**Data Exploration & Preparation**

This is a group based project using R programming language to analyse a specific problem in the following areas, such as Pharmacy, Library, Holiday Booking System, Medical Practice, Concert hall, Motor mechanic, Sales, Customers behaviour, Primary School, Role-playing game and manufacturing etc. You may choose data from any other category based on your interest from Kaggle (https://www.kaggle.com/) or UCI (<https://archive.ics.uci.edu/ml/index.php>) or any other repository. The dataset should have at least 10000 rows and 12 columns (for example, type of variables may be categorical, continuous and discrete) after cleaning and there is not any maximum limit for data records. Your group may use any data set and complete the following tasks as mentioned below

1) Identify which variables are categorical, discrete and continuous in the chosen data set and show using some visualization or plot. Explore whether there are missing values for any of the variables.

2) For each of the categorical variables, construct a bar chart of the variable, with an overlay of the target variable/ dependent variable (vertical axis). Normalize if necessary.

3) Report the mean, median, minimum, maximum, and standard deviation for each of

the numerical variables.

4) Apply Min-Max Normalization and Z-score Standardization on the numerical data

variables.

5) For each pair of numerical variables, construct a scatter plot of the variables. Discuss

your salient results.

6) Develop Data Exploratory analysis (EDA) report which shows that the numeric

variables in the data set and exhibit no obvious association.

7) Based on your EDA, identify interesting sub-groups of records within the data set that

would be worth further investigation.

8) Apply dummy encoding to categorical variables (at least one variable use from the data set) and discuss the benefits of dummy encoding to understand the categorical data.

9) Apply PCA with your chosen number of components. Write up a short profile of the

first few components extracted based on your understanding.

10) What is the purpose of dimensionality reduction? Suppose that you perform the PCA using three components. Considering the following information, which variable or variables might we be well advised to omit from the PCA, and why? If we really need all these variables in the analysis, then what should we do?

Your group will present the findings and defend the results in the report. The report should capture the following aspects that are relevant to your project.

i. Brief description of the domain problem of selected dataset

(10 marks)

ii.

Motivates for the problem and Challenges faced in Data Exploration project

(10 marks)

iv.

Characterisation of the data set: source URLS; size; number of attributes; has/does not have missing values; number of examples etc.

(10 marks)

v.

Application of Data Exploration methods and EDA visualizations along with explanation.

(30 marks)

vi. Use of dummy encoding and PCA for dimensionality reduction.

(20 marks)

vii.

Provide the explanation of code that will be used to solve the problem. Comments must be provided along with code.

(15 marks)

viii. Conclusions, Findings of data set and references (HARVARD style).

(10 marks)

Maximum Number of Words for the report (1500 words excluding diagrams, code and HARVARD style References).

Note: Clearly mentioned the contribution of each member of the team.