Author: **Shreyas Shashikant Vaishnav**
Purpose: **Assignment 1**

Basic statistics 1

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	**Discrete**
Number of kids	**Discrete**
Number of tickets in Indian railways	**Discrete**
Number of times married	**Discrete**
Gender (Male or Female)	**Discrete**

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	**Interval**
Weight	**Ratio**
Hair Color	**Ratio**
Socioeconomic Status	**Interval**

Fahrenheit Temperature	**Ratio**
Height	**Ratio**
Type of living accommodation	**Ordinal**
Level of Agreement	**Interval**
IQ(Intelligence Scale)	**Interval**
Sales Figures	**Interval**
Blood Group	**Ratio**
Time Of Day	**Interval**
Time on a Clock with Hands	**Interval**
Number of Children	**Interval**
Religious Preference	**Ratio**
Barometer Pressure	**Ratio**
SAT Scores	**Ratio**
Years of Education	**Nominal**

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

Sol:-

If three coins are tossed:

{HHH, HHT, HTH, THH, HTT, TTH, THT, TTT}

Total Result = 8

Two Heads & One Tail Result = 3 = { HHT, HTH, THH }

Hence, 3/8 = 0.375

Probability of getting Two Heads and One Tail is 37.5%

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

a) Equal to 1
b) Less than or equal to 4

c) Sum is divisible by 2 and 3

Sol:-

If two dice are rolled, then the total outcome will be {6*6 = 36}

a) There is no outcomes which corresponds sum is equal to one, i.e. 0/36.

Probability is 0.

b) There six possible outcome that are less than or equal to 4,

{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)} = **6**

6/36 = 1/6 or 16%

c) Probability of its sum is divisible by 2 and 3:

{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1), (6, 6)} = **6/36 = <u>16%</u>**

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?

Sol: - Total no. of balls = 2+3+2 = **7**

Let S be sample space.

n(S) = No. of ways of drawing 2 balls out of 7 = (7/2) x (6/1) = **21** [7 x 6 / 2 x 1]

Let E be the event of drawing 2 balls and none of them is blue.

n(E) = No. of ways of drawing 2 balls out of 5 balls excluding blue =

(5/2) x (4/1) = **10**

The Probability of drawing two balls and none of them is blue = **n(E) / n(S)** = **10/21 = 0.47=<u>47%</u>**

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

Sol: - Expected number of candies for a randomly selected child:

= 1*0.015 + 4*0.20 + 3 *0.65 + 5*0.005 + 6 *0.01 + 2 * 0.12 = 3.090

So, the Expected number of candies for a randomly selected child is <u>3.09</u>

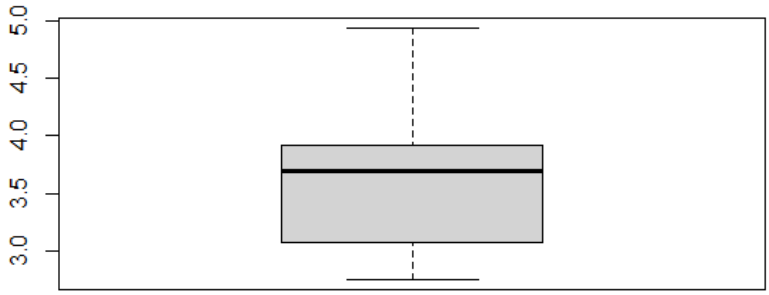
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

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

Sol:-

	Points	Score	Weight
Mean	3.596	3.217	17.848
Median	3.695	3.325	17.71
Mode	3.891	3.54	17.43
Variance	0.285	0.957	3.19
Standard Deviation	0.534	0.978	1.786

Inferences:

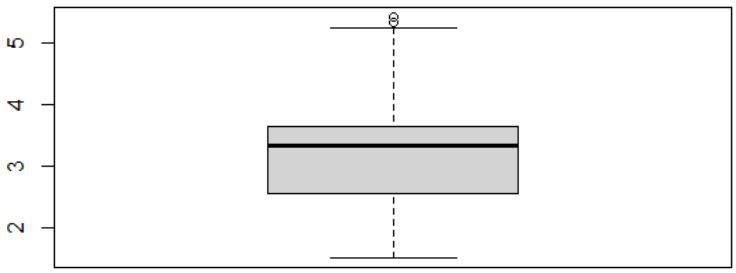


1) For Points dataset:

- The data is concentrated around Median.

- There are no Outliers.

- The distribution is Right skewed.

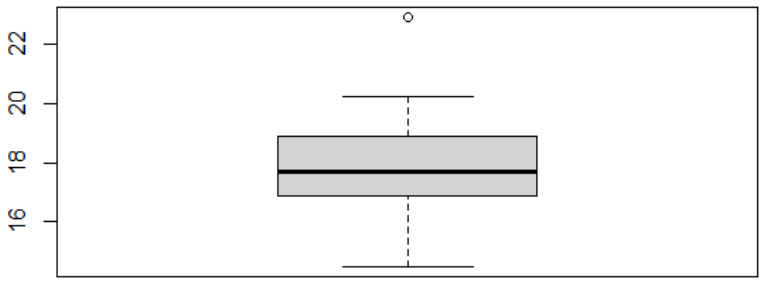


2) For Score dataset:

- The data is concentrated around Median.

- There are 3 Outliers: 5.250, 5.424, and 5.345.

- The distribution is slightly right skewed



3) For Weigh dataset:

- The data is concentrated around Median

- There is 1 Outlier: 22.90

- The distribution is Left skewed

Q8) Calculate Expected Value for the problem below.

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?

Sol:-

Expected Value = ∑ (probability * Value)

∑ P(x).E(x)

There are 9 patients

Probability of selecting each patient = 1/9

Ex 108, 110, 123, 134, 135, 145, 167, 187, 199

P(x) 1/9 1/9 1/9 1/9 1/9 1/9 1/9 1/9 1/9

Expected Value = (1/9) (108) + (1/9) 110 + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199

= (1/9) (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

= (1/9) (1308)

= 145.33

Expected value of the weight of that patient = 145.33

Scripting:

>>import numpy as np

>>np.array([108, 110, 123, 134, 135, 145, 167, 187, 199]).mean()

>>145.33333333333334

Q9) Calculate Skewness, Kurtosis & draw inferences on the following data - Cars speed and distance

Use Q9_a.csv

Sol: -

1) For cars speed and distance:-

Skewness for speed= -0.117510, skewness value is negative so it is left skewed. Since magnitude is slightly greater than 0 it is slightly left skewed.

	Speed	**Distance**
Skewness	-0.117510	0.806895
Kurtosis	-0.508994	0.405053

Skewness for distance= 0.806895, right skewed (Positive) slight magnitude to right**.**

Kurtosis Inference:

1. Speed distribution is ***platykurtic*** (negative kurtosis i.e. flatter than normal distribution)

2. Distance distribution is ***leptokurtic*** (positive kurtosis i.e. peaked than normal distribution)

2) For cars speed and weight

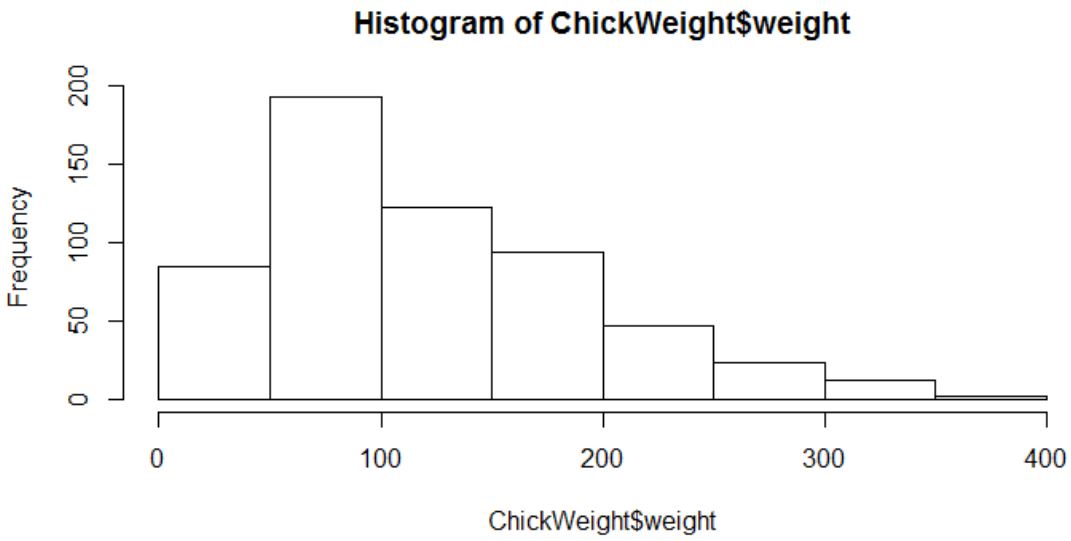
	SP	**WT**
Skewness	1.611450	-0.614753
Kurtosis	2.977329	0.405053

Skewness for speed = 1.611450, Skewness value is positive so it is right skewed
Skewness for weight = -0.614753 Weight distribution is left skewed (negative skewness)

Kurtosis Inference:

1. Speed distribution is ***leptokurtic*** (positive kurtosis i.e. peaked than normal distribution)
2. Weight distribution is ***leptokurtic*** (positive kurtosis i.e. peaked than normal distribution)

Q10) Draw inferences about the following box plot & histogram



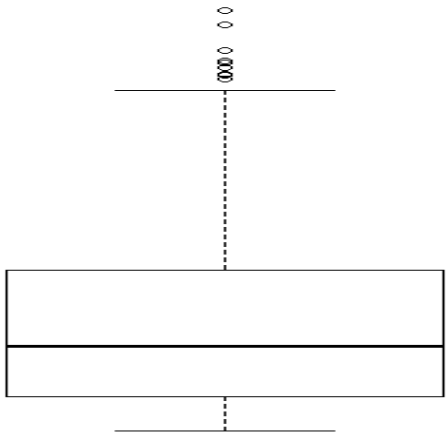
Sol:-

The most of the data points are concerted in the range 50-100 with frequency 200.

And least range of weight is 400 somewhere around 0-10.

So the expected value the above distribution is 75.

Skewness- we can notice a long tail towards right so it is heavily right skewed.



Sol:-

Median is less than mean, positively skewed and we have outliers on the right side of box plot and there is less data points between Q1 and bottom point.

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?

Sol:-

confidence interval	**Z value**	**Range**
confidence interval 94%	1.880794	198.74,201.26
confidence interval 96%	2.053749	198.62,201.38
confidence interval 98%	2.326348	198.43,201.56

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.
2) What can we say about the student marks?

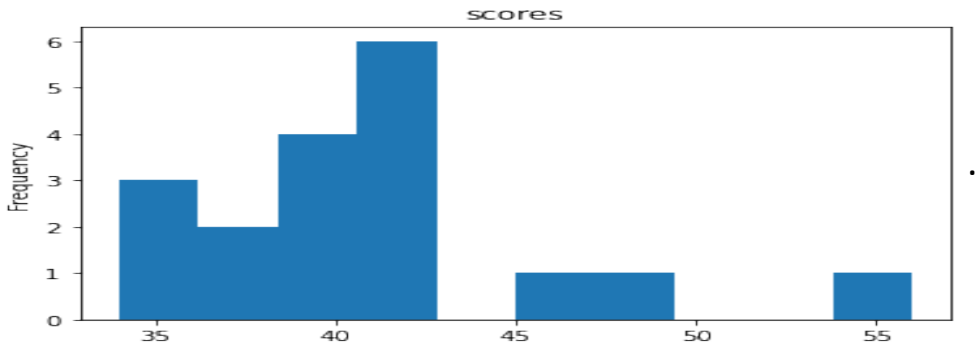
Sol:-

1)

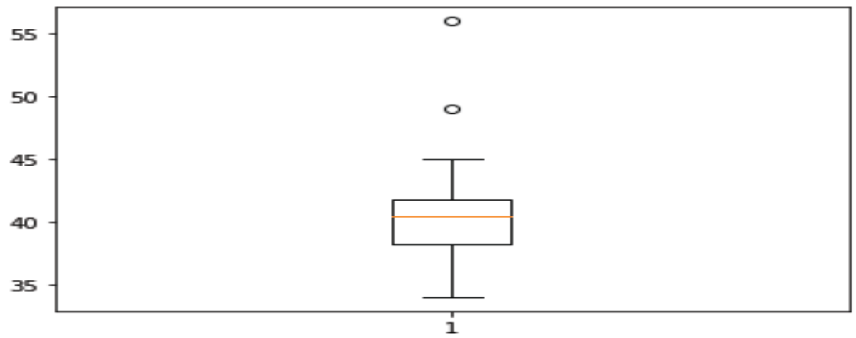
Mean	**41**
Median	**40.5**
Variance	**25.52**
Standard deviation	**5.05664**

Inference: There are 2 Outliers in Student's marks: 49 and 56

2) Mass of students marks between 38 – 42



Skewness is positive because mass of marks in right side of plot.



*Inference: 1. There are 2 Outliers in Student's marks: 49 and 56

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

Sol: - Symmetrical.

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

Sol: - Right Skewed.

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

Sol: - Left Skewed.

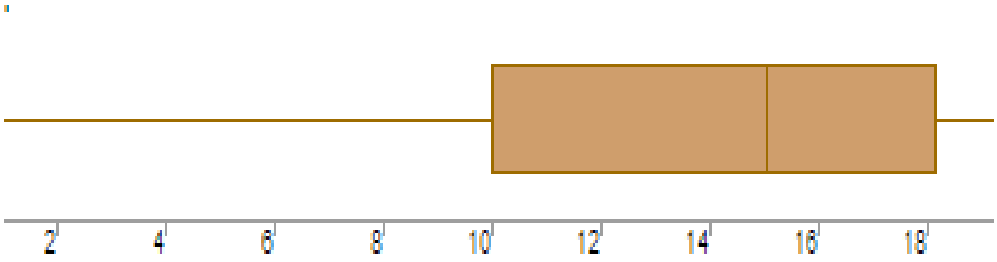
Q16) What does positive kurtosis value indicates for a data?

Sol: - The data is normally distributed and kurtosis value is 0.

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

Sol: - The distribution of the data has lighter tails and a flatter peaks than the normal distribution.

Q18) Answer the below questions using the below box plot visualization.



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

Sol:-

A) What can we say about the distribution of the data?

Sol: Let's assume above box plot is about ages of the students in a school.

50% of the people are above 10 yrs. old and remaining are less.

And students whose age is above 15 are approx. 40%.

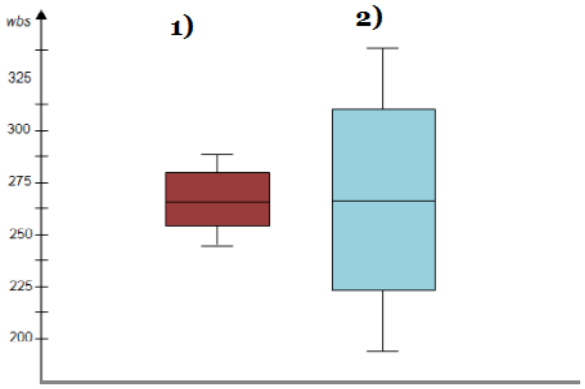
B) What is nature of skewness of the data?

Sol: Left skewed, median is greater than mean.

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

Sol: Approximately -8

Q19) Comment on the below Box plot visualizations?



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

Sol: By observing both the plots whisker's level is high in box plot 2, mean and median are equal hence distribution is symmetrical.

Q20) 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)

Sol)

A] 1 - stats.norm.cdf(x = 38, loc= 34.42, scale=9.13)

>> Probability of MPG of car less than 38 is 35%

B] stats.norm.cdf(x = 40, loc= 34.42, scale=9.13)

>> Probability of MPG of car more than 40 is 72%

C] stats.norm.cdf(x = 50 , loc= 34.42, scale=9.13) - stats.norm.cdf(x = 20 , loc= 34.42, scale=9.13)

>> Probability of MPG of car between 20 to 50 is 89%

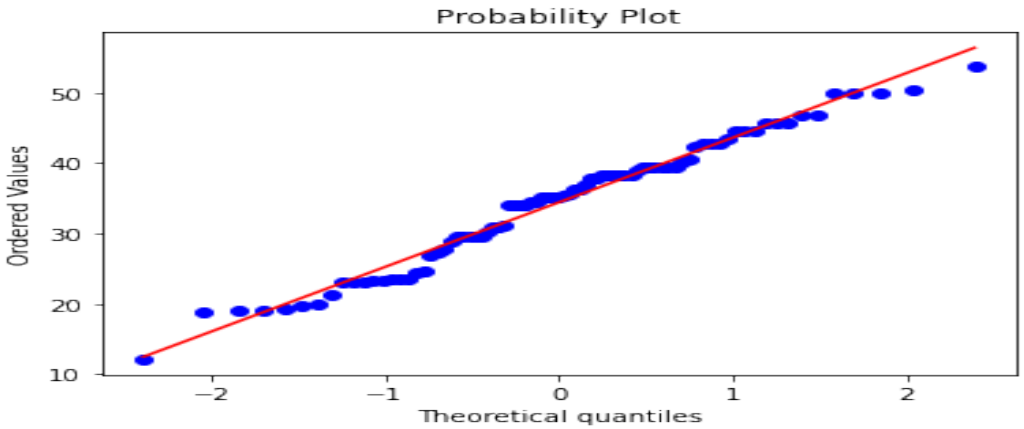
Q21) Check whether the data follows normal distribution
Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

Sol:-

Follows Normal distribution as indicated by ***QQ-plot.***

>> stats.probplot(cars_data.MPG, plot=plt)

>> plt.show()



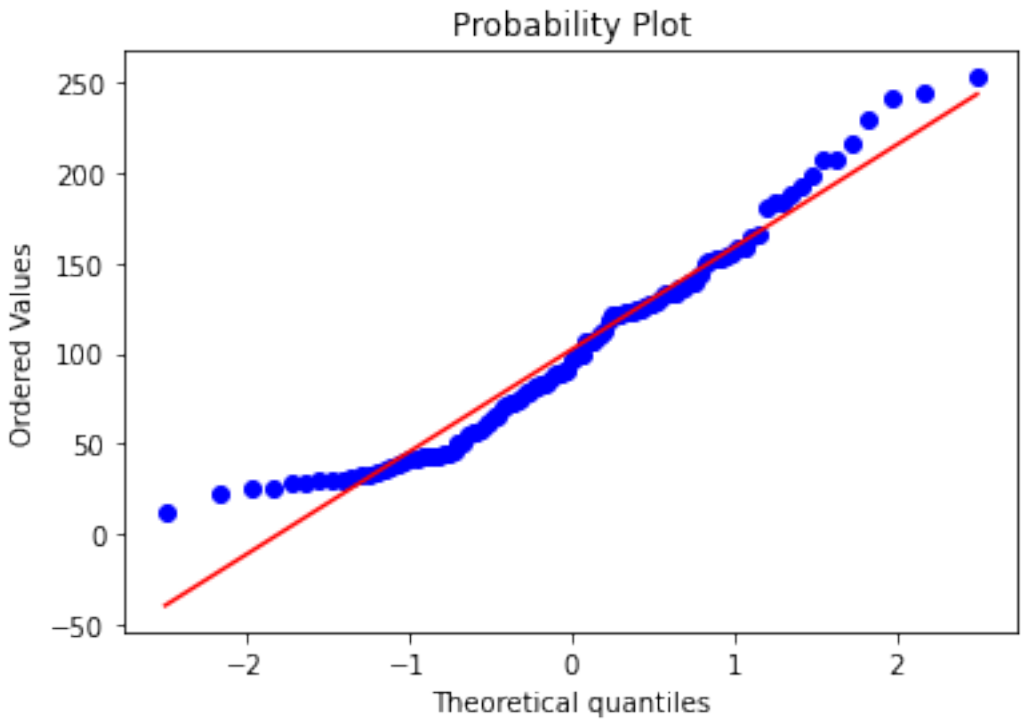
- Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution.
  Dataset: wc-at.csv

Sol:- For Adipose Tissue (AT)

Follows Normal distribution as indicated by ***QQ-plot.***

>> stats.probplot(wtat_data.AT, plot=plt)

>> plt.show()

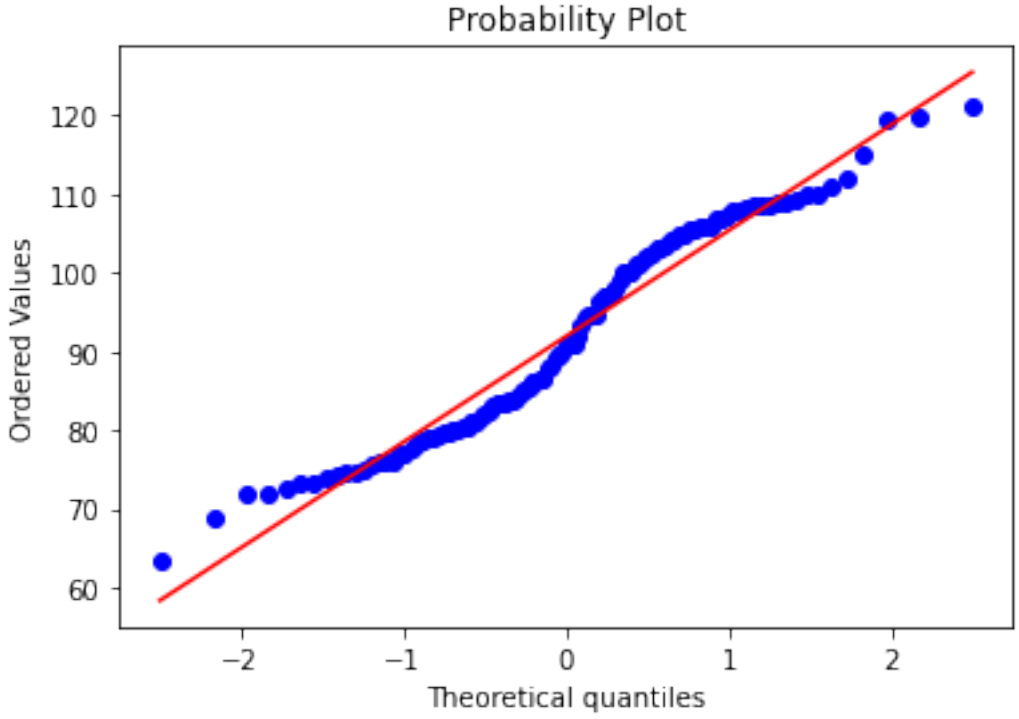


For Waist Circumference(Waist)

Follows Normal distribution as indicated by ***QQ-plot.***

>> stats.probplot(wtat_data.Waist, plot=plt)

>> plt.show()



Q22) Calculate the Z scores of 90% confidence interval, 94% confidence interval, 60% confidence interval

SOL:-

For 90% = >> norm.ppf(0.95) = 1.6448
For 94% = >> norm.ppf(0.97) = 1.8807
For 60% = >> norm.ppf(0.8) = 0.8416

Q23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

SOL:-

For 95% = >> stats.t.ppf(0.975,24) = 2.0638

For 96% = >> stats.t.ppf(0.98,24) = 2.1715

For 99% = >> stats.t.ppf(0.995,24) = 2.796

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

SOL:-

μ=270, $\underline{x}$=260, SD=90, n=18, df=n-1=18-1= 17

tscore= $\frac{\underline{x}-\mu}{s/\sqrt{n}}$ **=** $\frac{260-270}{90/\sqrt{18}}$ **= -10/21.23= -0.47**

Hence,
>> pt(-0.47,17) = 0.3221639
Required probability = 0.32=32%
