**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 returns 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/devkegovind/Assessment-Completion/blob/main/Python/Q.1.py**](https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.1.py)

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/devkegovind/Assessment-Completion/blob/main/Python/Q.2.py**](https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.2.py%20)

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.

**Answer.** [**https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.3.py**](https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.3.py)

Question 4: Write a program to download the data from the link given below and then read the data and convert the 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.

Excepted Output Data Attributes

● Name of Earth Meteorite - string id - ID of Earth

● Meteorite - int nametype - string recclass - string

● mass - Mass of Earth Meteorite - float year - Year at which Earth

● Meteorite was hit - datetime format reclat - float recclong - float

● point coordinates - list of int

**Answer.** [**https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.4.py**](https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.4.py)

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

Excepted Output Data Attributes -

id - int, url - string, name - string, season - int, number - int, type - string, airdate - date, format airtime - 12-hour time format, runtime - float, average rating - float, summary - string without html tags, medium image link - string, original image link – string.

**Answer.** [**https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.5.py**](https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.5.py)

Question 6 Using the data from Question 3, write code to analyze the data and answer the following questions

Note

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

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.

● Get all Pokemon who have more than two types of capabilities

**Answer.** <https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.6.ipynb>

Question 7 Using the data from Question 4, write code to analyze the data and answer the following questions

Note -

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

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/devkegovind/Assessment-Completion/blob/main/Python/Q.7.ipynb>

Question 8 Using the data from Question 5, write code the analyze the data and answer the following questions

Note -

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

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/devkegovind/Assessment-Completion/blob/main/Python/Q.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/devkegovind/Assessment-Completion/blob/main/Python/Q.9.ipynb](https://github.com/devkegovind/Assessment-Completion/blob/main/Python/Q.9.ipynb%20)

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/devkegovind/Assessment-Completion/blob/main/Python/Q.10.ipynb>

**Statistics**

Q-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. 1 Statistics - >

Q-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. 2 Statistics - >

Q-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. 3 Statistics - >

Q-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 erfect square (i.e., 1, 4, 9, or 16)?

Answer. 4 Statistics - >

Q-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?

Answer. 5 Statistics - >

Q-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. 6 Statistics - >

Q-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

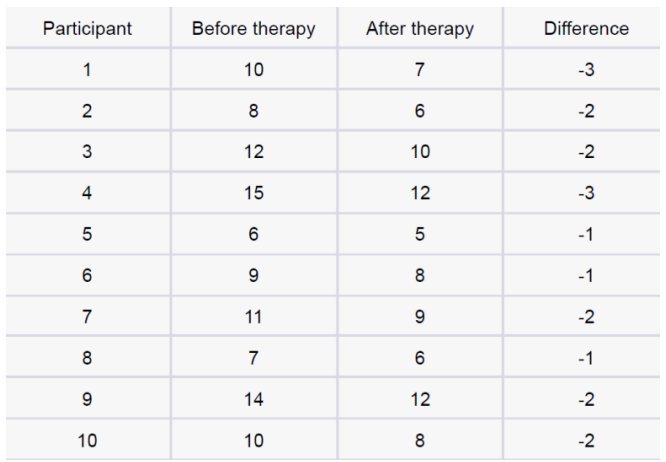
a. Variance of Y

b. Coefficient of determination of C and Y

c. Standard error of estimate of X on Y and of Y on X.

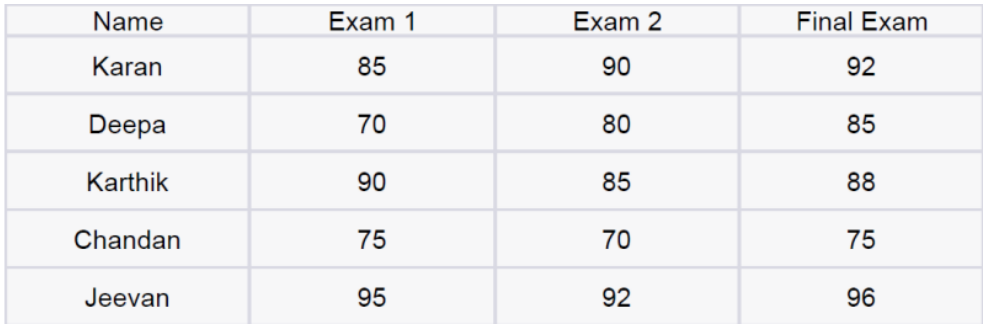
Answer.<https://github.com/devkegovind/Assessment-Completion/blob/main/Stats/Statistics_Q7.pdf>

Q-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. 8 Statistics - >

Q-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. <https://github.com/devkegovind/Assessment-Completion/blob/main/Stats/Q.9.py>

Q-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. 10 Statistics - >

Q-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. 11 Statistics - >

Q-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. 12 Statistics - >

**Machine learning**

Total Marks: 210 Each question 15 marks

**INTERMEDIATE QUESTIONS:**

Q-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. Dataset This is the Dataset You can use this dataset for this question.

Ans. <https://github.com/devkegovind/Assessment-Completion/blob/main/notebooks/instgram.ipynb>

Q-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.

Ans.

<https://github.com/devkegovind/Assessment-Completion/blob/main/notebooks/obesity.ipynb>

Q-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.

Ans.

<https://github.com/devkegovind/Assessment-Completion/blob/main/notebooks/news.ipynb>

Q-4. Imagine you working as a sale 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.

Ans.

<https://github.com/devkegovind/Assessment-Completion/blob/main/notebooks/onlineshop.ipynb>

Q-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.

Ans.

<https://github.com/devkegovind/Assessment-Completion/blob/main/notebooks/uber.ipynb>

Q-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.

Ans.

<https://github.com/devkegovind/Assessment-Completion/blob/main/notebooks/approval_loan.ipynb>

Q-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.

Ans.

<https://github.com/devkegovind/Assessment-Completion/blob/main/notebooks/music.ipynb>

Q-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.

Q-9. A cyber security agent wants to check the Microsoft Malware so need he came to you as a Machine learning Engineering 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. 1. An Ad- Agency analyzed 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.

Advance QUESTIONS :

Q-1. A Social Media Influencer collected data on Facebook friend requests and used a supervised algorithm to predict whether a user would accept a friend request or not. Dataset This is the Dataset You can use this dataset for this question. Note : Use only Dask and Use MLflow

Q-2. A chemist had two chemical flasks labeled 0 and 1 which consist of two different chemicals. He extracted 3 features from these chemicals in order to distinguish between them, you provided the results derived by the chemicals and your task is to create a model that will label chemical 0 or 1 given its three features and built-in docker and use some library to display that in frontend. Note : Use only pyspark Dataset This is the Dataset You can use this dataset for this question.

Q- 3. A company wants to predict the sales of its product based on the money spent on different platforms for marketing. They want you to figure out how they can spend money on marketing in the future in such a way that they can increase their profit as much as possible 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. Note: Use only Dask

Q-4. Take any 3 questions and deploy them to AWS using GitHub Actions and show a demo link

Q-5. Take any 3 questions and deploy them to AWS using Circle-CI and show a demo lin