The goal of this assignment is to understand the relationship between various factors affecting MBA students' academic performance, as represented by their GPA, using multiple regression analysis. You will explore the assumptions of multiple regression and validate them using Python.

# Dataset Description:

The dataset contains information on 100 MBA students, with the following variables:

GPA (Dependent Variable): Students' Grade Point Average.

Hours\_Studied\_Per\_Week (Independent Variable): The average number of hours a student studies in a week.

Group\_Projects\_Participated (Independent Variable): The number of group projects a student has participated in.

Internship\_Experiences (Independent Variable): The number of internships a student has completed.

Attendance\_Percentage (Independent Variable): The percentage of classes attended by a student.

Class\_Discussion\_Participation (Independent Variable): The level of participation in class discussions, measured on a scale from 0 to 10.

# Tasks:

Task 1: Data Preparation

Load the dataset into a pandas DataFrame.

Perform a preliminary analysis to understand the structure and characteristics of the data (e.g., describe(), info()).

Task 2: Multiple Regression Model

Fit a multiple regression model using GPA as the dependent variable and all other variables as independent variables.

Interpret the coefficients of the model and discuss the potential impact of each independent variable on the GPA.

Task 3: Assumption Checks

Conduct tests and provide visualizations to check the following assumptions of multiple regression:

Linearity: The relationship between the dependent and independent variables should be linear. Use scatter plots or Partial Regression plots to assess this.

No Multicollinearity: Independent variables should not be too highly correlated. Check this using Variance Inflation Factor (VIF) or correlation matrices.

Homoscedasticity: The residuals should have constant variance. Use scatter plots of residuals or statistical tests like Breusch-Pagan.

Normality of Residuals: The residuals of the model should be normally distributed. Use Q-Q plots or the Shapiro-Wilk test to assess this.

Independence of Residuals: There should be no autocorrelation in the residuals. This can be checked using the Durbin-Watson statistic.

Task : Model Evaluation

Evaluate the performance of your model using appropriate metrics (e.g., R-squared, Adjusted R-squared).

Discuss any potential improvements or alternative approaches that could be explored.