correlation	High positive correlation	Low Negative Correlation
hp	Negative Correlation	High Positive Correlation		Low negative Correlation	High Positive Correlation	Negative Correlation
drat	Positive Correlation	Negtive Correlation	Low Negative Correlation		Negative Correlation	No Correlation
wt	High Negative Correlation	High Positive Correlation	Positive Correlation	Negative Correlation		No Correlation
qsec	Low Positive Correlation	Low Negative Correlation	Negative Correlation	No Correlation	Low Negative Correlation	
	mpg	**disp**	**hp**	**drat**	**wt**	**qsec**

For Discrete Variables



For Discrete Variables, these are showing the frequencies and showing linear association between them.

Correlation Analysis

a) For Continuous Variables

For the EDA of all Continuous Variables taken together, we will use:

- **Pearson's Correlation and its significance**

Correlations	**THE HYPOTHESIS OF INTEREST :**

		mpg	disp	hp	drat	wt	qsec
mpg	Pearson Correlation	1	-.848**	-.817**	.681**	-.883**	.441*
	Sig. (2-tailed)		.000	.000	.000	.000	.013
	N	32	32	31	32	30	31
disp	Pearson Correlation	-.848**	1	.859**	-.710**	.858**	-.430*
	Sig. (2-tailed)	.000		.000	.000	.000	.016
	N	32	32	31	32	30	31
hp	Pearson Correlation	-.817**	.859**	1	-.508**	.679**	-.704**
	Sig. (2-tailed)	.000	.000		.004	.000	.000
	N	31	31	31	31	29	30
drat	Pearson Correlation	.681**	-.710**	-.508**	1	-.728**	.040
	Sig. (2-tailed)	.000	.000	.004		.000	.830
	N	32	32	31	32	30	31
wt	Pearson Correlation	-.883**	.858**	.679**	-.728**	1	-.234
	Sig. (2-tailed)	.000	.000	.000	.000		.222
	N	30	30	29	30	30	29
qsec	Pearson Correlation	.441*	-.430*	-.704**	.040	-.234	1
	Sig. (2-tailed)	.013	.016	.000	.830	.222	
	N	31	31	30	31	29	31

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

H_0 : Correlation is insignificant V/S
H_1 : Correlation is significant

If Sig. (2-tailed) > 0.05 then there is significant correlation between two variables else correlation is insignificant.

INFERENCE :

- From the above table we infer that correlation between DRAT & QSEC and WT & QSEC is insignificant at 5% los
- Correlation between the rest of the variables is significant at 5% los.

b) For Discrete / Categorical Variables

For the EDA of all Discrete Variables taken together, we will use:

- **Spearman's Rank Correlation and its significance**

Correlations

			cyl	vs	am	gear	carb
Spearman's rho	cyl	Correlation Coefficient	1.000	-.814**	-.522**	-.564**	.580**
		Sig. (2-tailed)	.	.000	.002	.001	.001
		N	32	32	32	32	32
	vs	Correlation Coefficient	-.814**	1.000	.168	.283	-.634**
		Sig. (2-tailed)	.000	.	.357	.117	.000
		N	32	32	32	32	32

THE HYPOTHESIS OF INTEREST :

H_0 : Correlation is insignificant V/S
H_1 : Correlation is significant

If Sig. (2-tailed) > 0.05 then there is significant correlation between two variables else correlation is insignificant.

INFERENCE :

- From the above table we infer that correlation between VS & AM, VS & GEAR, AM & CARB and GEAR & CARB is insignificant at 5% los.
- Correlation between the rest of the variables is significant at 5% los.

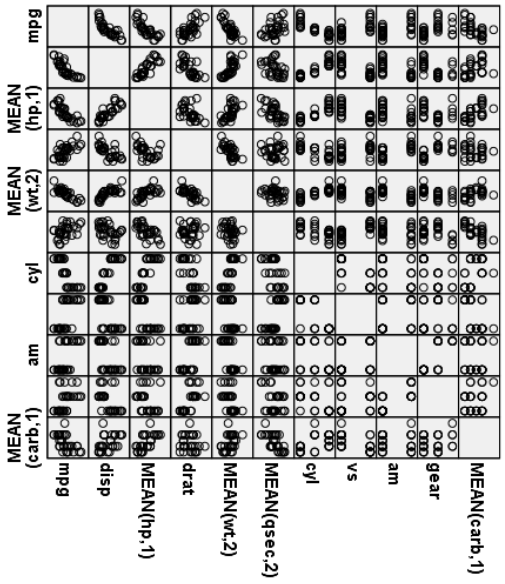
	am	Correlation Coefficient	-.522**	.168	1.000	.808**	-.064
		Sig. (2-tailed)	.002	.357	.	.000	.726
		N	32	32	32	32	32
	gear	Correlation Coefficient	-.564**	.283	.808**	1.000	.115
		Sig. (2-tailed)	.001	.117	.000	.	.531
		N	32	32	32	32	32
	carb	Correlation Coefficient	.580**	-.634**	-.064	.115	1.000
		Sig. (2-tailed)	.001	.000	.726	.531	.
		N	32	32	32	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

In the above table we analyze the data before outlier is removed. Now we analyze the data after outlier is removed.

For Combined Variables (Discrete and Continuous simultaneously)

Scatter Plot:



This is the Combined Scatter Plot (after outlier is removed) showing linear association between the variables.

Correlation Table:

Correlations

			mpg	disp	MEAN(hp,1)	drat	MEAN(wt,2)	MEAN(qsec,2)	cyl	vs	am	gear	MEAN(carb,1)
Spearman's rho	mpg	Correlation Coefficient	1.000	-.909**	-.881**	.651**	-.806**	.462**	-.911**	.707**	.562**	.543**	-.650**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.008	.000	.000	.001	.001	.000
		N	32	32	32	32	32	32	32	32	32	32	32
	disp	Correlation Coefficient	-.909**	1.000	.865**	-.684**	.808**	-.463**	.928**	-.724**	-.624**	-.594**	.540**
		Sig. (2-tailed)	.000	.	.000	.000	.000	.008	.000	.000	.000	.000	.001
		N	32	32	32	32	32	32	32	32	32	32	32
	MEAN(hp,1)	Correlation Coefficient	-.881**	.865**	1.000	-.539**	.698**	-.619**	.898**	-.752**	-.445*	-.437*	.690**
		Sig. (2-tailed)	.000	.000	.	.001	.000	.000	.000	.000	.011	.012	.000
		N	32	32	32	32	32	32	32	32	32	32	32
	drat	Correlation Coefficient	.651**	-.684**	-.539**	1.000	-.729**	.079	-.679**	.447*	.687**	.745**	-.120
		Sig. (2-tailed)	.000	.000	.001	.	.000	.667	.000	.010	.000	.000	.514
		N	32	32	32	32	32	32	32	32	32	32	32
	MEAN(wt,2)	Correlation Coefficient	-.806**	.808**	.698**	-.729**	1.000	-.245	.825**	-.532**	-.724**	-.637**	.403*
		Sig. (2-tailed)	.000	.000	.000	.000	.	.176	.000	.002	.000	.000	.022
		N	32	32	32	32	32	32	32	32	32	32	32
	MEAN(qsec,2)	Correlation Coefficient	.462**	-.463**	-.619**	.079	-.245	1.000	-.558**	.792**	-.162	-.164	-.655**
		Sig. (2-tailed)	.008	.008	.000	.667	.176	.	.001	.000	.376	.369	.000
		N	32	32	32	32	32	32	32	32	32	32	32
	cyl	Correlation Coefficient	-.911**	.928**	.898**	-.679**	.825**	-.558**	1.000	-.814**	-.522**	-.564**	.570**
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.001	.	.000	.002	.001	.001
		N	32	32	32	32	32	32	32	32	32	32	32
	vs	Correlation Coefficient	.707**	-.724**	-.752**	.447*	-.532**	.792**	-.814**	1.000	.168	.283	-.630**
		Sig. (2-tailed)	.000	.000	.000	.010	.002	.000	.000	.	.357	.117	.000
		N	32	32	32	32	32	32	32	32	32	32	32
	am	Correlation Coefficient	.562**	-.624**	-.445*	.687**	-.724**	-.162	-.522**	.168	1.000	.808**	-.090
		Sig. (2-tailed)	.001	.000	.011	.000	.000	.376	.002	.357	.	.000	.625
		N	32	32	32	32	32	32	32	32	32	32	32
	gear	Correlation Coefficient	.543**	-.594**	-.437*	.745**	-.637**	-.164	-.564**	.283	.808**	1.000	.085
		Sig. (2-tailed)	.001	.000	.012	.000	.000	.369	.001	.117	.000	.	.642
		N	32	32	32	32	32	32	32	32	32	32	32
	MEAN(carb,1)	Correlation Coefficient	-.650**	.540**	.690**	-.120	.403*	-.655**	.570**	-.630**	-.090	.085	1.000
		Sig. (2-tailed)	.000	.001	.000	.514	.022	.000	.001	.000	.625	.642	.
		N	32	32	32	32	32	32	32	32	32	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

H_0 : Correlation is insignificant V/S H_1 : Correlation is significant

If Sig. (2-tailed) > 0.05 then there is significant correlation between two variables else correlation is insignificant.

From the above table we observe that H_0 is accepted i.e Correlation is insignificant in the following cases

Drat & Qsec, Drat & Carb, Wt & Qsec, Qsec & Am, Qsec & Gear, Vs & Am, Vs & Gear, Am & Carb, Gear & Carb.

Note:

Further We can analyze Multi-Colliniarity, Autocorrelation, Homoscedasticity or Heteroscedasticity and fit linear regression model.