|  |  |
| --- | --- |
| 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 |

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 | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Ordinal |
| 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 | Interval |
| Years of Education | Ordinal |

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

Ans: If 3 coins are tossed, total\_outcomes = 2^3 (2 for each coin, 3 times tossed)

The possible combination formed with two heads and one tail are,

{H, H, T}, {T,H, H}, {H, T, H} which is 3.

So, P(Two heads and 1 tail) = 3/total\_outcomes => 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

Ans: Two dice rolled, total\_outcomes = 6^2 (6 for each dice, 2 dice rolled)

1. For sum equal to 1, minimum value for single die is 1, so adding two minimum values from two different rolled dice will be 2, hence it gives minimum sum of 2. So, there are no such combinations. So, probability is,

P(Sum Equal to 1) = 0/total\_outcomes = 0/36 => **0**

1. When rolling two dice the probability less than or equal to 4 is, So, the possible outcomes are, {1,1}, {1,2}, {1,3}, {2,1}, {2,2}, {3,1},

So,

P(Less than or Equal to 4) = 6/36 => **0.16666**

1. Possible combinations are,

Combination formed are => {1,5},{2,4},{3,3},{4,2},{5,1},{6,6}

So, P(Sum divisible by 2 and 3) = 6/36 => **0.16666**

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 number of outcomes 2 balls can be drawn is,

Combination => C(7,2) = 7! / (7-2)!\*2! => 5040/(240) => **21**

Drawing 2 balls without any blue balls, means we have to select from 2 red and 3 green balls. So total balls = 5, and 2 balls drawn.

Combination => C(5,2) = 5!/ (5-2)!\*2! => 120/12 => **10**

So, Probability is = 10/21 => **0.476**

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. To find expected number of candies, we have to multiply candies with the probability values and sum those values such as,

Expected number of candies = (1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.120) => 3.09 => **3**

**3** candies to be expected for randomly selected child.

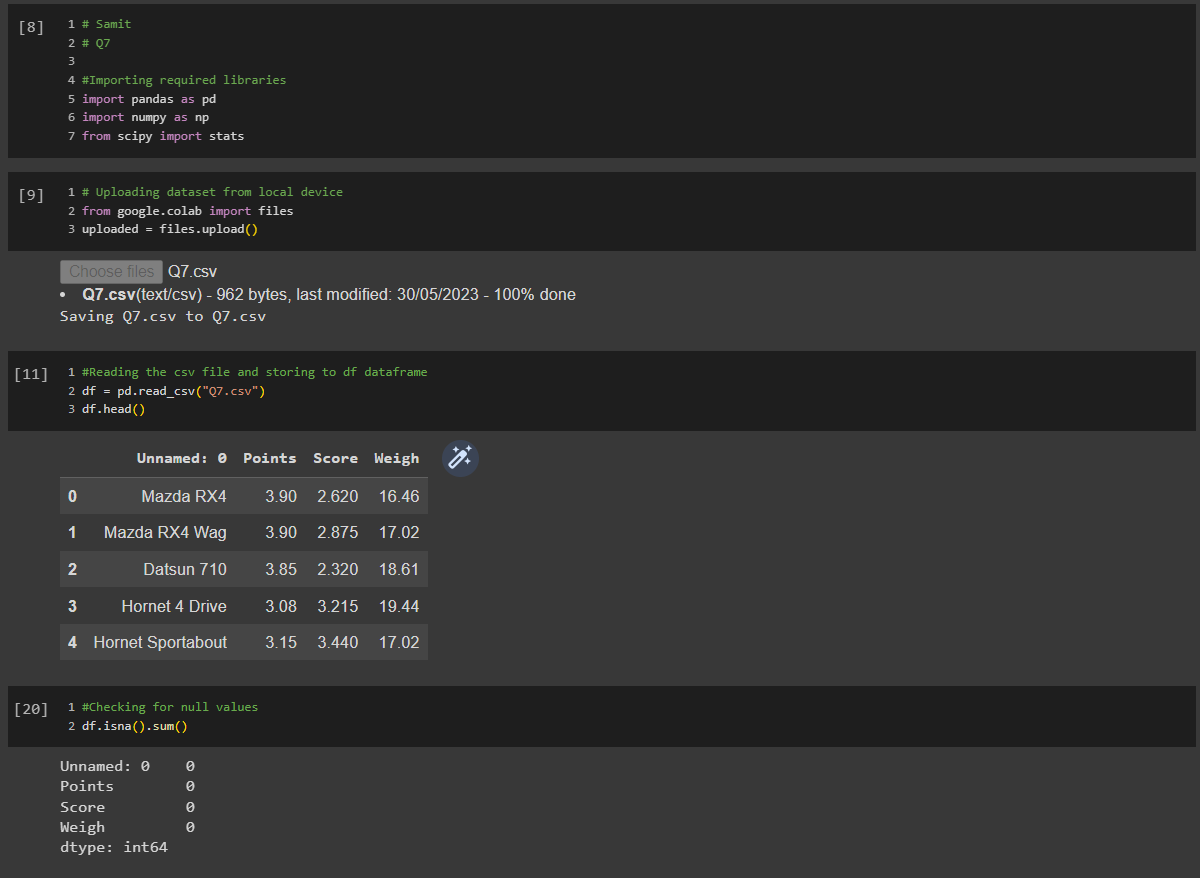
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points, Score, Weigh>

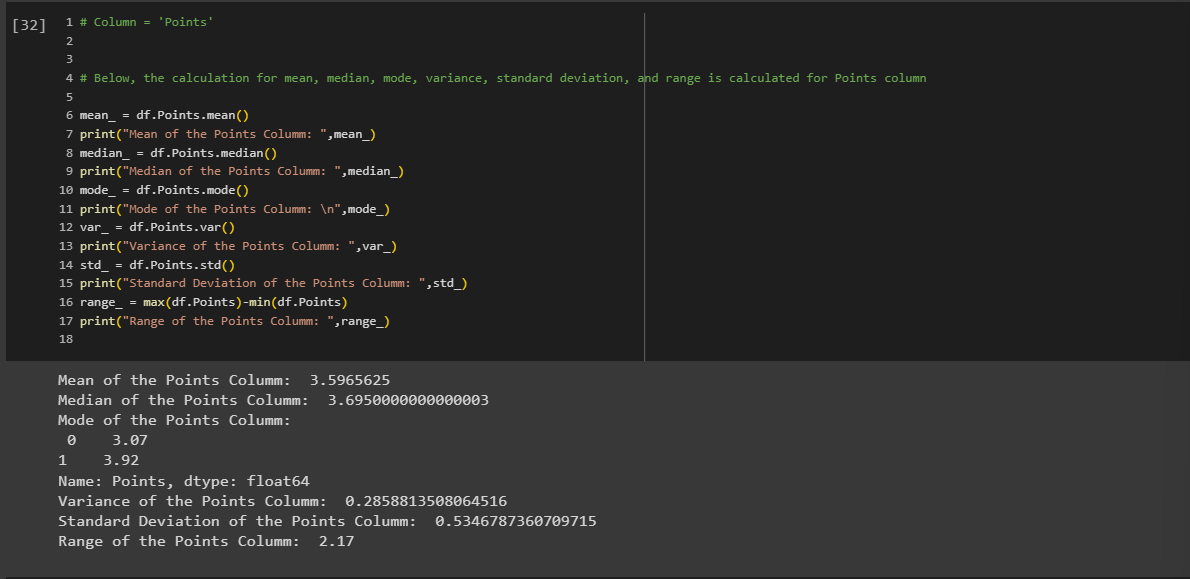
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

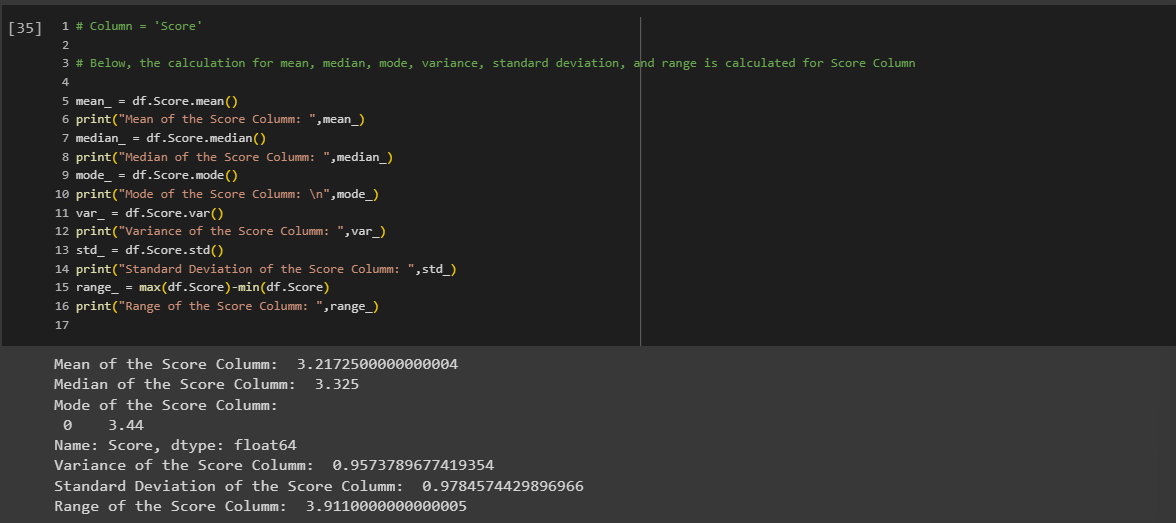
**Use Q7.csv file**

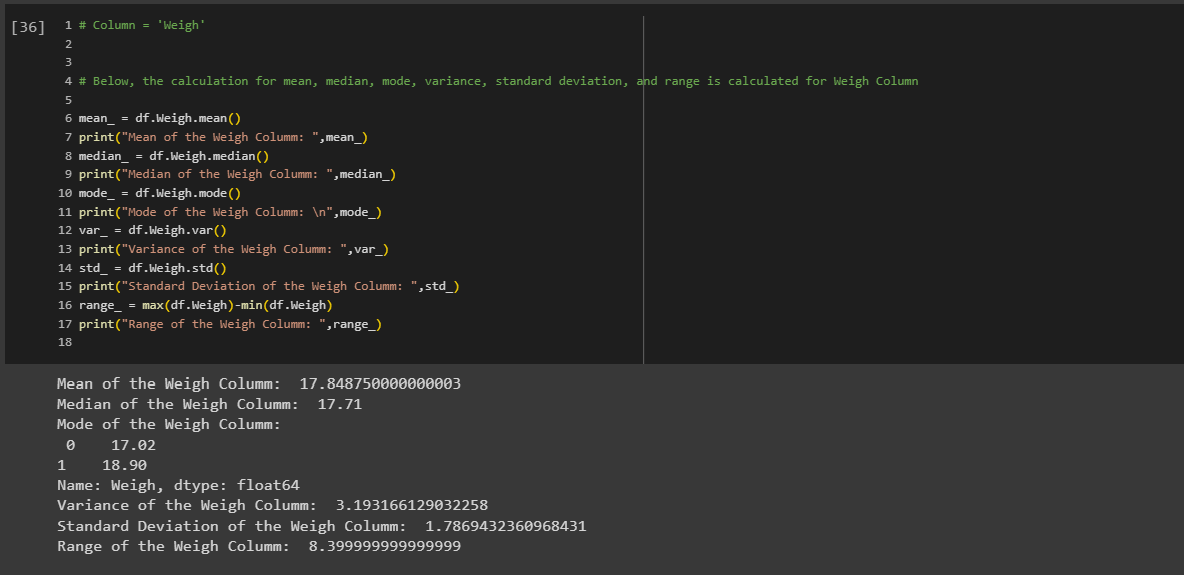
Ans.



* The value for mean is, mean(data) = sum(data all points)/len(data)
* The value for median is,
  + if len(data) is even,
    - median =value at ((len(data)/2) + ((len(data)/2)+1))/2
  + if len(data) is odd,
    - median = value at len(data)/2
* The value for mode is, mode(data) = more frequent item in the data. It can have more than one mode value
* variance = sum(datapoint – mean(datapoints))/len(datapoints)-1
* std. dev. = sqrt(sum(datapoint – mean(datapoints))/len(datapoints)-1)
* range = It is the difference between maximum value and minimum value of the dataset.

****

****

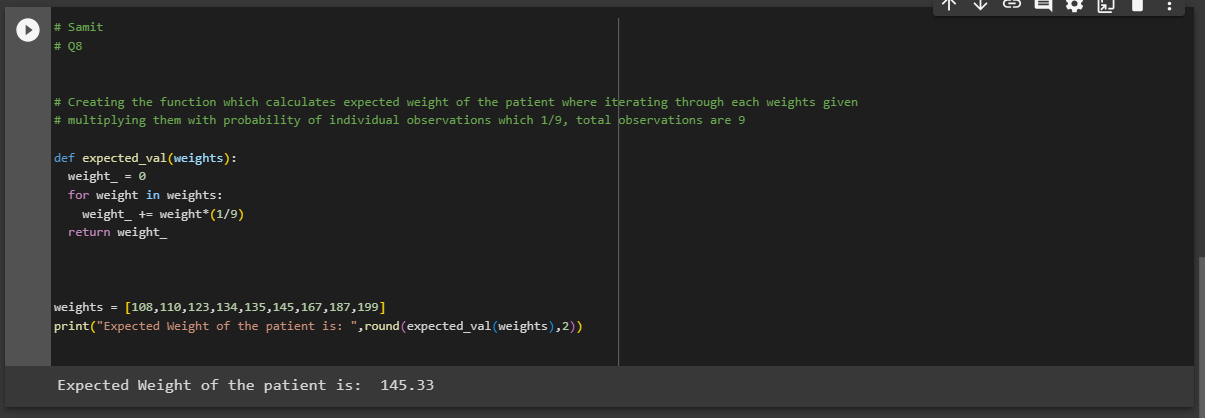
****

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

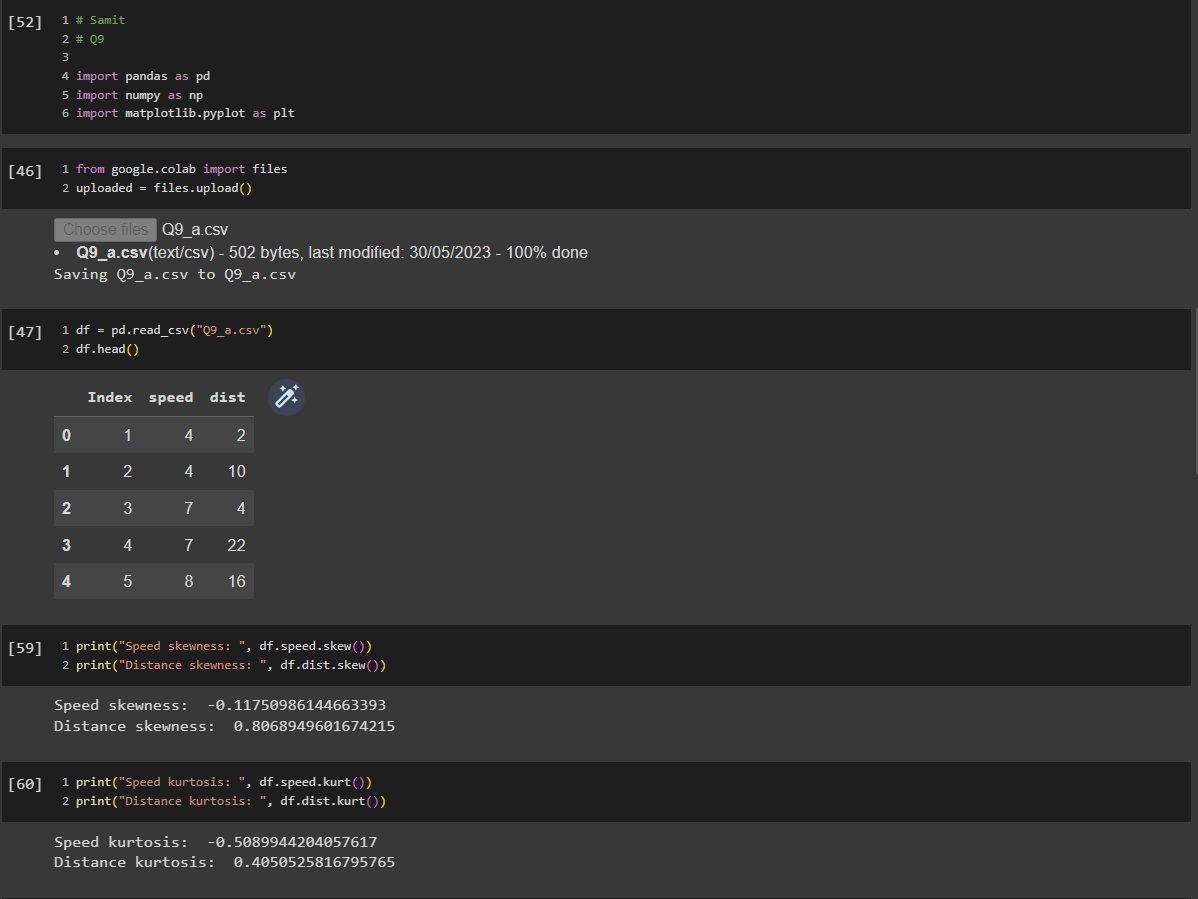
Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

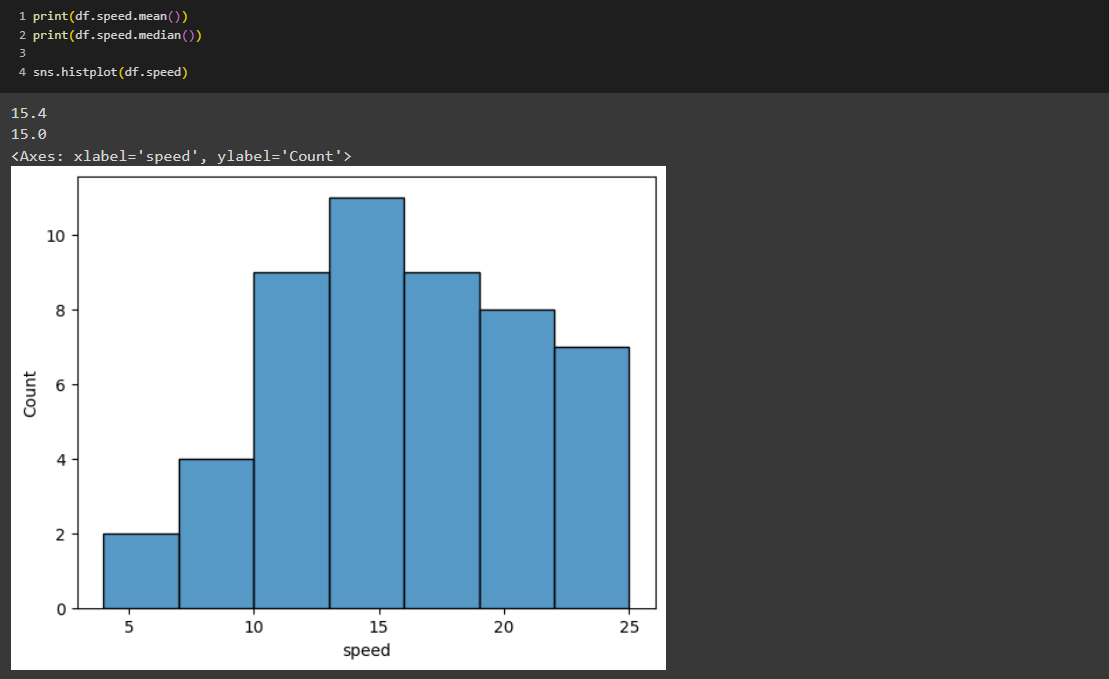


**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

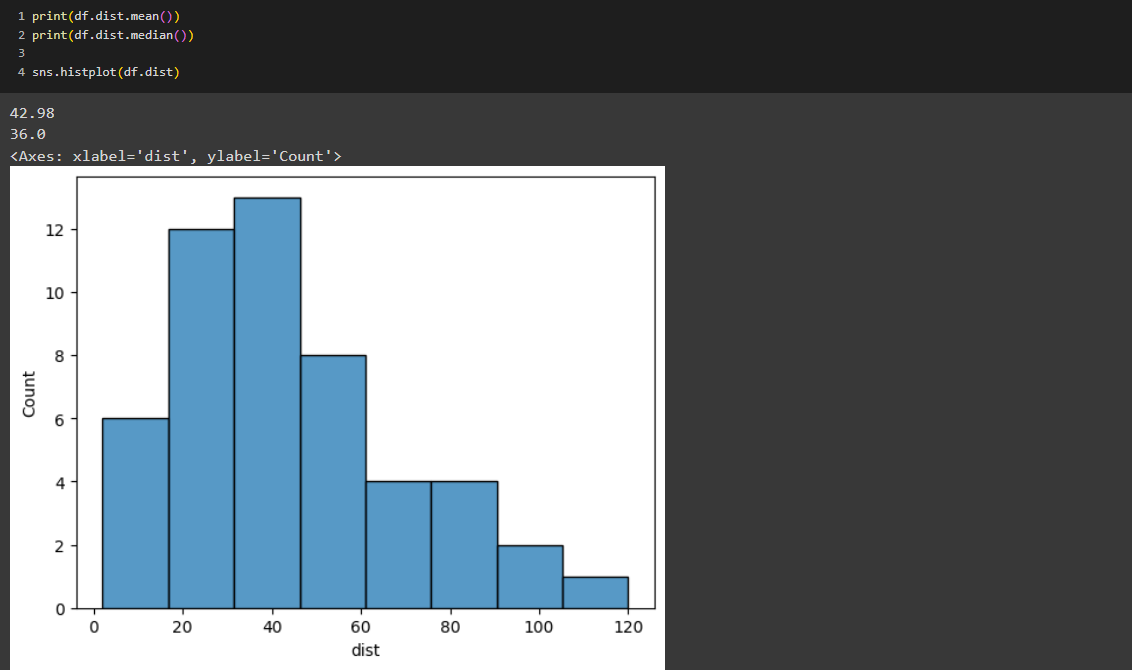
**Cars speed and distance**

**Use Q9\_a.csv**

****

****

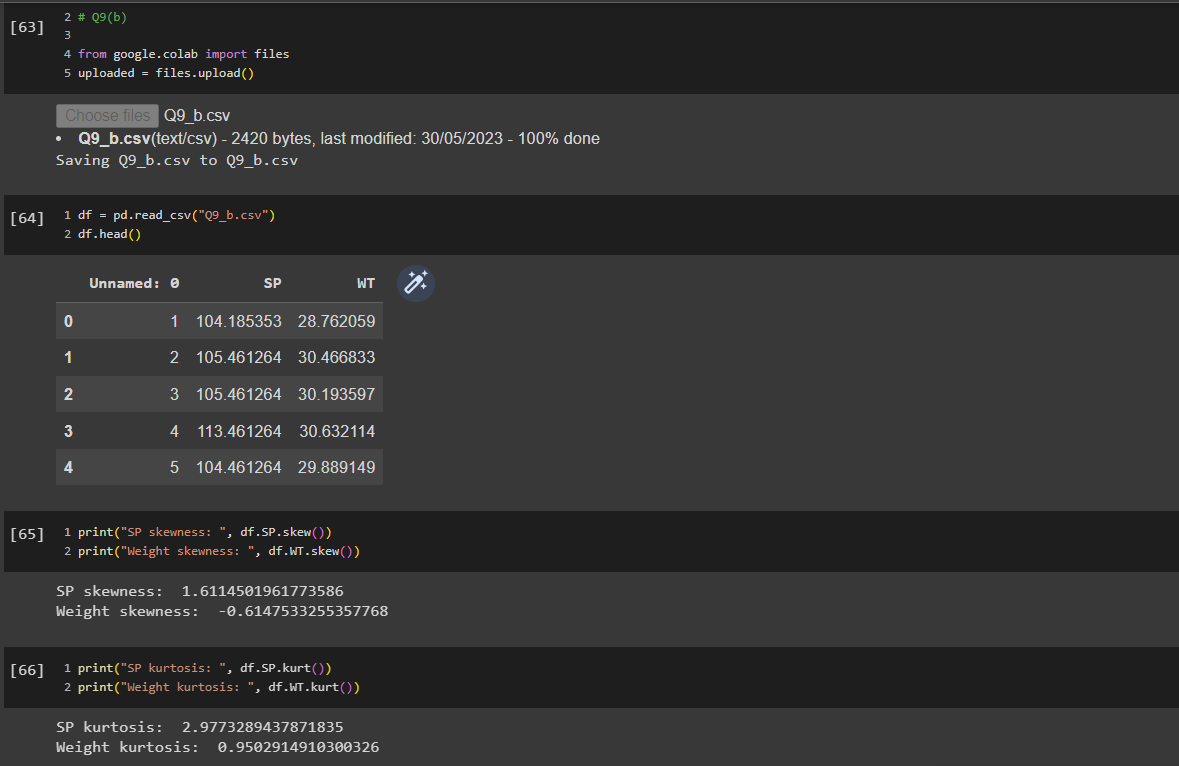
In the above plot we can see the symmetric or normal distribution where mean and median are also same and the kurtosis value indicates that the distribution has light tails and there are fewer extreme values in the distribution than in a normal distribution.

****

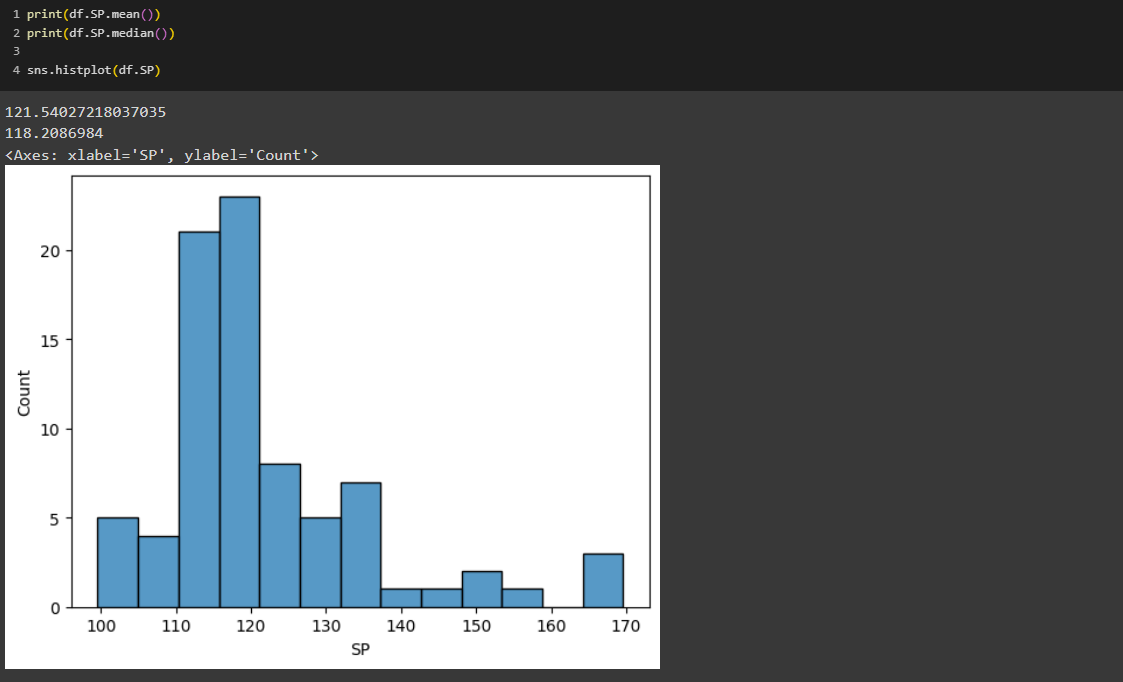
In the above plot we can see that the distribution is right skewed where the concentration of the lies towards the lower side. Whereas, kurtosis value indicated that it is slightly more peaked and has slightly heavier tails compared to normal distribution. Which means distribution has slightly sharper peak and more outliers in the tails compared to a normal distribution.

**SP and Weight (WT)**

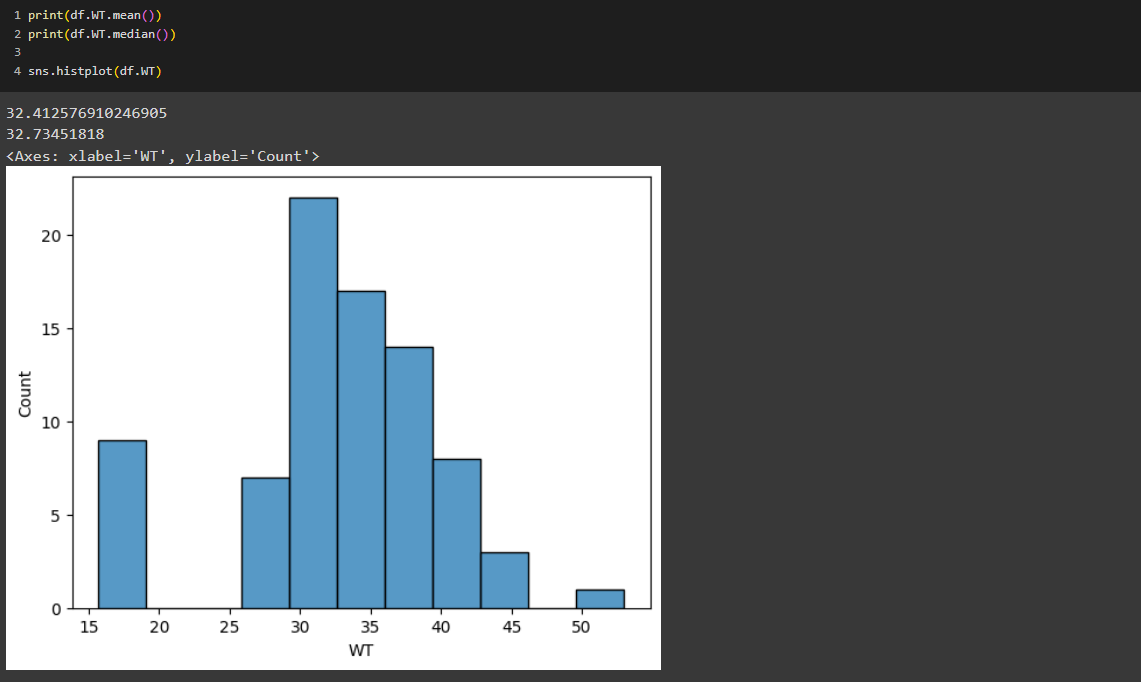
**Use Q9\_b.csv**

****

In the below plot we can see the distribution is right skewed which means the data is not normally distributed. However, the kurtosis value indicates that the distribution has heavier tails and a sharper peak compared to the normal distribution (which has a kurtosis value of 0). Positive value indicates that the distribution has more outliers or extreme values in the tails and a higher peak compared to a normal distribution.

****

In the below plot we can see the data is almost normally distributed where mean and median are almost same with little left skewness. Whereas, kurtosis value indicates that the shape is slightly more peaked and has slightly heavier tails.

****

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



In the above Histogram we can clearly see that it is Right skewed in which most of the ChickWeight lies towards the lower side of the data.



In above, box plot we can see that there are outliers which shows that there are over weighted ChickWeights.

**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:

Population = 3000000

Sample mean = 200

Sample std. dev. = 30

n = 2000

df = 1999

std. error = Sample std/sqrt(n) = 30/sqrt(2000) => **0.67082**

#Since population std. dev. is not given so sample std. dev. is used to calculate std. error.

Now,

t\_94% = 1.960 (from t-distribution table)

t\_98% = 2.326 (from t-distribution table)

t\_96% = 1.960 (from t-distribution table)

94% confidence interval when sample dev. is 30 and sample mean is 200

Lower limit = Sample mean – (t\_94%\*std. error) = 200-(1.960\*0.67082) => **198.6852**

Upper limit = Sample mean + (t\_94%\*std. error) = 200+(1.960\*0.67082) => **201.3148**

Conclusion: 94% of the time average weight of an adult male would be between 198.6852 and 201.3148.

98% confidence interval when sample dev. is 30 and sample mean is 200

Lower limit = Sample mean – (t\_98%\*std. error) = 200-(2.326 \*0.67082) => **198.4397**

Upper limit = Sample mean + (t\_98%\*std. error) = 200+(2.326\*0.67082) => **201.5603**

Conclusion: 94% of the time average weight of an adult male would be between 198.4397 and 201.5603.

96% confidence interval when sample dev. is 30 and sample mean is 200

Lower limit = Sample mean – (t\_96%\*std. error) = 200-(1.960\*0.67082) => **198.6852**

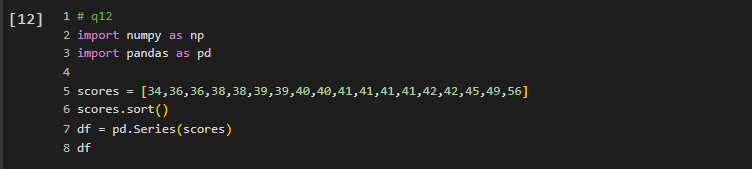
Upper limit = Sample mean + (t\_96%\*std. error) = 200+(1.960\*0.67082) => **201.3148**

Conclusion: 96% of the time average weight of an adult male would be between 198.6852 and 201.3148.

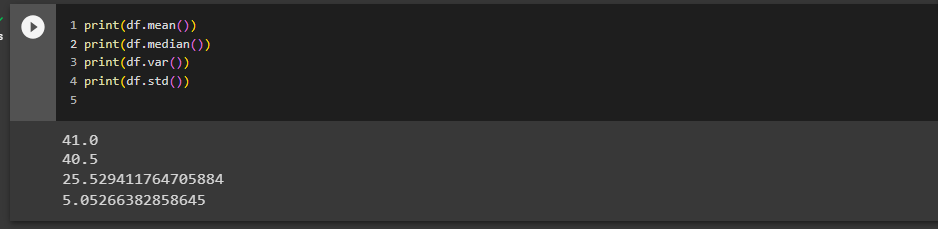
**Note:** If there are little variation then that will be because of the value taken from the t-distribution table which does not have exact match value in the t-distribution table.

**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.



1. What can we say about the student marks?

Ans: we can say that the average marks obtained by the students in the tests is 41 with marks deviation of 5.

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

Ans: When the mean and median of a dataset are equal, it indicates zero skewness, implying that the dataset is symmetric. This means that the data is not skewed towards either the right or the left. The concentration of the dataset lies to the center which forms bell shaped distribution.

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

Ans: When the mean is greater than the median, it suggests positive skewness, indicating that the dataset is right-skewed. In other words, the tail of the distribution extends towards the right side, while the majority of the data points are concentrated towards the left side.

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

Ans: When the median is greater than the mean, it suggests negative skewness, indicating that the dataset is left-skewed. In other words, the tail of the distribution extends towards the left side, while the majority of the data points are concentrated towards the right side.

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

Ans: The positive kurtosis indicates that there are more extreme values in the data set, both positive and negative due to which it forms higher peak in the distribution compare to normal distribution in which peak is flat.

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

Ans: The negative kurtosis indicates that there are fewer extreme values in the data set, both positive and negative. The peak is the point where the distribution is higher. However, in negative distribution the peak is lower than the normal distribution in which peak is flat.

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



What can we say about the distribution of the data?

Ans: The data is negatively skewed where the median value is close to the top of the box and we can see a longer whisker on the left side and small whisker on the right side. Concentration of data is more towards right side of distribution.

What is nature of skewness of the data?

Ans: The nature of the skewness of the data is Left skewed where we can see the concentration of the data is more towards the right side of the distribution.

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

Ans:

Q3 = 18

Q1 = 10

IQR = Q3-Q1 = 18-10 => **8**(Approximately)

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 above two Boxplot it seems to have a symmetric distribution. They differ each other by Four parameters only i.e., Minimum, Q1, Q3, and Maximum value the Median value is same for both the Boxplots.

Q20) Calculate probability from the given dataset for the below cases

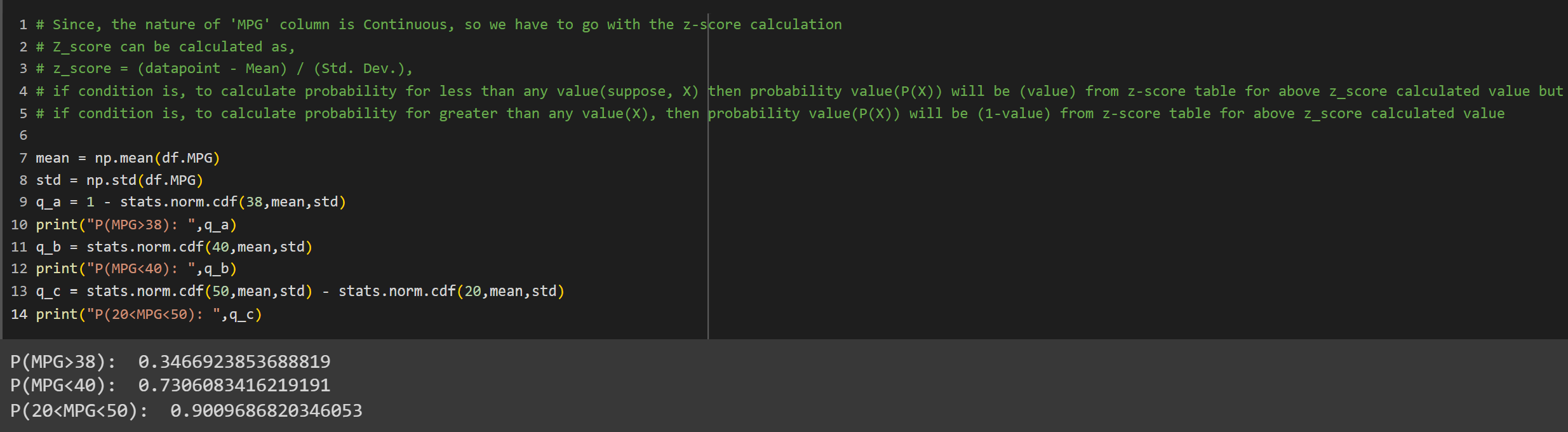
Data \_set: Cars.csv

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

MPG <- Cars$MPG

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

c. P (20<MPG<50)

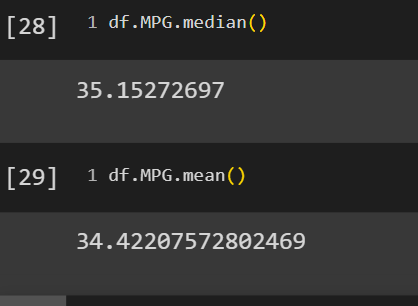


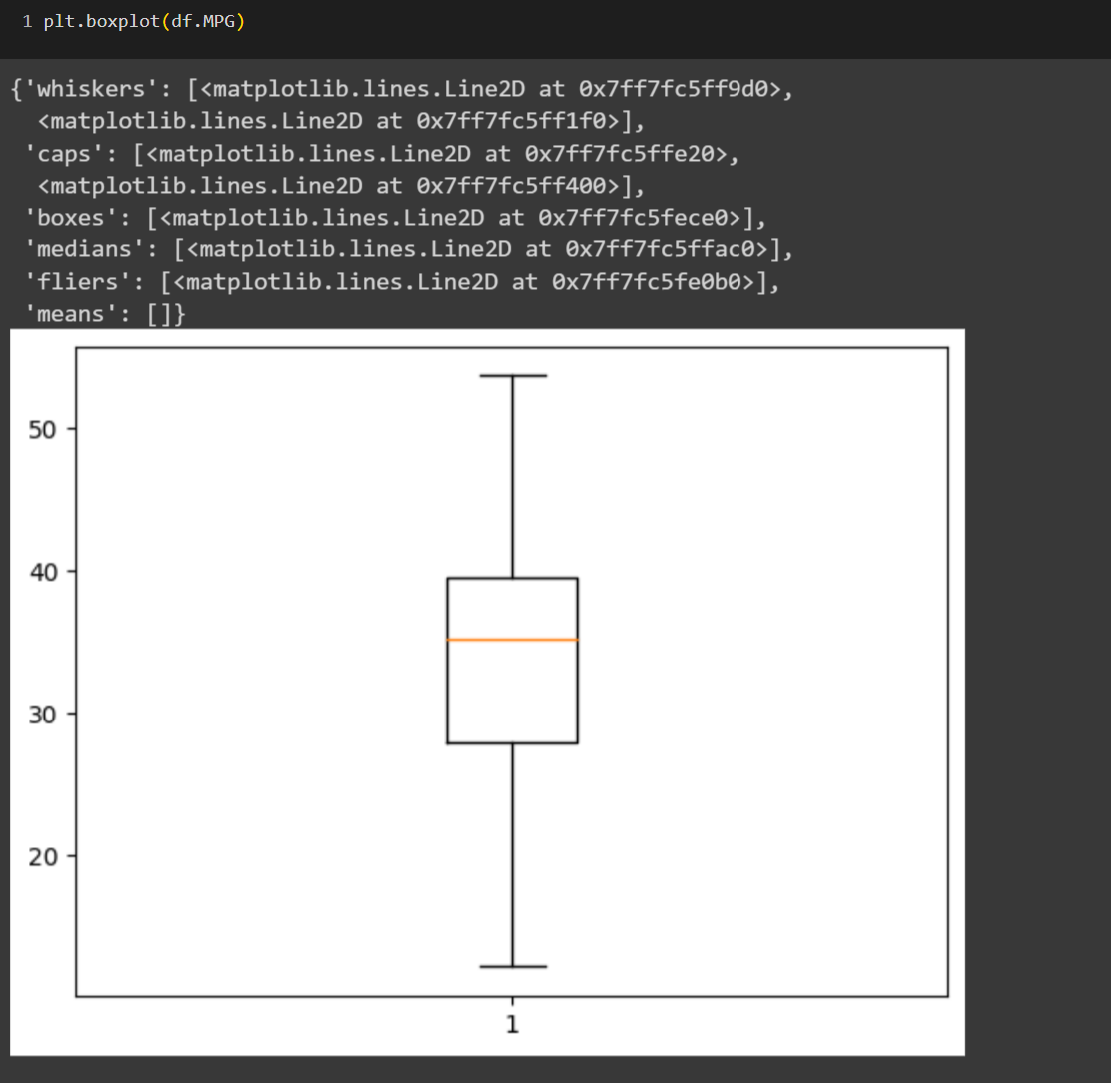
Q21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: No, It doesn’t follows Normal Distribution since it’s median value is higher than the mean value of MPG which shows the Negative skewness of the data which is skewed towards left and concentration is little more towards right of the distribution and median value is near to the top of the box which is shown below.

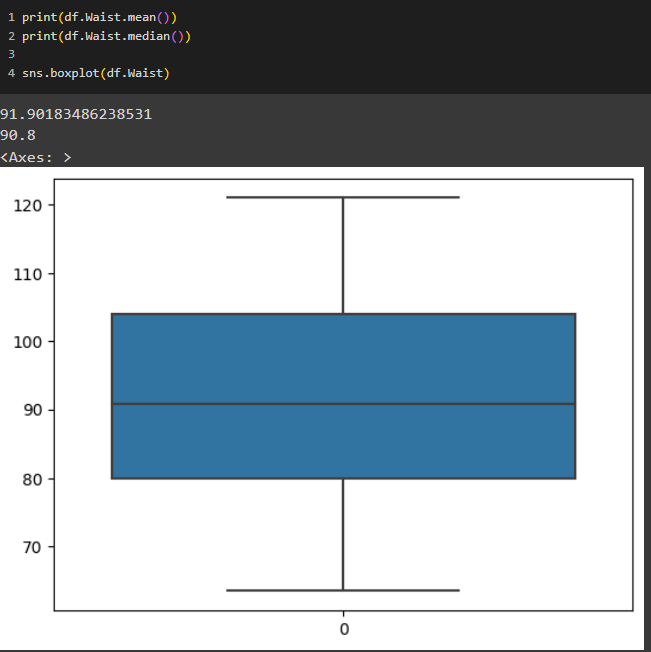
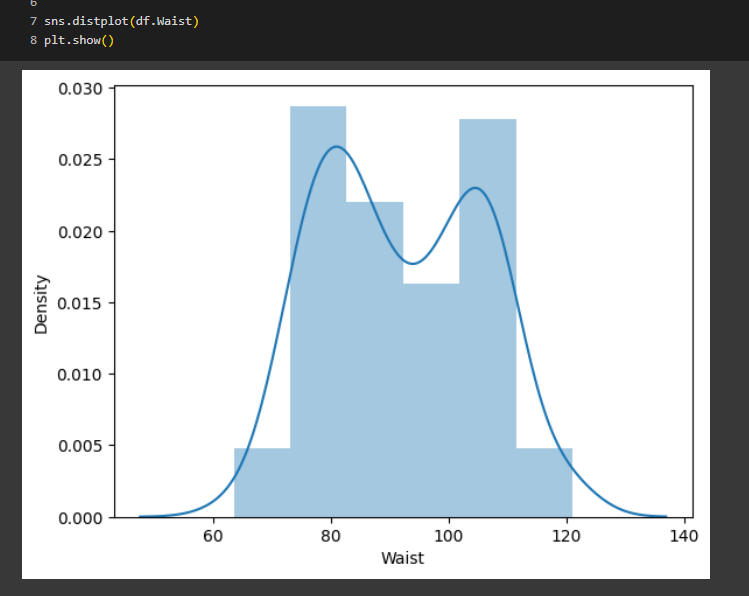




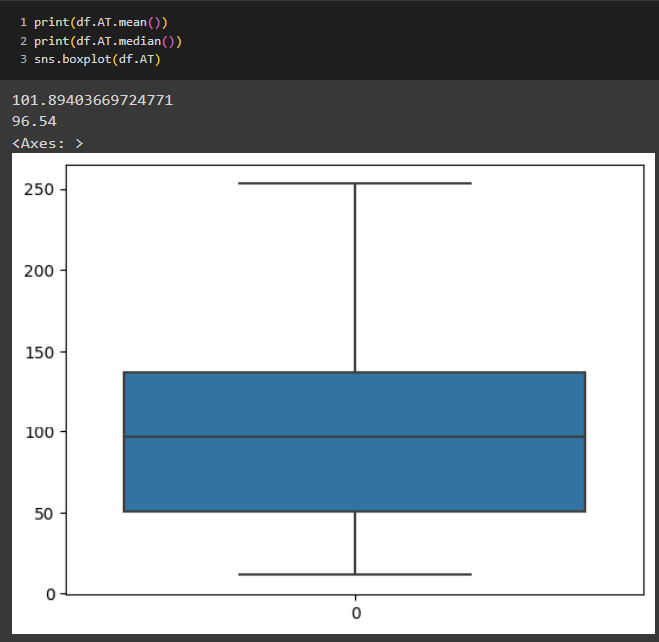
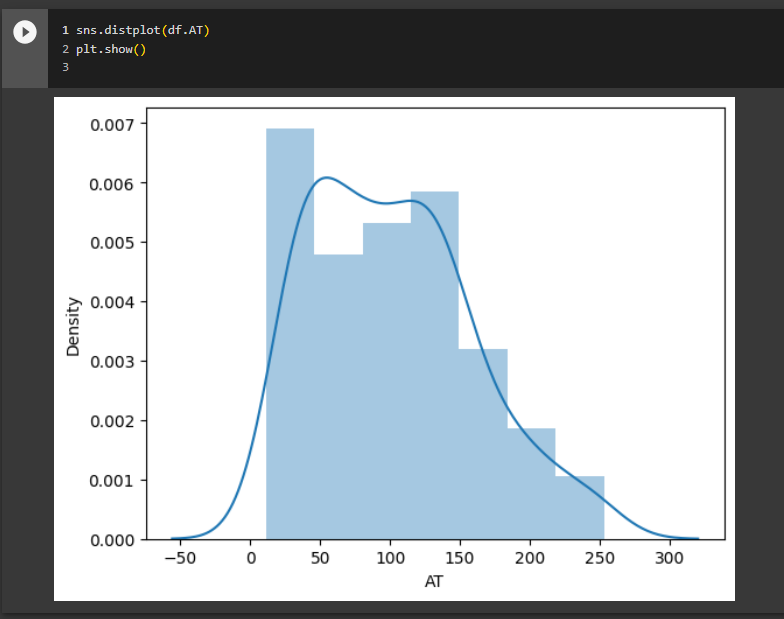
1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

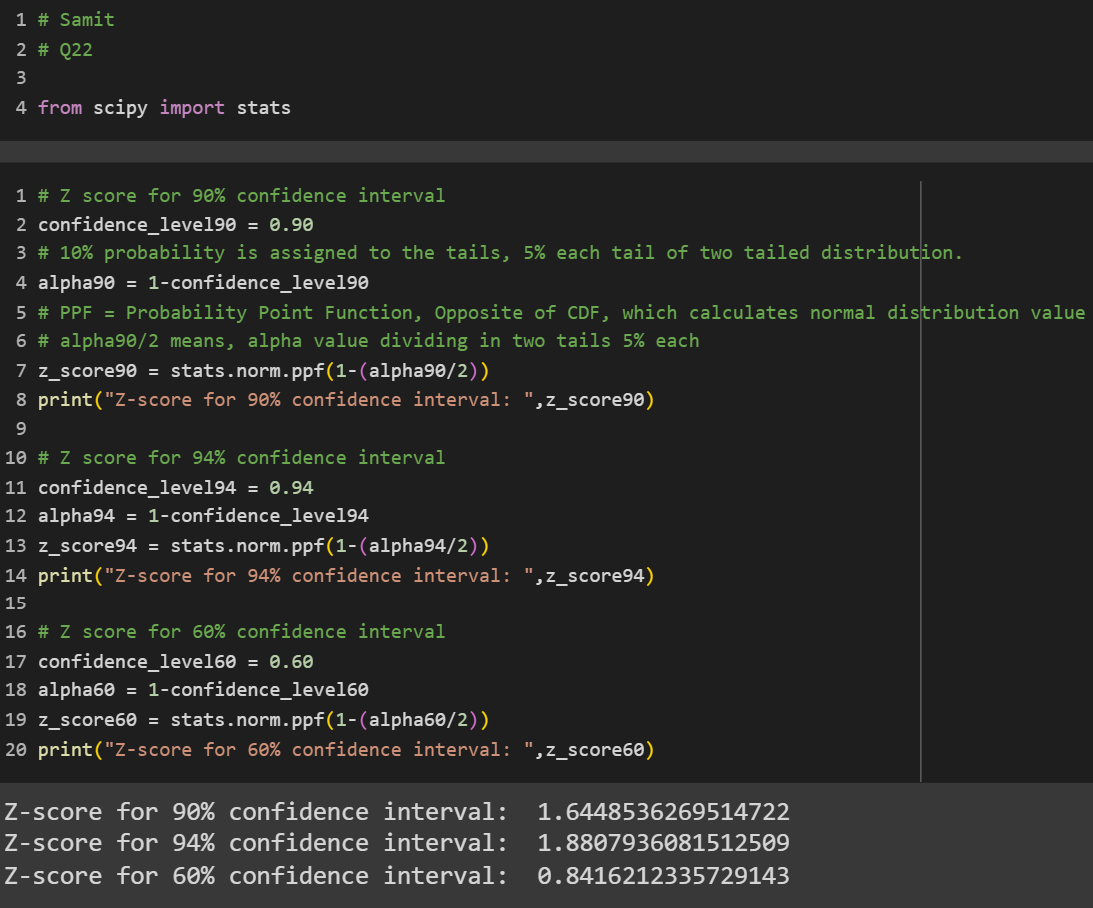
The column Waist Circumference seems to have the Normal Distribution which we can see from the below graph.



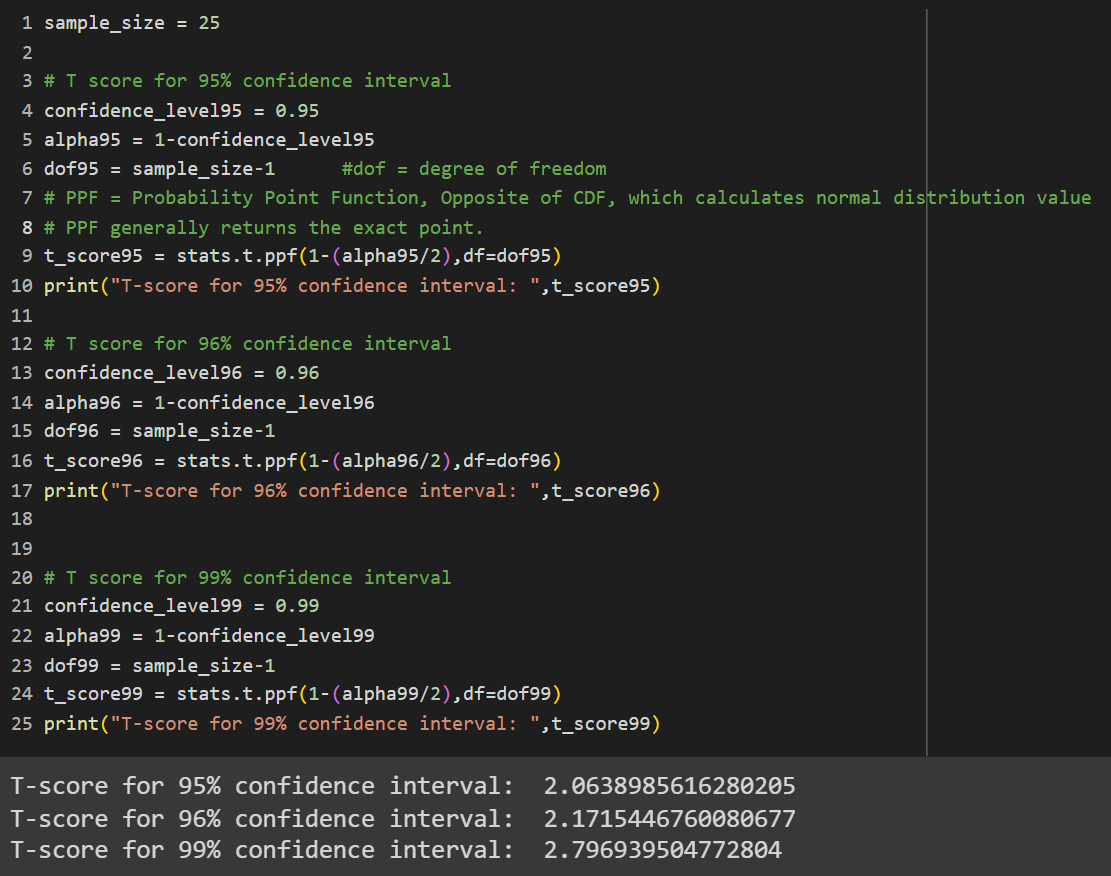
The column AT is right skewed that means AT column is not normally distributed which we can see from below graph.



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



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



Q24**)** 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

