1. Python

# Question 1: -

Write a program that takes a string as input, and counts the frequency of each word in the string, there might be repeated characters in the string. Your task is to find the highest frequency and return the length of the highest-frequency word.

Note - You have to write at least 2 additional test cases in which your program will run successfully and provide an explanation for the same.

Example input - string = “write write write all the number from from from 1 to 100”

Example output - 5

Explanation - From the given string we can note that the most frequent words are “write” and “from” and the maximum value of both the values is “write” and its corresponding length is 5.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question%201.ipynb>

# Question 2: -

Consider a string to be valid if all characters of the string appear the same number of times. It is also valid if he can remove just one character at the index in the string, and the remaining characters will occur the same number of times. Given a string, determine if it is valid. If so, return YES , otherwise return NO .

Note - You have to write at least 2 additional test cases in which your program will run successfully and provide an explanation for the same.

Example input 1 - s = “abc”. This is a valid string because frequencies are { “a”: 1, “b”: 1, “c”: 1 }

Example output 1- YES

Example input 2 - s “abcc”. This string is not valid as we can remove only 1 occurrence of “c”. That leaves character frequencies of { “a”: 1, “b”: 1 , “c”: 2 }

Example output 2 - NO

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question%202.ipynb>

# 

# Question 3: -

Write a program, which would download the data from the provided link, and then read the data and convert that into properly structured data and return it in Excel format.

Note - Write comments wherever necessary explaining the code written.

Link - <https://raw.githubusercontent.com/Biuni/PokemonGO-Pokedex/master/pokedex.json>

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question-3.ipynb>

# Question 4: -

Write a program to download the data from the link given below and then read the data and convert it into the proper structure and return it as a CSV file.

Link - <https://data.nasa.gov/resource/y77d-th95.json>

Note - Write code comments wherever needed for code understanding.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question-4.ipynb>

# Question 5: -

Write a program to download the data from the given API link and then extract the following data with proper formatting

Link - <http://api.tvmaze.com/singlesearch/shows?q=westworld&embed=episodes>

Note - Write proper code comments wherever needed for the code understanding

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question-5.ipynb>

# Question 6: -

Using the data from Question 3, write code to analyse the data and answer the following questions

Note 1. Draw plots to demonstrate the analysis for the following questions for better visualisations.

2. Write code comments wherever required for code understanding

Insights to be drawn -

● Get all Pokemons whose spawn rate is less than 5%

● Get all Pokemons that have less than 4 weaknesses

● Get all Pokemons that have no multipliers at all

● Get all Pokemons that do not have more than 2 evolutions

● Get all Pokemons whose spawn time is less than 300 seconds.

Note - spawn time format is "05:32”, so assume “minute: second” format and perform the analysis.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question-6.ipynb>

# Question 7: -

Using the data from Question 4, write code to analyse the data and answer the following questions Note -

1. Draw plots to demonstrate the analysis for the following questions for better visualisations

2. Write code comments wherever required for code understanding

Insights to be drawn -

● Get all the Earth meteorites that fell before the year 2000

● Get all the earth meteorites co-ordinates who fell before the year 1970

● Assuming that the mass of the earth meteorites was in kg, get all those whose mass was more

than 10000kg.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question-7.ipynb>

# Question 8: -

Using the data from Question 5, write code the analyse the data and answer the following questions Note -

1. Draw plots to demonstrate the analysis for the following questions and better visualisations

2. Write code comments wherever required for code understanding

Insights to be drawn -

● Get all the overall ratings for each season and using plots compare the ratings for all the seasons, like season 1 ratings, season 2, and so on.

● Get all the episode names, whose average rating is more than 8 for every season

● Get all the episode names that aired before May 2019

● Get the episode name from each season with the highest and lowest rating

● Get the summary for the most popular ( ratings ) episode in every season

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question-8.ipynb>

# Question 9: -

Write a program to read the data from the following link, perform data analysis and answer the following questions

Note -

1. Write code comments wherever required for code understanding

Link - <https://data.wa.gov/api/views/f6w7-q2d2/rows.csv?accessType=DOWNLOAD>

Insights to be drawn -

● Get all the cars and their types that do not qualify for clean alternative fuel vehicle

● Get all TESLA cars with the model year, and model type made in Bothell City.

● Get all the cars that have an electric range of more than 100, and were made after

2015

● Draw plots to show the distribution between city and electric vehicle type.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question%209.ipynb>

# Question 10: -

Write a program to count the number of verbs, nouns, pronouns, and adjectives in a given particular phrase or paragraph, and return their respective count as a dictionary.

Note -

1. Write code comments wherever required for code

2. You have to write at least 2 additional test cases in which your program will run successfully and provide an explanation for the same.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/Python/Question-10.ipynb>

1. Statistics

# Question 1:-

A university wants to understand the relationship between the SAT scores of its applicants and their college GPA. They collect data on 500 students, including their SAT scores (out of 1600) and their college GPA (on a 4.0 scale). They find that the correlation coefficient between SAT scores and college GPA is 0.7. What does this correlation coefficient indicate about the relationship between SAT scores and college GPA?

***Answer*:** The correlation coefficient measures the strength and direction of the linear relationship between two variables. In this case, the correlation coefficient of 0.7 between SAT scores and college GPA indicates a strong positive linear relationship between the two variables.

Here's what the correlation coefficient of 0.7 tells us about the relationship between SAT scores and college GPA:

1. Strength: The correlation coefficient ranges from -1 to +1. A correlation coefficient of 0.7 indicates a relatively strong relationship between SAT scores and college GPA. The closer the correlation coefficient is to 1 (either positive or negative), the stronger the relationship.

2. Positive Relationship: The positive sign of the correlation coefficient (+0.7) indicates a positive relationship. This means that as SAT scores increase, the college GPA tends to increase as well. Similarly, as SAT scores decrease, the college GPA tends to decrease.

3. Linear Relationship: The correlation coefficient of 0.7 suggests that the relationship between SAT scores and college GPA is approximately linear. This means that as SAT scores increase or decrease by a certain amount, the college GPA changes in a relatively consistent manner.

It's important to note that correlation does not imply causation. While the correlation coefficient indicates a strong relationship, it does not establish a cause-and-effect relationship between SAT scores and college GPA. Other factors could also be influencing college GPA, and further analysis would be required to determine the precise nature and strength of the relationship.

# Question 2:-

Consider a dataset containing the heights (in centimetres) of 1000 individuals. The mean height is 170 cm with a standard deviation of 10 cm. The dataset is approximately normally distributed, and its skewness is approximately zero. Based on this information, answer the following questions:

a. What percentage of individuals in the dataset have heights between 160 cm and 180 cm?

b. If we randomly select 100 individuals from the dataset, what is the probability that their average height is greater than 175 cm?

c. Assuming the dataset follows a normal distribution, what is the z-score corresponding to a height of 185 cm?

d. We know that 5% of the dataset has heights below a certain value. What is the approximate height corresponding to this threshold?

e. Calculate the coefficient of variation (CV) for the dataset.

f. Calculate the skewness of the dataset and interpret the result.

***Answer:***a. To find the percentage of individuals in the dataset with heights between 160 cm and 180 cm, we can use the properties of the normal distribution. Since the dataset is approximately normally distributed, we can calculate this by finding the area under the normal curve between these two heights.

Using the Z-table or a statistical software, we can find the z-scores corresponding to 160 cm and 180 cm, and then find the corresponding probabilities.

Let's calculate it step by step:

1. Calculate the z-score for 160 cm:

z1 = (160 - 170) / 10 = -1

2. Calculate the z-score for 180 cm:

z2 = (180 - 170) / 10 = 1

Using a Z-table, we can find the area to the left of -1 (z1) and the area to the left of 1 (z2) and subtract the smaller area from the larger one to find the area between them.

From the Z-table, the area to the left of -1 is approximately 0.1587, and the area to the left of 1 is approximately 0.8413. Thus, the area between -1 and 1 is:

area = 0.8413 - 0.1587 = 0.6826

Converting this to a percentage, we have:

percentage = 0.6826 \* 100 = 68.26%

Therefore, approximately 68.26% of individuals in the dataset have heights between 160 cm and 180 cm.

b. To find the probability that the average height of 100 randomly selected individuals from the dataset is greater than 175 cm, we can use the properties of the sampling distribution of the mean. Since the dataset is approximately normally distributed, the sampling distribution of the mean will also be approximately normal.

The mean of the sampling distribution of the mean will be the same as the mean of the original dataset, which is 170 cm. The standard deviation of the sampling distribution of the mean, often referred to as the standard error, can be calculated by dividing the standard deviation of the dataset by the square root of the sample size.

Standard error = standard deviation / √(sample size)

= 10 / √100

= 10 / 10

= 1

Now, we need to find the z-score for a height of 175 cm using the mean of the sampling distribution of the mean and the standard error.

z = (175 - 170) / 1

= 5 / 1

= 5

Using a Z-table or a statistical software, we can find the area to the right of 5 (z) under the standard normal distribution. The probability that the average height is greater than 175 cm is equal to the area to the right of 5.

From the Z-table, the area to the right of 5 is extremely close to 0. Therefore, the probability is approximately 0.

Thus, the probability that the average height of 100 randomly selected individuals from the dataset is greater than 175 cm is approximately 0.

c. To find the z-score corresponding to a height of 185 cm, we can use the formula:

z = (x - μ) / σ

where x is the height, μ is the mean, and σ is the standard deviation.

Substituting the given values:

z = (185 - 170) / 10

= 15 / 10

= 1.5

Therefore, the z-score corresponding to a height of 185 cm is 1.5.

d. To find the approximate height corresponding to a threshold of 5% in

# Question 3:-

Consider the ‘Blood Pressure Before’ and ‘Blood Pressure After’ columns from the

data and calculate the following

<https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_>

a. Measure the dispersion in both and interpret the results.

b. Calculate mean and 5% confidence interval and plot it in a graph

c. Calculate the Mean absolute deviation and Standard deviation and interpret

the results.

d. Calculate the correlation coefficient and check the significance of it at 1% level

of significance.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/stats/Question-3.ipynb>

# Question 4:-

A group of 20 friends decide to play a game in which they each write a number between 1 and 20 on a slip of paper and put it into a hat. They then draw one slip of paper at random. What is the probability that the number on the slip of paper is a perfect square (i.e., 1, 4, 9, or 16)?

***Answer***: To find the probability that the number on the slip of paper is a perfect square, we need to determine the number of favourable outcomes (slips of paper with perfect squares) and the total number of possible outcomes.

The favourable outcomes are the slips of paper with the numbers 1, 4, 9, and 16. There are 4 perfect squares between 1 and 20.

The total number of possible outcomes is 20 since there are 20 slips of paper in the hat.

Therefore, the probability is given by:

Probability = (Number of favourable outcomes) / (Total number of possible outcomes)

= 4 / 20

= 1 / 5

= 0.2

So, the probability that the number on the slip of paper is a perfect square is 0.2 or 20%.

# Question 5:-

A certain city has two taxi companies: Company A has 80% of the taxis and Company B has 20% of the taxis. Company A's taxis have a 95% success rate for picking up passengers on time, while Company B's taxis have a 90% success rate. If a randomly selected taxi is late, what is the probability that it belongs to Company A?

To solve this problem, we can use Bayes' theorem. Let's define the following events:

A: Taxi belongs to Company A

B: Taxi belongs to Company B

L: Taxi is late

We are given the following probabilities:

P(A) = 0.8 (Company A has 80% of the taxis)

P(B) = 0.2 (Company B has 20% of the taxis)

P(L|A) = 0.05 (Company A's taxis have a 95% success rate)

P(L|B) = 0.1 (Company B's taxis have a 90% success rate)

We want to calculate P(A|L), the probability that a late taxi belongs to Company A.

By Bayes' theorem, we have:

P(A|L) = (P(L|A) \* P(A)) / P(L)

To calculate P(L), we need to consider all possible cases where a taxi can be late, regardless of the company. This can be expressed as:

P(L) = P(L|A) \* P(A) + P(L|B) \* P(B)

Substituting the values, we have:

P(L) = (0.05 \* 0.8) + (0.1 \* 0.2) = 0.04 + 0.02 = 0.06

Now, we can substitute the values of P(L) and P(A) into the formula for P(A|L):

P(A|L) = (0.05 \* 0.8) / 0.06 = 0.04 / 0.06 ≈ 0.6667

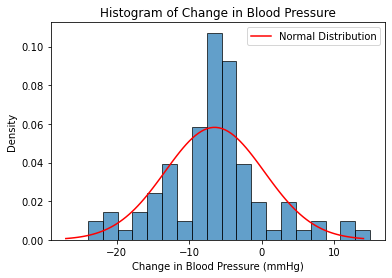
Therefore, the probability that a randomly selected taxi is late and belongs to Company A is approximately 0.6667 or 66.67%.

# Question 6:-

A pharmaceutical company is developing a drug that is supposed to reduce blood pressure. They conduct a clinical trial with 100 patients and record their blood pressure before and after taking the drug. The company wants to know if the change in blood pressure follows a normal distribution.

<https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_>

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/stats/Question-6.ipynb>



# Question 7:-

The equations of two lines of regression, obtained in a correlation analysis

between variables X and Y are as follows: and . 2𝑋 + 3 − 8 = 0 2𝑌 + 𝑋 − 5 = 0 The variance of 𝑋 = 4 Find the

a. Variance of Y

b. Coefficient of determination of C and Y

c. Standard error of estimate of X on Y and of Y on X.Q-7

Answer: a. To find the variance of variable Y, we need to know the values of Y or have additional information about the data. The equations provided represent the lines of regression for the relationship between X and Y, but they do not provide the actual values of Y.

The variance of Y can be calculated based on the deviations of the observed values of Y from their mean. If you have a dataset with the actual values of Y, you can calculate the variance using the following formula:

Variance of Y = Σ(Y - Ȳ)^2 / n

where Σ represents the sum, Y is the observed value of Y, Ȳ is the mean of Y, and n is the number of data points.

Without the actual values of Y or additional information about the data, it is not possible to calculate the variance of Y.

b. The coefficient of determination, denoted as R^2, is a measure that indicates the proportion of the variance in the dependent variable (Y) that can be explained by the independent variable (C) in a regression analysis.

To calculate the coefficient of determination, we need additional information such as the actual data points for C and Y, as well as the regression model used to obtain the equations. The equations provided in the question (2X + 3 - 8 = 0 and 2Y + X - 5 = 0) only represent the lines of regression and do not provide enough information to calculate the coefficient of determination.

In a regression analysis, the coefficient of determination can be obtained by squaring the correlation coefficient (r) between the independent variable and the dependent variable. If you have the correlation coefficient (r) between C and Y, you can calculate R^2 as follows:

R^2 = r^2

However, since the correlation coefficient (r) is not given in the question, it is not possible to calculate the coefficient of determination without additional information.

c. The standard error of estimate (also known as the standard error of regression or residual standard error) is a measure of the average distance between the observed values and the predicted values from a regression analysis. It represents the typical amount by which the observed values deviate from the regression line.

To calculate the standard error of estimate for X on Y and for Y on X, we need the actual data points for both X and Y, as well as the regression models used to obtain the equations.

Assuming you have the actual data points for X and Y and the regression models, you can calculate the standard error of estimate as follows:

Standard Error of Estimate (X on Y):

1. Fit the regression model of X on Y.

2. Calculate the residuals (differences between the observed values of X and the predicted values from the regression model).

3. Calculate the sum of squared residuals.

4. Divide the sum of squared residuals by the degrees of freedom (n - 2, where n is the number of data points) to obtain the mean squared error.

5. Take the square root of the mean squared error to obtain the standard error of estimate for X on Y.

Standard Error of Estimate (Y on X):

1. Fit the regression model of Y on X.

2. Calculate the residuals (differences between the observed values of Y and the predicted values from the regression model).

3. Calculate the sum of squared residuals.

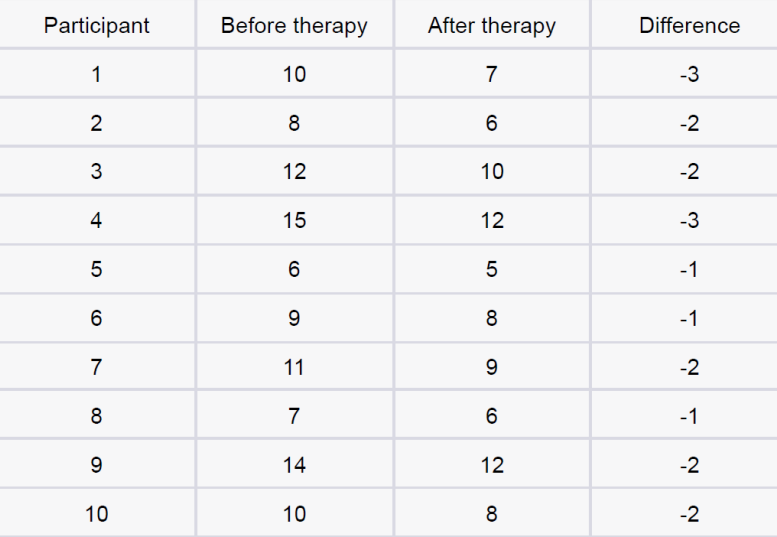
4. Divide the sum of squared residuals by the degrees of freedom (n - 2) to obtain the mean squared error.

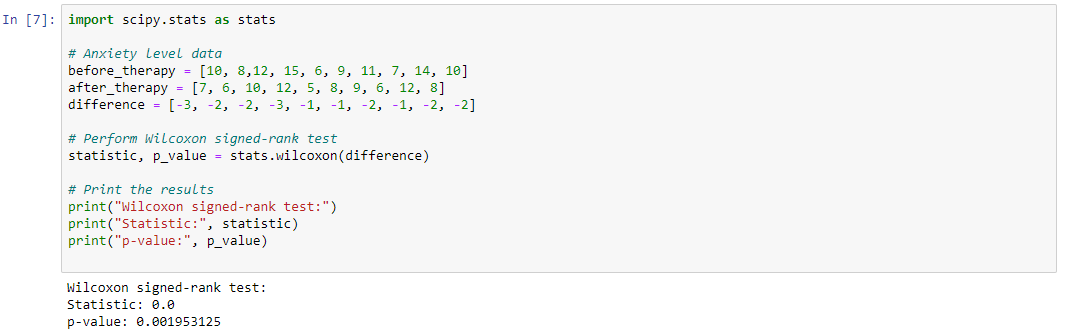
5. Take the square root of the mean squared error to obtain the standard error of estimate for Y on X.

Please note that without the actual data points and regression models, it is not possible to calculate the standard error of estimate for X on Y and Y on X.

# Question 8:-

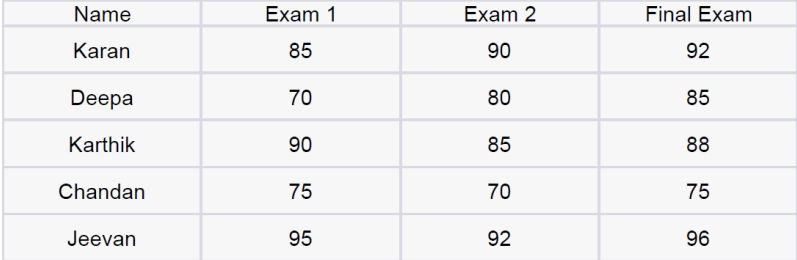
The anxiety levels of 10 participants were measured before and after a new therapy. The scores are not normally distributed. Use the Wilcoxon signed-rank test to test whether the therapy had a significant effect on anxiety levels. The data is given below: Participant. Before therapy After therapy Difference



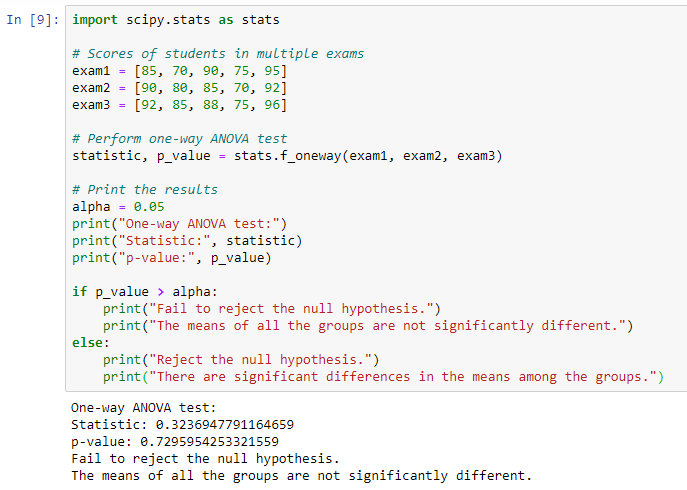
Answer: 

# Question 9:-

Given the score of students in multiple exams



Test the hypothesis that the mean scores of all the students are the same. If not, name the student with the highest score.

Answer: 

# Question 10:-

A factory produces light bulbs, and the probability of a bulb being defective is 0.05. The factory produces a large batch of 500 light bulbs.

a. What is the probability that exactly 20 bulbs are defective?

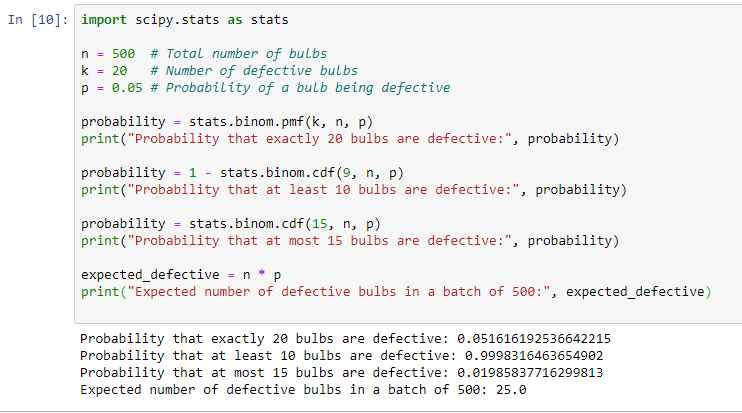
b. What is the probability that at least 10 bulbs are defective?

c. What is the probability that at max 15 bulbs are defective?

d. On average, how many defective bulbs would you expect in a batch of 500?

Answer:

1. Probability that exactly 20 bulbs are defective: 0.051616192536642215
2. Probability that at least 10 bulbs are defective: 0.9998316463654902
3. Probability that at most 15 bulbs are defective: 0.01985837716299813
4. Expected number of defective bulbs in a batch of 500: 25.0



# Question 11:-

Given the data of a feature contributing to different classes

<https://drive.google.com/file/d/1mCjtYHiX--mMUjicuaP2gH3k-SnFxt8Y/view?usp=share_>

a. Check whether the distribution of all the classes are the same or not.

b. Check for the equality of variance/

c. Which amount LDA and QDA would perform better on this data for classification and why.

d. Check the equality of mean for between all the classes.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/stats/Question-11.ipynb>

# Question 12:-

A pharmaceutical company develops a new drug and wants to compare its effectiveness against a standard drug for treating a particular condition. They conduct a study with two groups: Group A receives the new drug, and Group B receives the standard drug. The company measures the improvement in a specific symptom for both groups after a 4-week treatment period.

a. The company collects data from 30 patients in each group and calculates the mean improvement score and the standard deviation of improvement for each group. The mean improvement score for Group A is 2.5 with a standard deviation of 0.8, while the mean improvement score for Group B is 2.2 with a standard deviation of 0.6. Conduct a t-test to determine if there is a significant difference in the mean improvement scores between the two groups. Use a significance level of 0.05.

b. Based on the t-test results, state whether the null hypothesis should be rejected or not. Provide a conclusion in the context of the study.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/stats/Question-12.ipynb>

3. Machine Learning

# Question 1:-

Imagine you have a dataset where you have different Instagram features like username , Caption , Hashtag , Followers , Time\_Since\_posted , and likes , now your task is to predict the number of likes and Time Since posted and the rest of the features are your input features. Now you have to build a model which can predict the number of likes and Time Since posted.This is the Dataset You can use this dataset for this question.

Answer:

# Question 2:-

Imagine you have a dataset where you have different features like Age , Gender , Height , Weight , BMI , and Blood Pressure and you have to classify the people into different classes like Normal , Overweight , Obesity , Underweight , and Extreme Obesity by using any 4 different classification algorithms. Now you have to build a model which can classify people into different classes. Dataset This is the Dataset You can use this dataset for this question.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/ML/Question-2.ipynb>

# Question 3:-

Imagine you have a dataset where you have different categories of data, Now you need to find the most similar data to the given data by using any 4 different similarity algorithms. Now you have to build a model which can find the most similar data to the given data. Dataset This is the Dataset You can use this dataset for this question.

Answer:

# Question 4:-

Imagine you are working as a sales manager. Now you need to predict the Revenue and whether that particular revenue is on the weekend or not and find the Informational\_Duration using the Ensemble learning algorithm Dataset This is the Dataset You can use this dataset for this question.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/ML/Question-4.ipynb>

# Question 5:-

Uber is a taxi service provider as we know, we need to predict the high booking area using an Unsupervised algorithm and price for the location using a supervised algorithm and use some map function to display the data Dataset This is the Dataset You can use this dataset for this question.

Answer: <https://github.com/vbhatsaccnt/Ineuron_Interview_Based_Questions/blob/main/ML/Question-5.ipynb>

# Question 6:-

Imagine you have a dataset where you have predicted loan Eligibility using any 4 different classification algorithms. Now you have to build a model which can predict loan Eligibility and you need to find the accuracy of the model and built-in docker and use some library to display that in frontend Dataset This is the Dataset You can use this dataset for this question.

Answer:

# Question 7:-

Imagine you have a dataset where you need to predict the Genres of Music using an Unsupervised algorithm and you need to find the accuracy of the model, built-in docker, and use some library to display that in frontend Dataset This is the Dataset You can use this dataset for this question.

Answer:

# Question 8:-

Quora question pair similarity, you need to find the Similarity between two questions by mapping the words in the questions using TF-IDF, and using a supervised Algorithm you need to find the similarity between the questions. Dataset This is the Dataset You can use this dataset for this question.

# Question 9:-

A cyber security agent wants to check the Microsoft Malware so he came to you as a Machine learning Engineer with Data. You need to find the Malware using a supervised algorithm and you need to find the accuracy of the model. Dataset This is the Dataset

You can use this dataset for this question.

Answer:

# Question 10:-

An Ad- Agency analysed a dataset of online ads and used a machine learning model to predict whether a user would click on an ad or not. Dataset This is the Dataset You can use this dataset for this question.