**DSBDA Lab Problem statements and FAQs for Practice.**

**Instructions Common for all statements:**

1. Read the statement carefully and take note of the required dataset
2. The Datasets are uploaded on Google class room
3. Observe the variables in the dataset and implements basic Pandas operation to understand the dataset. (to get shape, size, datatypes, outliers, missing/Junk values, duplicates, Summary etc)
4. Specify comments to make your script more readable.
5. Save your work after each operation
6. Refer the notes and Tableaue tutorial and how tos available in Tableau public
7. Refer the FAQs for viva

**Statements on Data Analytical Operations-**

1. Perform the following operations in python on given dataset **[Iris.csv]**
2. **Subset** rows where petal length > 1.5 and species is "setosa".
3. **Merge** the Iris dataset with a custom species info table (e.g., color or habitat).
4. **Sort** flowers by sepal\_width and then by sepal\_length.
5. **Transpose** the dataset’s first 5 rows to view feature-wise comparisons.
6. **Reshape** the dataset using melt() and pivot() to transform features into long and wide formats.
7. Perform the following operations in python on given dataset [**StudentsPerformance.csv]**
8. **Subset** the dataset to find students with math score > 80 and who completed the test preparation course.
9. **Merge** this dataset with a demographic table (Subset) that includes socioeconomic status.
10. **Sort** by reading score and then writing score in descending order.
11. **Transpose** average scores by gender across all subjects.
12. **Reshape** the data to compare average scores across lunch types and test prep status using a pivot table.
13. Perform the following operations in python on given dataset [**housing.csv]**
14. **Subset** houses with median income > 5 and average rooms < 6.
15. **Merge** with a regional lookup table mapping latitude/longitude to regions.
16. **Sort** by median\_house\_value and population.
17. **Transpose** statistics summary to compare features.
18. **Reshape** the data to view average house value across income and housing age bins.
19. Perform the following operations in python on given wine quality dataset [**WnieQT.csv]**
20. **Subset** data for wines with quality ≥ 7 and alcohol > 10%.
21. **Merge** red and white wine datasets to form a complete dataset with a new type column.
22. **Sort** wines by citric acid and residual sugar.
23. **Transpose** summary statistics of chemical properties for different quality levels.
24. **Reshape** using pivot\_table() to show average values of key features by wine quality
25. Perform the following operations using Python on the given datasets **(Toyota.csv)**
    1. Create data subsets
    2. Merge Data
    3. Sort Data on any specified column values
    4. Transposing Data
    5. Shape and reshape Data
26. Perform the following operations in python on given dataset **[student\_info.csv and student\_scores.csv]**
    1. Clean both datasets to handle missing values, incorrect formats, and inconsistent entries.
    2. Remove invalid scores (e.g., -1 in Science).
    3. Convert Grade to string format like "Grade 10" instead of just a number.
    4. Perform an inner join and analyze which students are present in both datasets.
    5. Handle any mismatches (e.g., StudentID = 106 is not in student\_info.csv).
27. Perform the following operations in python on given dataset **[student\_info.csv and student\_scores.csv]**
28. Clean both datasets to handle missing values, incorrect formats, and inconsistent entries.
29. Identify and fill or remove missing values in Age, Email, and History.
30. Standardize email addresses and correct improperly formatted ones (e.g., "eva.email.com").
31. Create a new column for average score across subjects.
32. Add a binary column HighPerformer (1 if avg score > 85, else 0).
33. Perform the following operations in python on given dataset

**[patients.csv:** Patient demographic infoand **visits.csv:** Doctor visits and diagnosis codes**]**

1. Fill or drop missing diagnosis codes and ages.
2. Standardize gender values (e.g., “M”, “Male”, “F” → “Male”, “Female”).
3. Merge patient info with visits
4. Group data to get total visits and unique diagnoses per patient.
5. Correct out-of-range values (e.g., age > 120).
6. Perform the following operations in python on given dataset

[**applicants.csv:** Students who applied and **exam\_scores.csv:** Standardized test results]

* 1. Clean up inconsistent formatting in names and missing test scores.
  2. Join on ApplicationID to combine personal data with scores.
  3. Normalize test scores.
  4. Convert Admission\_Status to binary labels (1 = admitted).
  5. Remove duplicate applications and fix invalid test score entries.

1. Perform the following operations in python on given dataset

**[employee\_info.csv:** Contains employee demographic and department data

**performance.csv:** Contains performance review scores**]**

* 1. Clean invalid ages, join dates, and department names.
  2. Combine datasets using EmployeeID.
  3. Create performance average scores.
  4. Bucket performance into categories (Low/Medium/High).
  5. Correct mismatched or blank department entries.

1. Perform the following operations in python on **Titanic-Dataset.csv** Dataset
   1. Handle missing Age and Cabin values using appropriate imputation techniques.
   2. Convert Sex and Embarked columns into numeric form using encoding.
   3. Create a new feature FamilySize = SibSp + Parch.
   4. Bin Fare into price categories (e.g., Low, Medium, High).
2. Perform the following operations in python on **StudentsPerformance.csv** Datase**t**
   1. Check for and impute any missing test scores (math, reading, writing).
   2. Create an overall AverageScore column.
   3. Bucket students into performance bands (e.g., Excellent, Average, Poor).
   4. Check for inconsistent or duplicate student records.
   5. Encode gender, lunch, and test preparation course using Label Encoding.
3. Perform the following operations in python on **Heart Disease Dataset [ ]**
   1. Fill missing values in cholesterol, restecg, and thal columns.
   2. Encode categorical columns like sex, cp, thal using One-Hot Encoding.
   3. Create an AgeGroup column (young, middle-aged, elderly).
   4. Normalize features like chol, thalach, and oldpeak.
   5. Build a classification model to predict presence of heart disease (target).
4. Perform the following operations in python on given dataset **[Iris.csv]**
   1. Check for and handle any duplicated rows or missing values (insert some intentionally for practice).
   2. Combine with an external species characteristics table (e.g., color, blooming time).
   3. Normalize petal/sepal measurements.
   4. Add a size\_ratio = petal\_length / sepal\_length column.
5. Perform the following operations using Python on the given data sets **(Toyota.csv)**
   1. Data cleaning
   2. Data integration
   3. Data transformation
   4. Error correcting
6. Perform the following operations using Python on the **airquality dataset**
   1. Read data set and display summary
   2. Create data subsets having observations in range 11 to 49 and another subset having temperature value less than 60
   3. Merge observations of any two subsets
   4. Apply Sort Data on Temp values
7. Perform the following operations using Python on the **airquality dataset**
   1. Create data subsets selecting specified columns (Ozone, Solar.R, Wind and Temp only) and index range
   2. Rationally Replace nan values
   3. Data transformation- Apply Min-max Normalization on Solar.R
   4. Plot Month wise Temperature using Matplotlib/ seaborn library
8. Perform the following operations using Python on the given data sets **(Toyota.csv)**
   1. Remove missing values
   2. set Doors Value to uniform format
   3. Provide concise summary of all numeric variables
   4. Remove All duplicate records
   5. Get dummies for categorical data Fuel type (One hot Encoding)
9. Perform the following operations using Python on the given data sets **(Toyota.csv)**
   1. Sort observations on Price values order
   2. Create Subset by Selecting columns, selecting rows and columns,
   3. Create subset of cars data having Price greater than 15000 and Age less than 8
   4. Create subset of cars data consuming Petrol
   5. Apply decimal normalization on Price column
10. Perform the following operations using Python on the given data sets **(Toyota.csv)**
    1. Remove all missing values
    2. display datatypes and concise summary of all numeric variables
    3. Remove All duplicate records
    4. Apply Z-score Normalization on Price Column
    5. shape and reshape using pivot\_ table
11. Perform following operations on the given data set **(Toyota.csv)**
12. Get unique values of categorical ‘Doors’
13. Transform of all values in same format
14. Apply Decimal scaling normalization on HP column
15. Perform following operations on the given data set **(Toyota.csv)**
16. Display shape and summary and count of missing values in the dataset
17. Remove duplicate records
18. Clean the data set- Replace the missing values in each column with appropriate value
19. Convert the datatype of Metcolor and Automatic column in as object type
20. Perform following operations on the given data set **(Toyota.csv)**
21. Remove duplicate records from dataset and display concise summary
22. Create Subset selecting columns 'Price', 'Age', 'FuelType' and initial 10 records.
23. Transpose of this subset
24. Apply mean-max normalization on HP column
25. Perform following operations on the given data set **(Toyota.csv)**
26. Add a new column ‘Revised’ to the dataset specifying 5% increase in old Price.
27. Create subset of cars’ data having Price greater than 15000 and Age less than 8
28. Sort observations in descending order of Revised Price
29. Apply ZScore\_normalization on HP column
30. Perform following operations on the given data set **(Toyota.csv)**
31. Create Subset selecting columns 'Price', 'Age', 'FuelType', 'Automatic'
32. Create subset having records of all Petrol vehicles
33. Create subset of cars’ data having Price greater than 15000 and Age less than 8 years
34. Merge records the above two data subsets
35. Perform the following operations using Pyth on on the given data sets **(Toyota.csv)**
    1. Sort observations on Price values order
    2. Create Subset by Selecting columns, selecting rows and columns,
    3. Create subset of cars data having Price greater than 15000 and Age less than 8
    4. Create subset of cars data consuming Petrol
    5. Apply decimal normalization on Price column
36. Perform the following operations using Python on the given data sets **(Toyota.csv)**
37. Remove missing values
38. display datatypes and concise summary of all numeric variables
39. Remove All duplicate records
40. Apply Z-score Normalization on Price Column
41. shape and reshape using pivot\_ table
42. Perform the following operations using Python on the **Salaries.csv** data set

a. Create data subsets

b. Merge Data

c. Sort Data

d. Transposing Data

e. Shape and reshape Data

1. Perform the following operations using Python (Dataset: **Salaries.csv**)
2. get the rank and salary of all staff that does not belong to discipline A
3. get rank salary and years of service of all male staff and only female professor
4. get all Female staff who are either professor or earning more than 75000
5. get the rank and salary of all staff other than professors and who are serving from at least 10 years

**Statements on Data Visualization (using matplotlib and seaborn):**

1. Plot following using Python libraries **matplotlib / seaborn** **(Toyota.csv)**
2. Scatter plot- Car-Price by Age
3. Histogram on Cars data KM
4. Bar plot on counts of FuelType category (Petrol, Disel and CNG)
5. Plot following data visualization using Python **matplotlib** **library** by plotting histogram, scatter-plot and bar-plot **(Toyota.csv)**
   1. Age vs Price- Scatter plot
   2. Distribution of KM data - histogram
   3. FuelType wise CarsCount- Barplot
6. Visualize the data using Python **Seaborn** /**matplotlib** **library** by plotting histogram, scatter-plot and bar-plot **(Toyota.csv)**
7. Age vs Price- Scatter plot
8. Distribution of KM data - histogram
9. FuelType wise CarsCount- Barplot

**Statements on Data Visualization using Tableau:**

1. Perform the data visualization operations using tableau. **[Iris.csv]**
   1. Create a **scatter plot** comparing petal length and sepal length by species with color coding.
   2. Use **boxplots** or **violin plots** to compare the distribution of each numeric feature across species.
   3. Build a **heatmap** to visualize correlations between all four numeric features.
   4. Create a **bar chart** showing the average sepal length, sepal width, petal length, and petal width for each species.
   5. Use a **pie chart** or **donut chart** to show the proportion of each Iris species in the dataset.
   6. Build an interactive dashboard with filters for feature ranges, and visualizations that update to show the impact on species prediction.
2. Perform the data visualization operations using tableau.

**[Sales Dashboard for a Retail Store- Dataset**: Superstore (available in Tableau)]

1. Create an interactive dashboard showing total sales, profit, and quantity by region and category.
2. Use filters to drill down by **State**, **Segment**, and **Sub-Category**.
3. Build a map visualization of **sales by state** with color-coded profit margins.
4. Identify top 10 products by **profit** and by **sales volume**.
5. Perform the data visualization operations using tableau.

**[Student Performance Analysis- Dataset**: **StudentsPerformance.csv]**

* 1. Visualize score distributions in **Math**, **Reading**, and **Writing** using histograms.
  2. Compare performance by **gender**, **parental education**, and **lunch type**.
  3. Create a KPI dashboard showing average scores and pass/fail rates by gender.
  4. Identify if students who took the **test preparation course** performed better.

1. Perform the data visualization operations using tableau.

**[Real Estate Market Dashboard- Dataset**: Kaggle housing datasets **housing.csv]**

1. Compare **median listing prices** by region and time using line graphs.
2. Use a **scatter plot** to show price vs. square footage.
3. Filter views by **property type** and **city**.
4. Visualize **price trends** over time and identify areas with highest growth.
5. Perform the data visualization operations using Sales order dataset using tableau. (Use **Sample superstore Dataset**)
   1. Plot Sales by region
   2. Plot year, month, quarterwise Sale
   3. Plot yearwise sale vs profit
6. Perform the data visualization operations using Sales order dataset using tableau. **(Sample superstore Dataset)**
   1. Plot Sales by region
   2. Plot year, month, quarterwise Sale
   3. Plot yearwise sale vs profit
   4. Create dashboard
7. Perform the data visualization operations using tableau. (iris\_data\_sample)
   1. Scatter Plot of SepalLength by SepalWidth
   2. Show species wise distribution
   3. Visualize correlation of PetalLength and PetalWidth

Frequently Asked Questions in Oral -

1. Explain Big data and its characteristics
2. GFS components
3. Hadoop framework
4. What is cluster of Commodity hardware?
5. HDFS Components and their working- name node, SNN, Data node
6. YARN- Resource manager, Node Manager
7. Explain the Fault tolerance, data durability and parallel processing in Hadoop
8. About Map-reduce
9. How the data files are stored on HDFS
10. Hadoop Ecosystem tools- Hbase, Hive etc
11. Pandas Data Frame – Datatyps, Attributes and methods
12. Tableau Features- Project stages in DV project
13. information of different panes in tableaue
14. Measures and dimensions
15. Measure of Central tendency, mean, median, mode,
16. Why normalizing of values required? 3 Ways of normalization.
17. CAP theorem
18. NoSql Databases
19. What is data visualization? Objectives, Names of tools known to you- Tableau, Helical insight, MS PowerBI and other
20. Different types of charts and their purposes (Exmples)
21. What is Dashboard? how it is useful in getting insights and decision making any example?
22. What are the Phases of data science project?
23. Explain Big data 4 Vs as data processing challenges