Q1) Identify the Data type for the Following:

Activity	Data Type
Number of beatings from Wife	Discrete
Results of rolling a dice	Discrete
Weight of a person	Continuous
Weight of Gold	Continuous
Distance between two places	Continuous
Length of a leaf	Continuous
Dog's weight	Continuous
Blue Color	Categorical
Number of kids	Discrete
Number of tickets in Indian railways	Discrete
Number of times married	Discrete
Gender (Male or Female)	Categorical

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

Data	Data Type
Gender	Nominal
High School Class Ranking	Ordinal
Celsius Temperature	Ratio
Weight	Ratio
Hair Color	Nominal
Socioeconomic Status	Ordinal
Fahrenheit Temperature	Ratio
Height	Ratio
Type of living accommodation	Ordinal
Level of Agreement	Ordinal
IQ(Intelligence Scale)	Interval
Sales Figures	Ratio
Blood Group	Nominal
Time Of Day	Ordinal
Time on a Clock with Hands	Ratio
Number of Children	Nominal
Religious Preference	Nominal
Barometer Pressure	Ratio
SAT Scores	Interval
Years of Education	Interval

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

Ans: Let, H = Heads, T = Tails

Possible outcomes: (H,H,H), (H,H,T), (H,T,H), (H,T,T), (T,H,H), (T,H,T),
(T,T,H), (T,T,T)

Total number of outcomes = 8

Number of outcomes gives two heads and one tail = 3

Number of favorable outcomes = 3

P[E]=(Number of favorable outcomes) / (Total number of outcomes)

P[E] = 3/8.

Q4) Two Dice are rolled, find the probability that sum is

a) Equal to 1

Ans: The minimum sum is (1,1) = 2

Therefore the sum is equal to 1 = 0/36 = 0

b) Less than or equal to 4

Ans = {(1,1)(1,2)(1,3)(2,1)(2,2)(3,1)}

= 6/36

c) Sum is divisible by 2 and 3

Ans = {(1,5),(2,3)(3,2)(4,2)(5,1)(6,6)}

= 6/36

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

Ans: Let us consider a formula ${}^{n}C_{r} = \frac{n!}{r!(n-r)!}$

Total number of balls = 2+3+2 = 7

Let us consider A be the sample.

n(A) = number of ways of drawing 2 balls out of 7

$= {}^{7}C_{2}$

$= \frac{7!}{2!(7-2)!}$

$=\frac{7!}{2!(5)!}$

$= \frac{7\times6\times5\times4\times3\times2\times1}{2\times1(5\times4\times3\times2\times1)}$

$= \frac{5040}{240}$

= 21

Let E = evet of drawing 2 balls out of 7 none of which is blue = 7-2=5

n(E) = number of ways of drawing 2 balls out of 5

$={}^{5}C_{2}$

$= \frac{5!}{2!(5-2)!}$

$=\frac{5!}{2!(3)!}$

$= \frac{5\times4\times3\times2\times1}{2\times1(3\times2\times1)}$

$= \frac{120}{12}$

= 10

$P(E) = \frac{n(E)}{n(A)} = \frac{10}{21}$

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

CHILD	Candies count	Probability
A	1	0.015
B	4	0.20
C	3	0.65
D	5	0.005
E	6	0.01
F	2	0.120

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

Ans: Expected number of candies = $\sum(X*P(x))$

$= \sum(\text{Candies count}*\text{Probability})$

$= 1*0.05+4*0.20+3*0.65+5*0.005+6*0.01+2*0.120$

$= 3.09$

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

- For Points,Score,Weigh>
  Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

Use Q7.csv file

	Points	Score	Weigh
Mazda RX4	3.9	2.62	16.46
Mazda RX4 Wag	3.9	2.875	17.02
Datsun 710	3.85	2.32	18.61
Hornet 4 Drive	3.08	3.215	19.44
Hornet Sportabout	3.15	3.44	17.02
Valiant	2.76	3.46	20.22
Duster 360	3.21	3.57	15.84
Merc 240D	3.69	3.19	20
Merc 230	3.92	3.15	22.9
Merc 280	3.92	3.44	18.3
Merc 280C	3.92	3.44	18.9
Merc 450SE	3.07	4.07	17.4
Merc 450SL	3.07	3.73	17.6
Merc 450SLC	3.07	3.78	18
Cadillac Fleetwood	2.93	5.25	17.98
Lincoln Continental	3	5.424	17.82
Chrysler Imperial	3.23	5.345	17.42
Fiat 128	4.08	2.2	19.47
Honda Civic	4.93	1.615	18.52
Toyota Corolla	4.22	1.835	19.9
Toyota Corona	3.7	2.465	20.01

Dodge Challenger	2.76	3.52	16.87
AMC Javelin	3.15	3.435	17.3
Camaro Z28	3.73	3.84	15.41
Pontiac Firebird	3.08	3.845	17.05
Fiat X1-9	4.08	1.935	18.9
Porsche 914-2	4.43	2.14	16.7
Lotus Europa	3.77	1.513	16.9
Ford Pantera L	4.22	3.17	14.5
Ferrari Dino	3.62	2.77	15.5
Maserati Bora	3.54	3.57	14.6
Volvo 142E	4.11	2.78	18.6

Mean = $\frac{\text{Sum off all the values}}{\text{Total numbers of values present}}$

Median = It is the middle value of a given set.

Mode = Most repeated value in the variables set.

Range = It is lies between minimum value to maximum value i.e [min value , max value]

Variance = It measures the spread between numbers in a data set.

Formula :

$\sigma^2 = \sum (x_{r\,-}\,\mu)^2/ (n-1)$

X_r= r^{th} data point r =1 to n

μ = Mean of all data points

n = Number of data points

Standered diviation = Its tells about how much the data of a set is differ from the mean value of the data set.

Formula

- ❖ Mannual calculation

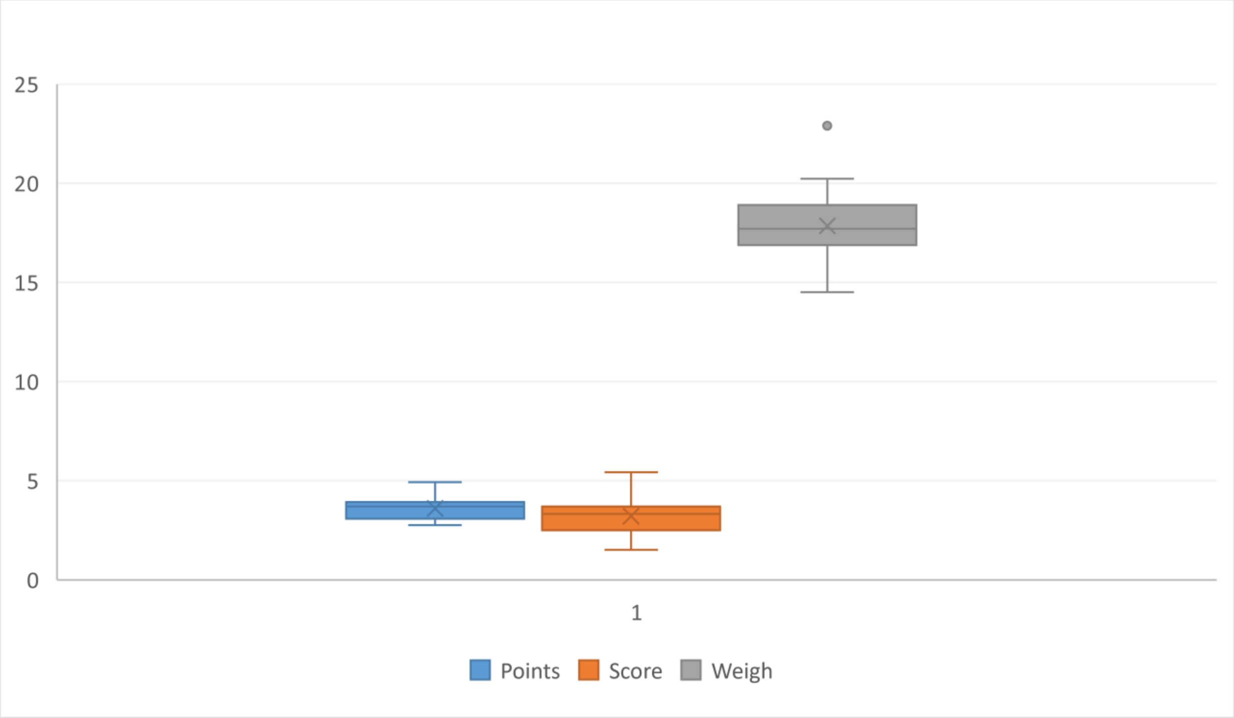
$\boldsymbol{\sigma} =$ $\sqrt{}$ $[(x_{r\,-}\,\mu)/ n]$

X_r= r^{th} data point r =1 to n

μ = Mean of all data points

n = Number of data points

Mean	3.596563	3.21725	17.84875
Median	3.695	3.325	17.71
Mode	3.92	3.44	17.02
Standered Diviation	0.534679	0.978457443	1.786943
Varience	0.285881	0.957378968	3.193166
Range	2.17	3.911	8.4



Q8) Calculate Expected Value for the problem below

a) The weights (X) of patients at a clinic (in pounds), are
108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

Ans: 145.3333

Q9) Calculate Skewness, Kurtosis & draw inferences on the following data

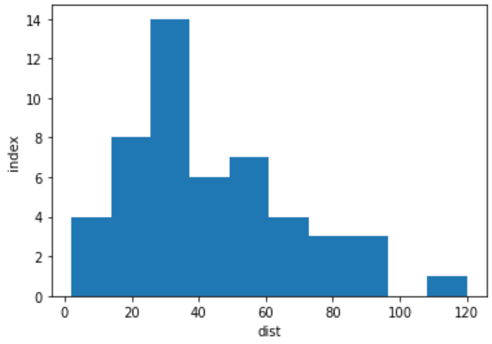
Cars speed and distance

Use Q9_a.csv

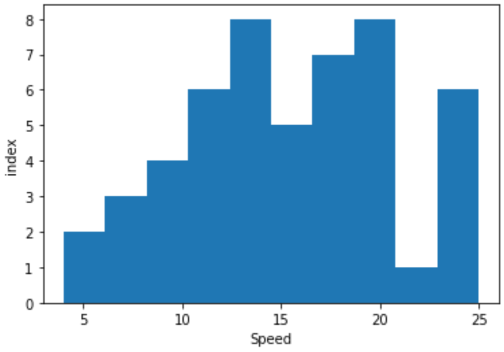
Index	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10
7	10	18
8	10	26
9	10	34
10	11	17
11	11	28
12	12	14
13	12	20
14	12	24
15	12	28
16	13	26
17	13	34
18	13	34
19	13	46
20	14	26
21	14	36
22	14	60
23	14	80
24	15	20
25	15	26
26	15	54
27	16	32
28	16	40
29	17	32
30	17	40
31	17	50
32	18	42
33	18	56
34	18	76
35	18	84
36	19	36
37	19	46
38	19	68
39	20	32
40	20	48

41	20	52
42	20	56
43	20	64
44	22	66
45	23	54
46	24	70
47	24	92
48	24	93
49	24	120
50	25	85

A) Cars distance inferences:



B) Cars speed covered



A) In car distance we observed skewness is occurred at the right side , so its means we found positive skewness . Most of the vales lies in the left side of the plot.

B) In car speed we observed skewness is occurred at the left side, so it means we found nagative skewness. Most of the vales lies in the right side of the plot.

C) In the car distance and car speed we observed that plots peak is more than the normal distribution.

	Car speed	Car distance
Skewness	-0.113955	0.7824835
Kurtosis	-0.6730924	0.119397

SP and Weight(WT)

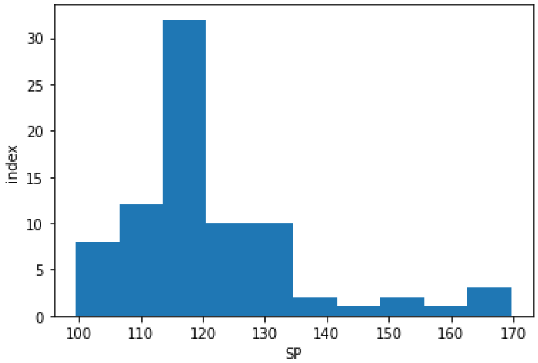
Use Q9_b.c

SP	WT
104.1854	28.76206
105.4613	30.46683
105.4613	30.1936
113.4613	30.63211
104.4613	29.88915
113.1854	29.59177
105.4613	30.30848
102.5985	15.84776
102.5985	16.35948
115.6452	30.92015
111.1854	29.36334
117.5985	15.75353
122.1051	32.81359
111.1854	29.37844
108.1854	29.34728
111.1854	29.60453
114.3693	29.53578
117.5985	16.19412
114.3693	29.92939
118.4729	33.51697
119.1051	32.32465
110.8408	34.90821
120.289	32.67583
113.8291	31.83712
119.1854	28.78173
114.5985	16.04317
120.7605	38.06282
119.1051	32.83507
99.56491	34.48321
121.8408	35.54936
113.4846	37.04235
112.289	33.23436
119.9211	31.38004

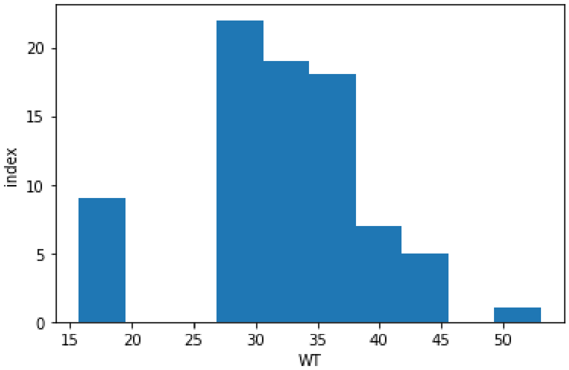
121.3926	37.57329
111.289	32.70164
115.0131	31.91122
114.0934	28.754
116.9094	27.87992
116.9094	28.6305
128.4613	30.11543
116.3926	37.39252
115.7488	35.02718
117.4613	30.52743
114.0934	28.34398
114.381	33.07863
117.1051	32.62192
118.2087	36.49862
116.4729	33.91006
127.9094	28.0706
118.289	33.45847
118.289	33.21395
118.289	33.43671
120.4043	40.39816
143.3926	37.62069
135.3926	37.25439
126.4043	40.58907
110.4613	30.14754
118.289	32.73452
112.6452	30.61528
115.5766	37.66287
130.2087	36.88815
117.6685	37.86041
126.0481	43.39099
125.3123	40.72283
128.1284	40.15948
126.5985	15.71286
132.4846	37.97996
133.6802	41.57397
133.3123	40.47204
158.3007	37.14173
164.5985	15.82306
133.416	44.01314
133.1401	43.35312
124.7152	52.99775

121.8642	42.6187
132.8642	42.77822
169.5985	16.13295
150.5766	37.92311
151.5985	15.76963
167.9445	39.4231
139.8408	34.94861

A) SP



B) WEIGHT



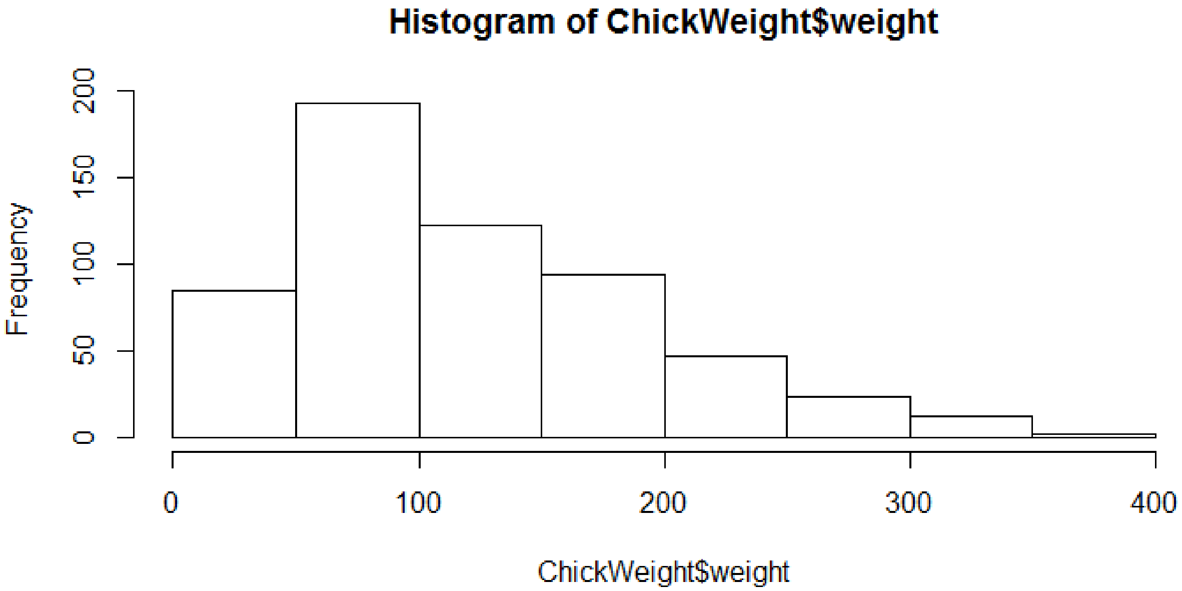
A) In SP we observed skewness is occurred at the right side , so its means we found positive skewness abd most of the data lies in the left side of the plot.

B) In weight we observed skewness is occurred at the left side, so it means we found negative skewness and most if the fata lies in the right side of the plot.

C) In SP and we observed that plots peak is more than the normal distribution but in Weight we observed that plot peak is nearly equal to normal distribution

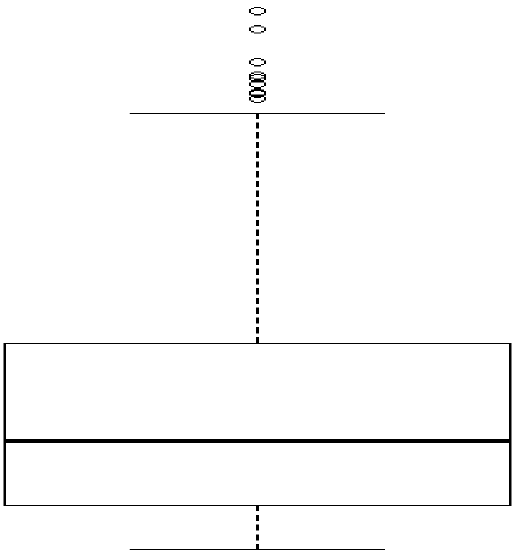
	SP	Weight(WT)
Skewness	1.581454	-0.6033099
Kurtosis	5.723521	3.819466

Q10) Draw inferences about the following boxplot & histogram



Inferences:

- Chick weight data is rightly skewed or it is having positive skewness. There is less concentration of chick weight is in the 300 to 400 grams
- More than 50% of Chick weight is lies between 50gm to 150gm range. And maximum chick weight is 100gm and frequency is 200.



Inferences:

- The data is rightly skewed.
- The outliers are present at the upper side.
- The median is less than the mean.

Q11) Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

Ans: Here we use the formula of Z test

i.e $\bar{x}$ (+ or -) $Z_{1-\alpha} = \frac{\boldsymbol{\sigma}}{\sqrt{n}}$

where **σ =** Standered diviation of population

$\bar{x}$ = Sample average or mean

n = Number of sample

For 94% of confidence interval (CL)

$\alpha = \frac{\mathbf{1+CL}}{2} = \frac{\mathbf{1+0.94}}{2} = 0.97$

similarly ,

for 98%, $\alpha = 0.99$

for 96%, $\alpha = 0.98$

after calculation we got Z values and interval or Range

Confidence interval	Z value	Range
94%	1.88079	134.850 , 265.149
96%	2.05374	122.651 , 277.348
98%	2.32634	130.153 , 269.846

Q12) Below are the scores obtained by a student in tests

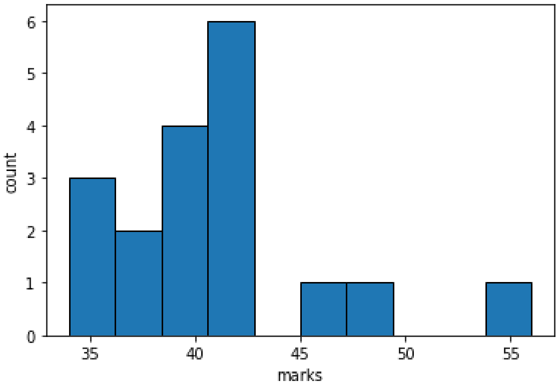
34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

1) Find mean, median, variance, standard deviation.

Ans: 1)

Mean	41
Median	40.5
Sandered diviation	5.052664
Varience	25.52941

2) What can we say about the student marks?



Here marks are not normally distributed. The person with mark 56 can be outlier in the data. The students marks is medicore. The average percentage of the students in the class is 42%.

Q13) What is the nature of skewness when mean, median of data are equal?

Ans: Data is normaly distributed and the skewness is symmetrical.

Q14) What is the nature of skewness when mean > median ?

Ans: When mean> median the skewness will be positive skewness. The skewness occurred at the right side of the plot and most of the data present in the left side of the plot. Skewness influence the mean.

Q15) What is the nature of skewness when median > mean?

Ans: When median > mean the skewness will be negative skewness. The skewness occurred at the left side of the plot and most of the data present in the right side of the plot.

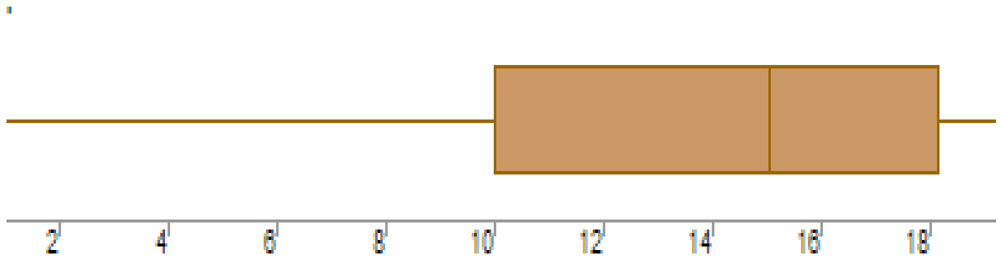
Q16) What does positive kurtosis value indicates for a data ?

Ans: Positive kurtosis value indicates that thinner peak and wider tails than the normal distribution.

Q17) What does negative kurtosis value indicates for a data?

Ans: Negative kurtosis value indicates that wider peak and thinner tails than the normal distribution..

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

Ans: Not normally distribuited, the outliers might be influencing to the data

What is nature of skewness of the data?

Ans: Negative skewness

What will be the IQR of the data (approximately)?

Ans: Q1 = 10

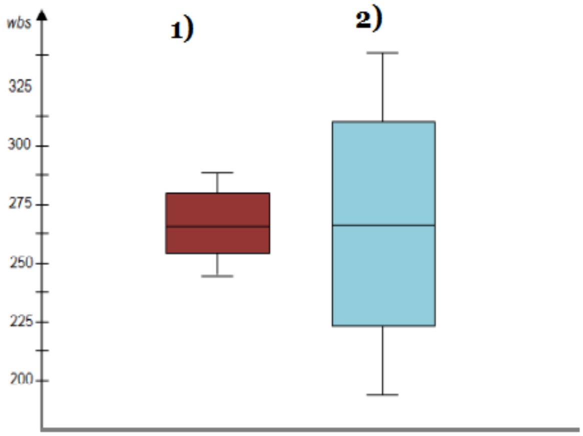
Q3 = 18

IQR = Q3 – Q1 = 18 – 10 = 8

Q19) Comment on the below Boxplot visualizations?

Here there is representation of two box plots. The data present in box plot of (2) is more and also its more distributed compare to box plot (1), the data present in box plot (1)is lightly less and distribution is also less.

In the diagram (2) the data is spread 100% across the values lies in the range 220 – 310. In diagram (1) the values lies in the range of 250 – 290.



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

- The spread of the data in box plot 1 is wide compare to box plot 2.
- The data range is varies high in box plot 2.
- The data range is varies slightly less in box plot 1 compare to box plot 2
- We can easily make prediction in box plot 1, in box plot 2 making of prediction is hard.
- The median in the two box plot are equal.
- The data spread in both box plots are symmetrical.

Q 20) Calculate probability from the given dataset for the below cases
Data _set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

a. P(MPG>38)
b. P(MPG<40)
c. P (20<MPG<50)

Ans: Using python

a) P(MPG>38)

= mean(MPG) = 34.12208 = loc

= sd(MPG) = 9.131445 = scale

= 1- stats.norm.cdf(x= 38, loc=34.12208, scale=9.131445)

= 0.34 = 34%

b) P(MPG<40)
= mean(MPG) = 34.12208 = loc

= sd(MPG) = 9.131445 = scale

= 1- stats.norm.cdf(x= 40, loc=34.12208, scale=9.131445)

= 0.270 = 27%

c) P (20<MPG<50)

=mean(MPG) = 34.12208 = loc

= sd(MPG) = 9.131445 = scale

= stats.norm.cdf(x= 20,loc=34.12208,scale=9.131445)

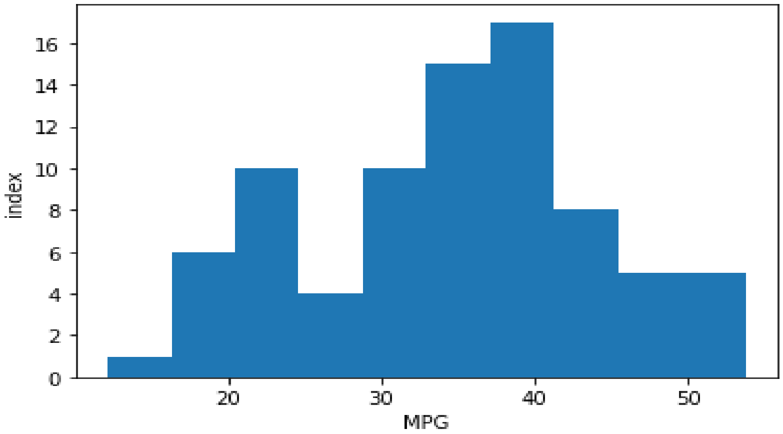
- stats.norm.cdf(x= 50, loc=34.12208, scale=9.131445)

= 0.89886

= 89%

Q 21) Check whether the data follows normal distribution

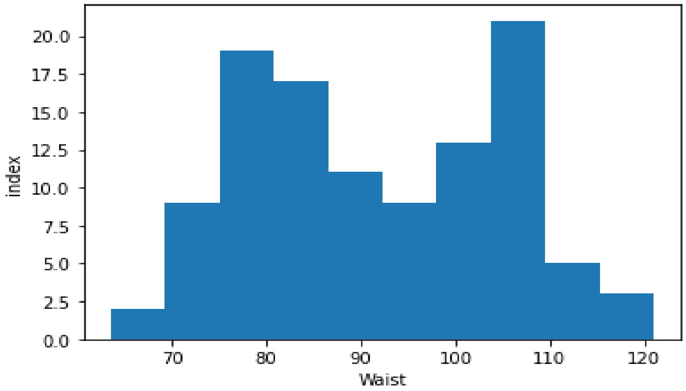
a) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

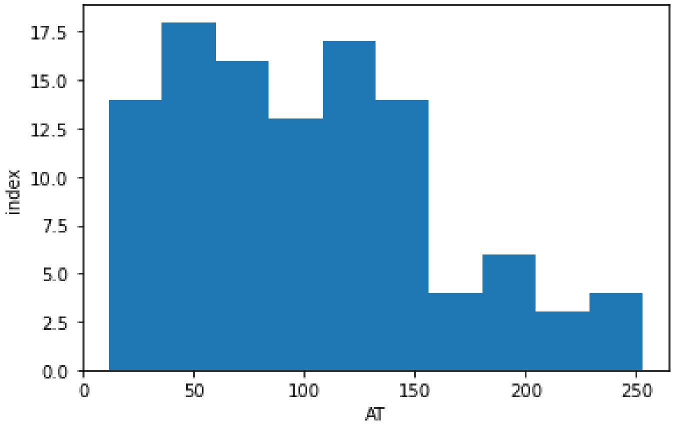


It follows normal distribution

b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution
Dataset: wc-at.csv

Ans:





Adipose Tissue(AT) and Waist Circumference(Waist) not follows normal distribution.

Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

Ans: Z scores

- 90% of confidance interval
  $\alpha = \frac{\mathbf{1+CL}}{2} = \frac{\mathbf{1+0.90}}{2} = 0.95$
  similarly,
  for 94%, α = 0.97
  60%, α = 0.80

Confidence interval	Z scores
90%	1.64485
94%	1.88079
60%	0.84162

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

Ans: t scores

Here n=25 => n-1 =24 =degrees of freedom

- 95% of confidence interval

  α= $\frac{\mathbf{1+CL}}{2}$ = $\frac{\mathbf{1-0.95}}{2}$ = 0.975
  similarly,
  for 96% , α=0.98
  for 99%, α=0.995

Confidence interval	t scores
95%	2.063899
96%	2.171545
99%	2.79694

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode → pt(tscore,df)

df → degrees of freedom

Ans: Here, sample size(n) = 18

Sample mean(X) = 260 days

Population Mean(µ) = 270 days

Standard Diviation(S) = 90 days

t = $\frac{\mathbf{X-\mu}}{s/\sqrt{n}}$

= -0.4714

Using stats.t.cdf(tscore, df)

We get probability as 0.3216=32.16%