**EXPLORATORY DATA ANALYTICS AND DATA PRE-PROCESSING TASK.**

**PART 1**

The dataset cancer.csv was used for this task. This dataset contains variables that have bearing in predicting whether a brain tumour is likely to be malignant (cancerous) or benign (non-cancerous). Using Python code try to achieve what is below.

a) Import the requisite libraries.

b) Load the dataset into a Python data frame.

c) Plot a bar graph that shows the prevalence (occurrences) of the Malignant (M) and Benign (B) cases of the diagnosis variable i.e., the frequency of the cancer stages.

d) On the diagnosis column, replace the (M) with 1 and the (B) occurrences with 0.

e) You resolve, as a data analyst that the id feature will not be required in future analysis. Write code snippet to drop the feature.

f) Upon discussing business needs with stakeholders, there is an agreement to limit the number of variables to the first 10 remaining variables after dropping the id variable. Write python code snippet to achieve this.

g) Randomly select 300 records from the resultant data frame in d) above.

h) Display a summary of the total number of missing values for each of the 10 features.

i) Use imputation to replace missing values with the most frequently occurring value of a column.

j) Display the mean, standard deviation, and the maximum values of the texture\_mean and the area\_mean.

k) Output the correlation of all the numerical variables of the data frame.

l) Draw a heatmap of the correlations of all numerical variables.

m) How does the heat map help to decide variables to drop, in a bid to reduce dimensionality?

n) Draw a scatter plot between the texture\_mean and the area\_mean.

o) Comment on the relationship between the two variables in k) above, comparing the scatter plot and the correlation coefficient obtained above.

**PART 2**

You currently own a house in Francistown, Botswana and want to put your house on the market. You’re not in a hurry to get rid of the house and would like to try selling it yourself. One way to determine a reasonable asking price of a house is to call one or more real estate agents and seek their advice. Another is to hire an appraiser; this approach would cost several thousands of Pula. You have been wondering if there might be an easier and cheaper way to understand what determines the selling prices in the area.

Your data analyst friend has determined that the variable that has the highest predictive power for the price of a house is its floor area in square meters. She has collected data of house prices alongside floor areas and saved it in a file called prices.csv.

Write Python code snippets to perform the following:

a) Import the requisite libraries.

b) Load the data into a data frame and separate the predictor variable and the target variable into different files.

c) Determine the correlation coefficient between the area and the price of houses and comment on the result.

d) Using an 80/20 ratio, split the dataset into training and test set.

e) Fit the linear regression algorithm into the training dataset.

f) Use the model to predict the prices of houses using the test set.

g) On the same visualization, compare the scatter plot of the training dataset (X\_train, y\_train) and a line graph of X\_train and the predicted price from the training set.

h) On the same visualization, compare the scatter plot of the test set (X\_test, y\_test) and a line graph of X\_train and the predicted price from the training set.

i) Print out the coefficient of determination R2.

j) What does the coefficient of determination say about the accuracy of the model?

k) Print out the slope of the linear regression model.

l) Interpret the value of the slope that you printed out.

m) Print out the intercept of the linear regression model.

n) Interpret the value of the intercept that you printed out.

o) The model that we developed could suffer from overfitting. What is your understanding of overfitting?

p) Provide a brief explanation of any three (3) factors that lead to model overfitting and suggest anything that could have affected our model.

q) Suggest one way of addressing the overfitting challenge on our model above.