Project Proposal – Exploring and Predicting CGPA of Bangladeshi Students Based on Academic and Non-Academic Factors

Team 2

1. Introduction

Our Exploratory Question is: What are the relationships between educational and non-educational factors and students' cumulative GPA (CGPA)?

Our Predictive Question is: Can we identify the key factors that allow us to accurately predict students' CGPA?

This work addresses the general problem of understanding what factors influence students academic success. Specifically, we examine how both academic and non-academic characteristics relate to cumulative GPA (CGPA), using data from CS and Engineering students in Bangladesh.

The specific problem we aim to solve is identifying which features have the strongest association with students' CGPA. Our goal is to build accurate predictions and uncover the main drivers behind academic performance.

Understanding which factors truly affect CGPA is important because it can guide more effective learning strategies, help students focus on what matters most, and enable institutions to offer better academic support.

Despite prior studies on academic success, this dataset hasn't been analyzed in this way. Its richness—covering multiple domains and focusing on a specific academic field and country—makes it a unique opportunity for more focused and original insights.

Our approach begins with exploring how each variable relates to CGPA, followed by data cleaning and feature engineering. We then build predictive models, assess their performance, examine which features are most influential, and iterate as needed to improve accuracy and extract insights.

2. Data

This dataset was collected from a private university in Bangladesh and includes responses from 1,195 Computer Science and Engineering students. It contains 31 variables capturing academic performance, study habits, demographics, and personal background, offering a comprehensive view of factors potentially affecting CGPA.

To support our CGPA prediction task, we organized the features into several groups:

- **Academic Features:**

  Variables that reflect academic status and institutional support, such as CGPA, SGPA, attendance, admission year, scholarship, transportation use, and probation/suspension history.

- **Study Habits and Learning Behavior Features:**

  Metrics describing how students study, including daily study hours, session frequency, learning mode, consultation with faculty, and time invested in skill development.

- **Technology Access Features:**

  Indicators of digital resource availability and usage patterns, such as smartphone ownership, access to a personal computer, and hours spent on social media.

- **Demographic and Personal Context Features:**

  Information about the students' background, including age, gender, family income, living situation, relationship status, and involvement in co-curricular activities.

- **Health and Personal Challenges Features:**

  Self-reported health issues and disabilities that may affect students' ability to perform academically.

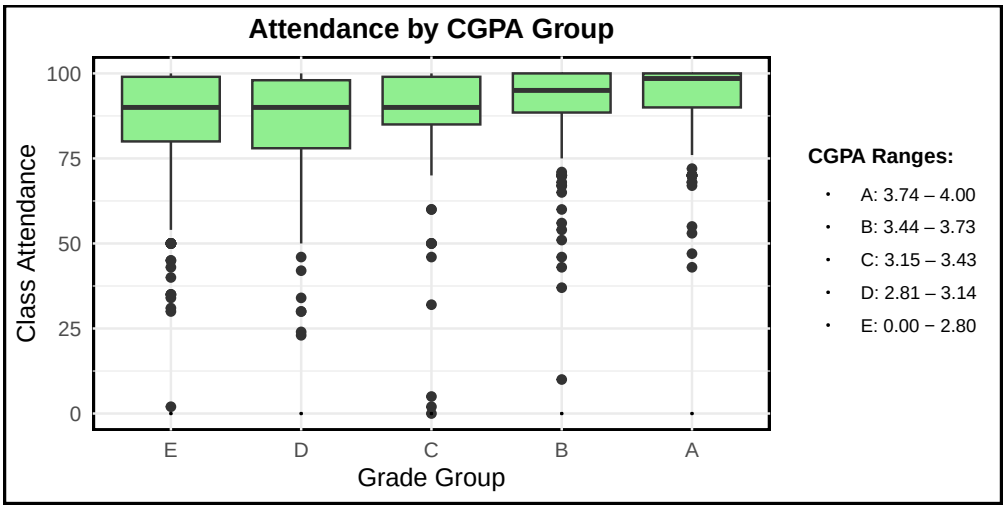
Since the data is self-reported by students, it may include limitations such as biased answers, inaccurate reporting, or subjective interpretations—for example, reporting gross rather than focused study hours. While these are common challenges in educational surveys, the dataset was designed to be representative. Still, some level of response bias or misreporting may exist.

3. Preliminary results

At this stage, we examined the relationships between selected features and CGPA to detect patterns that could inform model choice. The analysis revealed mainly non-linear relationships, supporting the decision to use non-linear models.

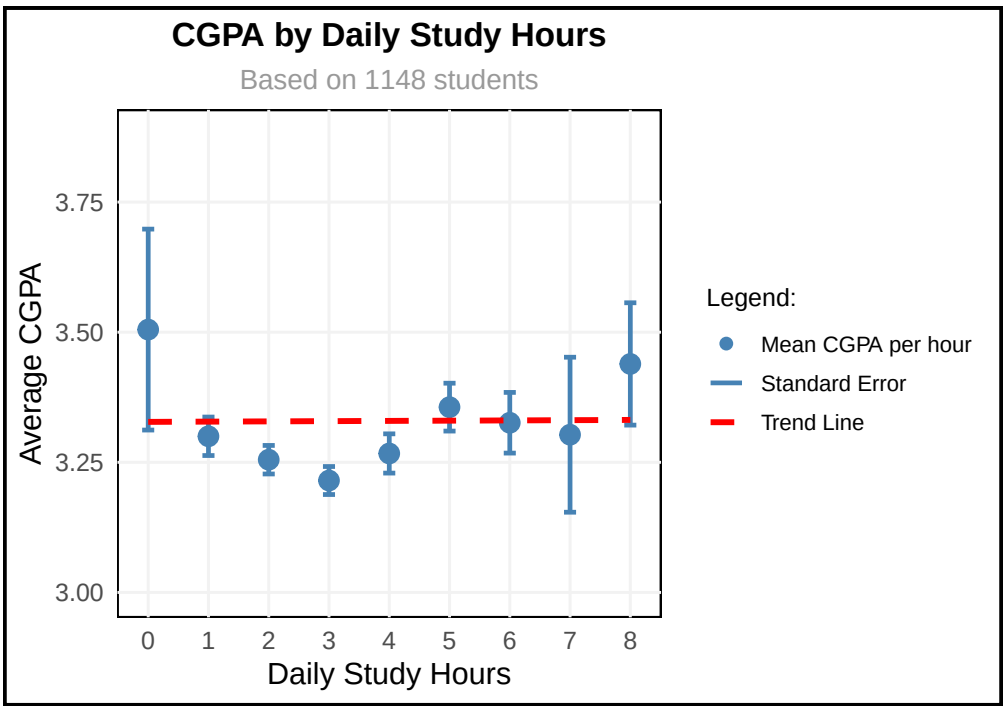
- **Relationship Between Class Attendance and CGPA:**

  Class attendance shows a clear trend across CGPA groups: higher CGPA is associated with higher and more consistent attendance, while lower CGPA groups display more variability and low-attendance outliers. This indicates a strong positive association between attendance and performance, making it a useful predictor—even if not necessarily causal.



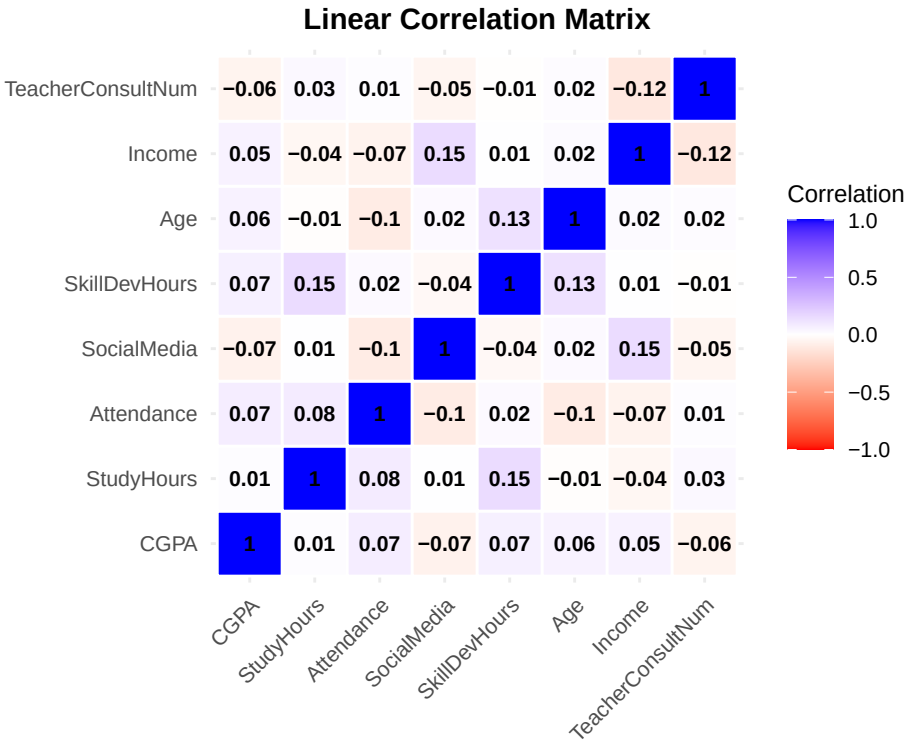
- **Study Hours and Academic Performance:**

  The scatter plot of study hours vs. CGPA reveals a non-linear and inconsistent relationship. Despite expectations, there is no clear linear relationship between study hours and CGPA. The flat trend line and the variability suggest that this feature does not exhibit a linear association with academic performance. This insight challenged our initial assumptions and led us to explore other relationships more thoroughly.



- **Linear Correlation Assessment Between Features and CGPA:**

  Based on earlier insights, we tested several features for linear correlation with CGPA, expecting meaningful relationships. However, the results showed weak correlations (mostly < 0.15), reinforcing the idea that linear connections are minimal. This further supports the need for non-linear modeling to better capture the data's structure.



- **Initial Examination of Predictive Models:**

  Given the weak linear correlations with CGPA, we tested the assumption of non-linearity by comparing a basic linear regression model with a non-linear alternative—Random Forest—using the same features and no advanced feature engineering.

  Model Performance Comparison:

Metric	Linear Regression	Random Forest Regression
R^2	0.1110	0.5471
R^2 Adjusted	0.0556	0.5161
MSE	0.5013	0.2590
RMSE	0.7080	0.5089
MAE	0.4582	0.3925

- **Model Interpretation and Key Insight:**

  The comparison confirms the benefit of non-linear modeling: linear regression explained only 11% of CGPA variance ($R^2 = 0.111$), while Random Forest reached 54.7% ($R^2 = 0.547$) with a 28% lower RMSE. This substantial improvement highlights that academic performance depends on non-linear relationships and complex feature interactions best captured by advanced models.

4. Data analysis plan

Our response variable (Y) is the students' cumulative GPA (CGPA), ranging from 0 to 4. The explanatory variables (X) are selected features from the dataset, covering learning habits, academic records, personal background, and family context.

Given the diversity of features, we first aim to evaluate how variables across different categories contribute to CGPA prediction. If needed, we will narrow the scope to focus on specific feature families, such as learning behaviors or personal characteristics.

Since Step 3 showed no strong linear relationships with CGPA, we now focus on non-linear regression models. Our goal is to identify the most suitable model through experimentation, while improving performance with feature selection and engineering. So far, we have tested Random Forest Regression and plan to explore additional non-linear methods.

Since we have currently chosen to focus on non-linear models, the results needed to support our hypotheses include:

- **Exploratory insights:** Identifying the features most associated with CGPA, based on model-derived importance and correlation analysis.
- **Predictive performance:** Achieving strong accuracy (high R^2, low RMSE/MAE), avoiding overfitting, and understanding which types of variables contribute most to model predictions.

Work Chronology and Team Roles: - Each pair of students will focus on a different non-linear model (e.g., Random Forest, XGBoost) to gain a deep understanding of its behavior and evaluate its suitability for both our predictive and exploratory objectives.

- One student will be responsible for cleaning and preparing the dataset, while the remaining team members will focus on feature engineering and selecting the most relevant variables for analysis.
- Model Evaluation:

  - One student will calculate and analyze performance metrics (R^2, RMSE, MAE) of the model.
  - A second student will handle overfitting analysis by comparing training and test predicted results.
  - Two students will focus on feature importance analysis and determine which features should be added or removed based on their impact.

- Model Optimization: If needed, we will return to Step 2 to refine the features and improve model performance based on the evaluation results.

- All team members will work together to interpret the results, assess how well the models address the research question, and compare the findings with previous studies. The team will also formulate actionable conclusions and suggest directions for future work.

Appendix

Data README

========== Team name 2

This Markdown file describes the data folder structure and organization of Academic performance dataset of Bangladeshi students:

1. "University Admission year" (chr): the range of enter the university is 2013-2023.
2. "Gender" (chr): Options- male/female.
3. "Age" (chr) the age range is 18-27.
4. "H.S.C passing year" (dbl): High School Certificate completion year the range is 2012-2022 (with an outlier in 2028).
5. "Program" (chr): all of them in Bachelor of Computer Science and Engineering.
6. "Current Semester" (dbl): values range 1-24.
7. "Do you have meritorious scholarship?" (chr): A merit-based financial award given to academically high-achieving students (Options- No/Yes).
8. "Do you use University transportation?" (chr): A service provided by the university to help students commute to campus (Options- No/Yes).
9. "How many hour do you study daily?" (chr): values range 0-13 hours.
10. "How many times do you seat for study in a day?" (dbl): values range 0-7 times
11. "What is your preferable learning mode?" (chr): Options- Offline/Online.
12. "Do you use smart phone?"(chr): Options- No/Yes.
13. "Do you have personal Computer?" (chr): Options- No/Yes.
14. "How many hour do you spent daily in social media?" (chr): values range 0-20 hours.
15. "Status of your English language proficiency" (chr): Options - Intermediate/Basic/Advanced.
16. "Average attendance on class" (chr): values range 0-100 percentage.

17. “Did you ever fall in probation?” (chr): A formal warning status assigned to students who fail to maintain minimum academic standards (Options- No/Yes).
18. “Did you ever got suspension?” (chr): A temporary removal from the university due to severe academic underperformance or rule violations (Options- No/Yes).
19. “Do you attend in teacher consultancy for any kind of academical problems?” (chr): Options- No/Yes.
20. “What are the skills do you have ?” (chr): Options are multiple skills. The most common: Programming (44.6%), Web development (21.0%), Networking (12.8%).
21. “How many hour do you spent daily on your skill development?” (chr): values range 0-12 hours
22. “What is you interested area?” (chr): Options are multiple area listed. The most common: Software (54.3%), Hardware (13.8%), Data Science (11.1%).
23. “What is your relationship status?” (chr): Options- Single/Relationship/Married/Engaged.
24. “Are you engaged with any co curriculum activities?” (chr): Options- No/Yes.
25. “With whom you are living with?” (chr): Options - Family/Bachelor.
26. “Do you have any health issues?” (chr): Options- No/Yes.
27. “What was your previous SGPA?” (chr): A measure of a student’s academic performance in their most recent completed semester. Values range - 0-4.
28. “Do you have any physical disabilities?” (chr): Options- No/Yes.
29. “What is your current CGPA?” (chr): The average of all grade points earned across all semesters to date. Values range - 0-4.
30. “How many Credit did you have completed?” (chr): values range 0-145.
31. “What is your monthly family income?” (chr): values range 4000-2,000,000.

Source code

```
#---------------- Import libraries -----------------------

library(knitr)
library(tidyverse)
library(broom)
library(htmltools)
library(readxl)
library(tidymodels)
library(rsample)
library(dplyr)
library(yardstick)
library(ranger)
library(ggplot2)
library(reshape2)
opts_chunk$set(echo=FALSE)

#---------------- Variable renaming and basic data cleaning -----------------------
```

```r
data <- read_excel("Students_Performance_data_set.xlsx")
glimpse(data)
data_model <- data.frame(
  CGPA = as.numeric(data$"What is your current CGPA?"),
  StudyHours = as.numeric(data$"How many hour do you study daily?"),
  Attendance = data$"Average attendance on class",
  SocialMedia = as.numeric(data$"How many hour do you spent daily in social media?"),
  StudyTimesPerDay = as.numeric(data$"How many times do you seat for study in a day?"),
  SkillDevHours = as.numeric(data$"How many hour do you spent daily on your skill development?"),
  Age = as.numeric(data$Age),
  Gender = as.factor(data$Gender),
  Income = as.numeric(data$"What is your monthly family income?"),
  LearningMode = as.factor(data$"What is your preferable learning mode?"),
  TeacherConsult = as.factor(data$"Do you attend in teacher consultancy for any kind of academical probl
  Smartphone = as.factor(data$"Do you use smart phone?"),
  PersonalComputer = as.factor(data$"Do you have personal Computer?"),
  UniversityTransport = as.factor(data$"Do you use University transportation?")
)

convert_range <- function(x) {
  if (is.na(x)) return(NA)
  if (grepl("-", x)) {
    bounds <- as.numeric(unlist(strsplit(x, "-")))
    return(mean(bounds, na.rm = TRUE))
  } else {
    return(as.numeric(x))
  }
}
data_model$Attendance <- sapply(data_model$Attendance, convert_range)



#---------------- Relationship Between Class Attendance and CGPA ------------------------

# Filter by income
data_model <- data_model %>%
  filter(as.numeric(Income) < 100000)

# Create CGPA group
data_model <- data_model %>%
  mutate(
    CGPA = as.numeric(CGPA),
    cgpa_group = case_when(
      CGPA <= 2.80 ~ "E",
      CGPA <= 3.14 ~ "D",
      CGPA <= 3.43 ~ "C",
      CGPA <= 3.73 ~ "B",
      CGPA <= 4.00 ~ "A",
      TRUE ~ NA_character_
    )
  )

# Order levels
```

```r
data_model$cgpa_group <- factor(data_model$cgpa_group, levels = c("E", "D", "C", "B", "A"))

# Define label with both letter and range
cgpa_legend <- data.frame(
  GradeGroup = factor(c("A", "B", "C", "D", "E"), levels = c("E", "D", "C", "B", "A")),
  RangeLabel = c(
    "A: 3.74 - 4.00",
    "B: 3.44 - 3.73",
    "C: 3.15 - 3.43",
    "D: 2.81 - 3.14",
    "E: 0.00 - 2.80"
  )
)

data_model <- data_model %>%
  left_join(cgpa_legend, by = c("cgpa_group" = "GradeGroup"))

# Plot
ggplot(data_model, aes(x = cgpa_group, y = Attendance)) +
  geom_boxplot(fill = "lightgreen") +

  # Dummy points to activate legend
  geom_point(
    data = cgpa_legend,
    aes(x = GradeGroup, y = 0, color = RangeLabel, shape = RangeLabel),
    size = 0, show.legend = TRUE
  ) +

  scale_color_manual(
    name = "CGPA Ranges:",
    values = rep("black", 5)
  ) +
  scale_shape_manual(
    name = "CGPA Ranges:",
    values = rep(12, 5)
  ) +

  labs(
    title = "Attendance by CGPA Group",
    x = "Grade Group",
    y = "Class Attendance"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 12, face = "bold"),
    legend.title = element_text(size = 9, face = "bold"),
    legend.text = element_text(size = 8),
    legend.key.height = unit(0.5, "cm"),
    legend.spacing.y = unit(0.2, "cm"),
    legend.position = "right",
    panel.background = element_rect(fill = "white", color = "black", size = 1),
    plot.background = element_rect(fill = "white", color = "black", size = 1)
  )
```

```r
#---------------- Study Hours and Academic Performance ------------------------

# Create processed data for the plot
study_performance <- data %>%
  mutate(
    study_hours_num = as.numeric(`How many hour do you study daily?`),
    cgpa_num = as.numeric(`What is your current CGPA?`)
  ) %>%
  filter(!is.na(study_hours_num) &
         !is.na(cgpa_num) &
         cgpa_num > 0 &
         study_hours_num >= 0) %>%
  group_by(study_hours = study_hours_num) %>%
  summarise(
    mean_cgpa = mean(cgpa_num, na.rm = TRUE),
    sd_cgpa = sd(cgpa_num, na.rm = TRUE),
    n = n(),
    se = sd_cgpa / sqrt(n),
    .groups = 'drop'
  ) %>%
  filter(n >= 5) %>%
  mutate(
    mean_cgpa = round(mean_cgpa, 3),
    se = round(se, 4)
  )

# Create improved plot
p1_improved <- ggplot(study_performance, aes(x = study_hours, y = mean_cgpa)) +
  geom_point(aes(color = "Mean CGPA per hour"), size = 3) +
  geom_errorbar(aes(ymin = mean_cgpa - se, ymax = mean_cgpa + se, color = "Standard Error"),
                width = 0.2, size = 0.8) +
  geom_smooth(aes(color = "Trend Line"), method = "lm", se = FALSE,
              linetype = "dashed", size = 1) +
  scale_color_manual(
    name = "Legend:",
    values = c("Mean CGPA per hour" = "steelblue",
               "Standard Error" = "steelblue",
               "Trend Line" = "red"),
    guide = guide_legend(
      override.aes = list(
        linetype = c("solid", "solid", "dashed"),
        shape = c(16, NA, NA),
        size = c(2, 1, 1)
      )
    )
  ) +
  scale_x_continuous(breaks = unique(study_performance$study_hours)) +
  scale_y_continuous(limits = c(min(study_performance$mean_cgpa - study_performance$se) * 0.95,
                                max(study_performance$mean_cgpa + study_performance$se) * 1.05)) +
  labs(
    title = "CGPA by Daily Study Hours",
```

```
    subtitle = paste("Based on", sum(study_performance$n), "students"),
    x = "Daily Study Hours",
    y = "Average CGPA"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 12, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 10, color = "gray60"),
    panel.grid.major = element_line(color = "gray95"),
    panel.grid.minor = element_blank(),
    panel.background = element_rect(fill = "white"),
    plot.background = element_rect(fill = "white", color = "black", size = 1),
    legend.position = "right",
    legend.title = element_text(size = 9),
    legend.text = element_text(size = 8),
    legend.key.size = unit(0.5, "cm"),
    legend.spacing.y = unit(0.2, "cm")
  )

# Display the plot
print(p1_improved)

# Save the plot
ggsave("cgpa_by_study_hours.png", plot = p1_improved,
       width = 10, height = 6, dpi = 300, bg = "white")



#---------------- Linear Correlation Assessment Between Features and CGPA ------------------------

# Convert attendance ranges to numeric averages (e.g., "80-90" → 85)
convert_range <- function(x) {
  if (is.na(x)) return(NA)
  if (grepl("-", x)) {
    bounds <- as.numeric(unlist(strsplit(x, "-")))
    return(mean(bounds, na.rm = TRUE))
  } else {
    return(as.numeric(x))
  }
}

data_model$Attendance <- sapply(data_model$Attendance, convert_range)

# Convert binary categorical variables from factors to 0/1
data_model$TeacherConsultNum <- ifelse(data_model$TeacherConsult == "Yes", 1, 0)
data_model$SmartphoneNum <- ifelse(data_model$Smartphone == "Yes", 1, 0)
data_model$PersonalComputerNum <- ifelse(data_model$PersonalComputer == "Yes", 1, 0)
data_model$UniversityTransportNum <- ifelse(data_model$UniversityTransport == "Yes", 1, 0)
data_model$GenderMale <- ifelse(data_model$Gender == "Male", 1, 0)
data_model$LearningOnline <- ifelse(data_model$LearningMode == "Online", 1, 0)

# Select only numeric columns for correlation analysis
numeric_cols <- c(
```

```r
  "CGPA", "StudyHours", "Attendance", "SocialMedia", "StudyTimesPerDay",
  "SkillDevHours", "Age", "Income",
  "TeacherConsultNum", "SmartphoneNum", "PersonalComputerNum",
  "UniversityTransportNum", "GenderMale", "LearningOnline"
)

# Filter to complete cases (no NAs)
correlation_data <- data_model %>%
  select(all_of(numeric_cols)) %>%
  na.omit()

# Compute correlation matrix
cor_matrix <- cor(correlation_data, use = "complete.obs")

# Select subset of variables to visualize (customizable)
selected_vars <- c(
  "CGPA", "StudyHours", "Attendance", "SocialMedia",
  "SkillDevHours", "Age", "Income", "TeacherConsultNum"
)
selected_cor_matrix <- cor_matrix[selected_vars, selected_vars]

# Convert to long format for ggplot
cor_melted <- melt(selected_cor_matrix)
names(cor_melted) <- c("Var1", "Var2", "Correlation")

# Create heatmap of correlations
heatmap_plot <- ggplot(cor_melted, aes(x = Var1, y = Var2, fill = Correlation)) +
  geom_tile(color = "white", size = 0.5) +
  scale_fill_gradient2(
    low = "red", high = "blue", mid = "white",
    midpoint = 0, limit = c(-1, 1), space = "Lab",
    name = "Correlation"
  ) +
  geom_text(
    aes(label = round(Correlation, 2)),
    color = "black", size = 2.8,
    fontface = "bold", vjust = 0.5, hjust = 0.5
  ) +
  theme_minimal() +
  labs(title = "Linear Correlation Matrix") +
  theme(
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1, size = 8),
    axis.text.y = element_text(size = 8),
    axis.title = element_blank(),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 11),
    legend.title = element_text(size = 9),
    legend.text = element_text(size = 8),
  ) +
  coord_fixed()

# Display the plot
print(heatmap_plot)
```

```r
cat(readLines('Data/README.md'), sep = '\n')
```

References

Source: Anwar, S. (2021). Academic performance dataset of Bangladeshi students. Mendeley Data, V1. https://doi.org/10.17632/dc3797vf3t.1