FINDINGS FROM EXPLORATORY DATA ANALYSIS

About the Project:

Academic performance is a key concern for students, as it directly impacts future opportunities. This study aims to analyze various factors that may influence students GPA, including program of study, gender, graduation year, and performance metrics such as Cumulative Grade Point Average (CGPA) and Semester Grade Point Average (SGPA). Understanding these factors can help students make informed decisions and improve their academic outcomes.

To achieve this, we employ a multivariate linear regression model to examine the relationships between these variables and overall academic performance. This model allows us to identify which factors have the most significant impact on GPA and which have little to no effect. By analyzing plots and test hypotheses on our model, we can provide insights into the key determinants of student success.

Dataset Overview:

- **Rows:** 3,046
- **Columns:** 10
- **Categorical Variables:** Prog Code, Gender
- **Numerical Variables:** CGPA, CGPA100, CGPA200, CGPA300, CGPA400, SGPA, YoG

The Target variable is **CGPA.**

VARIABLE DESCRIPTION:

ID No - Randomly generated number sequence
Prog Code - Program of Study
Gender - Gender
YoG - Year of Graduation
CGPA - Overall Cumulative Grade Point Average
CGPA100 - Cumulative Grade Point Average at the end of the first year
CGPA200 - Cumulative Grade Point Average at the end of the second year
CGPA300 - Cumulative Grade Point Average at the end of the third year
CGPA400 - Cumulative Grade Point Average at the end of the fourth year
SGPA - Secondary School Cumulative Grade Point Average

PROGRAM OF STUDY (Catagorical)
BCH - Biochemistry

BLD - Building technology
CEN - Computer Engineering
CHE - Chemical Engineering
CHM - Industrial Chemistry
CIS - Computer Science
CVE - Civil Engineering
EEE - Electrical and Electronics Engineering
ICE - Information and Communication Engineering
MAT - Mathematics
MCB - Microbiology
MCE - Mechanical Engineering
MIS - Management and Information System
PET - Petroleum Engineering
PHYE - Industrial Physics-Electronics and IT Applications
PHYG - Industrial Physics-Applied Geophysics
PHYR - Industrial Physics-Renewable Energy

SUMMARY OF THE VARIABLES:

ID.No Prog.Code Gender YoG CGPA CGPA100 CGPA200 CGPA300

Min. :23462 EEE : 418 Female:1093 Min. :2010 Min. :1.520 Min. :1.570 Min. :1.170 Min. :0.630

1st Qu.:42654 CIS : 342 Male :1953 1st Qu.:2011 1st Qu.:3.000 1st Qu.:3.180 1st Qu.:2.760 1st Qu.:2.810

Median :61759 MIS : 307 Median :2012 Median :3.560 Median :3.690 Median :3.340 Median :3.510

Mean :61083 ICE : 245 Mean :2012 Mean :3.495 Mean :3.636 Mean :3.322 Mean :3.419

3rd Qu.:79236 CEN : 237 3rd Qu.:2013 3rd Qu.:4.010 3rd Qu.:4.150 3rd Qu.:3.920 3rd Qu.:4.100

Max. :97563 CHE : 213 Max. :2014 Max. :4.990 Max. :5.000 Max. :5.000 Max. :5.000

(Other):1284

CGPA400 SGPA

Min. :0.000 Min. :1.46

1st Qu.:3.000 1st Qu.:2.66

Median :3.620 Median :3.06

Mean :3.533 Mean :3.12

3rd Qu.:4.150 3rd Qu.:3.57

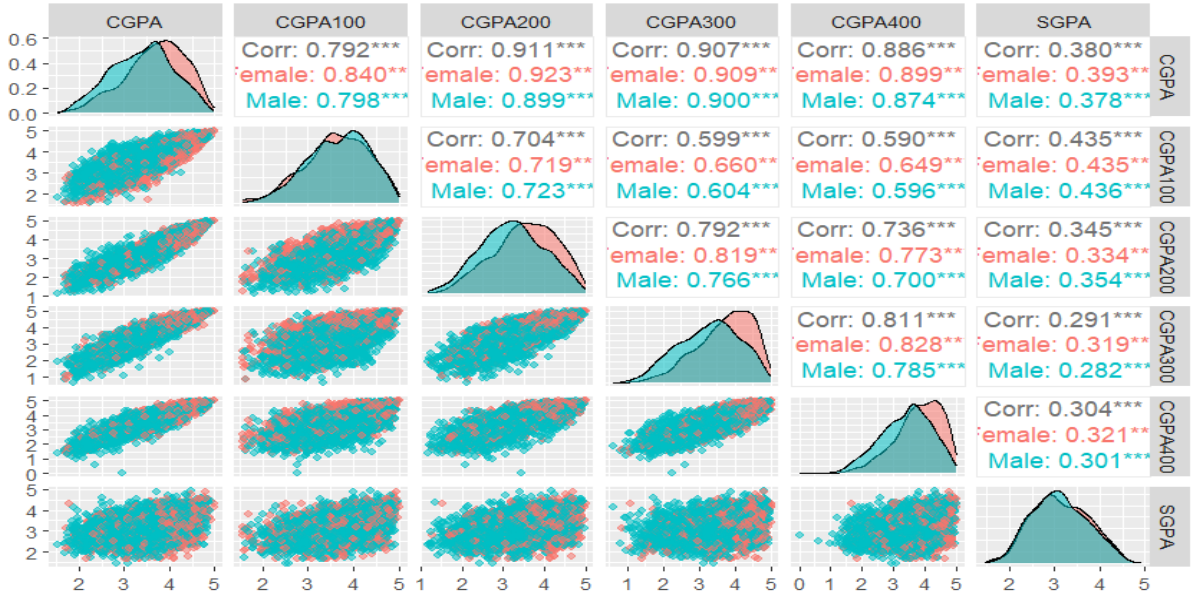
Max. :5.000 Max. :4.93

CHECK FOR MISSING VALUES:

ID.No	Prog.Code	Gender	YoG	CGPA	CGPA100	CGPA200	CGPA300	CGPA400	SGPA
0	0	0	0	0	0	0	0	0	0

There are no missing values.

VISUALIZING RELATIONSHIPS USING ggpairs()



- CGPA has strong positive correlation with CGPA100, CGPA200, CGPA300, CGPA400, the scatterplots for these variables suggest a **linear trend**, meaning a **linear regression model** is likely appropriate.
- CGPA has weak to moderate correlation with SGPA, the scatterplot appears **more spread out**, suggesting some **non-linearity** or additional factors influencing the relationship.
- Female students generally have slightly higher correlation values compared to male students.
- The difference is small but could indicate slight variations in how academic performance evolves over time between genders.
- The density plots on the diagonal show that most variables have a roughly normal distribution.

LITERATURE REVIEW

1. **"Predicting Students' Academic Performance Using Regression"**
   Source: https://pubs.sciepub.com/education/10/11/2/index.html
   Objective: Predict students' academic performance using multiple linear regression (MLR).

- **Research Questions:**
    1. Assess the level of students' academic performance based on enrollment data.
    2. Identify significant predictors of academic performance.
    3. Develop a model to predict academic performance based on enrollment data.
- **Methodology:**
    - Utilization of enrollment data to identify significant predictors.
    - Application of MLR to develop a predictive model.
- **Key Findings:** The study successfully identified significant predictors and developed a model to forecast academic performance.

2. **"A Regression Analysis for Predicting Student Academic Performance"**

   Source:
   https://www.researchgate.net/publication/382523383_A_Regression_Analysis_for_Predicting_Student_Academic_Performance

Objective: Identify factors that accurately predict academic performance and determine each factor's contribution to overall academic success.

Dataset: 21 attributes from 97 students at a Malaysian public university.

Methodology:

- Data preprocessing and feature selection to ensure data quality.
- Development of a regression model to predict CGPA using selected variables.

Key Findings: The study highlighted specific factors significantly impacting CGPA, providing insights into academic success predictors.