|  |  |
| --- | --- |
| 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 | Nominal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Nominal |
| 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 | Interval |
| SAT Scores | Interval |
| Years of Education | Ratio |

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

Ans) Chances of 2 heads and 1 tail= HHT, HTH, THH = 3

Total Outcomes from 3 coins= HHH, TTT, HTT, THT, TTH, THH, HTH, HHT=8

Probability= 3/8=0.375

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

1. Equal to 1

Ans:- 0, as even if both the dices have smallest number which is 1 the sum will still be 2 which is more than 1

1. Less than or equal to 4

Ans:- Total outcomes that we can get by rolling 2 dice are=2^6=36

Chances of sum less than or equal to 4 =(1,3), (1,1), (1, 2),(2,2),(2,1),(3,1)=6

Probability= 6/36=1/6

1. Sum is divisible by 2 and 3

Chances of sum of divisibility by 2 & 3=(3,3),(6,6), (1,5), (5,1), (2,4), (4,2)=6

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?

Total chances= 2R=2/7, 3G=3/7, 2B=2/7

Ball 1 Probability= 2/7+3/7=5/7 (considering its not blue)

Ball 2 Probability =4/6 (excluding the 1 ball taken out above)

Probability of both the balls=5/7+4/6=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:- 1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120=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.

Ans:- Using Q7.CSV Mean, median, mode, Variance & SD for Points, score & weigh are below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Mode** | **Variance** | **SD** |
| **Points** | 3.5965625 | 3.695 | 3.92 | 0.2858814 | 0.5346787 |
| **Score** | 3.2172500 | 3.325 | 3.44 | 0.957379 | 0.9784574 |
| **Weigh** | 17.8487500 | 17.71 | 17.02 | 3.193166 | 1.78694 |

**Use Q7.csv file**

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?

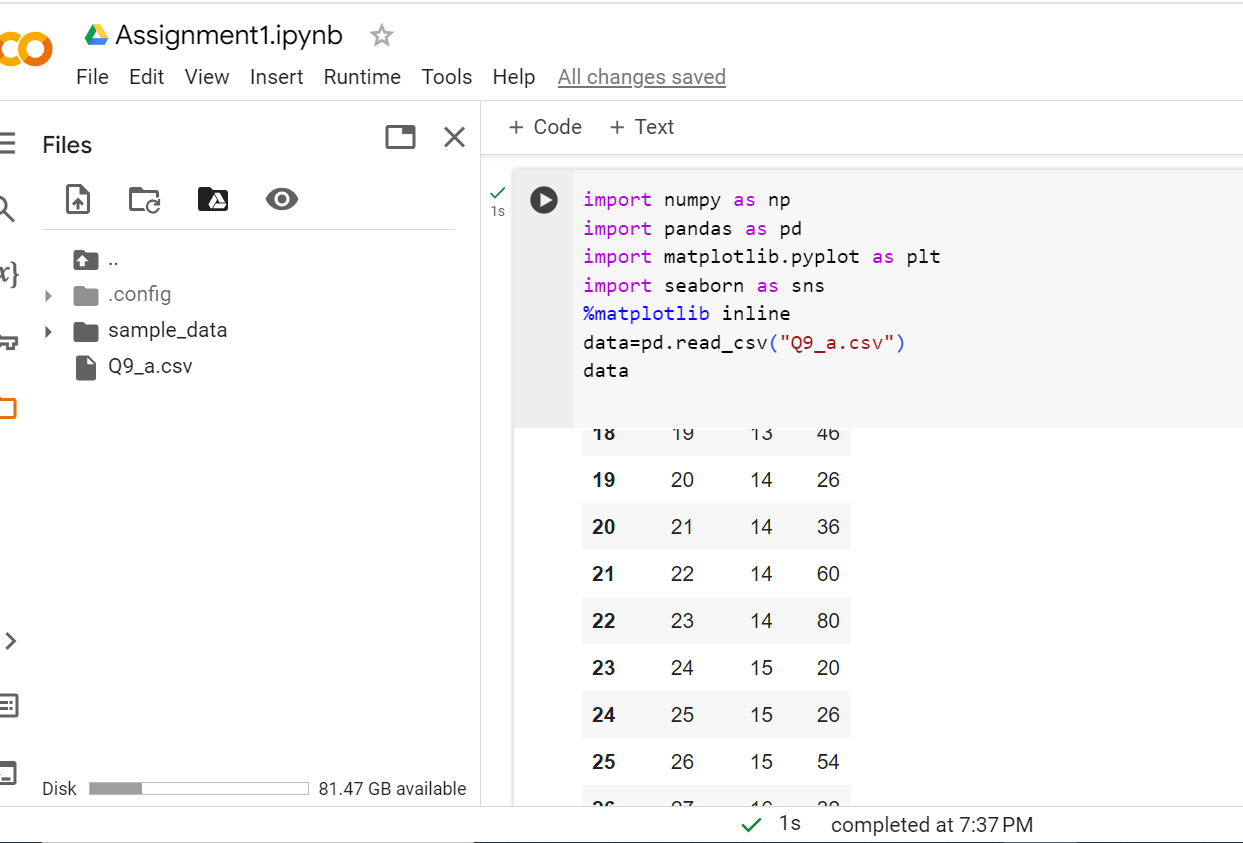
Ans:- Sum of X= 1308

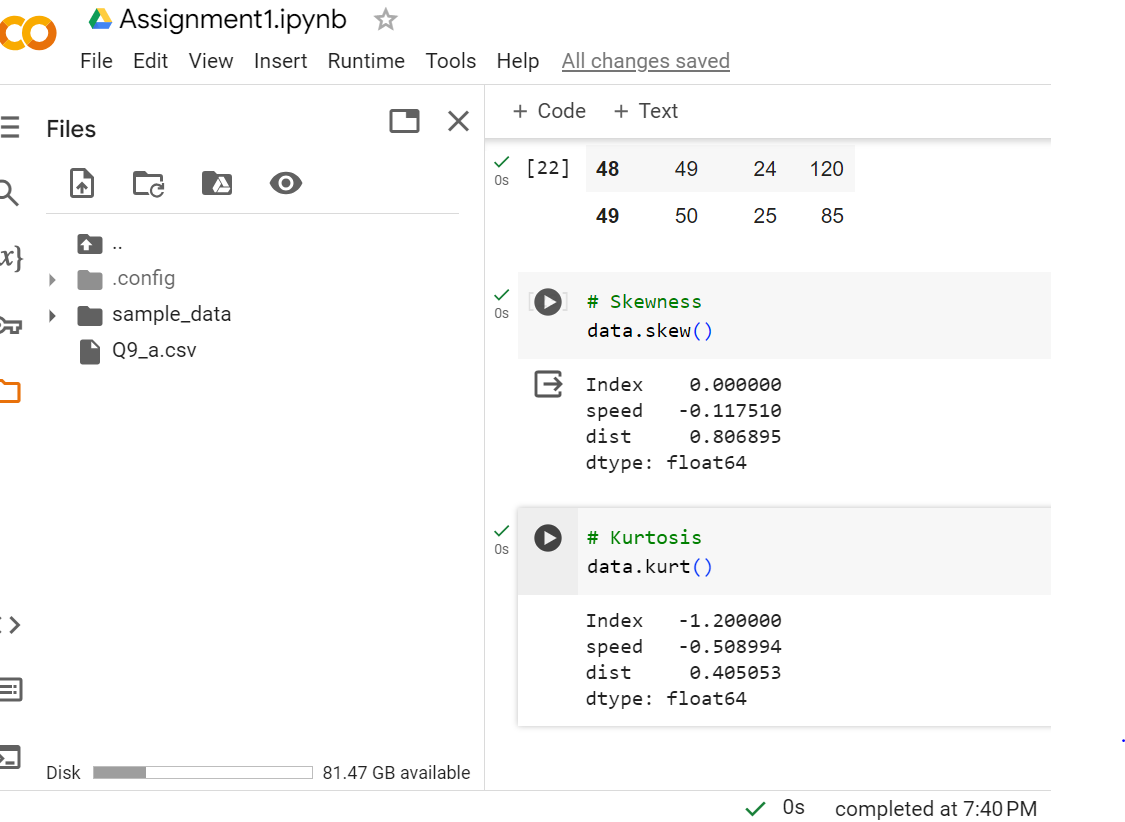
Expected value of the weight of the patient =1308/9=145.33 Pounds

**Q9) Calculate Skewness, Kurtosis & draw inferences on speed and distance**

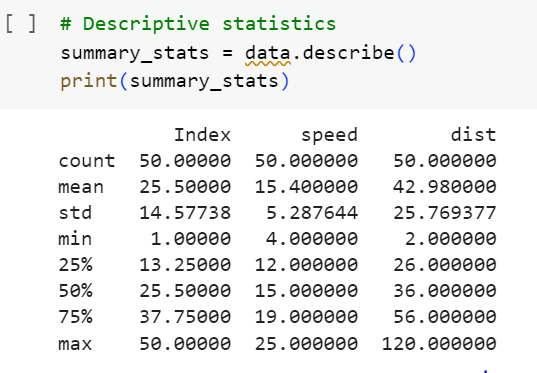
**Use Q9\_a.csv**

**Solution :**





Skewness for speed in above code is negative hence it is left skewed while skewness for distance is positive hence it is right skewed.



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



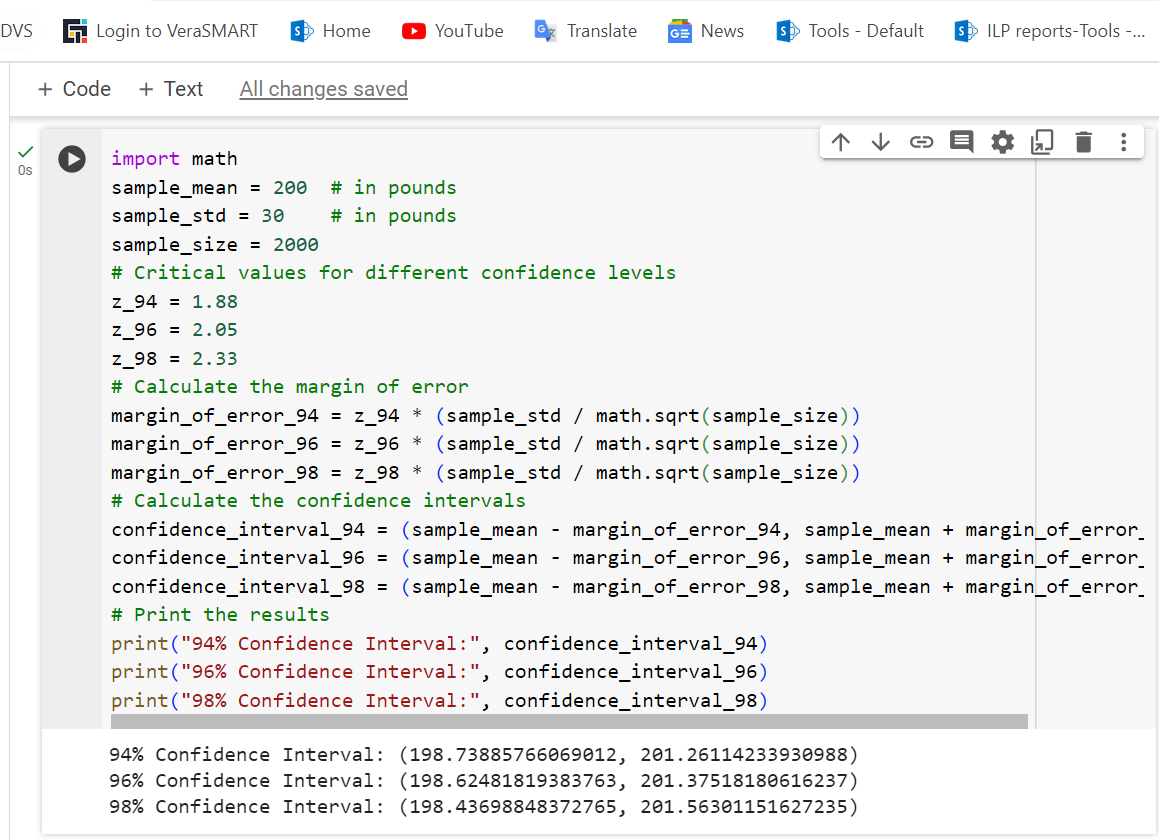
Solution: Most of the data points are concentrated around 50-100 with the frequency of 200.

And least range of weight is 400 somewhere around 0-10

Hence expected value from above distribution is 75

Median is less than mean. There is a long tail towards right, so it is right skewed and we have outliers on the upper side of box plot and there is less data points between q1 and bottom point.

**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?

Solution: 

**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?

Solution: 1) Mean=41 Median=40 Variance=24.111 Standard Deviation= 4.910

2)

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

Solution: Symmetrical

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

Solution: Right Skewed

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

Solution: Left Skewed

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

Solution: For a data positive kurtosis value indicates that Data is normally distributed, and Kurtosis Value is 0

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

Solution: Negative Kurtosis value for a data indicates that the distribution of the data has lighter tails and a flatter peak than the normal distribution.

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

What will be the IQR of the data (approximately)?   
Solution: We can say that the data is concentrated between 10 to 18 and median is greater than the mean also that the data is left skewed.

The IQR or interquartile range of the data is approximately -8

Q19) Comment on the below Boxplot visualizations?



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

Solution: When we observe both the boxplots, we observe that both the tails are higher in box plot 2. Both mean and median looks the same hence the distribution is symmetrical.

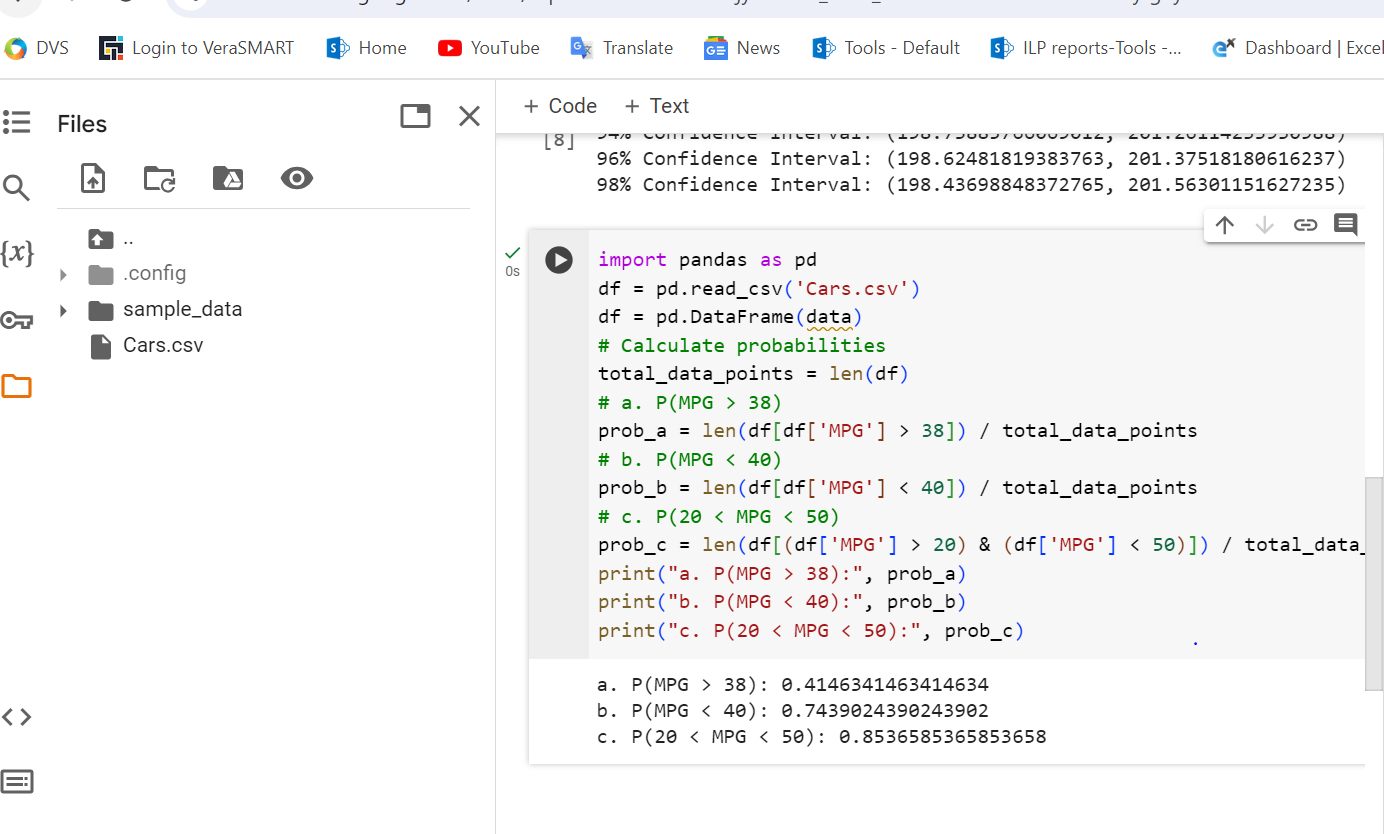
Q 20) Calculate probability from the given dataset for the below cases

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)

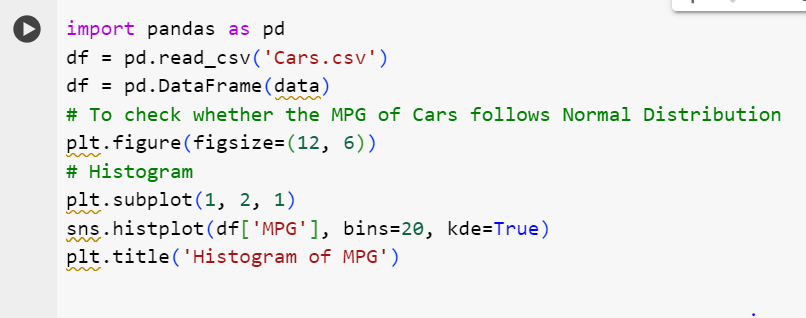
Solution: 

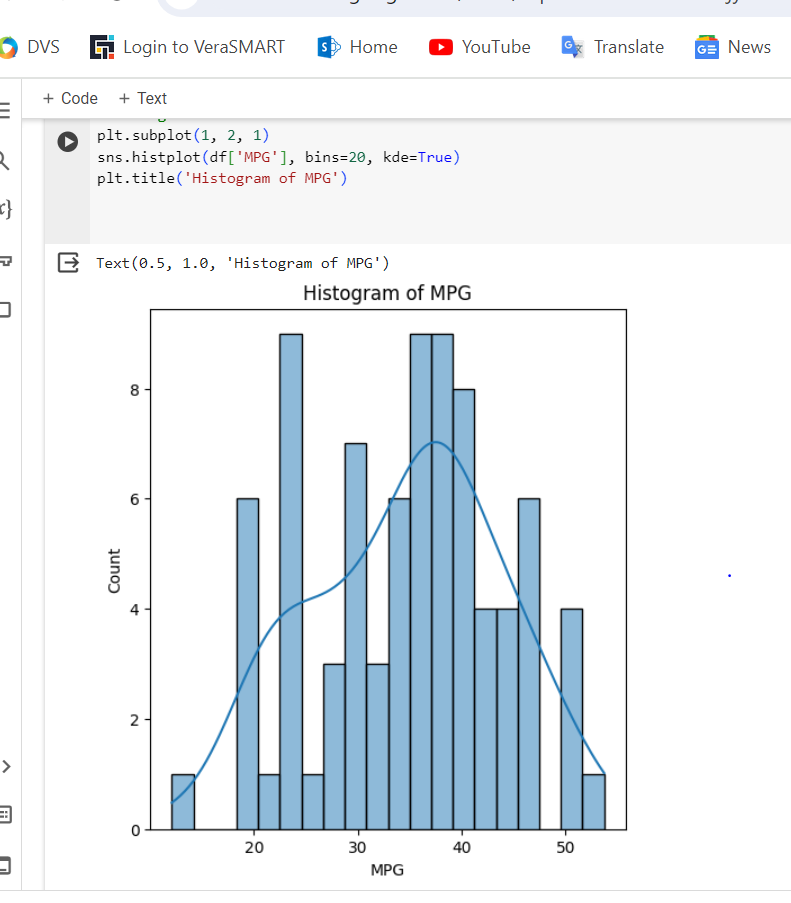
Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

To check whether the MPG of cars follows a normal distribution, we can do create a histogram to visualize the distribution





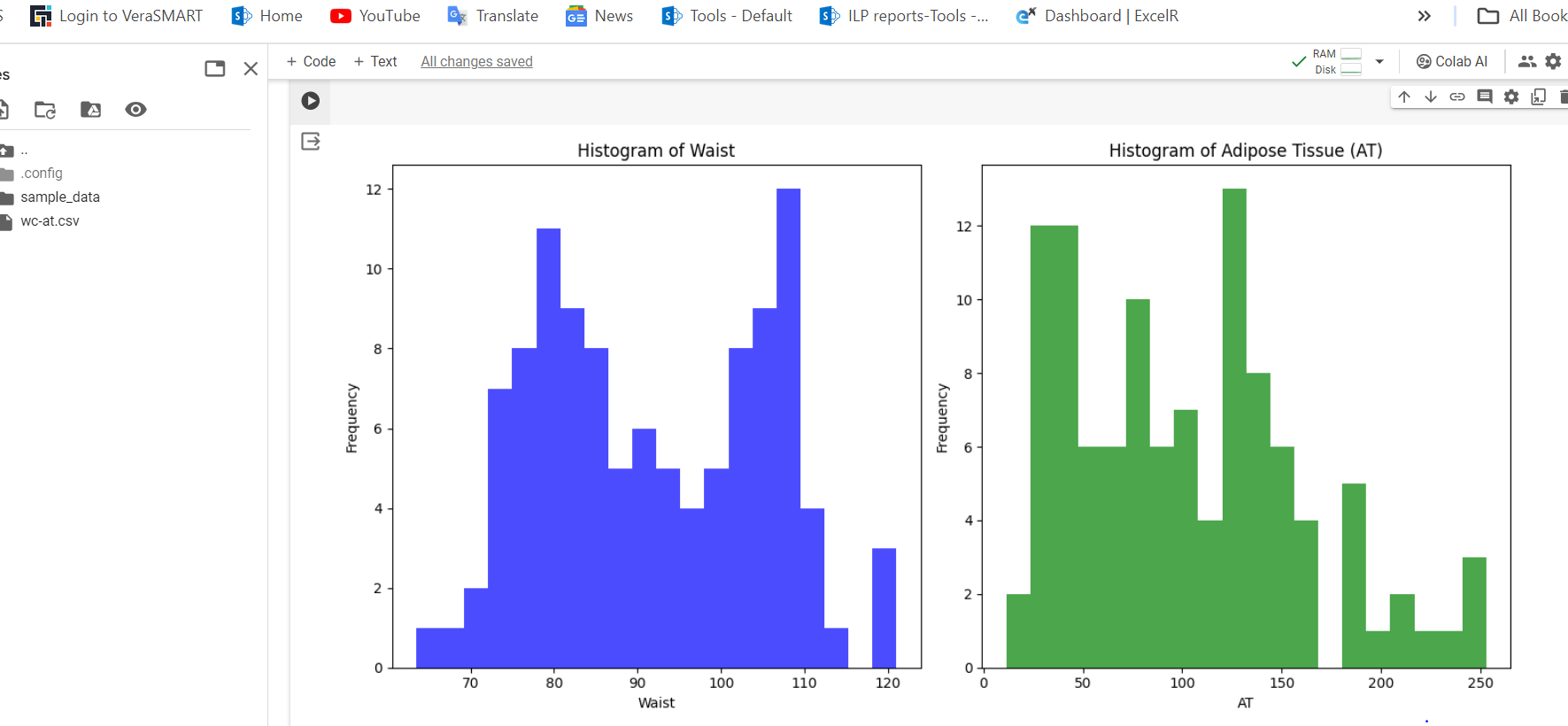
1. Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Solution: Variable waist circumference (waist) does not follow normal distribution.

Variable AT adipose tissue follow normal distribution.





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

Solution: To calculate the Z scores for different confidence intervals, you can use the standard normal distribution (Z distribution) and its percentiles. Here are the Z scores for the specified confidence intervals:

1. For a 90% confidence interval, you need to find the Z score corresponding to the middle 90% of the standard normal distribution. The Z score for a 90% confidence interval is approximately ±1.645.
2. For a 94% confidence interval, the Z score is approximately ±1.88.
3. For a 60% confidence interval, the Z score is approximately ±0.84.

These values are standard Z scores associated with the corresponding percentiles in a standard normal distribution. You can use these Z scores in various statistical calculations, such as constructing confidence intervals for sample means.

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

Solution: To calculate the t-scores for different confidence intervals with a sample size of 25, you need to use the t-distribution and its percentiles. The degrees of freedom for a t-distribution with a sample size of 25 would be 25 - 1 = 24 degrees of freedom.

Here are the t-scores for the specified confidence intervals:

1. For a 95% confidence interval with 24 degrees of freedom, the t-score is approximately ±2.064.
2. For a 96% confidence interval with 24 degrees of freedom, the t-score is approximately ±2.178.
3. For a 99% confidence interval with 24 degrees of freedom, the t-score is approximately ±2.797.

These values are based on the critical values associated with the t-distribution for the given degrees of freedom. You can use these t-scores to construct confidence intervals for sample means when the population standard deviation is unknown.

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

Solution: 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

t - statistics for the data is given as follows:

t=x- μ/(8/√n)

x = mean of the sample of bulbs = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

t=(260-270)/90 √18

t=-10/(90/3 √2)

t=-10/(30/ √2)

t=(-1\* √2)/3

t = - 0.471

For probability calculations, the number of degrees of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that **t < - 0.471 with 17 degrees of freedom** assuming the population mean is true, the t-value is less than the t-value obtained with 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of **0.3218** assuming the mean life of the bulbs is 300 days.