NAME: KANAKALA TARUN KUMAR

Email id: [kanakalatarunkumar0212@gmail.com](mailto:kanakalatarunkumar0212@gmail.com)

ASSIGNEMNT 1

Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discreate |
| Results of rolling a dice | Discreate |
| 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 | Discreate |
| Number of tickets in Indian railways | Discreate |
| Number of times married | Discreate |
| Gender (Male or Female) | Discreate |

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?

**Answer:**

N(t)=2^3=8

N(outcome)=3(HHT,HTH,THH)

P=3/8=0.375

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

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

**Answer:**

1. P(a)=0(minimum sum that can be obtained when rolling two dice is 2)
2. P(b)=6/36 ( total =36, outcome=(1,1),(1,2),(1,3),(2,1),(2,,2),(3,1)

= 1/6

= 0.16

1. P(c)= 6/36( total =36 outcome= (1,5), (2,4), (3,3), (4,2), (5,1), and (6,6))

= 1/6

= 0.16

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?

**Answer:**

P(outcome)= 5c2/7c2

= 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

**Answer:**

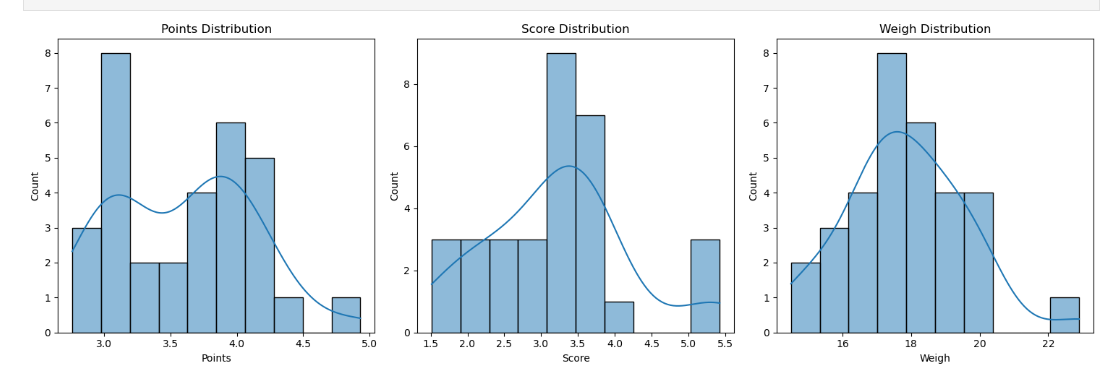
Expected number of candies = (1 x 0.015) + (4 x 0.20) + (3 x 0.65) + (5 x 0.005) + (6 x 0.01) + (2 x 0.120) = 3.205

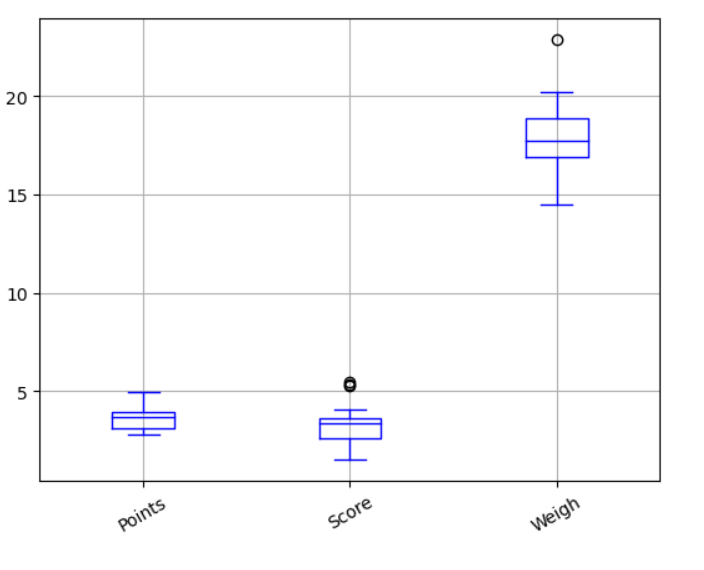
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

| **Functions** | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| Mean | 3.596563 | 3.217250 | 17.848750 |
| Median | 3.695 | 3.325 | 17.710 |
| Mode | 3.07 | 3.92 | 17.02 |
| Variance | 0.285881 | 0.957379 | 3.193166 |
| Standard Deviation | 0.534679 | 0.978457 | 1.786943 |
| Range | 2.170 | 3.911 | 8.400 |





points:

The mean and median are close, indicating a relatively symmetric distribution.

The mode is 3.07,3.92, suggesting that these values appear more frequently.

The variance and standard deviation are relatively low, indicating that the data points are clustered around the mean.

The range is 2.17, showing the spread of values within this column

Score:

The mean and median are relatively close, suggesting a somewhat symmetric distribution.

The mode includes two values, 3.44, indicating that this value are repeated.

The variance and standard deviation are moderate, indicating some dispersion of data points.

The range is 3.911, showing variability in this column.

weigh:

The mean and median are relatively close, suggesting a somewhat symmetric distribution.

The mode includes two values, 17.02 and 18.09, indicating that these values are repeated.

The variance and standard deviation are relatively high, indicating significant variation in the data.

The range is 8.4, indicating a relatively wide spread of values in this column.

**Use Q7.csv file**

Q8) Calculate Expected Value for the problem below

1. 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?

**Answer:**

X = {108, 110, 123, 134, 135, 145, 167, 187, 199}

Now, calculate the expected value (E[X]):

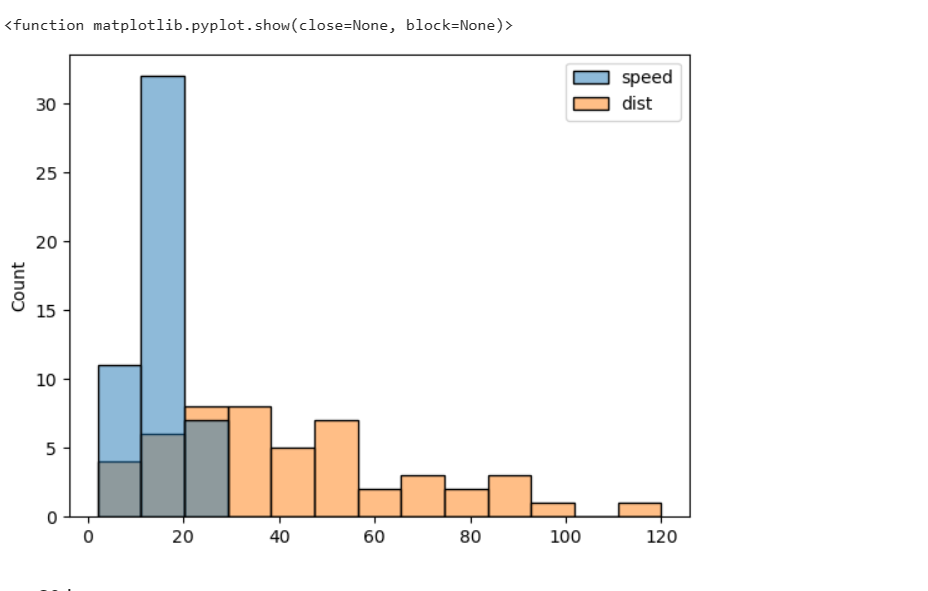
E[X] = (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199) / 9

≈ 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**



Skewness of Speed: -0.114

The skewness of speed is approximately -0.114, which is slightly negatively skewed. This suggests that the distribution of speeds is slightly skewed to the left, meaning that it has a longer tail on the left side.

Skewness of Distance: 0.782

The skewness of distance is approximately 0.782, which is positively skewed. This suggests that the distribution of distances is positively skewed, meaning that it has a longer tail on the right side.

Kurtosis of Speed: -0.577

The kurtosis of speed is approximately -0.577, which is less than 3. This indicates that the distribution of speeds has lighter tails (platykurtic) compared to a normal distribution. It has a flatter peak.

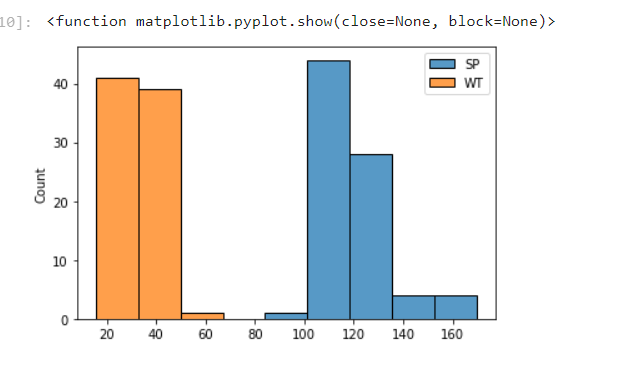
Kurtosis of Distance: 0.248

The kurtosis of distance is approximately 0.248, which is also less than 3. This indicates that the distribution of distances has lighter tails (platykurtic) compared to a normal distribution. It has a flatter peak.

In summary, both the speed and distance distributions are slightly skewed, with the speed distribution being slightly left-skewed and the distance distribution being right-skewed. Additionally, both distributions have lighter tails and flatter peaks compared to a normal distribution (platykurtic).

**SP and Weight(WT)**

**Use Q9\_b.csv**



**Speed (SP):**

Skewness of 1.5814536794423764: This positive skewness value indicates that the distribution of speed is right-skewed (positively skewed). It means that the tail of the distribution is longer on the right side, and there may be some higher-speed outliers.

Kurtosis of 2.7235214865269244: This positive kurtosis value suggests that the distribution of speed has heavier tails and potentially more outliers compared to a normal distribution. It indicates that the distribution is leptokurtic, meaning it has a higher peak and fatter tails compared to a normal distribution.

Weight (WT):

Skewness of -0.6033099322115126: This negative skewness value indicates that the distribution of weight is left-skewed (negatively skewed). It means that the tail of the distribution is longer on the left side, and there may be some lower-weight outliers.

Kurtosis of 0.8194658792266849: This positive kurtosis value suggests that the distribution of weight has tails that are less heavy than a normal distribution. It indicates that the distribution is platykurtic, meaning it has a flatter peak and lighter tails compared to a normal distribution.

**Q10) Draw inferences about the following boxplot & histogram**



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.



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 ouliers lie outside the boxplot. The median is situated closely towards 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?

[round(x,2) **for** x **in** st**.**norm**.**interval(alpha**=**0.98, loc**=**200, scale**=**30)]

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

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

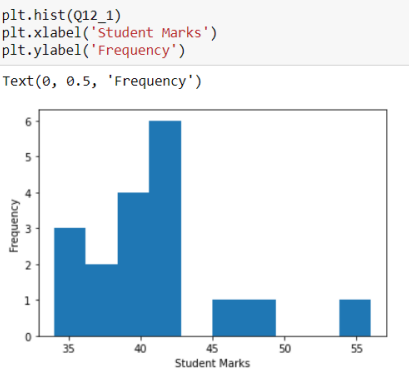
**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

Answer:

1) Mean = 41.00, Median= 40.5, Variance = 25.53, Standard Deviation = 5.052

2) Histogram plot of student marks



From the graph we can say that most of the students scored between 40 to 43 Marks.Few students managed to score between 45 and 50 marks,highest is 45, 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?

When the mean and median of a data are equal, the distribution is symmetric and has zero skewness.

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

If the mean is greater than the median, the distribution is skewed to the right.

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

If the median is greater than the mean, the distribution is skewed to the left.

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

A positive kurtosis value indicates that a distribution is leptokurtic, which means that it has a sharper peak and heavier tails compared to a normal distribution. This simply means that fewer data values are located near the mean and more data values are located on the tails.

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

A negative kurtosis value indicates that a distribution is platykurtic, which means that it has a flatter peak and thinner tails compared to a normal distribution. This simply means that more data values are located near the mean and less data values are located on the tails

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



What can we say about the distribution of the data?

**Answer:** It is not a Normal Distribution

What is nature of skewness of the data?

**Answer:** it is left skewed

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

**Answer:** IQR is 8

Q19) Comment on the below Boxplot visualizations?



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

The median for both the plots are almost same. The IQR in boxplot(2) is much greater than the IQR incase of boxplot(1).There is no outlier for both boxplot

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

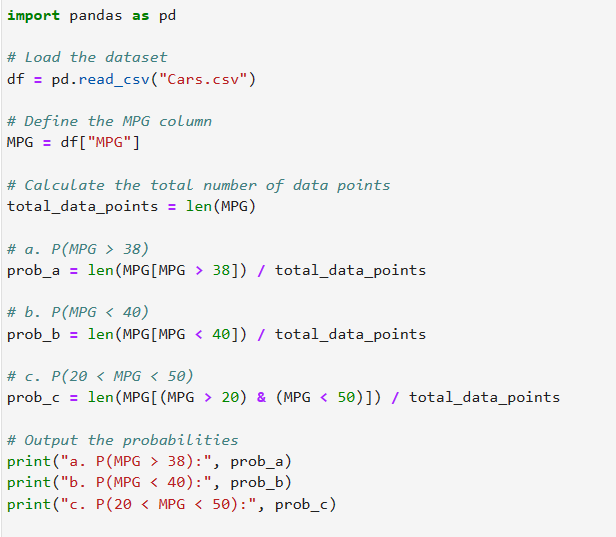
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

a. P(MPG > 38): 0.4074074074074074

b. P(MPG < 40): 0.7530864197530864

c. P(20 < MPG < 50): 0.8518518518518519

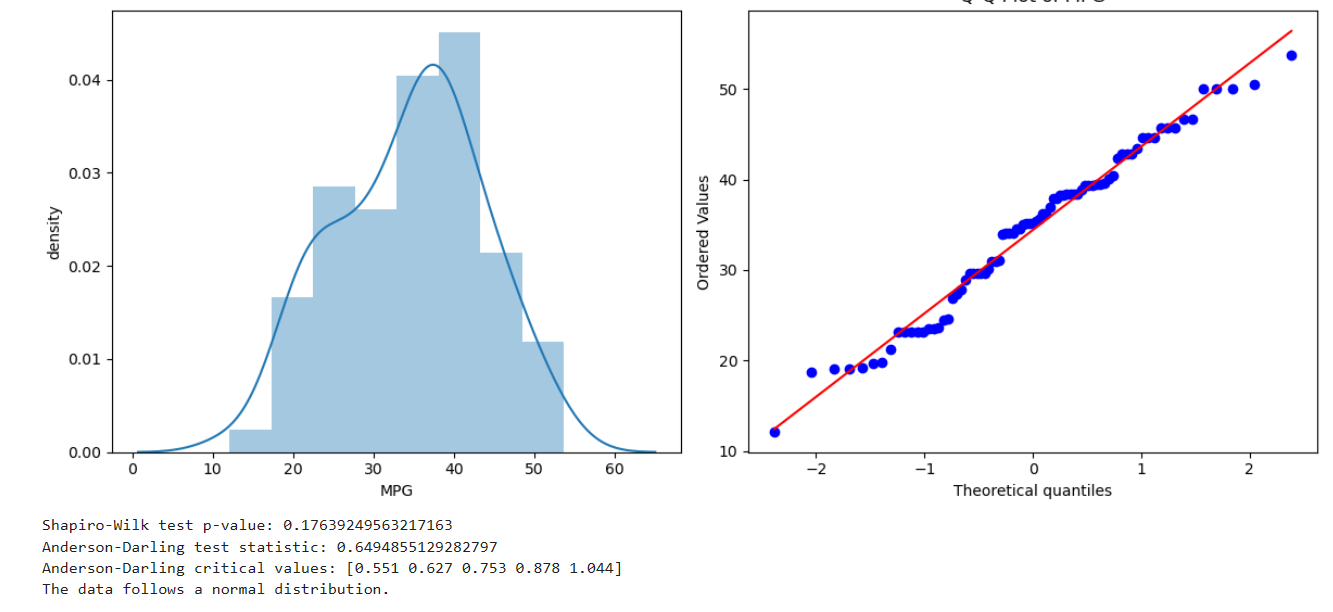


Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

MPG follows Normal Distribution

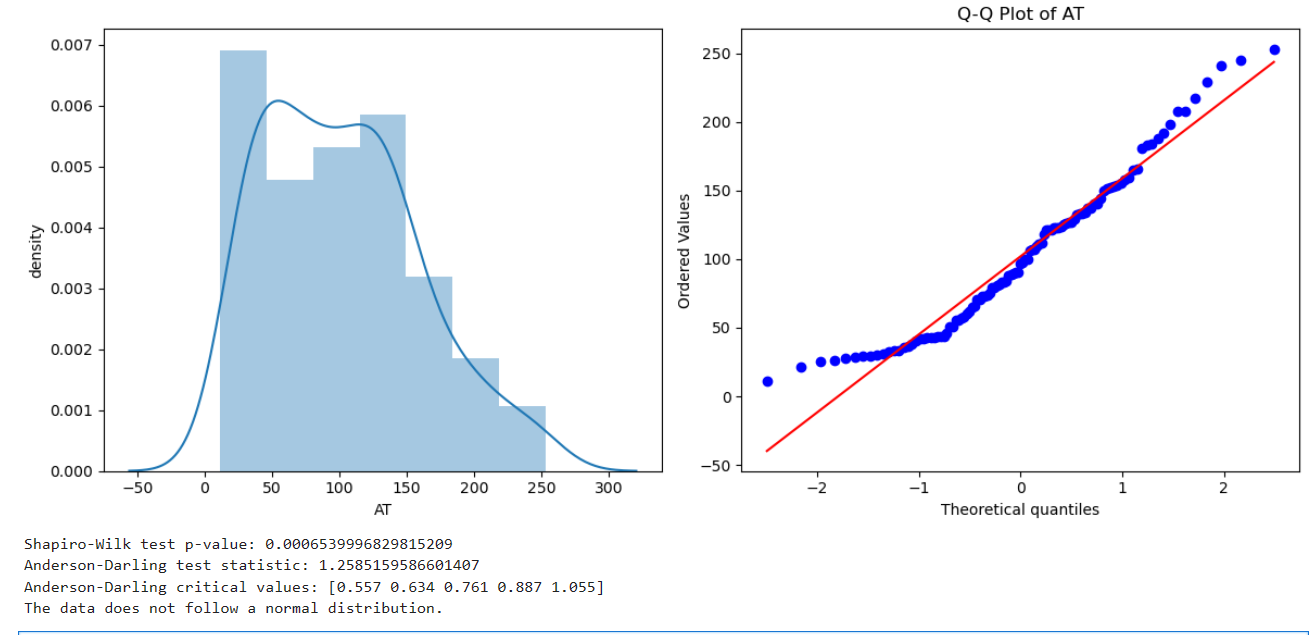


Yes it follows a normal distribution

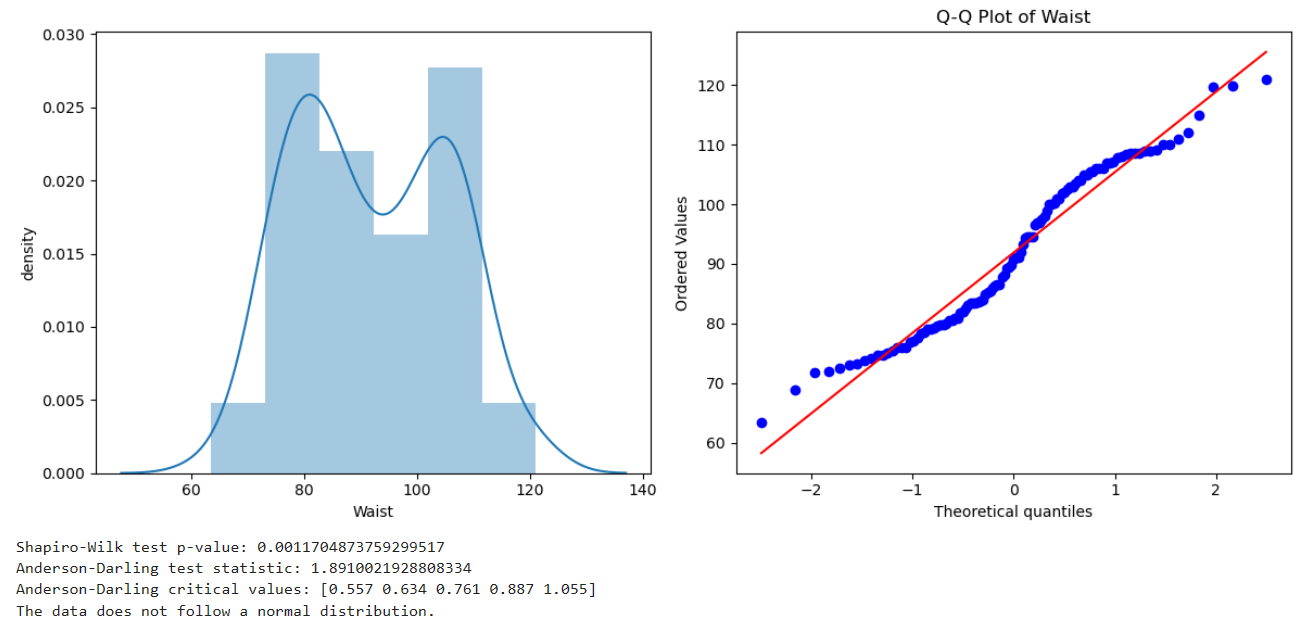
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

**AT** does not follow Normal distribution

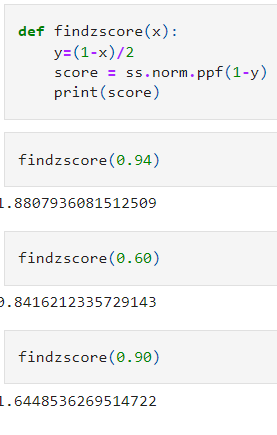


**Waist** does not follow Normal distribution



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

| **Confidence Interval** | **Z-Scores** |
| --- | --- |
| 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

| **Confidence Interval** | **T-Scores** |
| --- | --- |
| 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

µ=270, =260, SD=90, n=18, df=17

t-score = =-0.4714  
Degree of freedom = 17   
P(t) = 0.3216725