|  |  |
| --- | --- |
| 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) | Categorical |

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 | Nominal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Interval |
| Socioeconomic Status | Interval |
| Fahrenheit Temperature | Ratio |
| 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 | Interval |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Ordinal |

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

Ans- probability of three coins are tossed :- =HHH,HHT,HTH,HTT,TTT,THT,TTH,THH

=“HHT” , “HTH” , “THH”

=3/8 = 0.37 is the probability

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 are rolled

=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

=a):- 0

b):- 1-1 , 1-2 , 1-3 , 2-1 , 2-2 , 3-1 = 6/36 = 0.166

c):- 1-5 , 2-4, 3-3, 4-2 , 5-1, 6-6 =6/36 = 0.166

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 :- 2 balls out of 7 including blue balls n(S)= 7C2 = 21

2 balls out of 5 without blue balls n(S) =5C2 = 10

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 =  1 \* 0.015  + 4\*0.20  + 3 \*0.65  + 5\*0.005  + 6 \*0.01  + 2 \* 0.12

= 0.015 + 0.8  + 1.95 + 0.025 + 0.06 + 0.24

= 3.090

=  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**

**Ans** Mean:

Mean = (Sum of all the values)

Mean of Points = 3.590625

Mean of Score = 3.21725

Mean of Weigh = 19.8478125

Median:

To calculate the median, we need to arrange the values in ascending order and find the middle value.

Median of Points = (3.07+3.15) / 2 = 3.11

Median of Score = (3.29+3.44) / 2 = 3.365

Median of Weigh = (18.805+18.755) / 2 = 18.78

Mode:

To calculate the mode, we need to find the value that appears the most number of times in the dataset

The mode for Points, Score, and Weigh are:

Mode of Points = 3.92

Mode of Score = 3.44

Mode of Weigh = No mode

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 Probability of selecting each patient = 1/9

W(x) 108, 110, 123, 134, 135, 145, 167, 187, 199

P(x)  1/9  1/9   1/9  1/9   1/9   1/9   1/9   1/9  1/9

Expected Value  =  (1/9)(108) + (1/9)110  + (1/9)123 + (1/9)134 + (1/9)135 + (1/9)145 + (1/9(167) + (1/9)187 + (1/9)199

= (1/9) ( 108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

= (1/9)  (  1308)

= 145.33

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

**Cars speed and distance**

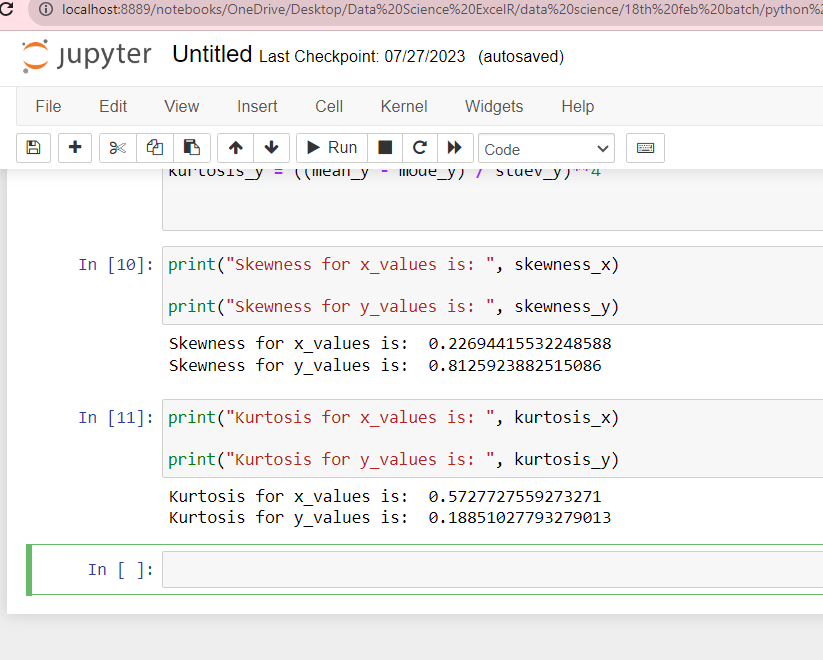
**Use Q9\_a.csv**

**Ans** Skewness for x\_values is: 0.22694415532248588

Skewness for y\_values is: 0.8125923882515086

Kurtosis for x\_values is: 0.5727727559273271

Kurtosis for y\_values is: 0.18851027793279013

****

**SP and Weight(WT)**

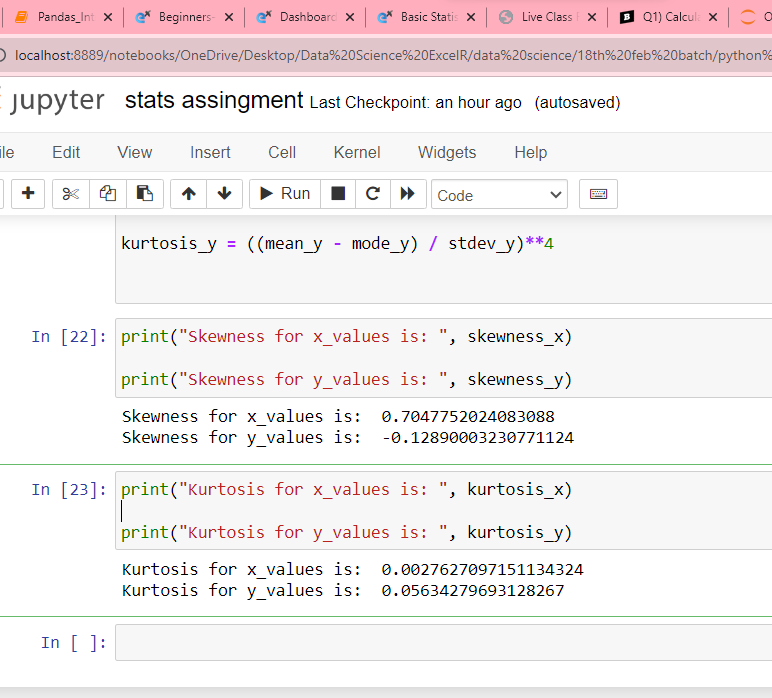
**Use Q9\_b.csv**

**Ans** Skewness for x\_values is: 0.7047752024083088

Skewness for y\_values is: -0.12890003230771124

Kurtosis for x\_values is: 0.0027627097151134324

Kurtosis for y\_values is: 0.05634279693128267

****

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



Ans -Chick weight data is right skewed or positively skewed.---- Yes

- More than 50% Chick Weight is between 50 to 150. ---- Yes

- Most of the chick weight is between 50 to 100. --- Yes



Ans - The data is right skewed.

- There areoutliers at upper 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?

Ans Using the **t-distribution**, it is found that:

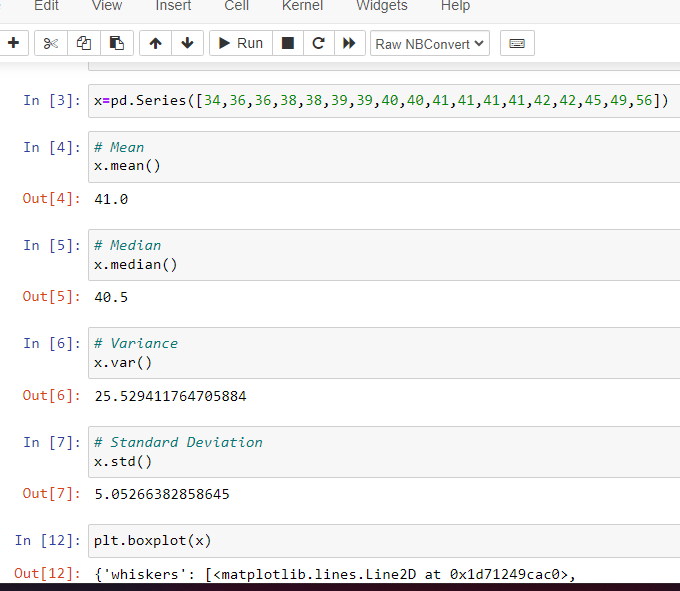
* The **94%** confidence interval is **(198.73, 201.27).**
* The **96%** confidence interval is **(198.61, 201.39).**
* The **98%** confidence interval is **(198.43, 201.57).**

**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**

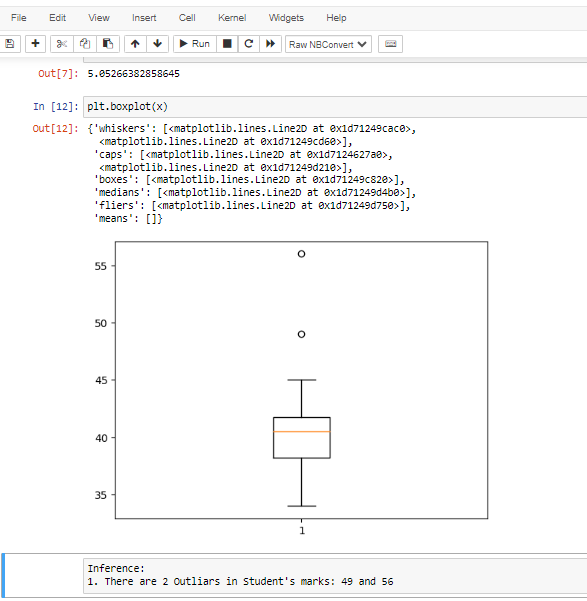
1. Find mean, median, variance, standard deviation.

Ans



1. What can we say about the student marks?

Ans



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

Ans The nature of skewness is zero, the distribution is symmetric, the mean equals the median, the skewness of the distribution is zero.

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

Ans The nature of skewness is positive, mean of positively skewed data will be greater than the median.

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

Ans The nature of skewness is negative ,mean of negatively skewed data will be greater then median

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

Ans Positive values of kurtosis indicate that distribution is peaked and possesses thick tails.

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

Ans negative values of kurtosis indicate that distribution is flatter and possesses thin tails

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 the median is towards the higher value

What is nature of skewness of the data?

Ans: The data is a skewed towards left. The whisker range of minimum value is greater than maximum

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

Ans: The Inter Quantile Range = 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 there are no outliers. Second both the box plot shares the same median that is approximately in a range between 275 to 250 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)

Ans: Prob\_MPG\_greater\_than\_38 = np.round(1 - stats.norm.cdf(38, loc= q20.MPG.mean(), scale= q20.MPG.std()),3)

print('P(MPG>38)=',Prob\_MPG\_greater\_than\_38)

P(MPG>38)= 0.348

* 1. P(MPG<40)

Ans: prob\_MPG\_less\_than\_40 = np.round(stats.norm.cdf(40, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<40)=',prob\_MPG\_less\_than\_40)

P(MPG<40)= 0.729

* 1. P (20<MPG<50)

Ans: prob\_MPG\_greater\_than\_20 = np.round(1-stats.norm.cdf(20, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('p(MPG>20)=',(prob\_MPG\_greater\_than\_20))

p(MPG>20)= 0.943

prob\_MPG\_less\_than\_50 = np.round(stats.norm.cdf(50, loc = q20.MPG.mean(), scale = q20.MPG.std()),3)

print('P(MPG<50)=',(prob\_MPG\_less\_than\_50))

P(MPG<50)= 0.956

prob\_MPG\_greaterthan20\_and\_lessthan50= (prob\_MPG\_less\_than\_50) - (prob\_MPG\_greater\_than\_20)

print('P(20<MPG<50)=',(prob\_MPG\_greaterthan20\_and\_lessthan50))

P(20<MPG<50)= 0.013000000000000012

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans

1. 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

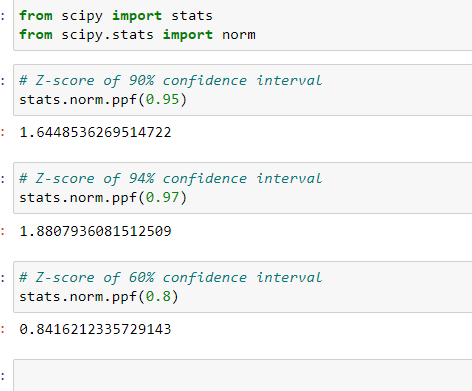
Ans: Adipose Tissue (AT) and Waist does not follow Normal Distribution





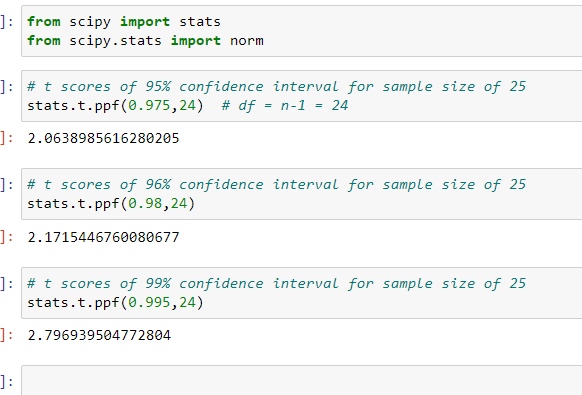
Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

Ans:



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

Ans



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: import numpy as np

Import scipy as stats

t\_score = (x - pop mean) / (sample standard daviation / square root of sample size)

(260-270)/90/np.sqrt(18))

t\_score = -0.471

stats.t.cdf(t\_score, df = 17)

0.32 = 32%