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

Q1) Identify the Data type for the Following:

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 | ratio |
| Height | ratio |
| Type of living accommodation | ordinal |
| Level of Agreement | nominal |
| IQ(Intelligence Scale) | ratio |
| Sales Figures | ratio |
| Blood Group | nominal |
| Time Of Day | ordinal |
| Time on a Clock with Hands | interval |
| Number of Children | ratio |
| Religious Preference | nominal |
| Barometer Pressure | interval |
| SAT Scores | ordinal |
| Years of Education | ratio |

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

Answer:-

When 3 coins are tossed, the possible outcomes are TTT, TTH, THT, HTT, THH, HTH, HHT, HHH which is 8.

so no. of times two heads and one head occur is 3.

Probability = no of times two heads and one tail occurred/ total outcomes

So required probability is 3/8

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. Probability that sum is equal to 1 is zero because when we roll a dice, it beings with (1,1)
2. Probability that sum is less than or equal to 4:

Sum of outcomes that are less than or equal to 4 are.

(1,1) (1,2) (1,3) (2,1) (2,2) (3,1). therefore, the no of outcomes is = 6

Probability = number of out comes that are less than or equal to 4 / total outcomes = 6/36

The required probability is 6/36 (i.e.) 1/6

1. Probability that sum is divisible by 2 and 3:

The total number of outcomes when 2 dice are rolled is 36 (i.e.)

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)

(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)

(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)

(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)

(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)

(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

As sum should be divisible by both 2 and 3

The favorable outcomes are = (1,5) (2,4) (3,3) (4,2) (5,1) (6,6)

Therefore, the number of favorable outcomes = 6

Probability of the event = no of favorable outcomes / no of total outcomes

So, the probability of the event is = 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?

Answer:-

Non blue color balls are about 5 which is, 2 red and 3 green balls. So, none of the balls drawn is blue, it happens only when the two balls picked randomly are either red and green.

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

Number of ways of drawing 2 balls out of 7 = 7C2=(7x6)/(2x1) = 42/2 = 21 Number of ways of drawing 2 balls out of 5 = 5C2=(5x4)/(2x1) = 20/2 = 10 Therefore, required probability = 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)

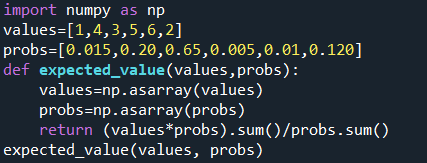
|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

Answer:-

Code-



Output:-



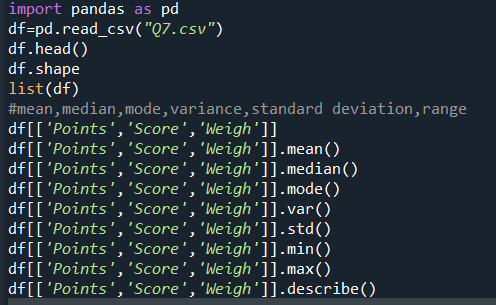
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* 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**

Answer:-



Output:-

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.596563 | 3.217250 | 17.848750 |
| Median | 3.695 | 3.325 | 17.710 |
| Mode | 3.07 | 3.44 | 17.02 |
| Variance | 0.285881 | 0.957379 | 3.193166 |
| Standard deviation | 0.534679 | 0.978457 | 1.786943 |
| Range(Min-Max) | 3.59-4.93 | 3.21-5.42 | 17.84-22.9 |

Comment:-

we can see the from the above table Points and score values have median greater than mean so they are left skewed. Where as in weigh value the mean is greater than median so they are right skewed. Not only by mean and median, by observing the standard deviation we can say that they are left skewed or right skewed.

Q8) Calculate Expected Value for the problem below

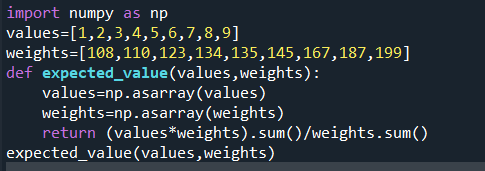
1. The weights (X) of patients at a clinic (in pounds), are

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:-

Code:-



Output:-



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

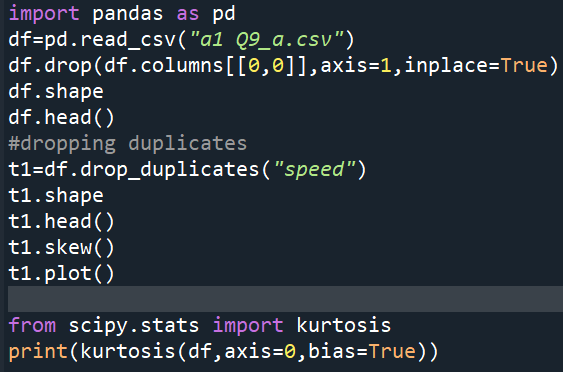
**Cars speed and distance**

**Use Q9\_a.csv**

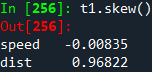
**Answer:-**

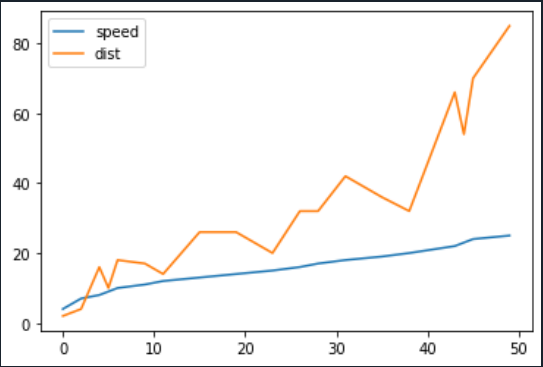
1. **Skewness**

**Code-**



**Output:-**





From the above graph we can see the skewness of speed and distance are negatively skewed.

1. Kurtosis:-

Code:-



Output:-



Kurtosis generally determines whether a distribution is heavily tailed in respect of the normal distribution and provides information about the shape of a frequency distribution. As kurtosis value of index, speed, distance is less than 3, which is also called playkurtic.

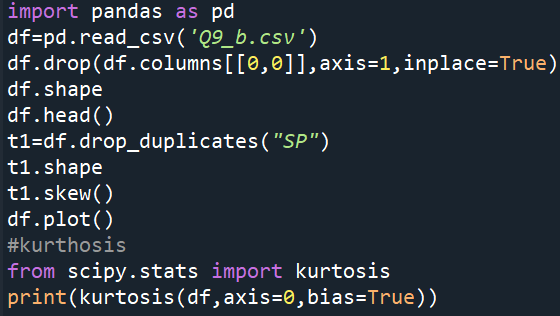
**SP and Weight(WT)**

**Use Q9\_b.csv**

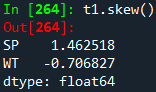
**Answer:-**

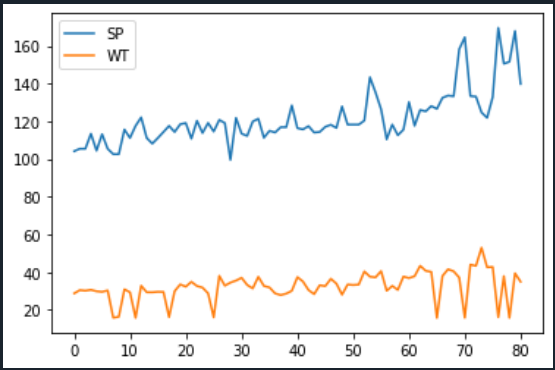
1. **Skewness-**

**Code:-**



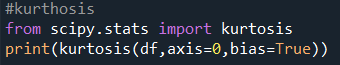
Output:-



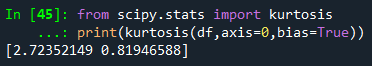


1. Kurtosis:-

Code:-



Output:-



The above values are positive kurtosis and zero kurtosis . Positive kurtosis is called Laplace distribution which has fatter tails as the normal distribution and the outliers are said to be more frequent. Zero kurtosis is called mesokurtic which has the similar tail as the normal distribution and the outliers occur same as normal distribution.

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



Answer:-

By observing the histogram firstly, we can say that it is left skewed since the peak is on the left side where as the tail is on the right which means the mean( the average) is tend to be greater than the median. The outliers might be on the highest side.

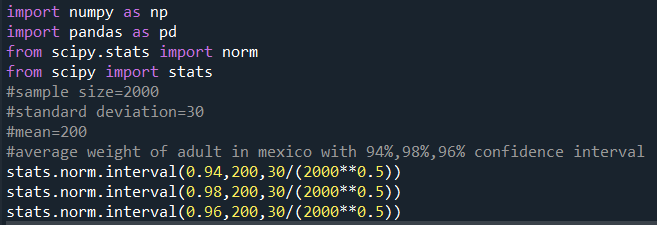


Answer:-

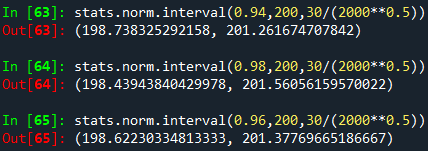
By observing the box plot we can say that it has more number of outliers on the maximum side.

**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?

Answer:-



Output:-



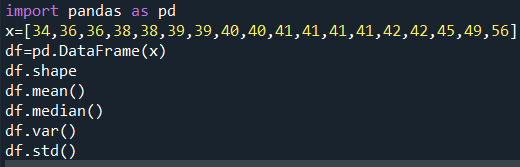
**Q12)** Below are the scores obtained by a student in tests

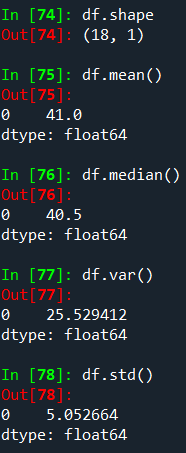
**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

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

**Answer:-**

**1)**





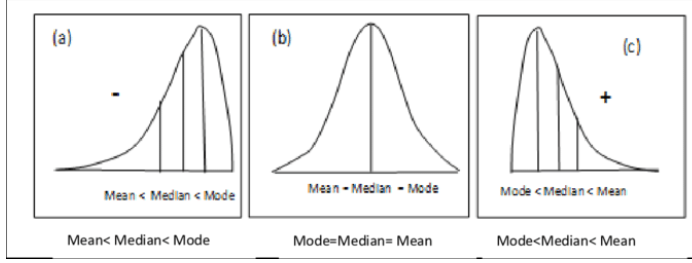
|  |  |
| --- | --- |
|  | Results |
| Mean | 41 |
| Median | 40.5 |
| Variance | 25.529412 |
| Standard deviation | 5.052664 |

**2)Comment:-**

Mean is greater than median which means it is positively skewed.Basically mean, median, variance, standard deviation measures can provide insight into different aspects of a dataset’s distribution, such as its central tendency, spread, and skewness. By the above data’s statistical measure, we can say that the data is highly positively skewed and as the variance is high, they are more spread out. Mean and median values are high which means that there’re outliers present and hence the central tendency gets affected.

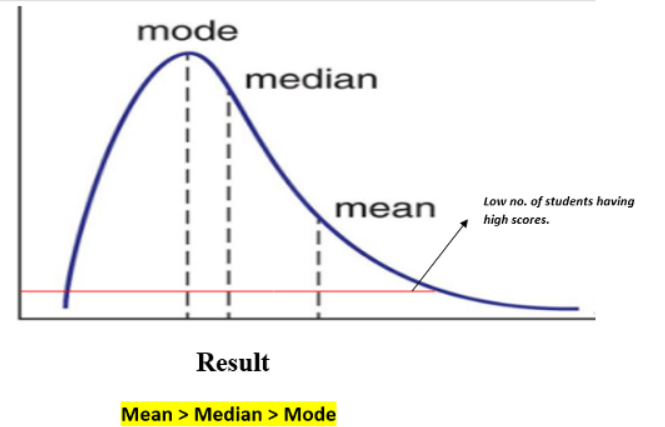
Q13) What is the nature of skewness when mean, median of data are equal?

**Answer:-**When mean and median of the data is equal, the nature of the skewness is said to be in a symmetrical shape. We can expect the symmetrical shape, when the data is said to be continuous data.



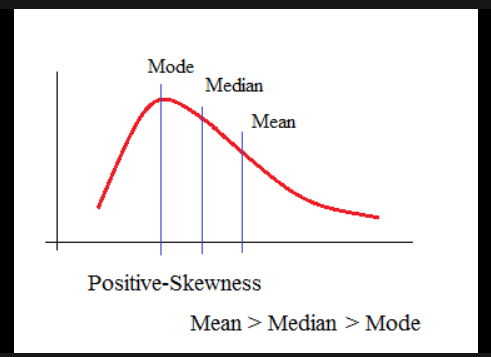
Q14) What is the nature of skewness when mean > median ?

**Answer:-**when mean is greater than median, the nature of the skewness is positively skewed, and it happens when more frequencies are towards the left side.



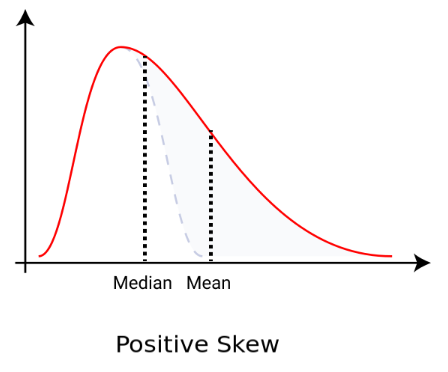
Q15) What is the nature of skewness when median > mean?

**Answer:-**when the median is greater than mean, the nature of the skewness is negative in which the frequencies are towards right side.



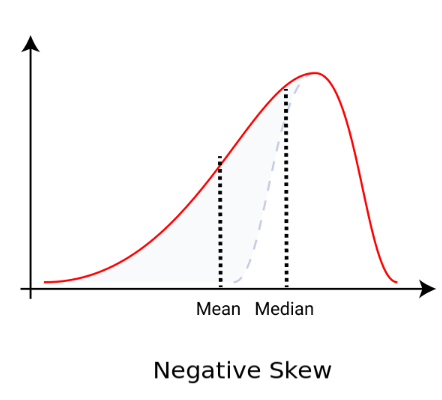
Q16) What does positive kurtosis value indicates for a data ?

**Answer:-**Positive kurtosis value indicates that the distribution has heavier tails than the normal distribution. It means that the data is more peaked and possesses thick tails.



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

**Answer:-**A negative kurtosis value indicates that the distribution has lighter tails than the normal distribution. Which means that the distribution is flatter than a normal curve with the same mean and standard deviation.



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



What can we say about the distribution of the data?

**Answer:**-Then the distribution is negatively skewed (skewed left) because the median is closer to the top of the box, and the whisker is shorter on the upper end of the box.

What is nature of skewness of the data?

**Answer:-**From the above boxplot visualization, the nature of the skewness of the given data is said to be negatively skewed.

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

**Answer:-** The IQR (Inter Quartile Range) can be calculated by subtracting Q3( median of the Upper half of the data) and Q1( median of the lower half of the data). Approximately let us assume the data measures from 10 to 18 which is also arranged in ascending order.

Lower whisker upper whisker  
 10 11 12 13 14 15 16 17 18

Median of 16,17,18= 17=Q3

median

Median of 10,11,12= 11=Q1

IQR(Inter Quartile Range)= Q3-Q1=17-11=6

Q19) Comment on the below Boxplot visualizations?



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

**Answer:-**

* Both the boxplots share the same median which might approximately range in between 250 and 275.
* By observing boxplot 1 and boxplot 2 we can say that boxplot 2 consists of more data when compared to boxplot 1 and there are no outliers.
* They are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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

Data \_set: Cars.csv

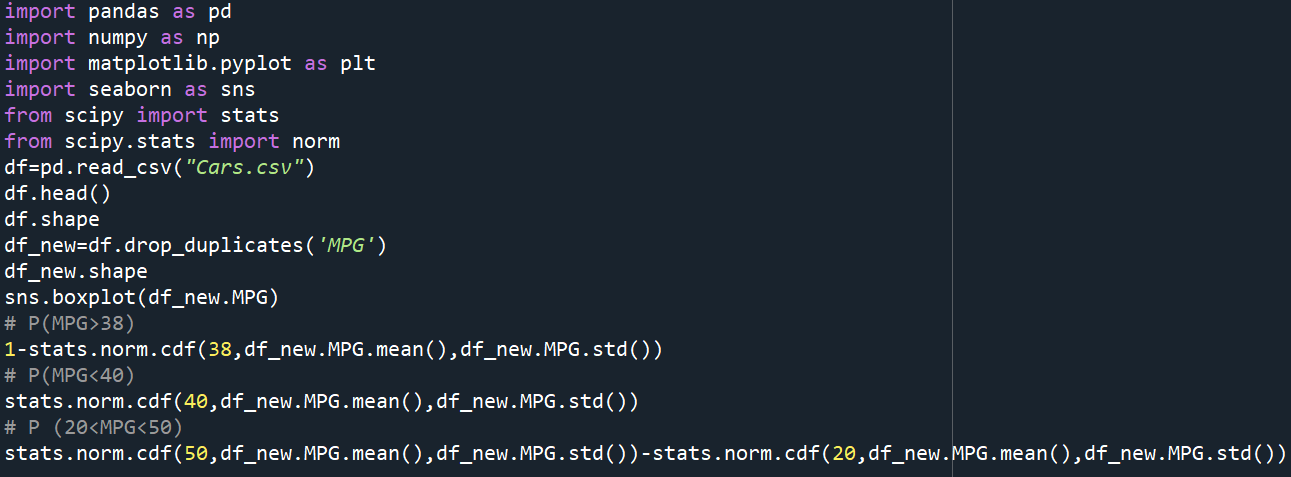
Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

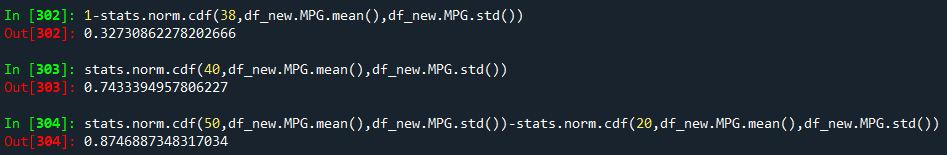
* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

Answer:-

**Code:-**



**Output:-**



Q 21) Check whether the data follows normal distribution

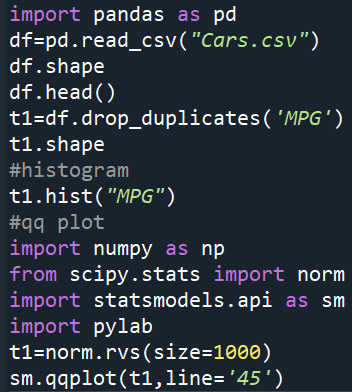
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

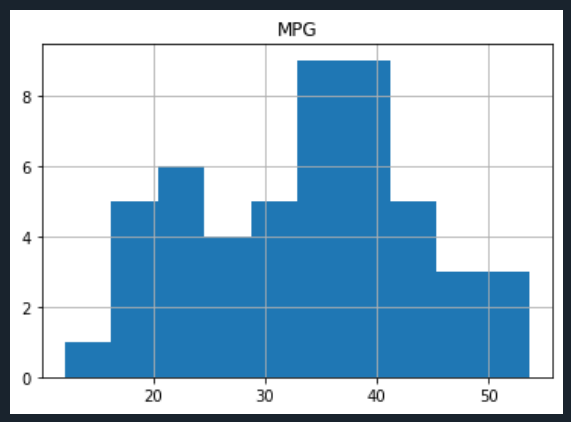
Answer:-

Yes, MPG from the Cars data set follows the normal distribution after removing the duplicates from it. Below is a code to find the Histogram, Box Plot and QQ plot which is used to see precisely if it follows the normal distribution or not.

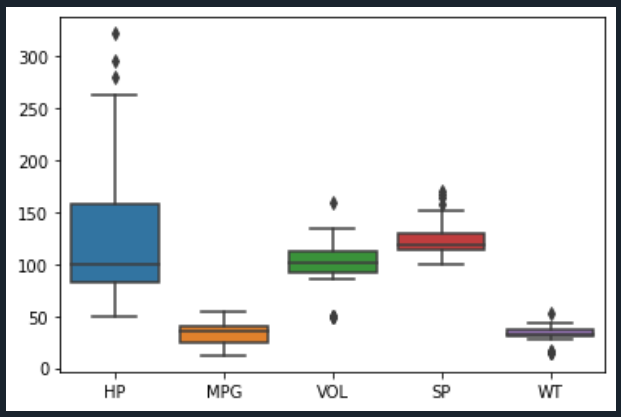
**Code:-**



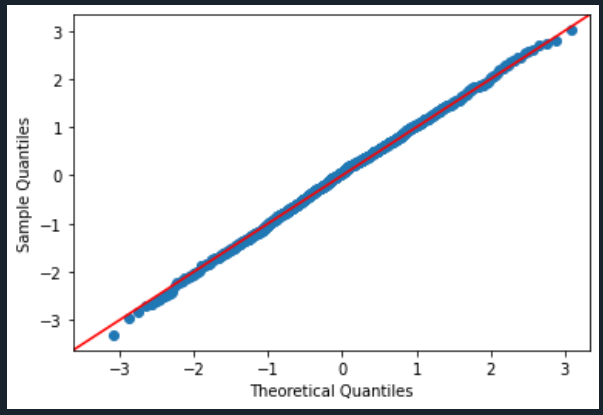
**Histogram:-**



**Boxplot:-**



**QQ-plot:-**

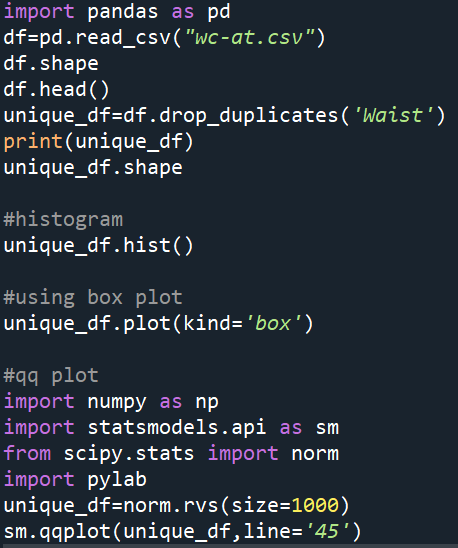


1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

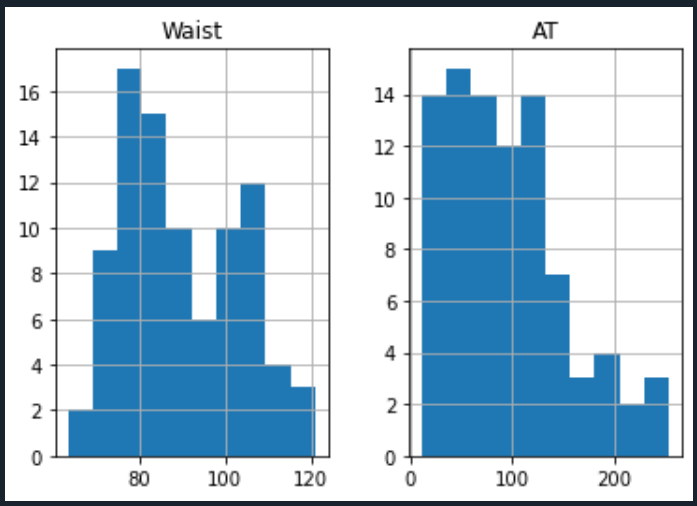
Dataset: wc-at.csv

**Answer:-**

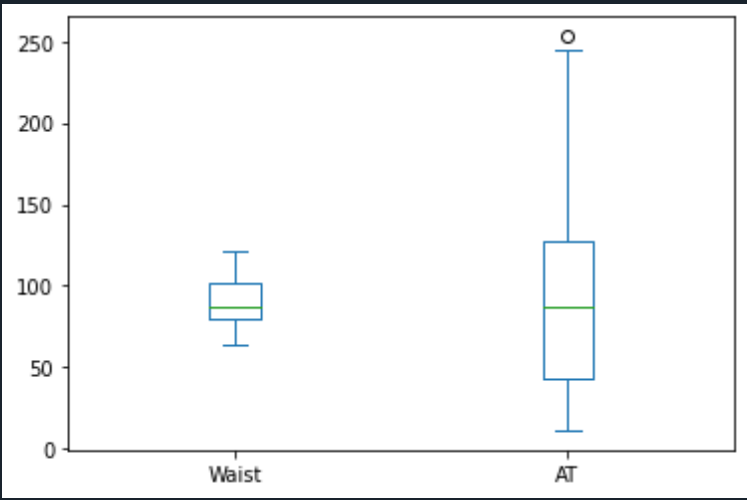
**Code:-**



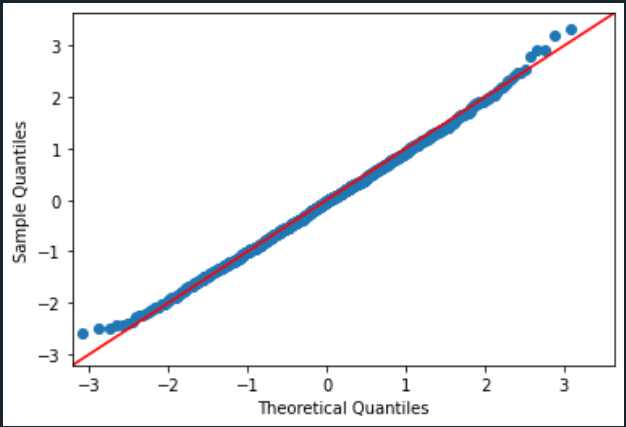
**Histogram:-**



**Boxplot:-**

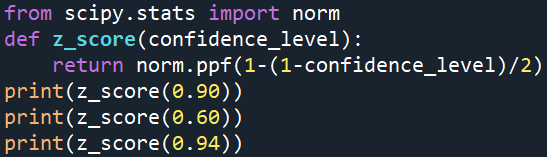


**QQ-Plot:-**

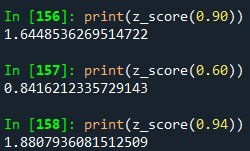


Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

**Answer:-**



**Output:-**

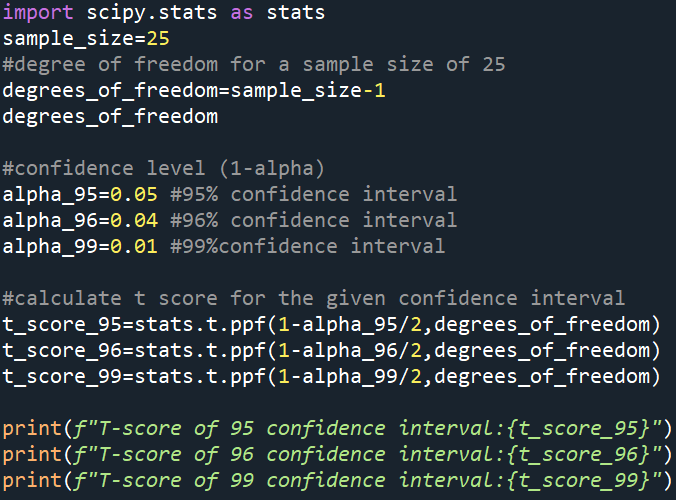


* z score of 90% confidence interval:- 1.644
* Z score of 94% confidence interval:- 1.880
* Z score of 60% confidence interval:- 0.841

As all the above values are positive it means that the data points are above the distribution’s mean (average).

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

**Answer:-**



**Output:-**



* T score for 95% confidence interval:- 2.063
* T score for 96% confidence interval:- 2.171
* T score for 99% confidence interval:- 2.796

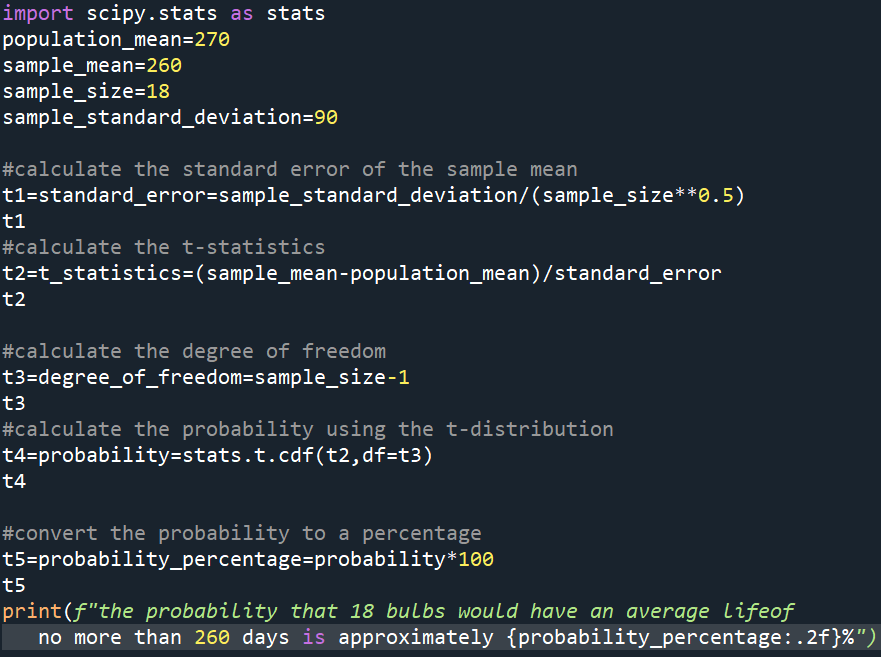
Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs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

Hint:

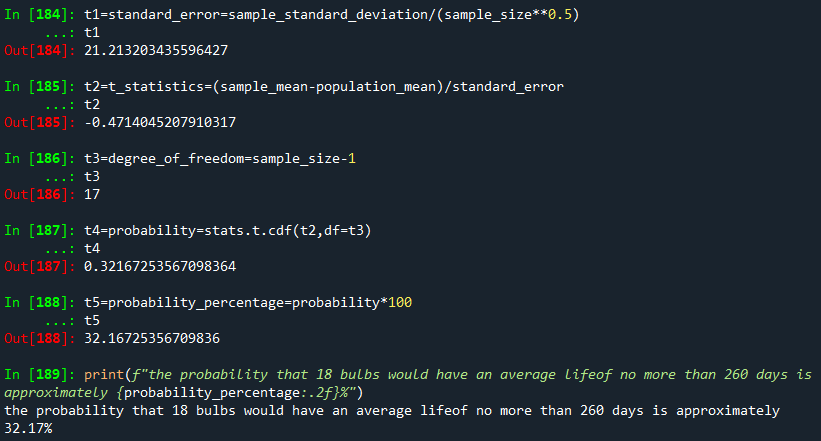
rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

**Answer:-**



**Output:-**



The probability is 32.17% (approx)