Q1: Using the BEML stock data perform the following operation using python(ETL operation) Load the stock data from the given CSV file in pandas frame

a) Display the first five observation of it
b) Display the last five observation of it
c) Display the first five observations with only two columns
d) List all the attributes name(that is the meta data)
e) Select the any three attributes(Data, close and any other) and move into a new data frame
f) Perform calculation of new variable called gain and add it to new data frames
g) Add a comment on result of all attributes
h) Accordingly take a necessary action
i) Display the time plot of close attribute
j) Create a CVS file of the new data frame

Q1: Using the GLAXO stock data perform the following operation using python(ETL operation) Load the stock data from the given CSV file in pandas frame

a) Display the first five observation of it
b) Display the last five observation of it
c) Display the first five observations with only two columns
d) List all the attributes name(that is the meta data)
e) Select the any two attributes(Data, close) and move into a new data frame suing drop method
f) Convert time attribure to index
g) Accordingly take a necessary action
h) Display the normal distribution plot of close attribute
i) Create a CVS file of the new data frame

1. Write a python program to perform binary classification of the given data, selecting 'survived' as the dependent variable.
2. Try implementing the program using different sampling techniques (random sampling, stratified sampling).
3. Comment on the performance of the model with a possible impact of the sampling techniques.

[Additionally, do different preprocessing techniques, and also try manual feature engineering to find the best attribute that can be used for building the model]

Q1: Prepare a classification model using Naive Bayes Theorem on breast cancer patient's data and perform the following

a) Separate feature and target variable
b) Display the number of samples, number of features and number of outcome present in the target attribute
c) Display the data from array form to dataframe
d) Perform some basic data exploration operation on data
e) Prepare classification model for the data using Naïve Bayes theorem
f) Perform metric preparation of the same(Accuracy, precision, recall confusion matrix etc..)
g) Save the model in pickle format

Q2: Perform an inferencing operation using the created model in another notebook.

1. Write a python code for the given data distribution:

Studies show colour blindness affects about 8% of men. A random sample of 10 men is taken.
Find the probability that:
a. All 10 men are color blind.
b. No men are color blind.
c. Exactly 2 men are color blind.
d. At least 2 men are color blind.

2. Write a python code for the given data distribution:

The number of calls arriving at a call center follows a Poisson distribution at 10 calls per hour.
a. Calculate the probability that the number of calls will be maximum 5.
b. Calculate the probability that the number of calls over a 3 hour period will exceed 30 calls.

Q. Perform Qualitative Analysis using Normal Distribution

Datasets: Glaxo.csv, BEML.csv

1. Load the datasets and perform basic descriptive analytics.
2. Selection of Attributes: Find out the attributes that are needed for qualitative analysis, and decide which ones to keep.
3. Basic Visualization of the selected attribute, to understand its nature.
4. With the opinion of experts, calculate a new attribute for performing the qualitative analysis.
5. Take necessary actions to remove the anomalies that have happened, due to the addition of the new attribute.
6. Make a comparative normal distribution plot of the 'gain' variable for both the stocks and comment on the results (quality of the variable, variance etc.)
7. Calculate the exact values of mean and standard deviation of the 'gain' variable for both the sticks data. REpeat the same with the help of interval function.
8. Finally, give a conclusion report of the qualitative analysis.

ASSIGNMENT 8:

Perform the various Hypotheses Test

- Q1: A passport office claims that the passport applications are processed within 30days of submitting the application form and all necessary documents. The file *passport.csv* contains processing time of 40 passport applicants. The population standard deviation of the processing time is 12.5 days. Conduct a hypothesis test at significance level ∝ =0.05 to verify the claim made by the passport office.
- Q2: Aravind Productions (AP) is a newly formed movie production house based out of Mumbai, India. AP was interested in understanding the production cost required for producing Bollywood movie. The industry believes that the production house will require INR 500 million on average. It is assumed that the Bollywood movie production cost follows a normal distribution. The production costs of 40 Bollywood movies in millions of rupees are given in *bollywoodmovies.csv* file. Conduct and appropriate hypothesis test at ∝ =0.05 to check whether the belief about average production cost is correct One
- Q3: A company claims that children who drink their health drink will grow taller than the children who do not drink that health drink. Data in the file *healthdrink.xlsx* shows average increase in height over one-year period from two groups: one drinking the health drink and the other not drinking the health drink. At ∝ =0.05, test whether the increase in height for the children who drink the health drink is different than those who do not drink health drink
- Q4: The file *breakup.csv* contains alcohol consumption before and after breakup. Conduct a paired t-test to check whether the alcohol consumption is same after the breakup at 95% confidence interval(or the level of confidence ∝ =0.05).
- Q5: Ms Rachael Khanna the brand manager of ENZO detergent powder at the “one-stop” retail was interested in understanding whether the price discounts have any impact on the sales quantity of ENZO. To test whether the price discounts had any impact, price discounts of 0%, 10%, and 20% were given on randomly selected days. The quantity of ENZO sold in a day under different discount levels is shown in Table 3.1. Conduct a one way ANOVA to check whether discount had any significant impact on the average sales quantity at ∝ = 0.05
- Q6: Hanuman Airlines (HA) operated daily flights to several Indian cities. One of the problems HA faces is the food preferences by the passengers. Captain Cook, the operations manager of HA, believes that 35% of their passengers prefer vegetarian food, 40% prefer non-vegetarian food, 20% low calorie food, and 5% request for diabetic food. A sample of 500 passengers was chosen to analyze the
  food preferences and the observed frequencies are as follows:
    1. Vegetarian: 190
    2. Non-vegetarian: 185
    3. Low-calorie: 90
    4. Diabetic: 35

    Conduct a chi-square test to check whether Captain Cook’s belief is true at ∝=0.05.

Create separate notebooks for each of the tests.

ASSIGNMENT 10
TITLE: FACTOR ANALYSIS ON BFI DATA

Q1. Mention the name of the library to be used for Factor Analysis (FA).

Q2. Perform the necessary data preparation operations, to be done for the FA of the given data.
2.1 Identification of necessary attributes. (A1-A5, C1-C5, E1-E5, N1-N5, O1-O5)
2.2 Check for the missing data and address it.
2.3 Convert non-numeric data to numeric data.

Q3. Perform the implementation of Step 1: Data Suitability Assessment, of the algorithm.
3.1 Check the data to have the necessary correlation using calculate_bartlett_sphericity().
3.2 Check for the adequacy of the sample size using calculate_kmo().

Q4. Implement Step 2: Factor Extraction, of the Algorithm.
4.1 Make the screen plot of the eigen values of the data against the no. of factors.
4.2 Make a decision regarding the no. of factors to be selected based on some conditions. Also define those conditions.

Q5. Implement Step 3: Factor Rotation, of the algorithm. i.e., perform the factor analysis of the given data.
5.1 Using the loading concept, perform FA without rotation. i.e., make a subset of data.
eg. Mapping of variables with factors based on the correlation values.
5.2 Repeat 5.1 using rotation.
5.3 Comment on the results of 5.1 and 5.2

Q6. Write a few lines of codes to create a new .csv file, which contains our data with respect to the created factors.

Q7. Give a better/technical name to the created factors.

Q8. Identify the Latent Variables given in the symbols (in the PPT) and write their names.

ASSIGNMENT 11

Q1. Identify a simple Business application for SWOT Analysis.
Q2. Perform implementation of Step 1 of SWOT Analysis, i.e. gathering information, performing internal and external assessments, followed by analysis.
Q3. Perform implementation of Step 2 of the SWOT analysis, i.e. using a suitable criteria to convert the analysis into action.

ASSIGNMENT 12

TITLE: Perform Data Preparation, followed by Model Prediction, followed by Decision Analytics Operation on the given diabetes.csv dataset (link is available in PPT).

Q1. Load the diabetes dataset and explore using text and visual analysis to perform identification of attributes.

Q2. Prepare a classification model using different procedures and different sampling techniques.

Q3. Evaluate the performance of the model by preparing a framework and by calculating the Expected Value. During this calculation, assume the cost matrix(in $) as

$$\begin{bmatrix} 99 & -1 \\ 0 & 45 \end{bmatrix}$$

Boxplot Scatterplot

Assignment -

Q) Load a dataset which contains four variable (2 numerical & 2 categorical). Find covariance between two variables, also find co-relation between two variables. Write a section of code which handles the outliers present in the data.

Assignment -

Q1 Load the dataset which contains 4 variables (2 numerical & 2 categorial), perform the following operations-

(a) Univariant text analysis: In this

(i) perform calculation of basis statistical calculation of four variable using describe method?

(ii) Repeat the basic statistical calculation of 4 variables using mean, median & mode method?

Q (b) Univariant Visual analysis of all 4 variables (Box plot)

Q (c) Bivariant text Analysis:

(i) Calculate mean of score 1 by 'city'

(ii) Calculate mean of score 2 by 'name'

Q (d) Bivariant Visual Analysis:

(i) Perform scatter plot visual betn 2 variables 'score1' & 'score2'.

(ii) Perfor Heatmap visual betn name & city.