Instructions for project submission

The participants are expected to work on a project on the topic of 'Regression' and submit their analyses. They are required to perform regression analysis on data and interpret the results obtained by providing business insights.

Important points to remember:

1. Please consider a case study, and clearly mention the **motivation**, **objectives** and **outcomes** of the study. Emphasis the business or management side of the problem by providing appropriate business context.
    a. This should be ideally one-page write-up specifying the following:
        i. Description of the problem or the business scenario
        ii. Intent/purpose/objective of the analysis with business contexts
        iii. How could regression analysis help you to attain your objective
2. Gather relevant data for the case study. Consider a data set with many observations and variables. Make sure some predictors are quantitative and some are categorical.
    a. For public data or data obtained from an open-source database, please explicitly mention the source and the URL.
    b. For private data, it is recommended to use a portion of the data as opposed to the full data to avoid any copyright issues, and that too with a written permission/consent from the concerned authority or organization.
3. Clearly describe the variables in the data set that you have used in the analyses. Specify the response variable and predictors/features.
    a. For example, if there is a variable 'Age' in the data set, then you must provide a description as shown.
    b.

NAME	TYPE	DETAILS
AGE	Continuous	Measures the age of a person in years

    b. If there is a derived variable 'Age' that is required for analyses, then you must specify how this derived variable is calculated:

NAME	TYPE	DETAILS
DATE	Date/Time	Today's date in DD-MM-YYYY
DOB	Date/Time	Date of Birth in DD-MM-YYYY
AGE	Continuous/derived	Measures age of a person in years; AGE=ROUND((DATE-DOB)/365.25)

4. Perform descriptive analysis on the response variable (Y) based on:
    a. Histogram, box plot, scatter plot (response vs. an important predictor)
    b. Measure of center (mean, median, mode), measures of dispersion (SD, CV), measures of position (max, min, 25^{th} and 75^{th} percentiles) – in a single table.
5. Fit a multiple linear regression (MLR) model. Make sure the response variable is continuous.
    a. Perform a variable selection and choose an optimum subset of predictors. If the variable selection method rejects a specific variable that could be important from a business point of view, decide by judgement and proper reasoning what is to be done (include/exclude) with that predictor.
    b. Present the MLR model with the optimal subset of predictors.
    c. Interpret the partial regression coefficients based on at least one continuous and one categorical predictor and discuss their impacts on the response variable.
    d. Comment and interpret R^2 and Adjusted R^2 values. Point out the difference between them, and specify the reason for the discrepancy in their values.
    e. Check and explain which variables are significant individual predictors in the model. Clearly mention which test (H0 and H1) is performed to check this and mention the corresponding sampling distribution.
    f. Check and explain if the overall model is significant in explaining or predicting the response variable. Clearly mention which test (H0 and H1) is performed to check this and mention the corresponding

sampling distribution.

g. Validate the model assumptions using appropriate graphs and hypothesis tests. Check whether residuals are normal, independent and homoscedastic. For non-normal errors, try out transformations like Y^{α} where α is any real number (positive or negative) and see if this makes the error normal. In case it does not, please go ahead with the usual analysis considering Y with no transformation (but remember, theoretically you should not do this)

h. Check for multicollinearity in the data using variance inflation factor (VIF). Remove the predictor with the highest VIF > 10, and refit the model. Continue checking for VIF values until all VIFs are less than 10.

i. Obtain the final MLR based predictive model considering variable transformation (if any), multicollinearity adjusted and variable selection technique applied.

6. Conclusion: Finally, conclude your analysis by commenting on how this model is going to beuseful to fulfil the objective of the study and help in the business. Write a paragraph on this.

7. Presentation:
   a. The project should be written in a word doc file (just like a report) and submitted as .pdf file.
   b. All output (including tables and figures/plots) should have proper headings and titles. All output should be numbered, e.g., Table 1, Table 2 or Figure 1, Figure 2 etc.
   c. Axes in plots/graphs should be properly labelled and titled.
   d. Please avoid the use of a variety of colours for graphs unless absolutely necessary. Simple black, blue, red and green are fine.

8. R-codes can be submitted separately and should be commented at every step.

9. Please adhere strictly to all the requirements/guidelines mentioned in points 1 – 9 to obtainfull marks.