

PDS Project FAQ's

1. What is the difference between size and dimension? **(Question-3 & 4)**
Ans: Size refers to the total number of elements (rows*columns) in dataset and dimension refers to the number of rows & columns in a dataset
2. How are Question-5 and Question-10 different? **(Question-5 and 10)**
Ans: In Q-5, one has to use the original dataset and in Q-10, one has to use the dataset with no missing value
3. do I need to check the number of missing values for the entire dataset or specific columns? **(Question-7)**
Ans: You need to check the total number of missing values in the entire dataset
4. Should we drop NA values completely from the original dataset? **(Question-9)**
Ans: You should first copy the data in “df” object and then drop NA values from the “df”.
5. Column 'PURPOSE*' has the most NAN values (503) in the dataset. The other columns do not have any missing values. Is it fine to drop the “PURPOSE” Variable completely instead of dropping the rows? **(Question-9)**
Ans: In this question you should drop rows containing missing values instead of dropping the “PURPOSE*” variable
6. In START* and STOP* columns, there is a value as R?walpindi and Rawalpindi. While analysing the unique values, these are treated as different values. Is it ok to proceed treating these as two different values? **(Question-11)**
Ans: For this project, please treat R?walpindi and Rawalpindi as different entities. You will learn how to clean this type of issue in the future courses
7. Do we need to display the entire list or only one entry which has maximum value? **(Question 15 &16)**
Ans: You can choose to display the entire output in the descending order (Entry with the maximum value at the top) or only the most popular point
8. Do we need to display the entire list of most frequent routes? **(Question-17)**
Ans: You may choose to display only the most frequent route

9. List the most frequent route taken by Uber drivers. What is the meaning of the most frequent route and an approach to solve this question? **(Question-17)**

Ans: Most Frequent route means the route with most frequent START and STOP combination.

Example

Start : Stop

Hyderabad: Delhi

Delhi :Chennai

Chennai: Bengaluru

Hyderabad: Bengaluru

Delhi: Chennai

Delhi: Bengaluru

The expected answer:

Frequent Route is Delhi -> Chennai

You may choose this approach to solve this question- Use 'groupby()' function on data-frame for the START* & STOP* columns {groupby will give you the grouping of START* & STOP* common points , Calculate size of the transformed data-set which will give the count/number, Then perform max() operation accordingly to find out the most frequent routes.

10. Plot a bar graph of Purposes vs Miles (Distance). In this question can I take the X-axis as Purposes as the title names are overlapping? **(Question-19)**

Ans: Yes, you are free to choose either Purpose or Miles on X-axis

11. Display a dataframe of Purposes and the distance travelled for that particular Purpose. What is expected from this question? **(Question-20)**

Ans: Display the total distance (using Miles variable) for each given purpose

12. Plot number of trips vs Category of trips. - What is expected from this question? **(Question-21)**

Ans: You are expected to plot a countplot or a barplot which should display the count of trips which are made in the given 2 categories. Here you can have number of trips on the Y-axis and the different categories on the X-axis

13. What is the difference between a proportion and percentage? **(Question-22)**

Ans: Proportion is basically a ratio and to if we multiply proportion with 100, we will have value in percentage