ASSIGNMENT 1

ATHUL SHAN N A

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

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

ANS: Probability dist. of 3 coins = {HHH,HHT,HTH,THH,HTT,TTT,TTH,THT}

Favorable cases = {HHT,HTH,THH}

Probability = **3/8**

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

ANS:

a) Sum = 1 is not possible. Therefore, probability is **0**

b) Favorable cases: {(1,1),(1,2),(1,3),(2,1),(2,2)(3,1)}
   Total cases = 36
   Probability = 6/36 = **1/6**

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: Total ways of drawing 2 balls of 7 = 7C2 = 21 ways

Total ways of drawing 2 balls that are not blue = 5C2 = 10 ways

Probability that none of the balls drawn are blue = **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 :

CHILD	Expected Value

A	**0.015**
B	**0.8**
C	**1.95**
D	**0.025**
E	**0.06**
F	**0.24**
SUM	**3.09**

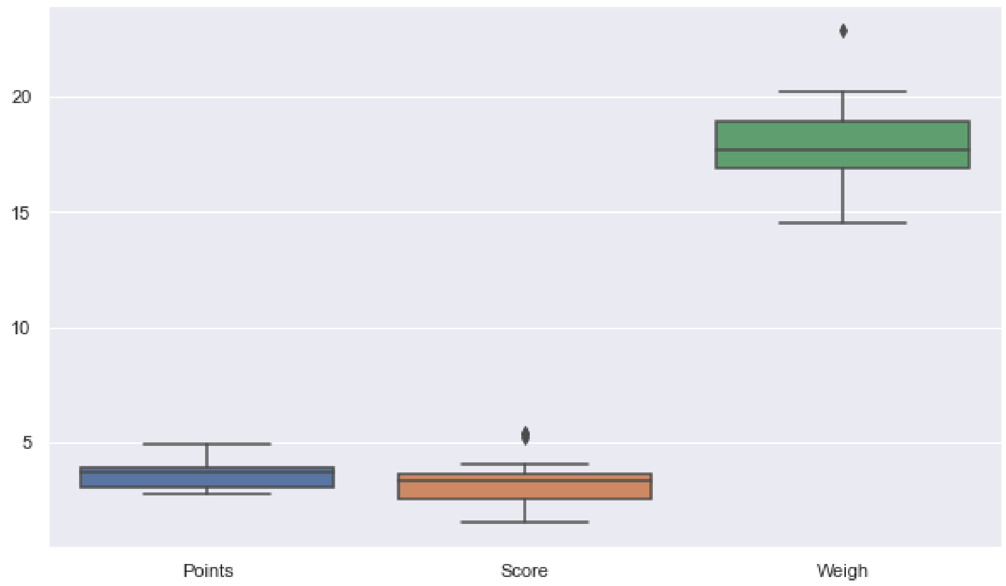
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

- 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

Function	Points	Score	Weight
Mean	3.596563	3.21724	17.848750
Mode	3.07	3.44	17.02
Median	3.695	3.325	17.71
Standard Deviation	0.534679	0.978457	1.786943
Variance	0.285881	0.957378	3.193166
Range	2.170	3.911	8.4

The above values are vintage car collections and its ratings. The Points of the cars lie between 4.93 (Honda Civic) and 2.76 (Dodge Challenger) with an average of 3.6. The scores have an average of 3.21 and its values are in between 5.42 (Lincoln Continental) and 1.51 (Lotus Europa). The Weights of the cars Have an average of 17.84 and range between 22.9 (Merc 230) and 14.5 (Ford Pantera L).



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 : Expected Value = 145.33

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

Cars speed and distance

Use Q9_a.csv

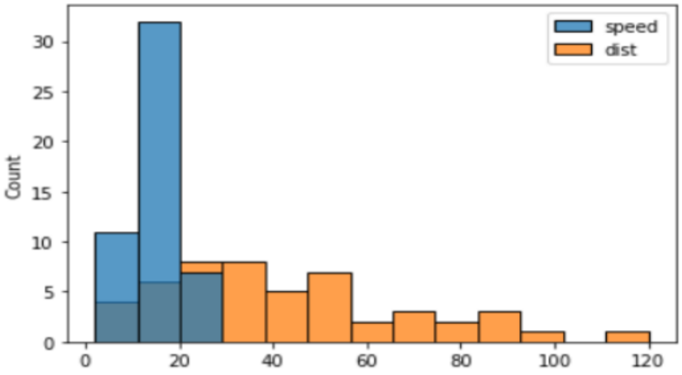
ANS :

Parameters	Skewness	Kurtosis
Speed	0.11395477	-0.57714742
Distance	0.78248352	0.24801866

Inference:

Speed has negative values for Skewness and Kurtosis which implies that it is left tailed with most values towards left. Kurtosis Values reveals that it is more or less situation around 14 or 15 range.

Distance has positive values for both Skewness and Kurtosis which implies that it is right tailed with extreme values to the right. Positive kurtosis reveals that there are extreme values in the distribution.



SP and Weight(WT)

Use Q9_b.csv

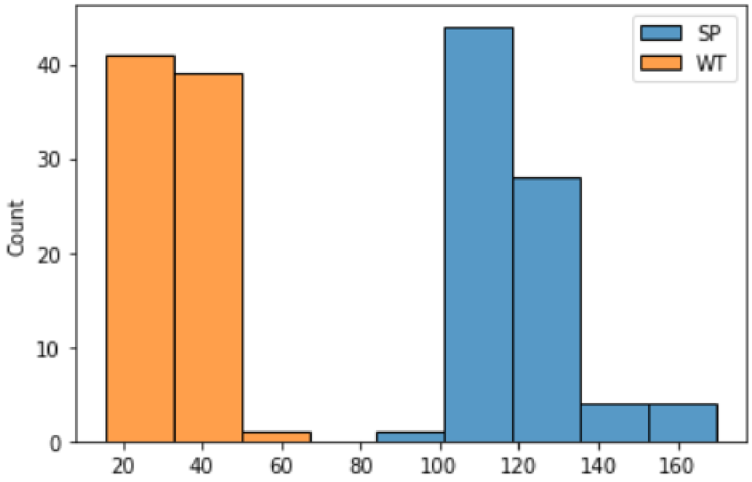
ANS:

Parameters	Skewness	Kurtosis
Speed	1.58145368	2.72352149
Weight	-0.60330993	0.81946588

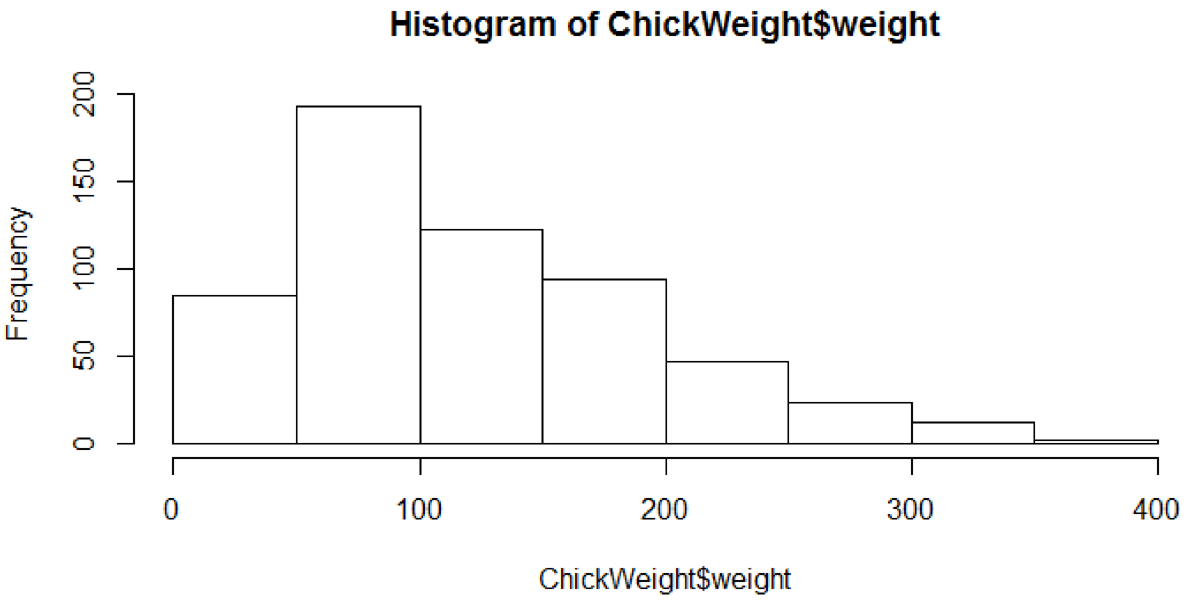
Inference:

SP has large positive values for both Skewness and Kurtosis which implies that it is right tailed with most values towards it right. Large Kurtosis values reveals that it contains extreme values.

WT has negative Skewness and positive Kurtosis which implies that it is left tailed with extreme values to the left. Positive kurtosis reveals that there are extreme values in the distribution in the range 20 to 35.

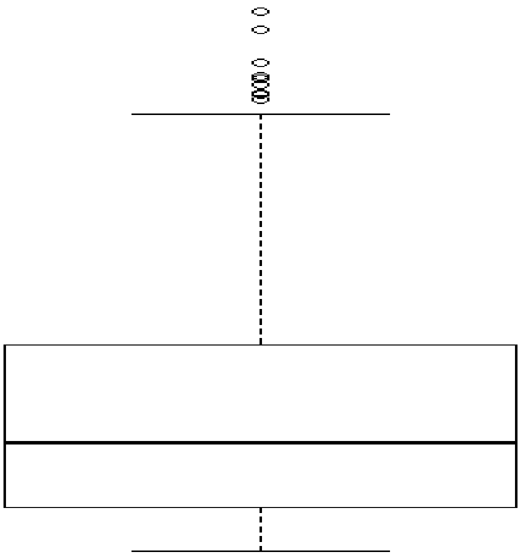


Q10) Draw inferences about the following boxplot & histogram



Inference: The distribution is a right tailed histogram with both positive kurtosis and skewness. The range 50-100 is the most observed chickweight. The graph gradually decreases for higher chickweights. There are a number of extreme values towards the right tail of the graph. Those values have very low frequency.

The mean in this case would be less than the median since most values are concentrated towards the left side.



Inference :

The boxplot distribution is closer to the lower limit of the graph. There are a number of extreme values (outliers) towards the upper limit of the graph. The outliers lie ouside of the boxplot. The median is close to the lower limit of the graph.

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:

import numpy as np

import scipy.stats as st

```
[round(x,2) for x in st.norm.interval(alpha=0.94, loc=200, scale=30)]
```

Confidence Interval	Interval
94%	143.58, 256.42
96%	130.21, 269.79
98%	138.39, 261.61

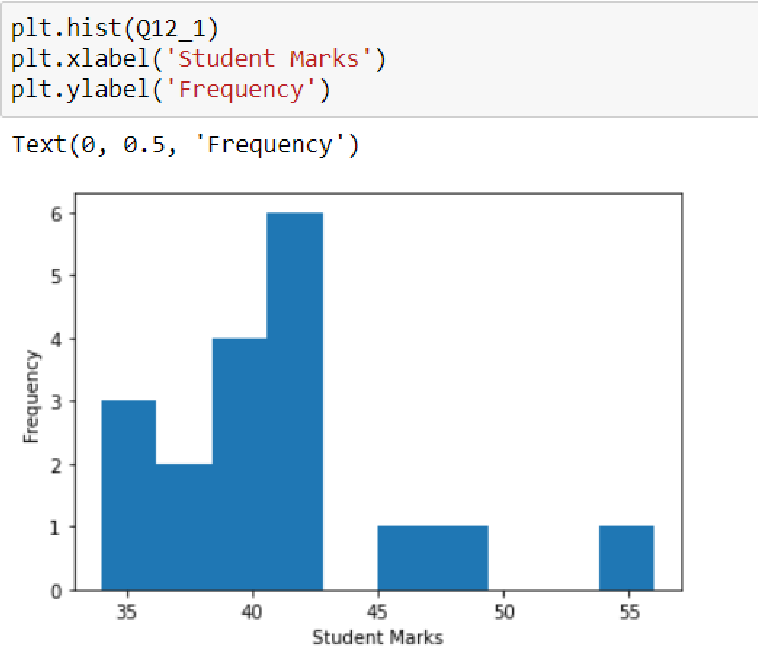
Q12) Below are the scores obtained by a student in tests

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?

ANS:

1) Mean = 41.00, Median= 40.5, Variance = 25.53, Standard Deviation = 5.052
2) Histogram plot of student marks



```
plt.hist(Q12_1)
plt.xlabel('Student Marks')
plt.ylabel('Frequency')
```

```
Text(0, 0.5, 'Frequency')
```

Inference: The graph reveals that most of the students scored between 40 to 43 Marks. Highest marks scored was 55 and lowest marks were below 35. Few students managed to score between 45 and 50 marks. Remaining all students scored below 43.5 marks making the mean and median centered closely around 40.

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

Ans: The data is symmetric and normalized. Therefore, there will be no skewness.

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

Ans: When mean is greater, it implies that the data is concentrated towards the right side and therefore will get a left tailed distribution- Negative skewness.

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

Ans: The data is concentrated towards the left side of the distribution and we get a positive skewness right tailed distribution.

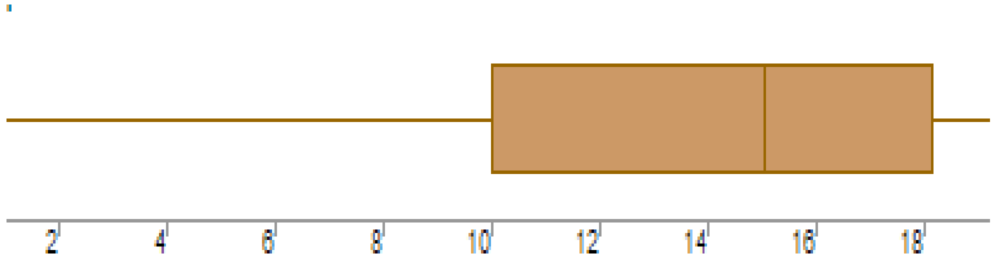
Q16) What does positive kurtosis value indicates for a data ?

Ans: It indicates wideness of tail and thin sharp peak of distribution.

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

Ans: It indicates thin tail and wide peak of data distribution.

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



What can we say about the distribution of the data?

Ans: The distribution of data is uneven and therefore not normally distributed.

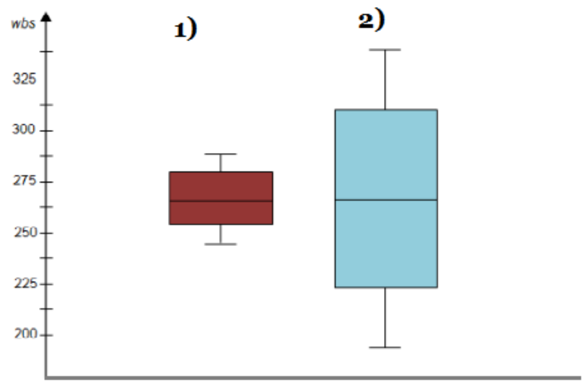
What is nature of skewness of the data?

Ans: Left tailed => Negative Skewness

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

Ans: Interquartile range will be 10 to 18

Q19) Comment on the below Boxplot visualizations?



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

Ans: In the case of boxplot(1), the max and min values lies approximately between 240 and 280 range whereas in boxplot(2), its between 200 and 330. The median for both the plots are almost same. The IQR in boxplot(2) is much greater than the IQR incase of boxplot(1).

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: The mean and standard deviation are required to calculate the probability using scipy.stats.norm.

Mean = 34.422076 and Std dev = 9.131445. The corresponding CDF is found out for each of the following:

a. P(MPG>38)

Ans: round(1 - ss.norm.cdf(38, 34.422076,9.131445), 4)

= 0.3476

b. P(MPG<40)

Ans: round(ss.norm.cdf(40, 34.422076,9.131445), 4)

= 0.7293

c. P(20<MPG)

Ans: round(ss.norm.cdf(50, 34.422076,9.131445) - (1 - ss.norm.cdf(20, 34.422076,9.131445)), 4)

= 0.0131

Q 21) Check whether the data follows normal distribution

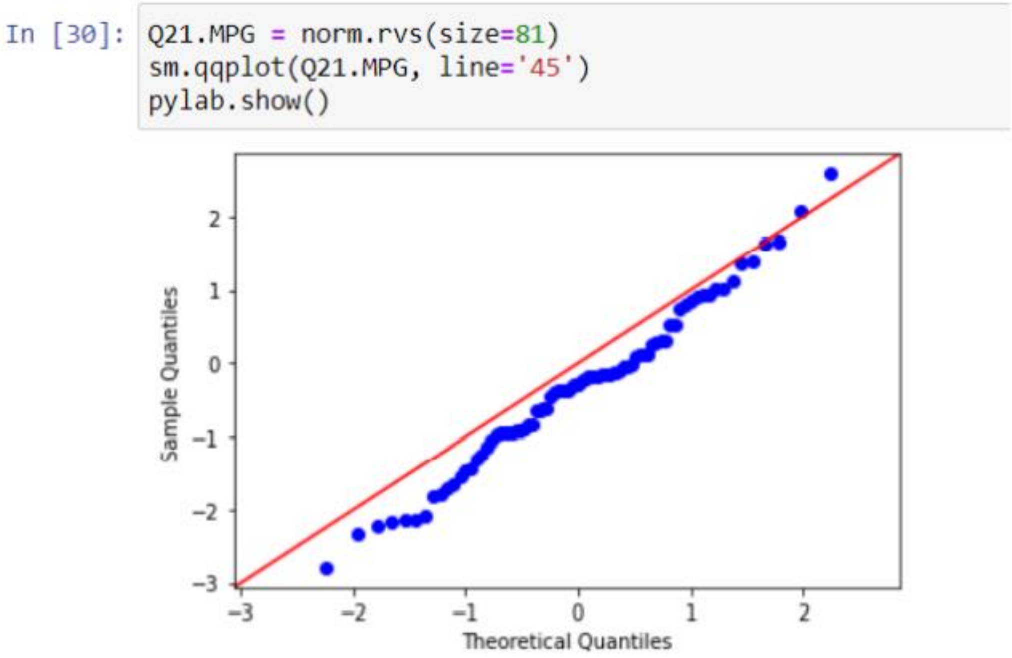
Using Shapiro Wilk test, we can check for normal dist. scipy.stats.shapiro(x)

- If the P-Value of the Shapiro Wilk Test is larger than 0.05, we assume a normal distribution.
- If the P-Value of the Shapiro Wilk Test is smaller than 0.05, we do not assume a normal distribution

Using QQ plot we can verify our result if the graph is a straight line through origin at 45 Deg.

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

Ans: scipy.stats.shapiro(MPG) => Pvalue: 0.2856500446796417 > 0.05
Therefore, MPG follows Normal Distribution.



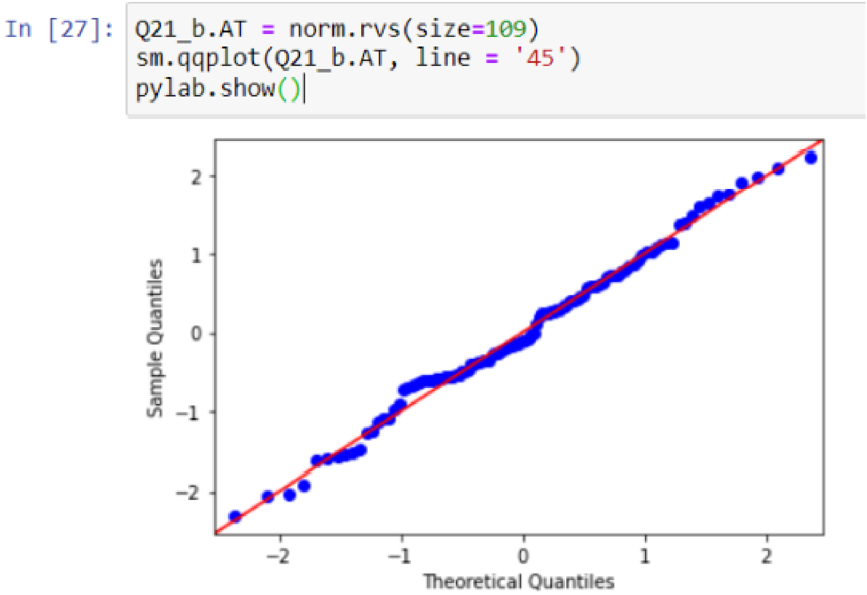
```
In [30]: Q21.MPG = norm.rvs(size=81)
         sm.qqplot(Q21.MPG, line='45')
         pylab.show()
```

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

Dataset: wc-at.csv

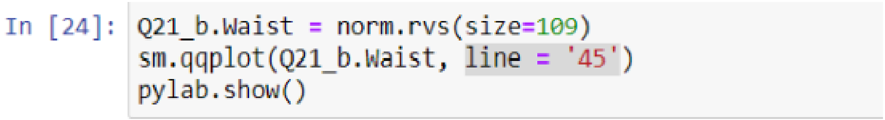
Ans: scipy.stats.shapiro(AT) => Pvalue: 0.1867150068283081 > 0.05

Therefore, AT follows Normal Distribution

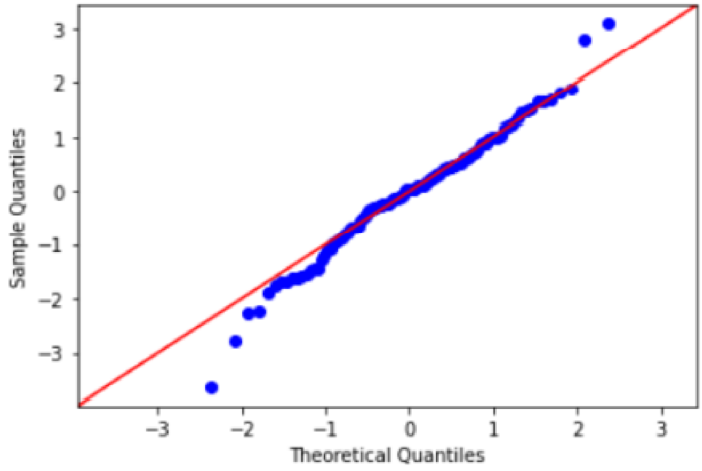


```
In [27]: Q21_b.AT = norm.rvs(size=109)
         sm.qqplot(Q21_b.AT, line = '45')
         pylab.show()
```

scipy.stats.shapiro(Waist) => Pvalue: 0.2651873826980591 > 0.05
Therefore, Waist also follows Normal Distribution.



```
In [24]: Q21_b.Waist = norm.rvs(size=109)
         sm.qqplot(Q21_b.Waist, line = '45')
         pylab.show()
```



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

ANS:

Confidence Interval	Z-Score
60%	0.8416212335729143
94%	1.8807936081512509
90%	1.6448536269514722

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

ANS:

Confidence Interval	T-Score
95%	2.0638985616280205
96%	2.1715446760080677
99%	2.796939504772804

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:

It is required to calculate the probability of n sample bulbs that have life < 260

Population Mean = 270

Sample Mean = 260, Sample Std Dev = 90, n = 18 ➔ t-score should be calculated.

Df = n -1 = 17

T-score = (Sample Mean – Pop Mean)/(SSD/(sqrt18))

T-score = -0.4714045207910317

Using T dist table, corresponding probability P➔

P = 0.32167253567098364

P is approx. 32.16%