|  |  |
| --- | --- |
| .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 | Ratio |
| 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 | Ordinal |
| Time on a Clock with Hands | Interval |
| 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: P (two heads and one tail) = Number of Event (two head and one tail) /Total Number of event (Three coins tossed)

= 3/8

= 0.375

= 37.5%

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

1. Equal to 1

b) Less than or equal to 4

c) Sum is divisible by 2 and 3

Ans: Number of possible outcomes for the above event is

Number of event (two dice rolled) =6^2 =36

1. P (sum is equal to 1)

= Number of outcomes / Total number of outcomes

= 1/36.

1. P (sum is less than or equal to 4)

= Number of outcomes / Total Number of outcomes

= 10/36

= 5 / 18

1. P (sum is divisible by 2 and 3)

= Number of outcomes / Total number of outcomes

= 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:

There are of total 7 balls including a blue ball. We need to exclude a blue ball so that we have 5 balls and 2 blue balls so we can consider it as

1. Probability of drawing the first ball that is not blue:

P (non-blue on first draw) = number of non-blue balls / total number

of balls

= 2(red balls) + 3(green) /7

= 5/7

1. Probability of drawing the second ball that is not blue after the first draw:

Non blue ball (2 red + 2 green)

P (non- blue ball on second draw/ non-blue on first draw)

= number of remaining non-blue balls/total number

of remaining balls

= 4/6

= 2/3

1. Probability of both events happening together:

P (both draws are non-blue) = P (non-blue on first draw) × P (non-blue on second draw /non-blue on first draw)

= 5/7 × 2/3

= 10/21

Therefore, the probability that none of balls drawn is 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: The expected number of candies for a randomly selected child is 3.315

Explanation: The formula for calculating the expected value(mean) is given by:

E(X) = Ei (Xi\*Pi)

E(X) = (1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.0005) + (6\*0.01) + (2\*0.120)

= 0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.24

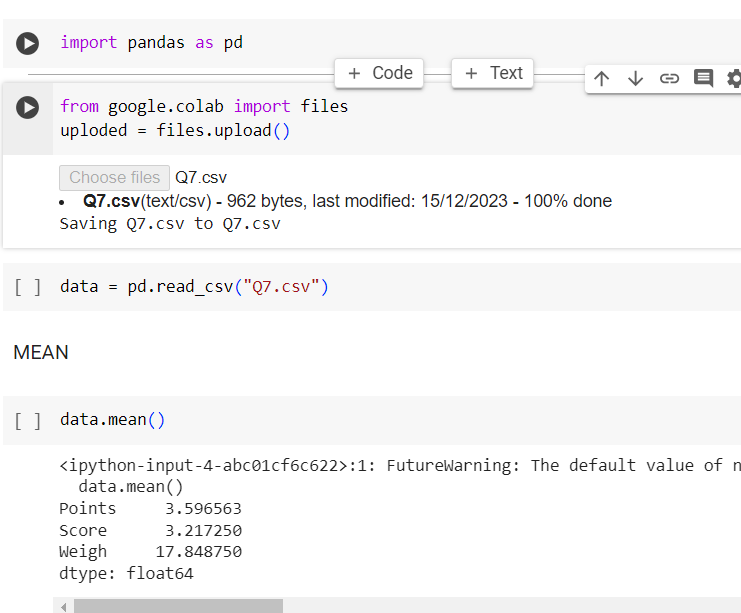
E(X) = 3.09

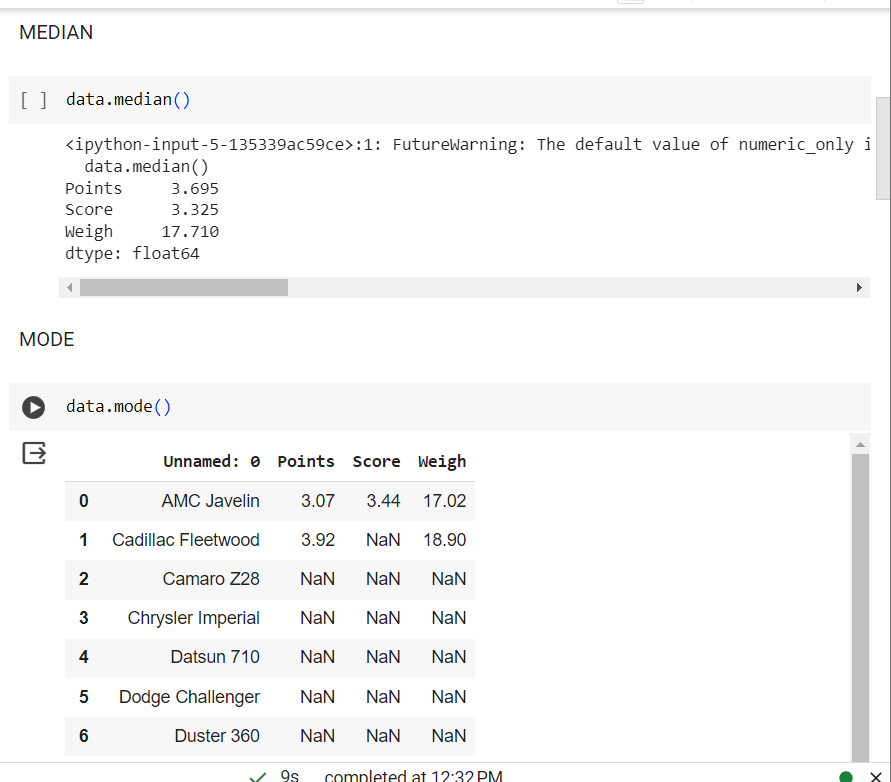
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

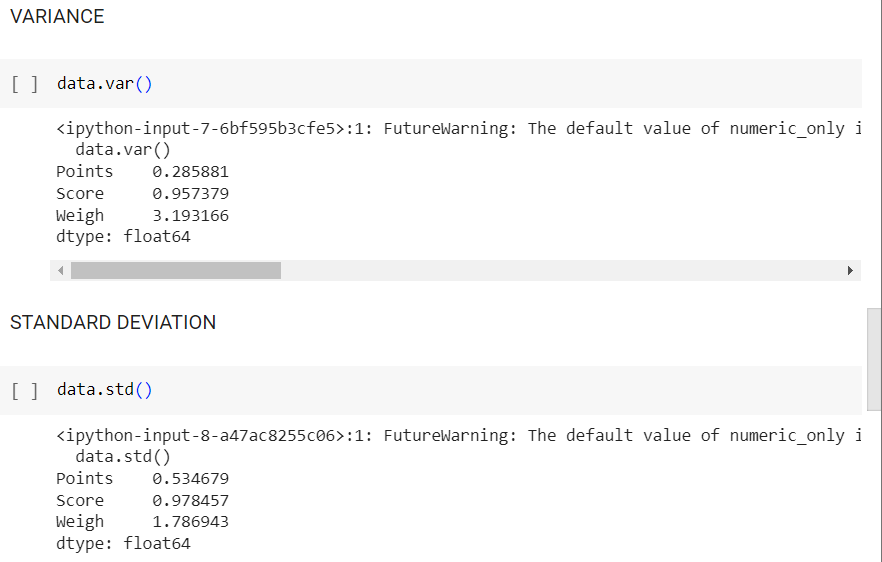
* For Points, Score, Weigh>

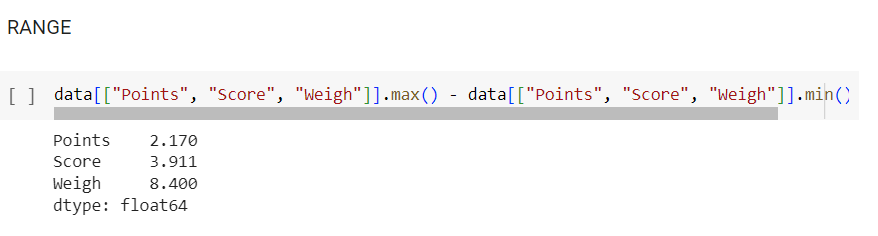
Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Use Q7.csv file**

****

****

****

****

**Inference:**

Mean

* Points = 3.59
* Score = 3.21
* Weigh = 17.84

Median

* Points = 3.68
* Score = 3.32
* Weigh = 17.71

Mode

* Points = 3.07
* Score = 3.44
* Weigh = 17.02

Variance

* Points = 0.28
* Score = 0.95
* Weigh = 3.19

Standard Deviation

* Points = 0.53
* Score = 0.97
* Weigh = 1.78

Range [Max - Min]

* Points = [3.59 – 4.93]
* Score = [3.21 – 5.42]
* Weigh = [17.84 – 22.9]

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?

Ans: Expected Value is also known as mean or average.

Formula of Mean

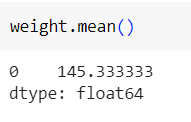
E(X) = ∑(Xi)/n

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

= 1308/9

E(X) = 145.33

So, the expected value of the weight of a randomly chosen patient is approximately 145.33 pounds



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

**Use Q9\_a.csv**

**Ans**: Skewness for car speed and distance

Speed= -0.11

Distance = 0.80

Kurtosis for car speed and distance

Speed = -0.50

Distance = 0.40



**SP and Weight (WT)**

**Use Q9\_b.csv**

Kurtosis for car speed and distance

Speed = -0.50

Distance = 0.40

****

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



Ans : The above histogram is right skewed histogram , also known as postivily skewed or right tail of the distribution when compared to the left side. In other words we can say that themajority of data points are concentrated on the left side, and there are fewer data points on the right side creating a tail that will extends to the right.

* The mean is typically greater than the median.
* The majority of data points are clustered on the left side of the histogram.

Boxplot:

Here, the box plotshows the outliers in upperbound. It means that identifying data points that are significantly higher than the majority of the dataset.

**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: To calculate the confidence intervals for the average weight of adult males in Mexico,

Confidence Interval (CI) = ˉx ± (*Zσ/√n*)

Where,

x (sample mean) = 20

σ (standard deviation) = 30

n (sample size) = 2000

Z = Z-score

The Z-score for 94%, 96% and confidence level are approximately 1.88, 1.96 and 2.33 respectively.

Let’s calculate the confidence intervals:

1. 94% Confidence interval:

Margin of Error = 1.88 × 30/√2000

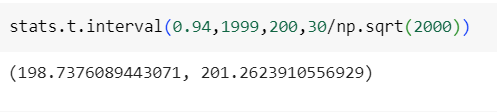
= 1.261

Confidence Interval = 200 - 1.261

= 198.739

Confidence Interval = 200 + 1.261

= 201.261



1. 96% Confidence Interval:

Margin of Error = 1.96 × 30/√2000

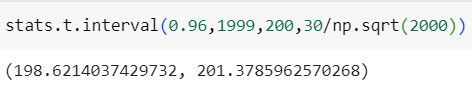
= 1.314

Confidence Interval = 200 - 1.314

= 198.621

Confidence Interval = 200 + 1.314

= 201.378



1. 98% Confidence Interval:

Margin of Error = 2.33 × 30/√2000

= 1.563

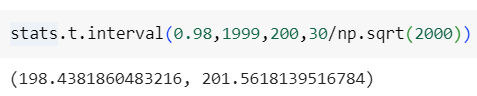
Confidence Interval = 200 - 1.563

\,

‘/ = 198.437

Confidence Interval = 200 + 1.563

= 201.563



Therefore, the estimated average weight of adult meals in Mexico with 94%, 96%, and 98% confidence are (198.739 - 201.261), (198.686 - 201.314) and (198.437 - 201.563) respectively.

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

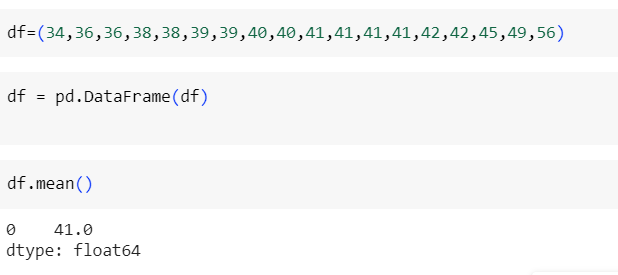
**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

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

Ans: I) Let’s calculate the mean, median, variance and standard deviation for the set of scores:

Scores: 34, 36, 36, 38, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

1. Mean (Average):



Mean = ∑(xi)/N

=34+36+36+38+38+39+39+40+40+41+41+41+41+42+

42+45+49+56

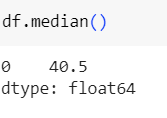
= 738/18

= 41

1. Median:

Generally Median is the middle value. If there is an even number of observations, the median is the average of the two values.

Median = 40.5



1. Variance:

Variance = ∑(xi-x)2/N

= (34-41)2+(36-41)2+(36-41)2+(38-41)2

+(38-41)2+(39-41)2+(39-41)2

+(40-41)2+(40)2+(41-41)2

+(41-41)2+(41-41)2+(41-41)2

+(42-41)2+(42-41)2+(45-41)2

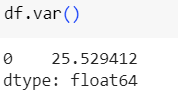
+(49-41)2+(56-41)2/18

= 14+ 25+ 25+ 9+ 9+ 4+ 4+ 1+ 1+ 0+ 0+ 0+ 0+1+ 1+

+16+ 64+ 225/18

= 474/18

Variance= 26.33

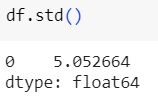


1. Standard Deviation:

Standard Deviation=√Variance

= √26.33

= 5.13



Therefore, let’s interpret the results:

Mean: The average score is 41.

Median: The median score is 40.5.

Variance: The variance score is 26.33

Standard Deviation: The standard deviation is 5.13

II) Interpretation:

1. The mean and median are close, suggesting that the data is not heavily skewed.
2. The variance and standard deviation indicate the spread or dispersion of the scores. A higher standard deviation implies more variability in the scores.
3. Mean is greater than median.

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 that the distribution is approximately symmetric. In a symmetric distribution, the data is evenly distributed on both sides of the center, and there is a balance between the left tails and the right tails.

Mean=Median then it is a perfectly symmetric distribution, the skewness is Zero.

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

Ans: When mean is greater than median in a distribution, it indicates that the distribution is positively skewed. Positive skewness means that the tail on the right side of the distribution is longer or fatter than the left side, and the majority of the data points are concentrated on the left side.

The nature of positively skewness distribution:

* The mean is pulled to the right by the presence of relatively larger values in the right tail.
* The median is generally closer to the left because it is less affected by extreme values.

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

Ans: When the median is greater than the mean in a distribution, it indicates that the distribution is negatively skewed. Negative skewness means that the tail on the left side of the distribution is longer or fatter than the right side, and the majority of the data points are concentrated on the right side.

The nature of negatively skewed distribution:

* The mean is pulled to the left by the presence of relatively smaller values in the left tail.
* The median is generally closer to the right because it is less affected by extreme values.

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

Ans: A positive kurtosis value indicates that a dataset has heavy tails and a relatively pronounced peak near the mean, compared to a normal distribution. Kurtosis measures the “tailedness” of a probability distribution, reflecting whether the data are heavy-tailed or light-tailed relative to a normal distribution.

Positive kurtosis indicates leptokurtosis, meaning that the tails of the distribution are fatter than those of a normal distribution. In a leptokurtic distribution:

* The data have more extreme value (outlines) than would be expected in a normal distribution.
* The distribution has a more peaked or taller central region compared to a normal distribution.

Positive kurtosis (greater than 0): Indicates leptokurtosis, heavy tails, and a pronounced peak.

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

Ans: A negative kurtosis value indicates that a dataset has lighter tails and a flatter peak compared to a normal distribution. Kurtosis measures the “tailedness” of a probability distribution and reflects whether the data are heavy-tailed or light-tailed relative to a normal distribution.

Negative kurtosis indicates platykurtosis.

* The tails of the distribution are less extreme the those of a normal distribution.
* The central peak is flatter or shorter compared to a normal distribution.

Negative kurtosis (less than 0): Indicates lighter tails, and a flatter peak.

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



What can we say about the distribution of the data?

Ans: The above Boxplot is not normally distributed because the median is towards the higher values

What is nature of skewness of the data?

Ans: The data is skewed towards left.

The whisker range of minimum value is greater than the maximum

What will be the IQR of the data (approximately)?   
 Ans: The Inter Quantile Range (IQR) = Q3(upper quartile) – Q1(lower quartile)

= 18 – 10

= 8

Q19) Comment on the below Boxplot visualizations?



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

Ans: First boxplot there is no outliers.

Second, both the box shares the equal portion of median this is approximately in the range between 250 to 275 and 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

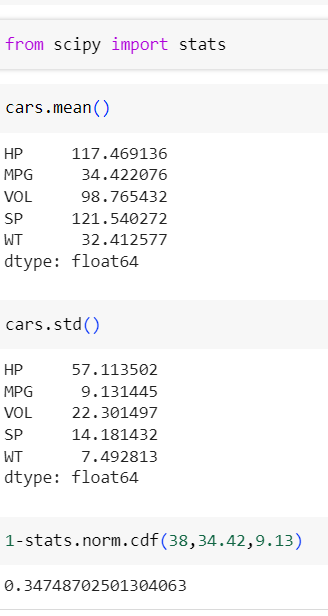
Data \_set: Cars.csv

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

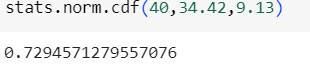
MPG <- Cars$MPG

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

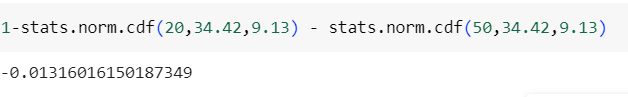
Ans: a) P(MPG>38) = 0.348



1. P(MPG>38) = 0.729



1. P(20<MPG<50) = 0.0131



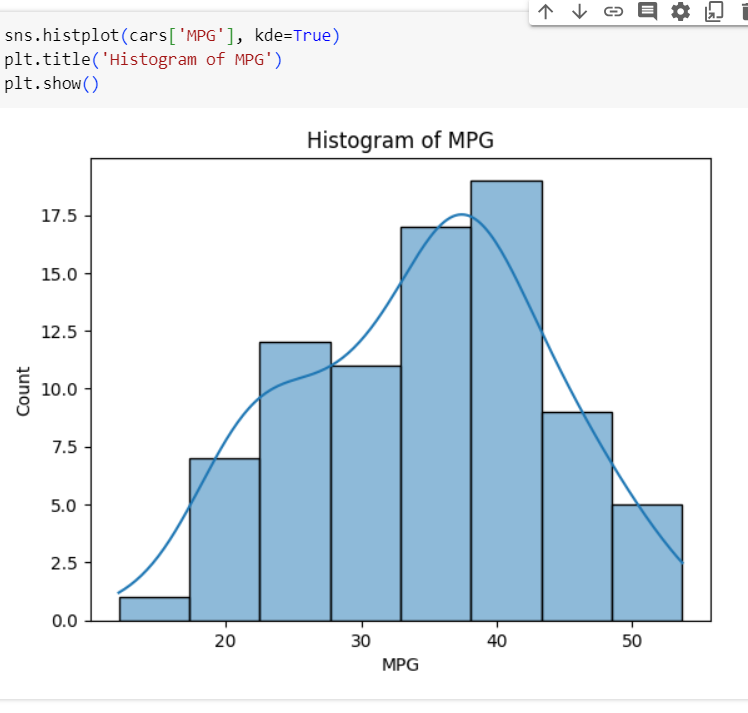
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans: MPG of cars follows normal distribution.



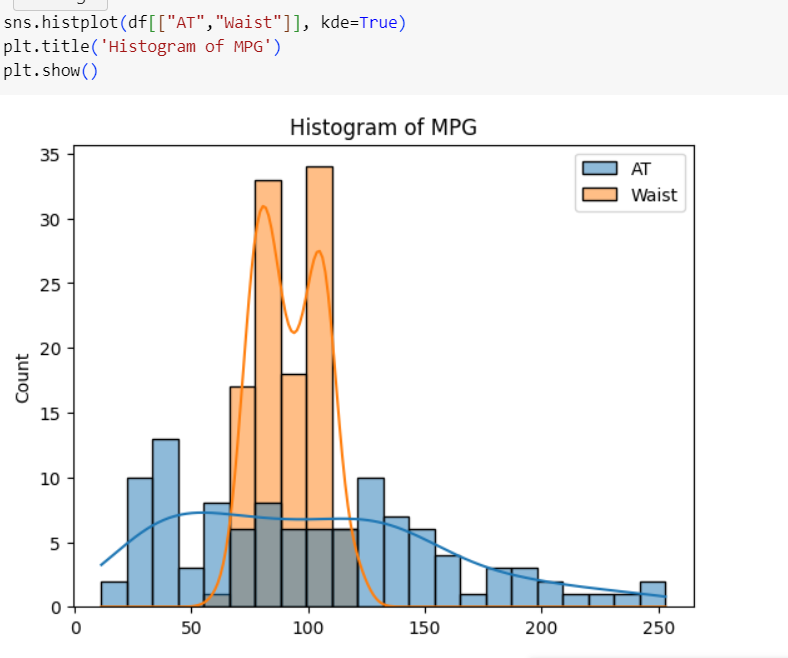


1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Ans:

Adipose Tissue (AT) and Waist Circumference (Waist) not follow Normal distribution.



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

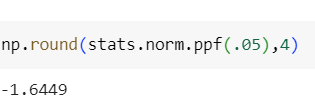
Ans: The Z-score represents the number of standard deviations a data point is from the mean in a normal distribution.

The formula for calculating the Z-score for a confidence interval is:

Z = ±Z value for the desired confidence interval

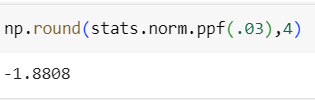
1. 90% confidence interval (two-tailed):

* The central region of a standard normal distribution for a 90% confidence interval of a Z-value of ± 1.6449



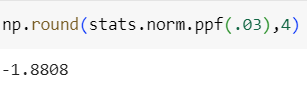
1. 94% confidence interval(two-tailed):

* The central region for a94% confidence interval corresponds to a Z-value of ±1.8808



1. 60% confidence interval (two-tailed):

* The Z-value for a 60% confidence interval, approximate value is ± 0.8416



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

Ans: Given,

n = 25

Formula of T-test:

t = ±t

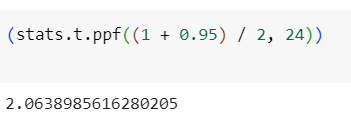
df = n-1

= 25-1

= 24.

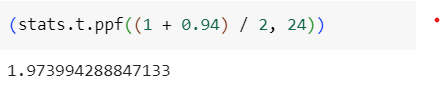
1. 95% confidence interval:

T-value for a 95% confidence interval = -2.0639



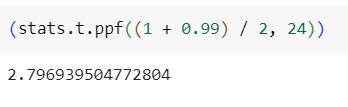
1. 94% confidence interval:

T-score for 94% Confidence Interval = -1.973



1. 99% confidence interval:

T-score for 99% Confidence interval = -2.7969



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:

Given:

* Population mean (µ) = 270 days
* Sample mean(ˉ*x*ˉ) = 260 days
* Sample standard deviation(s) = 90 days
* Sample size(n) = 18 bulbs



t\_score = -0.471

stats.t.cdf ( -0.471, 17) = 0.32 =32%