Problem Statement No. 01

Consider the “Academic performance” dataset of students (Academic_Performance_Dataset.csv) and perform the following operations using Python.

a) Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal with them.

b) Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them.

c) Apply data transformations on categorical variables to convert it into numerical variables.

Reason and document your approach properly.

Problem Statement No. 02

Perform the following operations on Age-Income dataset (Age-Income-Dataset.csv)

Provide summary statistics (mean, median, minimum, maximum, standard deviation) for numeric variables with and without using any library functions.

Provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorical variable.

Problem Statement No. 03

Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of ‘Iris-setosa’, ‘Iris-versicolor’ and ‘Iris-virginica’ of iris.csv dataset.

Calculate the measures of variability. Calculate and provide the visualization of the Correlation among the variables.

Problem Statement No. 04

Consider the Bangalore House Price Data. Perform following operations.

a) Find and replace null values in the data using appropriate technique.

b) Transform the ‘Size’ column to numerical values. For Example: 2 BHK to be converted as 2

c) Transform the ‘total_sqft’ column to contain numerical values on same scale. If the range is given average value of the range to be taken.

d) Calculate and add one more column as ‘Price_Per_Sqft’

e) Remove the outliers from Price_Per_Sqft and BHK Size column if any.

f) Apply the Linear Regression model to the data and display the training and testing performance measures as Mean Squared Error and Accuracy

---

Problem Statement No. 05

Write a Scala Program to process a log file of a system and perform following analytics on the given dataset.

(I) Display the list of top 10 frequent hosts.

(II) Display the list of top 5 URLs or paths

(III) Display the number of unique Hosts

Problem Statement No. 06

Write a Scala Program to process a log file of a system and perform following analytics on the given dataset.

(I) Display the count of 404 Response Codes

(II) Display the list of Top Twenty-five 404 Response Code Hosts

(III) Display the number of Unique Daily Hosts

Problem Statement No. 07

Perform the following operations using Python on any open source dataset (e.g., data.csv)

1. Import all the required Python Libraries.

2. Load the Dataset into pandas data frame.

4. Data Preprocessing: check for missing values in the data to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.

5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.

6. Turn categorical variables into quantitative variables in Python. In addition to the codes and outputs, explain every operation that you do in the above steps and explain everything that you do to import/read/scrape the data set.

Use NBA.CSV, Dirtydata.csv

Problem Statement No. 08

1. Implement logistic regression using Python/R to perform classification on Social_Network_Ads.csv dataset.

2. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset
Use :Social_Network_Ads.csv

Problem Statement No. 09

1. Implement Simple Naïve Bayes classification algorithm using Python/R on iris.csv dataset.

2. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset

Use iris.csv Diabetes.csv

Problem Statement No. 10

1. Use the dataset 'titanic'. Use the Seaborn library to see if we can find any patterns in the data.

2. Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.

Problem Statement No. 11

1. Use the dataset 'titanic'. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names : 'sex' and 'age')

2. Write observations on the inference from the above statistics. Use voilin plot and others plots from seaborn package

Use :Titanic.csv

Problem Statement No. 12

Scan the Iris dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.

2. Create a histogram for each feature in the dataset to illustrate the feature distributions.

3. Create a boxplot for each feature in the dataset. 4. Compare distributions and identify outliers.

Problem Statement No. 13

Write a code in SCALA for a simple WordCount application that counts the number of occurrences of each word in a given input set using the APACHE SPARK MapReduce framework on local-standalone set-up.

Problem Statement No. 14

Write a code in SCALA for a simple Sort application that sort the occurrences of each word in a given input set using the APACHE SPARK MapReduce framework on local-standalone set-up.

Problem Statement No. 15

Consider the Amazon Alexa Reviews Dataset. This dataset consists of a nearly 3000 Amazon customer reviews (input text), star ratings, date of review, variant and feedback of various amazon Alexa products like Alexa Echo, Echo dots, Alexa Firesticks etc. Perform following operations on this dataset.

(I) Plot a graph of Positive and Negative Feedback (1 = Positive Feedback, 0 = Negative Feedback)

(II) Convert the review text into lowercase.

(III) Remove all punctuations from review text.

(IV) Remove emoticons and emojis from the text

(V) Tokenize the review text into words.

(VI) Remove the Stopwords from the tokenized text.

Problem Statement No. 16

Consider the Amazon Alexa Reviews Dataset. This dataset consists of a nearly 3000 Amazon customer reviews (input text), star ratings, date of review, variant and feedback of various amazon Alexa products like Alexa Echo, Echo dots, Alexa Firesticks etc. Perform following operations on this dataset.

(I) Remove all punctuations from review text.

(II) Tokenize the review text into words.

(III) Remove the Stopwords from the tokenized text.

(IV) Perform stemming & lemmatization on the review text.

(V) Perform the word vectorization on review text using Bag of Words technique.

(VI) Create representation of Review Text by calculating Term Frequency and Inverse Document Frequency (TF-IDF)