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

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale | Ordinal |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Ordinal |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ratio |
| Years of Education | Ordinal |

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

A) Three coins tossed, so sample data = 2^3 = 8

Total outcomes:

H, H, H T, T, T

H, H, T T, T, H

H, T, H T, H, T

T, H, H H, T, T

So, the probability for obtaining two heads and one tail = 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

A) Two Dice are rolled, so sample data = 6^2 = 36

Total outcomes = (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)

1. The probability is equals to 0 (zero) as there is no scope for getting “sum is equal to 1” when two dice are rolled.
2. The total outcomes for getting “sum is less than or equal to 4” are 6, so the probability = 6/36 = 1/6
3. The total outcomes for getting “sum is divisible by 2 and 3” are 6, so the 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?

1. Total no.of balls = 7

Balls to be drawn together = 2

Neither should be blue, so 2 red + 3 green = 5 balls

i.e 2c2 +3c2 / 7c2

therefore, ncr = n! / (n-r)! r!

probability = 5\*4\*3\*2\*1 / (5-2)! 2! 120 / 12

= = 10 / 21 = 0.476

7\*6\*5\*4\*3\*2\*1 / (7-2)! 2! 5040 / 240

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

A screenshot of a computer code

Description automatically generated

A white background with black and green text

Description automatically generated

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.**

import pandas as pd #importing pandas library

data = pd.read\_csv("C:\\Users\\samee\\OneDrive\\Desktop\\DS assingments\\data\\Q7.csv") #copying the path where the requried csv file there and pasting to get the data

data.shape #gives the total number of rows and columns in our file

#calculations of "points"

data["Points"].mean()

data["Points"].median()

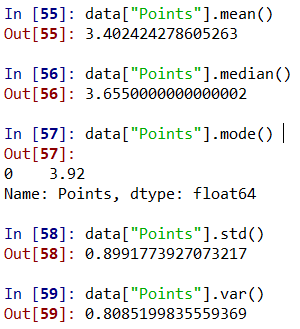
data["Points"].mode()

data["Points"].std()

data["Points"].var()

rangee = data["Points"].max() - data["Points"].min() #range = max value - min value

range

A computer code with text

Description automatically generated with medium confidence

#calculations of "Score"

data["Score"].mean()

data["Score"].median()

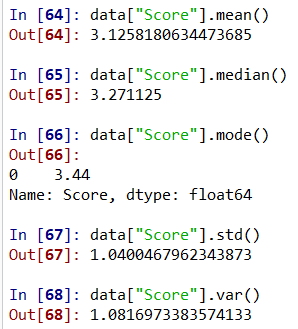
data["Score"].mode()

data["Score"].std()

data["Score"].var()

rangee = data["Score"].max() - data["Score"].min() #range = max value - min value

rangee

A close-up of a computer code

Description automatically generated

#calculations of "Weigh"

data["Weigh"].mean()

data["Weigh"].median()

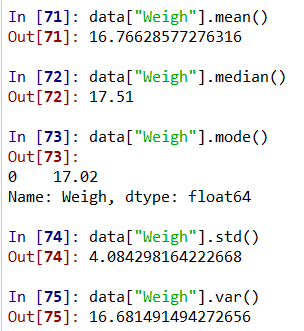
data["Weigh"].mode()

data["Weigh"].std()

data["Weigh"].var()

rangee = data["Weigh"].max() - data["Weigh"].min() #range = max value - min value

range

A computer screen shot of a computer screen

Description automatically generated

Inference:

1. The mean of points 3.402, score 3.125, weigh 16.766
2. The median of points 3.655, score 3.271, weigh 17.51
3. The mode of points 3.92, score 3.44, weigh 17.02
4. The SD of points 0.899, score 1.040, weigh 4.084
5. The variance of points 0.808, score 1.081, weigh 16.681
6. The Range of points 4.644, score 4.466, weigh 21.113

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?

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

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

**Cars speed and distance**

**Use Q9\_a.csv**

A graph with numbers and text

Description automatically generatedA computer code with numbers and symbols

Description automatically generated

Inference: so, the cars speed data values are negatively skewed and have less peakedness which is considered as platy kurtosis.

A computer screen with text

Description automatically generated

A graph with numbers and lines

Description automatically generated A close-up of a computer code

Description automatically generated

Inference: so, the cars distance data values are positively skewed and have slight peakedness which is considered as lefto kurtosis.

**SP and Weight (WT)**

**Use Q9\_b.csv**

A computer screen shot of a computer

Description automatically generated

A graph with numbers and lines

Description automatically generatedA close-up of a computer code

Description automatically generated

Inference: so, the SP data values are positively skewed and have high peakedness which is considered as lefto kurtosis.

A computer screen shot of a computer

Description automatically generated

A graph with numbers and a bar

Description automatically generated A close-up of a code

Description automatically generated

Inference: so, the WT data values are negatively skewed and have slight peakedness which is considered as lefto kurtosis.

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



Histogram:

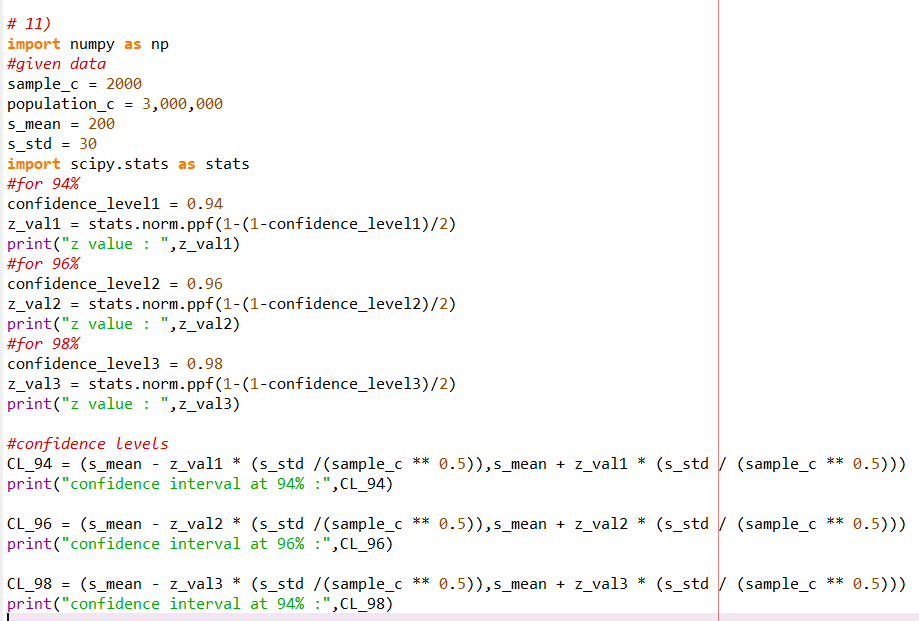
* We build histograms for continuous data.
* And Histogram is constructed based on based on the frequency table.
* The above histogram is between chick weights and weight frequencies.
* The highest frequency is between 150 – 200.
* The lowest frequency is between 0 – 50.
* And the above histogram is positively skewed.
* As it is positively skewed the data may contain outliers.
* Outliers need to be removed, so by using boxplot we can remove outliers.



Boxplot:

* The above graph is a boxplot in which we can identify the outliers.
* The boxplot considers where it has an average number of datapoints.
* So, in boxplot we calculate IQR (Inter quartile range)
* IQR consists of Q1, Q2, Q3.
* Q1 is considered as 25th percentile, Q2 is considered as 50th percentile, Q3 is considered as 70th percentile.
* In this boxplot we have data points above upper whisker length, so we need to calculate upper whisker length to get number of outliers.
* And we need to remove outliers using boxplot to get correct analysis of the data.

**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?



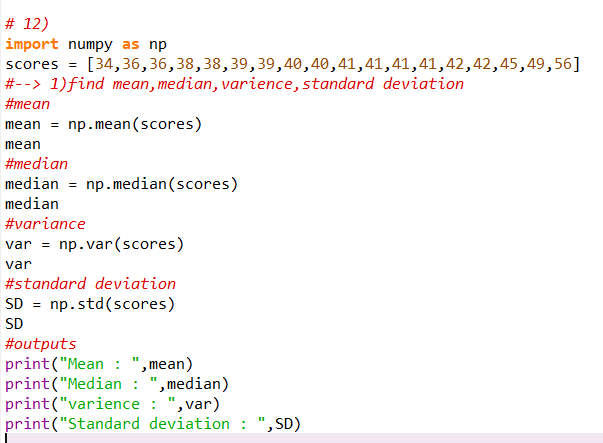
A screenshot of a computer code

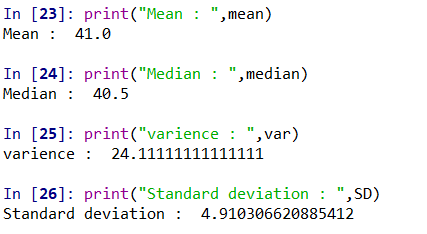
Description automatically generated

**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?





2) mean: mean is the average of all the student marks, here the mean is 41.0 , so it tells us that average marks are close to 41.0

Median: median means it is the middle value of the complete data, when it is in ascending order. so here the median is 40.5, so the marks of half students are above 40.5 and half students’ marks are below 40.5

Variance: variance measures spread of data points from the mean, higher variance suggests that the data points are more spread out

Standard deviation: standard deviation is deviation of the data points from the center. And this is the square root of variance, in this case the SD is 24.11

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

When the mean, median of data are equal then it tells us that the data distribution is symmetric. In this case the skewness of the distribution is close to zero or very small.

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

When the mean of the data is greater than the median, it tells us that the distribution is positively skewed.

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

When the median of the data is greater than the mean, it tells us that the distribution is negatively skewed.

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

A positive kurtosis value indicates that it has more peaked central distribution compared to normal distribution. Kurtosis is a statistical measure that calculates peakedness.

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

A negative kurtosis value indicates that it has lesser peaked central distribution compared to normal distribution. Kurtosis is a statistical measure that calculates peakedness.

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



What can we say about the distribution of data?

* In the above boxplot the complete range is between 2 and 18 data points.
* We have the midpoint of the data between 14 and 16 data points.
* So as from this data the left side from the midpoint is considered as MIN (lower whisker length) and right side from the midpoint is considered as MAX (upper whisker length).
* From this we can say Q1 = 10, Q2 = between (14 – 16), Q3 = above(18)

What is the nature of skewness of the data?

Based on the above boxplot, Q1 is 10 and Q3 is above 18 and considering Q2 is slightly closer to 14 (data point) which is slightly closer to Q1, so we can consider that there is slightly positive skewness.

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

The IQR is the difference between Q3 and Q1.

Here, Q1 = 10 and Q3 = exceeds 18 -> let’s say it is around 18.5

Then, IQR = Q3 – Q1 = 18.5 – 10 = 8.5

Q19) Comment on the below Boxplot visualizations?



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

* In the above image we have two different boxplots, the range for both the boxplots is between 200 and 325.
* So, from boxplot 1 we can say that the box plot range starts between 225 and 250, ends between 275 and 300.
* The midpoint of the box plot is between 250 and 275.
* The range between Q1 and Q3 is from approximately 225 to 300. This range

Contains the middle 50% of data.

* Median which is between 250-270, this value separates the lower whisker length and upper whisker length.
* Since median is closer to the lower end of the box, we may consider that the data could be positively skewed.
* So, from boxplot 1 we can say that the box plot range starts between 175 and 200, ends at 325.
* The midpoint of the box plot is between 250 and 275.
* The range between Q1 and Q3 is from approximately 175 to 325. This range

Contains the middle 50% of data.

* Median which is between 250-270, this value separates the lower whisker length and upper whisker length.
* Since median is closer to the lower end of the box, we may consider that the data could be positively skewed.
* The data points falling beyond upper whisker length are considered as outliers.

Inference between both boxplots:

* Boxplot 2 has a wider range (175-325) compared to boxplot 1 (225-300), this tells us that the spread of data in boxplot 2 is larger.
* IQR of Boxplot 2 is larger than boxplot 1.
* Medians of both the boxplots were same.

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

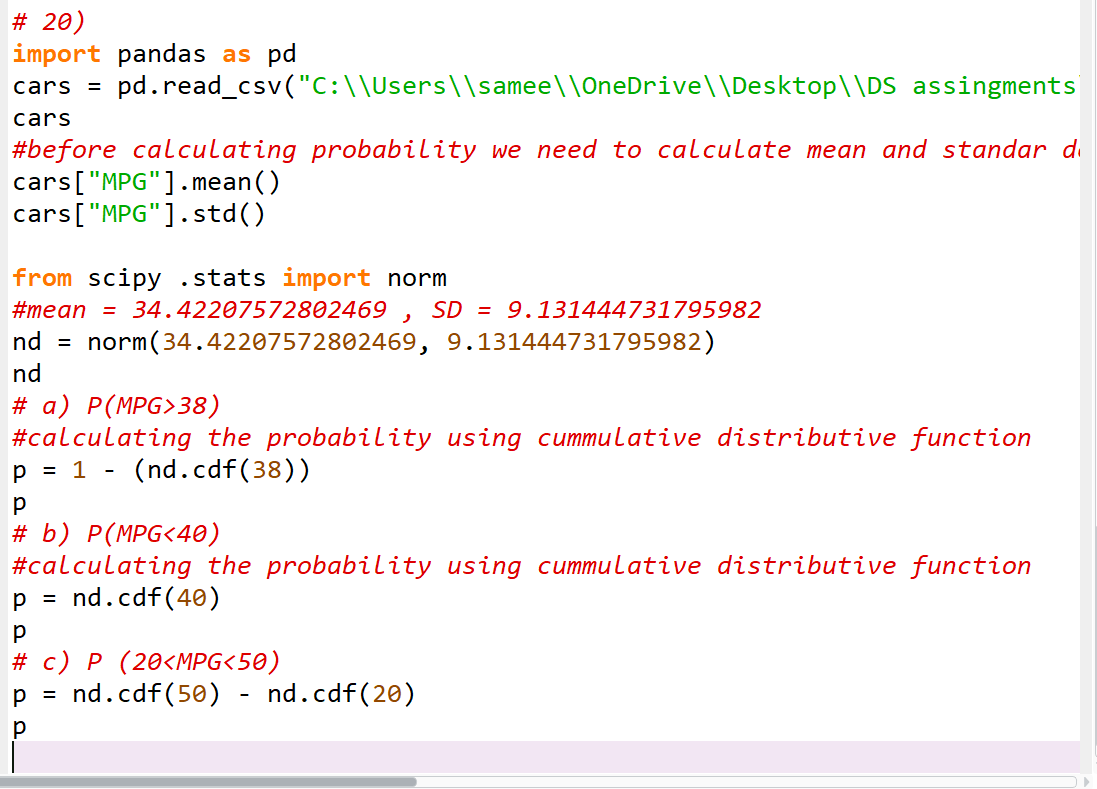
Data \_set: Cars.csv

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

MPG <- Cars$MPG

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

c. P (20<MPG<50)



A screenshot of a computer code

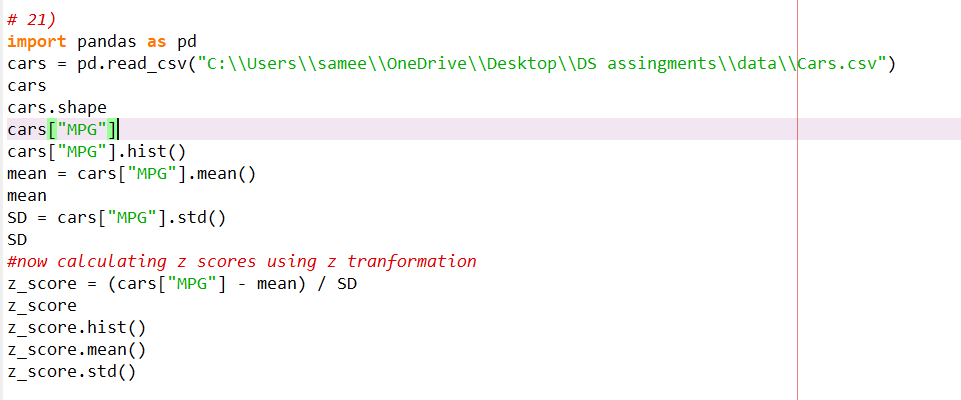
Description automatically generatedA screenshot of a computer

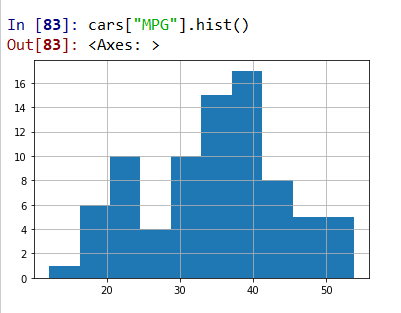
Description automatically generated

Q 21) Check whether the data follows normal distribution

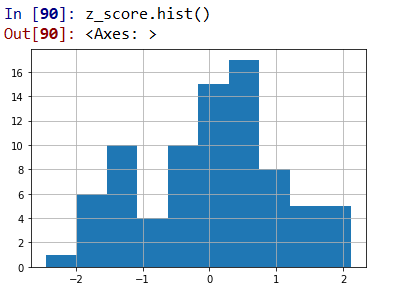
1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv



 A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

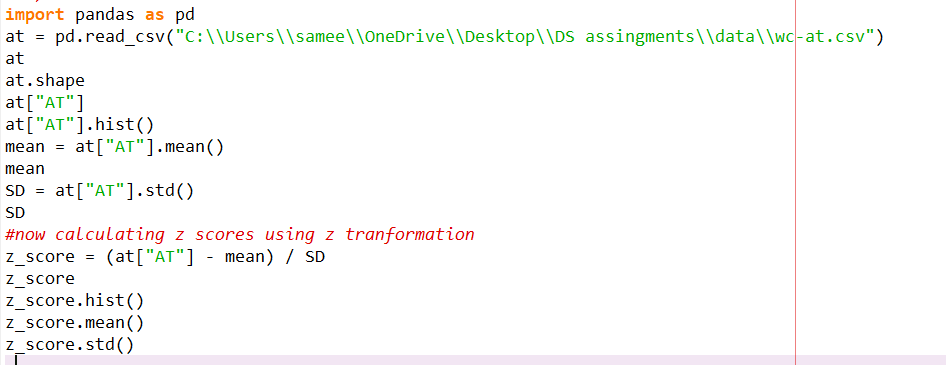
As the z\_score value equals to 1 and mean is approximately equals to 0

It follows normal distribution.

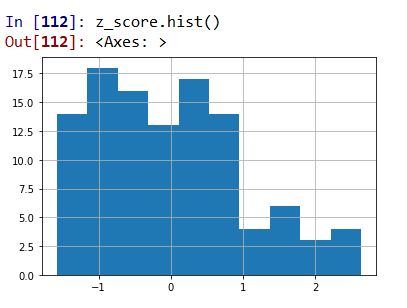
1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

#Adipose tissue



A graph with numbers and lines

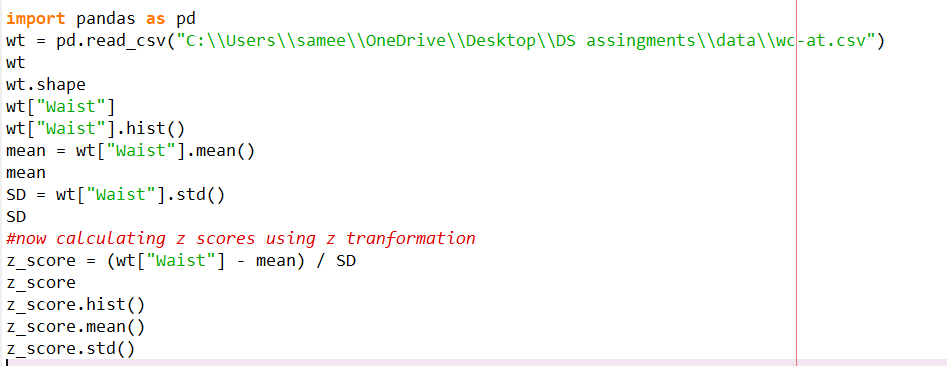
Description automatically generated

A screenshot of a computer code

Description automatically generated

The adipose tissue (AT), mean != 0 and SD != 1 ,so it does not follow normal distribution.

#waist



A graph with blue squares

Description automatically generated A graph with numbers and a bar

Description automatically generated

A black text on a white background

Description automatically generated

The Waist Circumference (WT), mean != 0 and SD != 1 ,so it does not follow normal distribution.

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

We calculate z score for sample size greater than 30, We can use z distribution and its percentiles.

The formula to calculate z score = Z α/2

90% confidence level:

90%, so α = 1 – 0.90 = 0.10, half of this α is 0.05

Look up the z score corresponding to the cumulative probability of

1 – 0.05 = 0.95

In standard normal distribution table Zscore = 1.645

94% confidence interval:

94%, so α = 1 – 0.94 = 0.06, half of this α is 0.03

Look up the z score corresponding to the cumulative probability of

1 – 0.03 = 0.97

In standard normal distribution table Zscore = 1.880

60% confidence interval:

60%, so α = 1 – 0.60 = 0.40, half of this α is 0.20

Look up the z score corresponding to the cumulative probability of

1 – 0.20 = 0.80

In standard normal distribution table Zscore = 1.842

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

We calculate t score when sample size is less than 30

Here, we use the T distribution instead of standard deviation.

t = tα/2

where tα/2 is corresponding t score to the desired significance alpha and the df is the degrees of freedom i.e df = n – 1

95% of confidence interval, so α = 1 – 0.95 = 0.05

Now, α/2 = 0.05 / 2 = 0.025.

Now degree of freedom of a sample of size 25 (n) = n – 1 = 25 – 1 = 24

Look up the t score corresponding to the cumulative probability is 1 – 0.025 = 0.975 and df = 24 in the t distribution table.

Therefore, t score = 2.064

96% of confidence interval, so α = 1 – 0.96 = 0.04

Now, α/2 = 0.04 / 2 = 0.02

Now degree of freedom of a sample of size 25 (n) = n – 1 = 25 – 1 = 24

Look up the t score corresponding to the cumulative probability is 1 – 0.02 = 0.980 and df = 24 in the t distribution table.

Therefor t score = 2.398

99% of confidence interval, so α = 1 – 0.99 = 0.01

Now, α/2 = 0.01 / 2 = 0.005

Now degree of freedom of a sample of size 25 (n) = n – 1 = 25 – 1 = 24

Look up the t score corresponding to the cumulative probability is 1 – 0.005 = 0.995 and df = 24 in the t distribution table.

Therefore, t score = 2.797

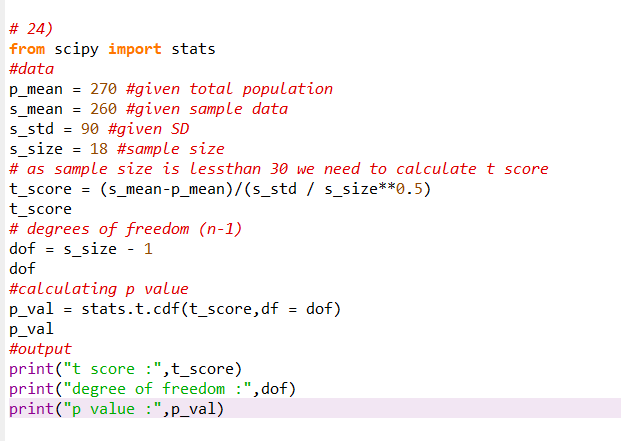
Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs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

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

from



Output:

A computer code with numbers and symbols

Description automatically generated

Therefore,

Hypothesis:

1)Null hypothesis (H0): the average bulb life is 270 days CEO’s claim

2)Alternative hypothesis (H1): the average bulb life is less than 270 days

Significance level:

α = 5% = 0.05

Calculation of t score:

The t score measures how many errors , the sample mean is away from the population mean

i.e T = sample mean – population mean

sample standard deviation / (n)^1/2

P\_value:

Find the p\_value and compare the p\_value with the significance level

Here, our p\_value = 0.321

Therefore p\_value < significance level

So, H0 is accepted and H1 is rejected.