EXPLORATORY DATA ANALYSIS OF ACADEMIC PERFORMANCE DRIVERS USING R FOR A U.S. COMMUNITY COLLEGE

1. Background

A community college in North Carolina observed declining course completion rates among first-year students, particularly in hybrid and online sections. Administrators wanted data-driven insights into what behavioral, demographic, and access-related factors were associated with GPA outcomes and dropout risk.

We were brought in to conduct an exploratory data analysis (EDA) using R, focusing on attendance, online access, study behavior, and student background. The objective was to identify early warning signals and performance patterns to inform proactive interventions.

2. Objective

- To explore academic performance patterns using structured EDA techniques in R
- To identify student subgroups at higher risk of underperformance
- To provide a foundation for developing predictive and advising models

3. Data Used

Source: Student Information System + Learning Management System logs (2022–2023)

Dataset Details:

- 1,120 first-year students enrolled in General Education courses
- Fields included:
    - Student_ID, Age, Gender, First_Generation_Flag, Access_Type (Online/Offline/Mixed), Attendance_Rate, Study_Hours_Per_Week, Assignment_Completion_Rate, GPA, Support_Service_Used

Preprocessing Steps:

- Cleaned and joined datasets using dplyr
- Imputed missing attendance with median; encoded access types using model.matrix()

- Normalized Study_Hours and Assignment_Completion_Rate using scale()

4. EDA Methodology

4.1 Univariate Analysis

- Histograms for GPA, Study_Hours_Per_Week, and Attendance_Rate
- Bar plots for Access_Type and Support_Service_Used

4.2 Bivariate Relationships

- Boxplots of GPA by Access_Type and First_Generation_Flag
- Scatterplots of Attendance_Rate vs. GPA with linear smoothing
- Correlation plot (cor() + corrplot) for numeric features

4.3 Group Comparisons

- Mean GPA by service usage status
- T-tests for GPA differences across gender and access types
- Quartile segmentation of GPA to profile low-performing students

5. Key Findings

Focus Area	Insight
Online Access	Fully online students had ~0.42 lower average GPA than hybrid students
Attendance	Strong positive correlation with GPA (r = 0.61)
Study Behavior	Students studying <5 hrs/week had 2.6× likelihood of GPA < 2.0
First-Gen Gap	First-gen students had 0.35 lower GPA on average; 72% never used support services
Gender Distribution	No significant GPA difference by gender (p = 0.39)

6. Interpretation and Recommendations

- **Encourage hybrid or mixed learning** for students with lower academic history
- Flag **low attendance** and **low study hours** early in the semester

- Promote targeted outreach to **first-generation students** with no support center interaction
- Integrate attendance and LMS engagement data into **student advising dashboard**

7. Reporting Output

- **R Markdown Report (PDF, 20 pages)**:
    - Visual summaries of study time, GPA distribution, and correlations
    - Clustered charts showing subgroup behaviors (e.g., online vs. hybrid learners)
    - Summary tables for student support planning
- **Excel File**:
    - GPA quartile labels per student
    - Flags for high-risk profiles based on EDA
    - Attendance and completion rate summaries
- **R Shiny Prototype**:
    - GPA and risk segmentation viewer by department
    - Student filter tool for advisors (beta version)

8. Institutional Impact

- **Support center participation increased by 24%** after targeted outreach using this analysis
- EDA framework adopted in annual **student success review cycle**
- Data-driven advising pilot launched for **Fall 2024 first-gen cohort**
- Insights used to design the college's first **study habit intervention workshop**