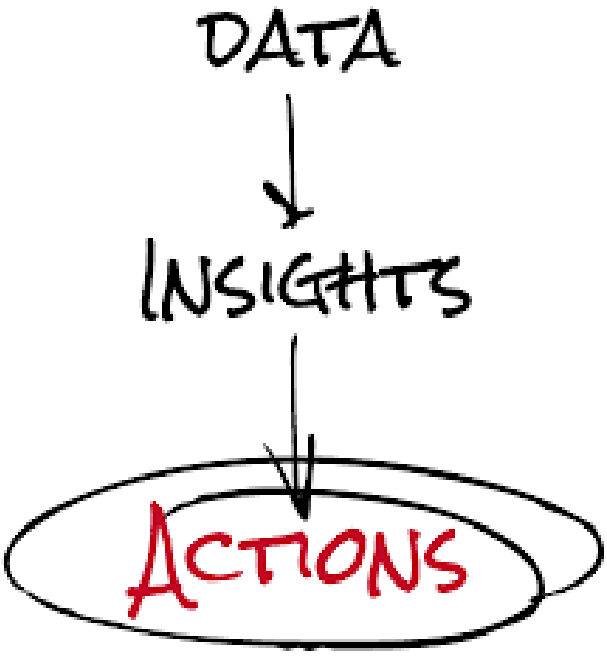
SPSS®

PRACTICAL FILE



RAJ KUMAR

Msc. Statistics (F)

Pr.III (Group. A)

CASE STUDY:

EXPLORATORY DATA ANALYSIS (EDA)

Due Date: 24th August, 2015

Date of Submission: 22nd August,2015

Supervisor's Remarks

Late Submission:

Plagiarism:

Completeness:

Quality of Content:

Results and Interpretations:

Additional Remarks:

Contents

About EDA:

“EDA” is a critical first step in analyzing the data.

Here are the main reasons we use EDA:

- Detection of mistakes
- Checking of assumptions
- Preliminary selection of appropriate models
- Determining relationships among the explanatory variables, and
- Assessing the direction and rough size of relationships between explanatory and outcome variables.

Any method of looking at data that does not include formal statistical modelling and inference falls under the term exploratory data analysis.

Dataset: “Mtcars”

Description

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

Each line of mtcars represents one model of car, which we can see in the row names. Each column is then one attribute of that car, such as the miles per gallon (or fuel efficiency), the number of cylinders, the displacement (or volume) of the car’s engine in cubic inches, whether the car has an automatic or manual transmission, and so on.

Source

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

This dataset is taken from R with 32 observations on following 11 variables (column 1 tells the brand of car).

[1] mpg Miles/(US) gallon
[2] Cyl Number of cylinders
[3] disp Displacement (cu.in.)
[4] Hp Gross horsepower
[5] drat Rear axle ratio
[6] Wt Weight (lb/1000)
[7] qsec 1/4 mile time
[8] Vs V/S
[9] Am Transmission (0 = automatic, 1 = manual)
[10] gear Number of forward gears
[11] carb Number of carburettors

We will discuss each of the above variables in detail in this study.

First, let's discuss type of data we have in this mtcars dataset. Classifying all the variables of given dataset into following formats:

Continuous data type: mpg, disp, hp, drat, wt, qsec.

Tools used in this study for EDA of continuous data type :

- **Descriptives** : Mean, Median, Mode and measures of dispersion like range, variance etc .
- **Histogram** : For distribution of data and comparing its proximity with the normal distribution
- **Q-Q Plot, KS Test and S-W Test** : To test whether the data is Normally distributed or not

  H_0 : Sample comes from a normal population

  H_1 : Sample does not comes from a normal population

  Tools like Histogram, Kutosis and SE can also be used to comment about normality.

- **Box Plot** : For detection of outliers in the data
- **Stem and Leaf Plot** :It is similar to a histogram, to assist in visualizing the shape of a distribution . It is important to note that no numbers are skipped, even if it means that some stems have no leaves. . 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.The leaves are listed in increasing order in a row to the right of each stem.

Discrete/categorical (nominal) data type: cyl, vs, am, gear, crab.

Tools used in this study for EDA of discrete /categorical data type :

- **Frequency Table** : To get the frequency of each data point
- **Descriptive Statistics :** Mean, Median, Mode and measures of dispersion like range, variance etc.
- **Bar plot** : Represents frequency distribution of data.
- **Stem and Leaf Plot**

Nominal variable : car_name

First step in EDA:

a) Summarization :

Let's divide this into two categories further:

- Measures of Central tendency: Mean, Median, Mode.
- Measures of Dispersion: Range, Variance, stddev etc.

Continuous variables: Measures of Central Tendency

		mpg	disp	hp	drat	wt	Qsec
N	Valid	32	32	32	32	32	32
	Missing	0	0	0	0	0	0
Mean		20.091	230.722	146.69	3.5966	3.21725	17.8488
Median		19.200	196.300	123.00	3.6950	3.32500	17.7100
Mode		10.4[a]	275.8	110[a]	3.07[a]	3.440	17.02[a]

a. Multiple modes exist. The smallest value is shown

Continuous Variables: Measures of Dispersion

		mpg	disp	hp	drat	wt	Qsec
N	Valid	32	32	32	32	32	32
	Missing	0	0	0	0	0	0
Std. Error of Mean		1.0654	21.9095	12.120	.09452	.172968	.31589
Std. Deviation		6.0269	123.9387	68.563	.53468	.978457	1.78694
Variance		36.324	15360.800	4700.867	.286	.957	3.193
Range		23.5	400.9	283	2.17	3.911	8.40
Minimum		10.4	71.1	52	2.76	1.513	14.50
Maximum		33.9	472.0	335	4.93	5.424	22.90

Discrete / categorical variables :

Measures of Central Tendency

		Cyl	vs	Am	gear	Carb
N	Valid	32	32	32	32	32
	Missing	0	0	0	0	0
Mean		6.19	.44	.41	3.69	2.81
Median		6.00	.00	.00	4.00	2.00
Mode		8	0	0	3	2[a]

a. Multiple modes exist. The smallest value is shown

Measures of Dispersion

		cyl	vs	am	gear	carb
N	Valid	32	32	32	32	32
	Missing	0	0	0	0	0
Std. Error of Mean		.316	.089	.088	.130	.286
Std. Deviation		1.786	.504	.499	.738	1.615
Variance		3.190	.254	.249	.544	2.609
Range		4	1	1	2	7
Minimum		4	0	0	3	1
Maximum		8	1	1	5	8

Skewness and Kutosis

		cyl	vs	am	gear	carb
	Valid	32	32	32	32	32
	Missing	0	0	0	0	0
Skewness		-.192	.265	.401	.582	1.157
Std. Error of Skewness		.414	.414	.414	.414	.414
Kurtosis		-1.763	-2.063	-1.967	-.895	2.020
Std. Error of Kurtosis		.809	.809	.809	.809	.809

Frequency Tables:

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	

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	

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	

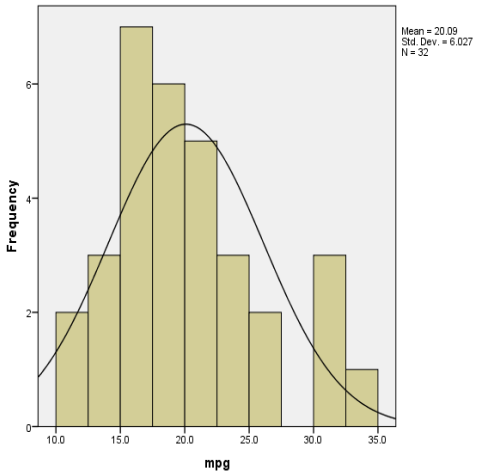
gear

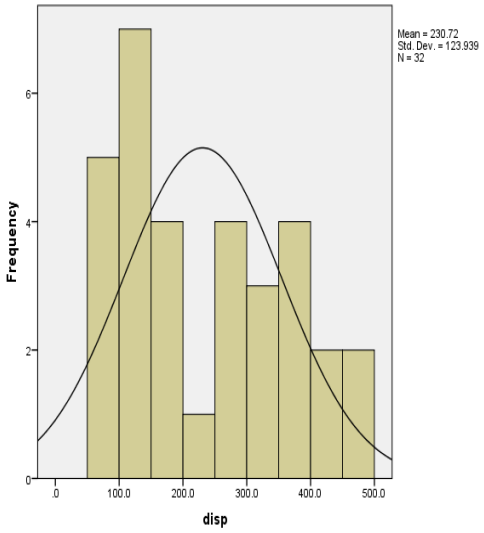
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	15	46.9	46.9	46.9
	4	12	37.5	37.5	84.4
	5	5	15.6	15.6	100.0
	Total	32	100.0	100.0	

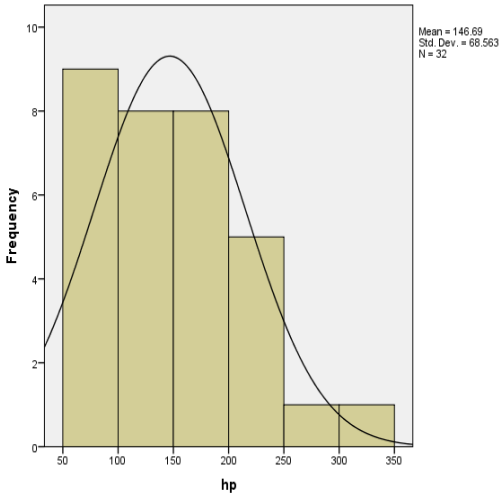
carb

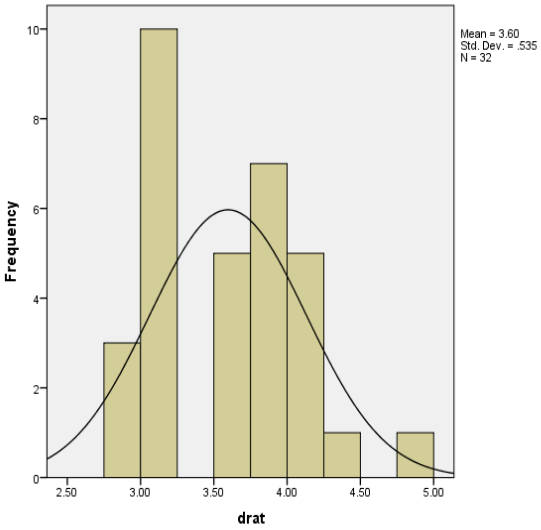
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	21.9	21.9	21.9
	2	10	31.3	31.3	53.1
	3	3	9.4	9.4	62.5
	4	10	31.3	31.3	93.8
	6	1	3.1	3.1	96.9
	8	1	3.1	3.1	100.0
	Total	32	100.0	100.0	

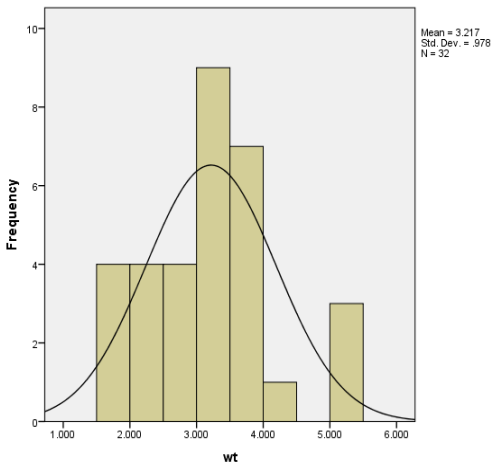
b) Data Visualisation:

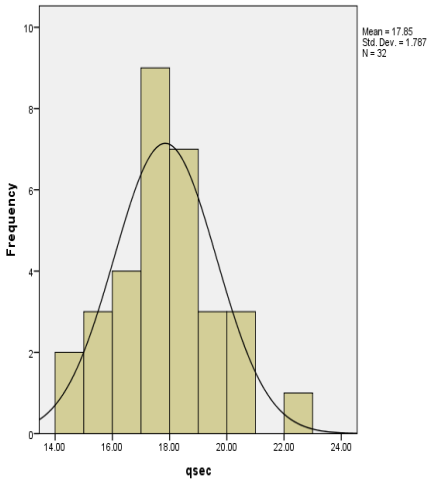












Inference :

From the histograms above , it can be infer that :

- “mpg” is almost normally distributed.
- “disp” is not normally distributed.
- “hp” is not normally distributed.
- “drat” is not normally distributed.
- “wt” is not normally distributed.
- “qsec” is almost normally distributed.

Stem-and-Leaf Plots

mpg

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)

disp

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
Each leaf: 1 case(s)

hp

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)

Stem width: 100
Each leaf: 1 case(s)

drat

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)

wt

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)

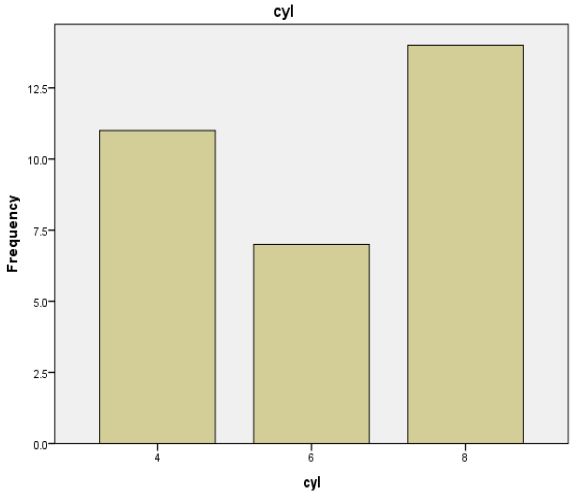
Stem width: 1.000
Each leaf: 1 case(s)

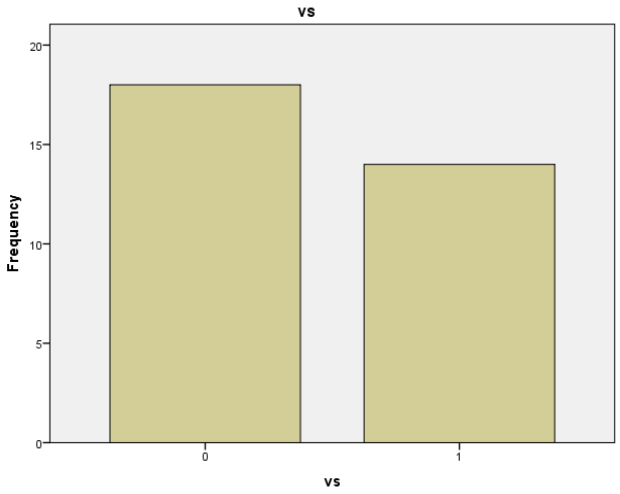
qsec

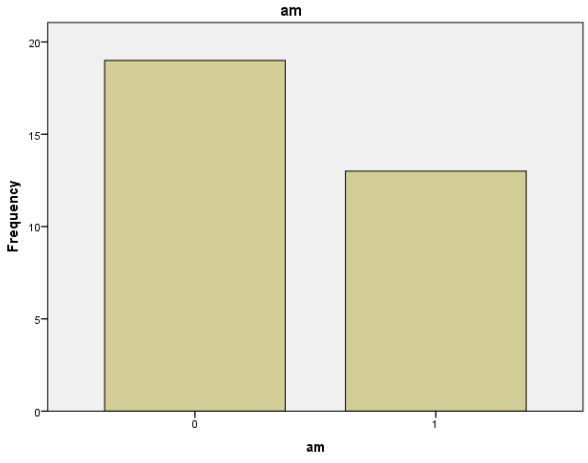
Frequency	Stem &	Leaf
2.00	14 .	56
3.00	15 .	458
4.00	16 .	4789
9.00	17 .	000344689
7.00	18 .	0356699
3.00	19 .	449
3.00	20 .	002
1.00 Extremes		(>=22.9)

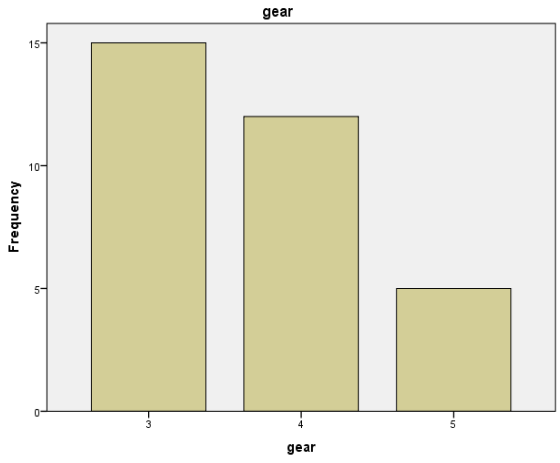
Stem width: 1.00
Each leaf: 1 case(s)

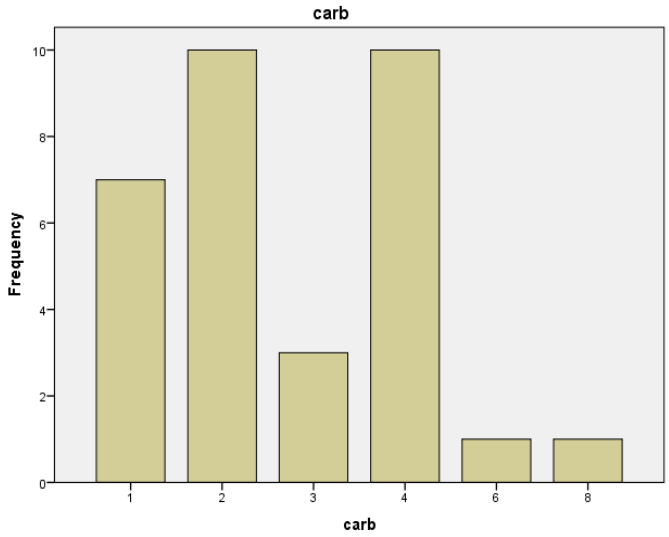
Discrete / categorical variables: Bar plot











Stem-and-Leaf Plot

cyl

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)

vs

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)

am

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)

gear

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)

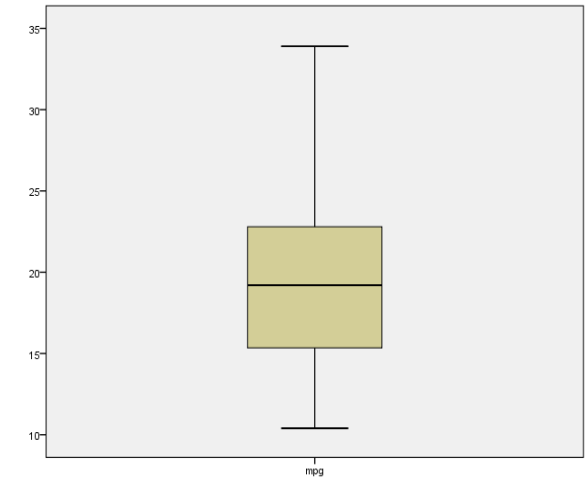
carb

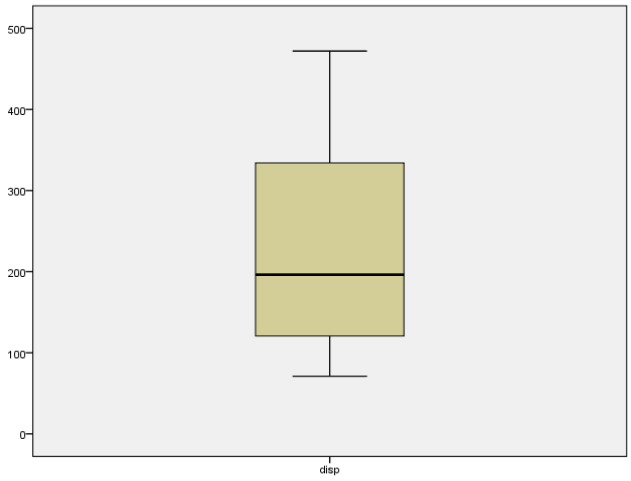
Frequency	Stem &	Leaf
7.00	1 .	0000000
10.00	2 .	0000000000
3.00	3 .	000
10.00	4 .	0000000000
.00	5 .	
1.00	6 .	0
1.00 Extremes		(>=8)

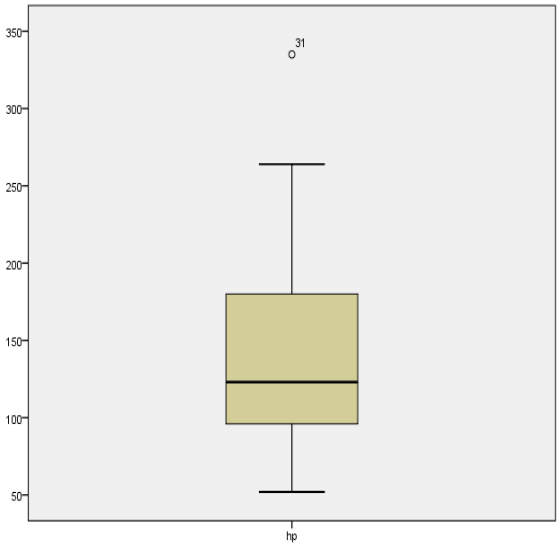
Stem width: 1
Each leaf: 1 case(s)

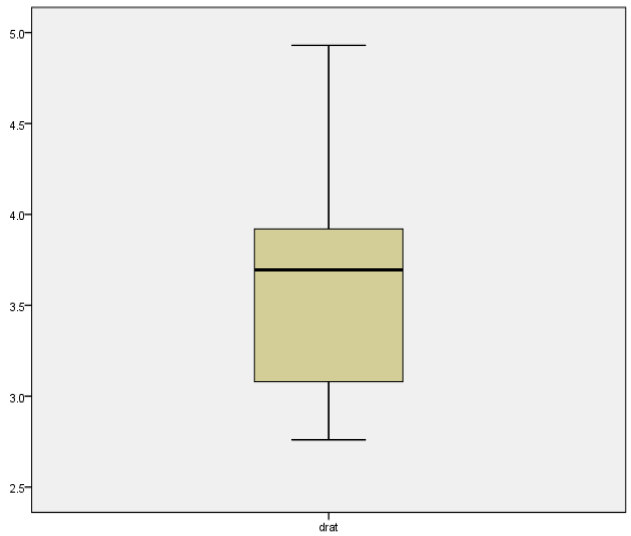
c) Outlier Detection: Box plot

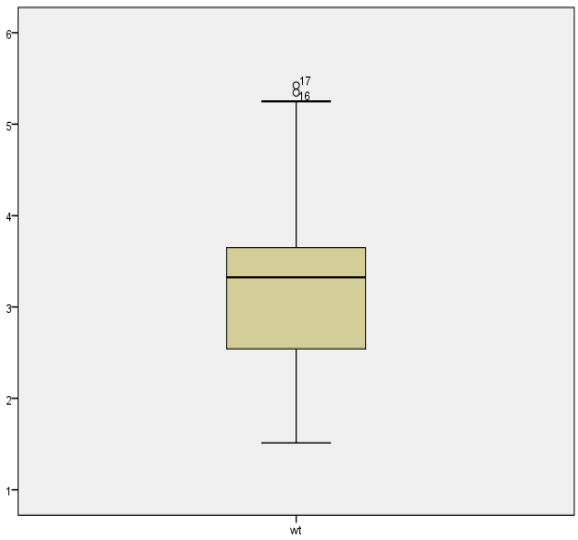
Continuous:

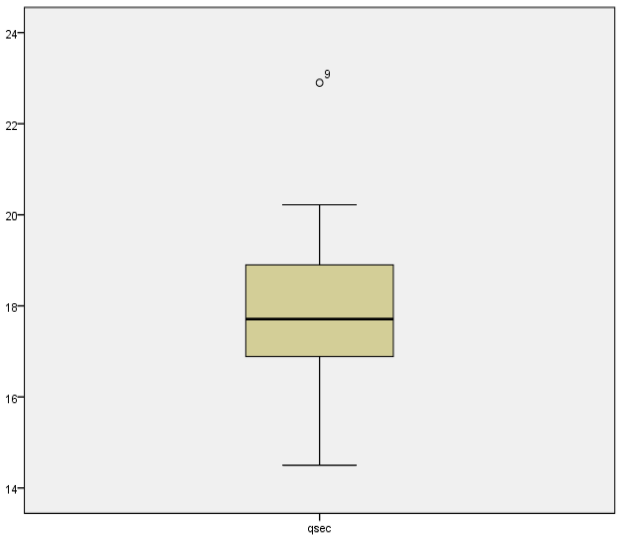












Inference:

- "mpg" does not contain any outlier.
- "disp" does not contain any outlier.
- "hp" contains an outlier as 31st observation.
- "drat" does not contain any outlier.
- "wt" contains 2 outliers which are 16th and 17th observation.
- "qsec" contain an outlier as 9th observation.

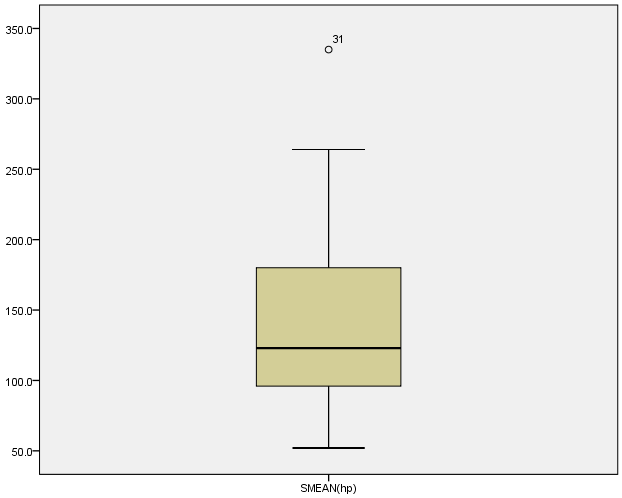
Actions:

Start treating the outliers observed in "hp","wt" and "qsec" as missing value .Next step is Replacing the outlier in by using Replacing Missing value function.

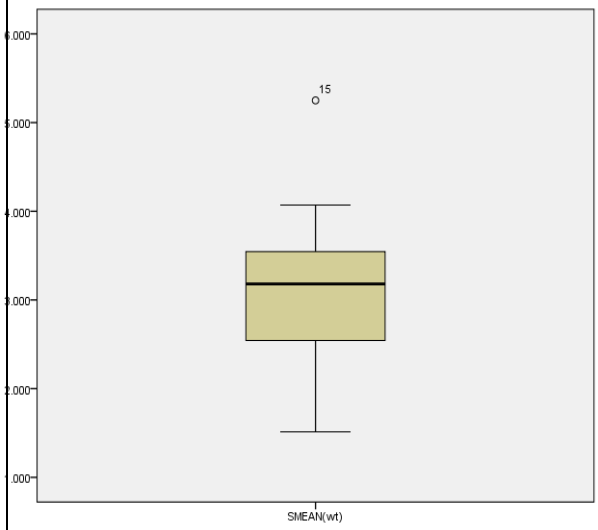
Now,

- 31st observation is replaced by 335 in "hp".
- 16th and 17th observation are replaced by 3.073 in "wt".
- 9th observation in "qsec" is replaced by 17.69.

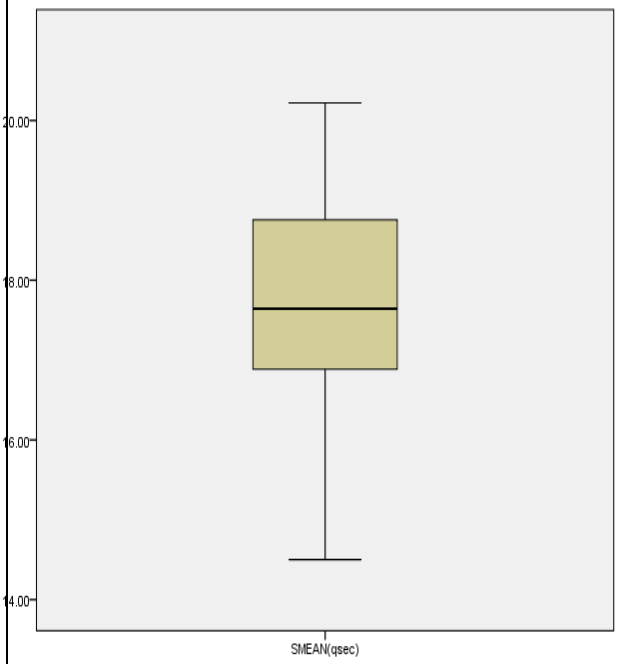
To check a sense of doing this, we plotted a boxplot for each of the variable.



We infer that boxplot still gives a outlier which ia again, after we replaced the earlier boxplot outlier using series mean method. So, we will stop eliminating outliers here and continue our study.

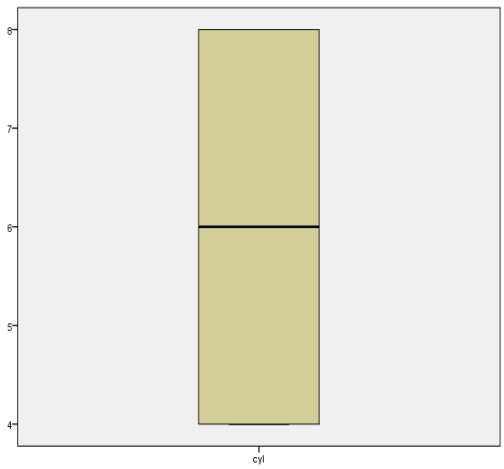


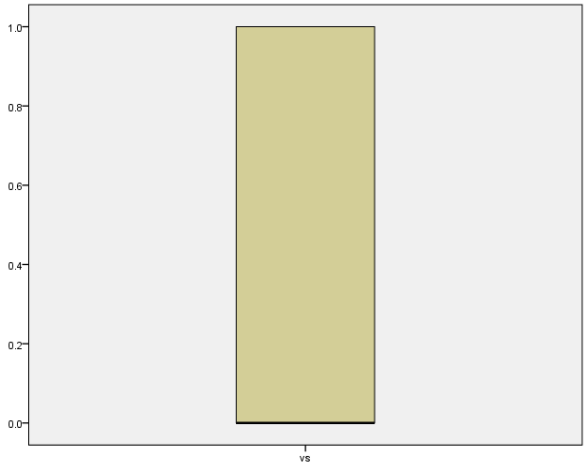
We infer that boxplot still gives a outlier which is now 15th observation, after we replaced the earlier boxplot outlier using series mean method. So, we will stop eliminating outliers here and continue our study.

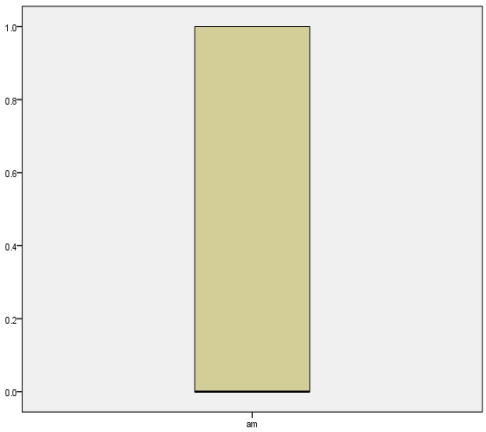


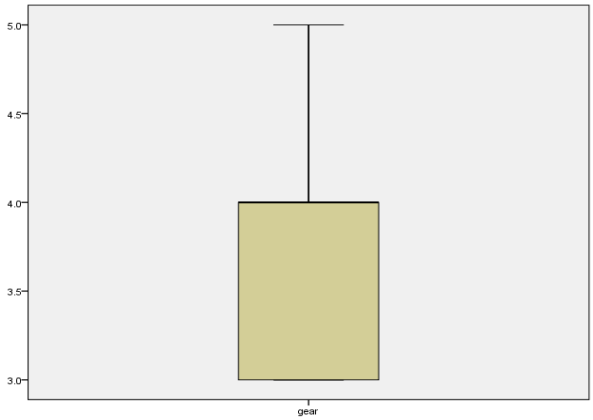
We infer that boxplot outlier has been removed after replacing the outlier using series mean method and this will help in evaluating the data more easily.

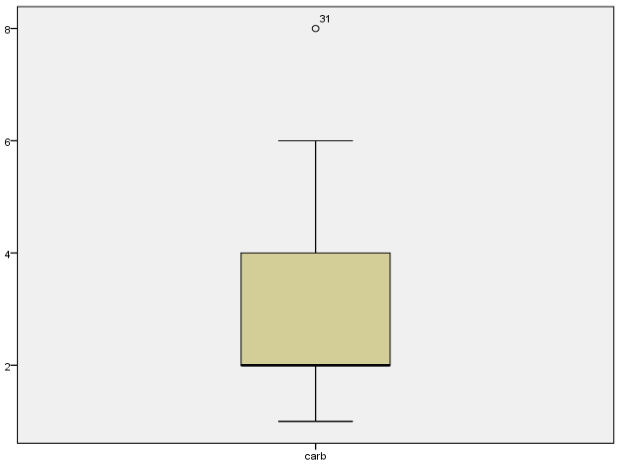
Discrete:











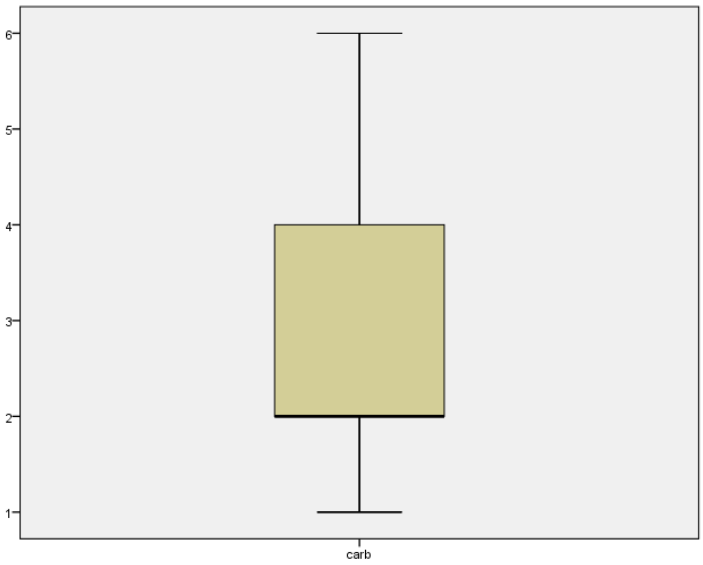
Inference:

- 31st observation in “carb” is the only outlier.

Actions:

Start treating this outlier as missing value .Next step is Replacing the outlier in “carb” by using Replacing Missing value function.

Now, 31st observation is replaced by 2.6 in “carb”. To check a sense of doing this, we plotted a boxplot.



We infer that boxplot outlier has been removed after replacing the outlier using series mean method and this will help in evaluating the data more easily.

d) Tests for Normality:

1. K-S test and S-W tests

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
Mpg	.126	32	.200*	.948	32	.123
Disp	.195	32	.003	.920	32	.021
Hp	.166	32	.024	.933	32	.049
Drat	.160	32	.037	.946	32	.110
Wt	.136	32	.142	.943	32	.093
Qsec	.073	32	.200*	.973	32	.594

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

a. Lillie fors Significance Correction

Inference :

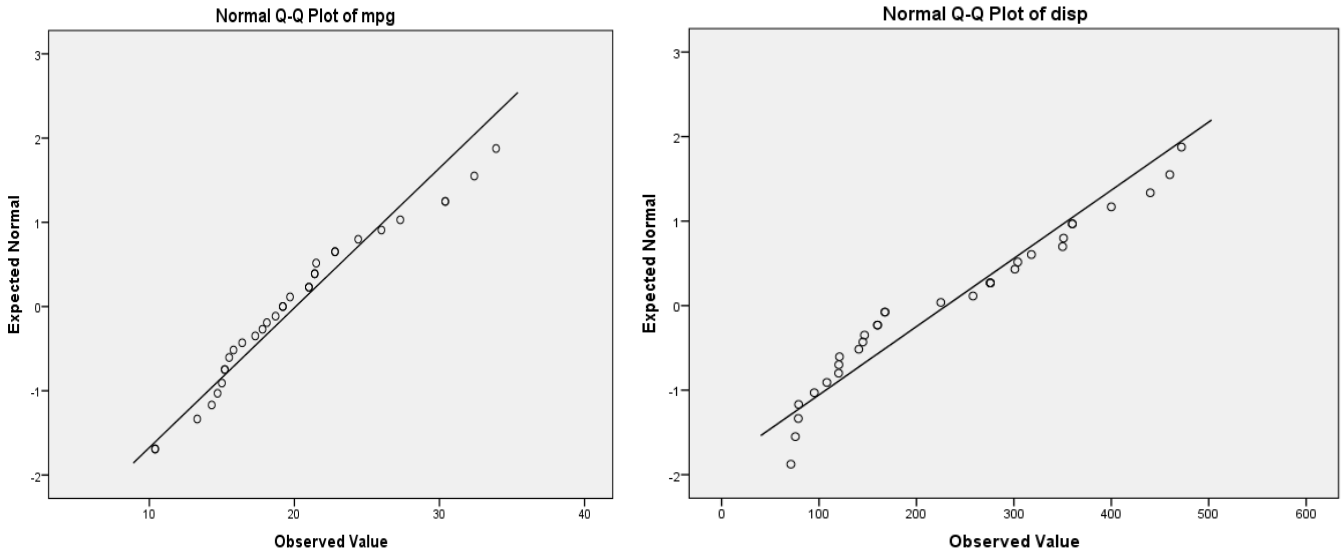
From Kolmogorov-Smirnov Test :

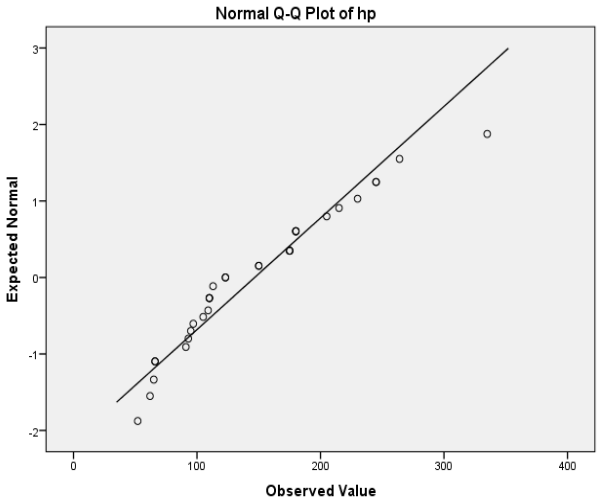
- mpg with Sig. = 0.20 (>0.05), we may conclude that at 5% l.o.s. “mpg” is normally distributed. From Shapiro-Wilk Test with
- disp with Sig. = 0.003 (<0.05), we may conclude that at 5% l.o.s. “disp” is not normally distributed.
- hp with Sig. = 0.024 (<0.05), we may conclude that at 5% l.o.s. “hp” is not normally distributed.
- drat with Sig. = 0.037 (<0.05), we may conclude that at 5% l.o.s. “drat” is not normally distributed
- wt with Sig. = 0.142 (>0.05), we may conclude that at 5% l.o.s. “wt” is normally distributed.
- qsec with Sig. = 0.20 (>0.05), we may conclude that at 5% l.o.s. “qsec” is normally distributed.

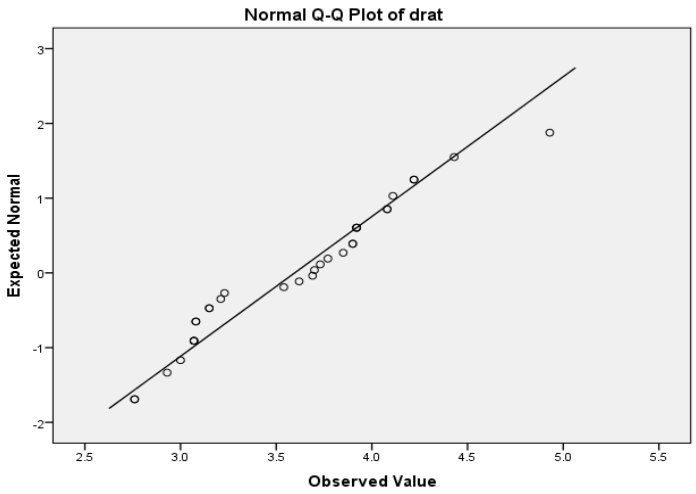
From Shapiro-Wilk Test :

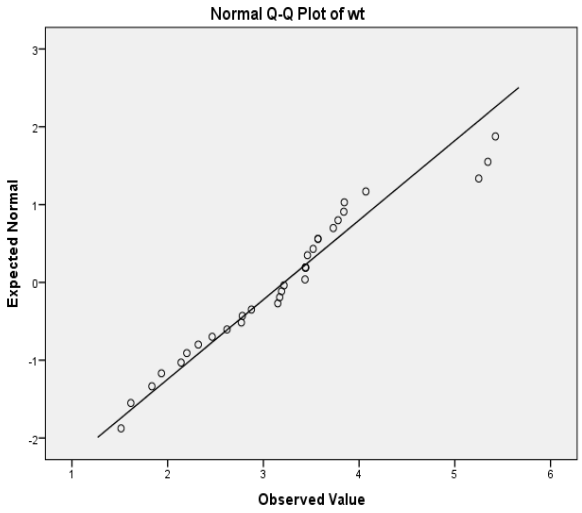
- mpg with Sig. = 0.123 (>0.05), we may conclude that at 5% l.o.s. “mpg” is normally distributed.
- disp with Sig. = 0.021 (<0.05), we may conclude that at 5% l.o.s. “disp” is not normally distributed.
- hp with Sig. = 0.049 (<0.05), we may conclude that at 5% l.o.s. “hp” is not normally distributed.
- drat with Sig. = 0.110 (>0.05), we may conclude that at 5% l.o.s. “drat” is normally distributed.
- wt with Sig. = 0.093 (>0.05), we may conclude that at 5% l.o.s. “wt” is normally distributed.
- qsec with Sig. = 0.594 (>0.05), we may conclude that at 5% l.o.s. “qsec” is normally distributed.

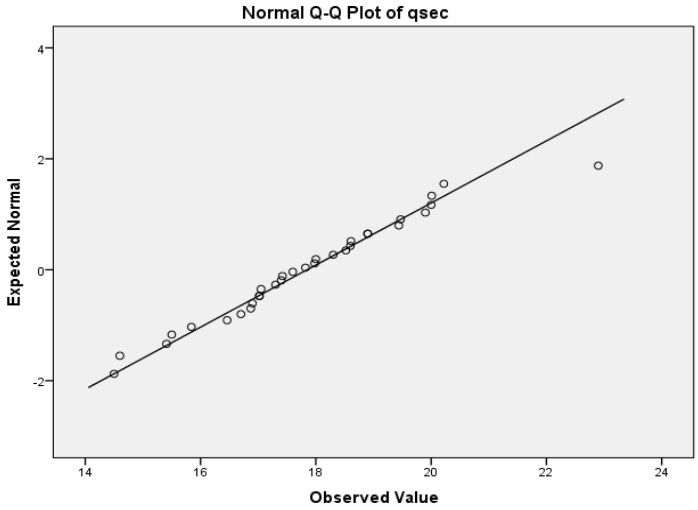
2. Q-Q Plots











Inference :

From the above Q-Q plots , we may infer here that :

- “mpg” is almost normally distributed.
- “disp” is not normally distributed.
- “hp” is not normally distributed.
- “drat” is not normally distributed.
- “wt” is almost normally distributed.
- “qsec” is almost normally distributed.

2. Using Skewness and Kurtosis

Statistics

		Mpg	disp	hp	drat	wt	Qsec
N	Valid	32	32	32	32	32	32
	Missing	0	0	0	0	0	0
Skewness		.672	.420	.799	.293	.466	.406
Std. Error of Skewness		.414	.414	.414	.414	.414	.414
Kurtosis		-.022	-1.068	.275	-.450	.417	.865
Std. Error of Kurtosis		.809	.809	.809	.809	.809	.809

Inference:

2*S.E for skewness for :

mpg,disp,hp,drat,wt and qsec are 1.344,0.84,1.598,.586,.932 and .812 respectively.

We may infer that coeffecient is insignificant for all variables since its absolute value is less than twice it's standard error.

Ratio of Kurtosis to its S.E for :

mpg,disp,hp,drat,wt and qsec are -0.0272, -1.988,.334, -0.5562,0.5154 and 1.0692 respectively.

Since the values lie between -2 and 2 , therefore we accept normality for all the variables listed above.

EDA for multiple variables :

a) Continuous variables:

Correlations

		mpg	disp	hp	drat	wt	Qsec
Mpg	Pearson Correlation	1	-.848**	-.776**	.681**	-.868**	.419*
	Sig. (2-tailed)		.000	.000	.000	.000	.017
	N	32	32	32	32	32	32
Disp	Pearson Correlation	-.848**	1	.791**	-.710**	.888**	-.434*
	Sig. (2-tailed)	.000		.000	.000	.000	.013
	N	32	32	32	32	32	32
Hp	Pearson Correlation	-.776**	.791**	1	-.449**	.659**	-.708**
	Sig. (2-tailed)	.000	.000		.010	.000	.000
	N	32	32	32	32	32	32
Drat	Pearson Correlation	.681**	-.710**	-.449**	1	-.712**	.091
	Sig. (2-tailed)	.000	.000	.010		.000	.620
	N	32	32	32	32	32	32
Wt	Pearson Correlation	-.868**	.888**	.659**	-.712**	1	-.175
	Sig. (2-tailed)	.000	.000	.000	.000		.339
	N	32	32	32	32	32	32
Qsec	Pearson Correlation	.419*	-.434*	-.708**	.091	-.175	1
	Sig. (2-tailed)	.017	.013	.000	.620	.339	
	N	32	32	32	32	32	32

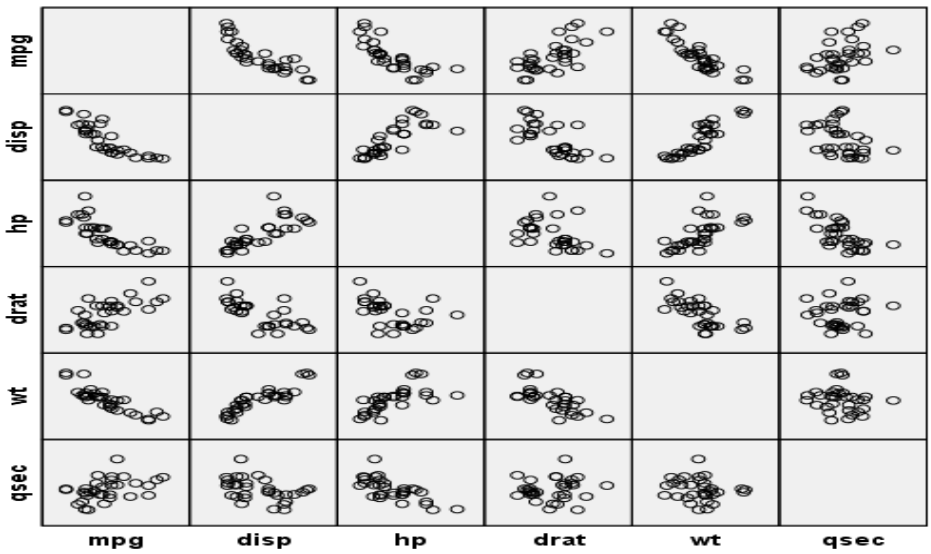
**. Correlation is significant at the 0.01 level (2-tailed).

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

Inference :

We can infer that except the correlation between pairs of variables (drat ,qsec) and (wt,qsec) , rest of the pair-wise correlations are significant at 5% l.o.s. .

Pairwise Scatter Plot :



2. Discrete/categorical variables :

Non-parametric Correlations

Correlations

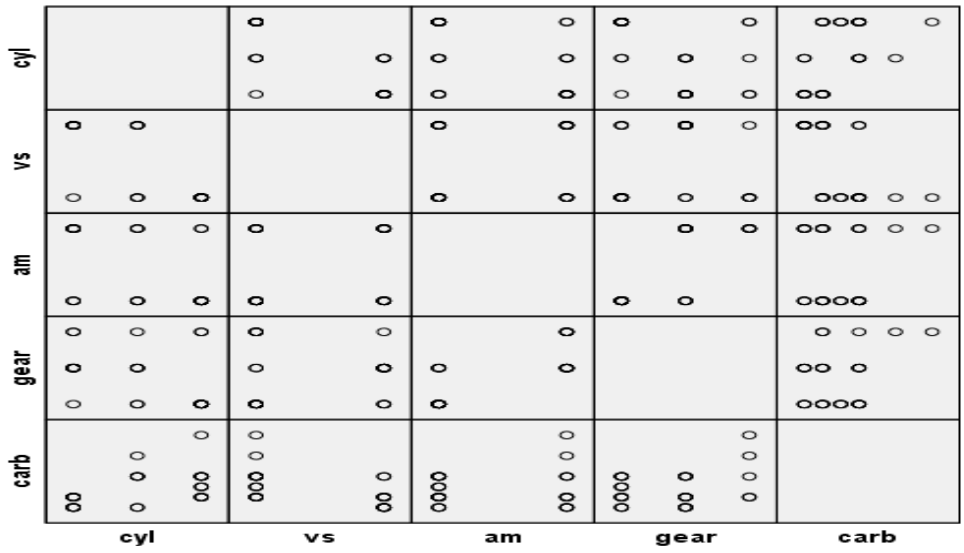
			cyl	vs	am	Gear	Carb
Spearman's rho	cyl	Correlation Coefficient	1.000	-.814**	-.522**	-.564**	.580**
		Sig. (2-tailed)	.	.000	.002	.001	.001
		N	32	32	32	32	32
	vs	Correlation Coefficient	-.814**	1.000	.168	.283	-.634**
		Sig. (2-tailed)	.000	.	.357	.117	.000
		N	32	32	32	32	32
	am	Correlation Coefficient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

Inference :

We can infer that except the correlation between pairs of variables (vs ,am) ,(vs,gear) ,(am,carb) and (gear ,carb) rest of the pair-wise correlations are significant at 5% l.o.s. .

Pairwise Scatter Plot :



Missing Value Imputation (mpg):

Values for mpg corresponding to 21st entry is 21.5.

Suppose 21st entry in dataset corresponding to mpg was missing initially in our data set . Using **Replace Missing Value** by **series mean method** we got the following results:

21st entry of mpg replaced by = 20.05.

Descriptive Statistics for mpg with original values

mpg

N	Valid	32
	Missing	0
Mean		20.091
Std. Error of Mean		1.0654
Median		19.200
Mode		10.4[a]
Std. Deviation		6.0269
Variance		36.324
Skewness		.672
Std. Error of Skewness		.414
Kurtosis		-.022
Std. Error of Kurtosis		.809
Range		23.5
Minimum		10.4
Maximum		33.9

a. Multiple modes exist. The smallest value is shown

Descriptive Statistics after Replacing Missing Value of mpg

SMEAN(mpg)

N	Valid	32
	Missing	0
Mean		20.0452
Std. Error of Mean		1.06445
Median		19.2000
Mode		10.40[a]
Std. Deviation		6.02146
Variance		36.258
Skewness		.698
Std. Error of Skewness		.414
Kurtosis		.013
Std. Error of Kurtosis		.809
Range		23.50
Minimum		10.40
Maximum		33.90

a. Multiple modes exist. The smallest value is shown

Hence , we may infer that there is not much significant difference between original and replaced value but a significant change is observed in Kurtosis which changes from negative to positive which indicates observations cluster more and have longer tails than those in normal distribution.

Sign-off note :

Exploratory data analysis is a powerful tool.

"*A picture is worth a thousand words*"

Where this can go next ?

EDA provides a great opportunity to test your simple hypotheses and hunches before jumping into a rigorous model building.The intuition we gained from EDA can now help us to start moving towards modelling and analysis. We'll want to build some of the intelligence/intuition we built from the Exploratory Data Analysis into our models .

Next Step could be Regression Analysis of mtcars dataset.