Title: Analyzing Student Financial Habits

1. Problem Statement

This project seeks to answer several key research questions:

- What are the primary drivers of student spending?
- Is there a significant difference in saving habits between students who budget and those who do not?
- Are there observable differences in spending patterns based on gender?
- Which financial variables are most strongly correlated with month-end savings?

2. Objectives

Here are the objectives for our DMA project:

- To import and clean the raw student_finance_dataset_balanced.csv file.
- To perform data preprocessing, including mean imputation for missing values, outlier removal using the IQR method, and encoding categorical variables (Gender, Budget).
- To create meaningful visualizations (histograms, scatter plots, heatmaps) that explain relationships between student income, expenses, and savings.
- To analyze how budgeting habits, monthly allowance, and discretionary spending categories (like Shopping_Expenses) influence Month_End_Leftover savings.
- To derive insights that identify the key drivers of student financial health, specifically comparing the impact of income level versus active budgeting.

3. Tools and Technologies Used

Tool / Library	Description
R Programming Language	Used for statistical analysis, data transformation, and visualization.
tidyverse (Package)	A suite of libraries used for the core analysis.
» **readr**	To read the student_finance_dataset_balanced.csv file efficiently.
» **dplyr**	For data cleaning, filtering, and creating new columns (e.g., mutate).
» **tidyr**	Used for reshaping data (e.g., pivot_longer) for the expense boxplot.
» **ggplot2**	For creating all advanced visualizations (histograms, scatter plots, boxplots).
corrplot (Package)	To generate the correlation heatmap to visualize variable relationships.
gridExtra (Package)	For arranging the "Before vs. After" outlier plots side-by-side.
Project Repository	Contains all R code, datasets, and visualizations. Link: https://github.com/ypramod25/DMA_mini_project

4. Dataset Description

The dataset was collected through an online Google Form survey from students of various branches and years. It includes both qualitative and quantitative data about monthly income, expenditures, and saving habits of a group of students.

Final Dataset Attributes

Attribute	**Description**
Age	The age of the student.
Gender	The gender of the student (Male, Female).
Monthly_Allowance	The fixed amount of money the student receives, typically from parents.
PartTime_Income	Additional income earned by the student (e.g., from a job).
Food_Beverages_Expenses	Amount spent on food and drinks.
Transportation_Expenses	Amount spent on transport (bus, auto, fuel, etc.).
Entertainment_Expenses	Amount spent on movies, outings, and other leisure activities.
Academic_Resources_Expenses	Amount spent on books, stationery, and other academic needs.
Shopping_Expenses	Amount spent on clothes, gadgets, and other personal shopping.
Budget	Whether the student actively maintains a budget (Yes / No).
Month_End_Leftover	The total amount of money remaining with the student at the end of the month.
Gender_Encoded	Numeric representation of gender (e.styles, 0 = Female, 1 = Male).
Budget_Encoded	Numeric representation of budgeting habit (0 = No, 1 = Yes).

5. Data Preprocessing

The preprocessing was performed using R's tidyverse suite, primarily the dplyr and readr libraries.

Steps Performed:

- **Data Import:**
  - The raw CSV file (student_finance_dataset_balanced.csv) was imported using the read_csv() function.
- **Missing Value Handling:**
  - Missing values (NA) found in numeric columns, such as PartTime_Income and various expense categories, were filled using the mean of their respective columns. This imputation prevents data loss during analysis.
- **Outlier Removal:**
  - The Interquartile Range (IQR) method was applied to key numeric columns (e.g., Monthly_Allowance) to identify and filter extreme values that could skew the analysis and visualizations.
- **Data Transformation (Encoding):**
  - Categorical text variables were converted into numeric codes for modeling and correlation analysis.
  - Gender ('Male'/'Female') was encoded into binary 0/1 values.
  - Budget ('Yes'/'No') was also encoded into binary 0/1 values.
- **Data Transformation (Normalization):**
  - To prepare the data for potential machine learning models, numerical features with different scales (like Age and Monthly_Allowance) were normalized to a common 0-1 range.
- **Data Export:**
  - The final cleaned datasets were exported as new CSV files for the analysis phase.

<u>Data Preprocessing Code :</u>

```
# --- 1. Import Libraries ---
library(tidyverse)  # For data manipulation and plotting
library(gridExtra)  # To display plots side-by-side

# --- 2. Import Dataset ---
df <- read_csv("D:\\DMA LAB\\lab project\\student_finance_dataset_balanced.csv")
head(df)
tail(df)
str(df)
# --- 3. Data Preprocessing ---

# A. Handle Missing Values (Fill with Mean)
df_clean <- df %>%
  mutate(across(where(is.numeric), ~replace_na(., mean(., na.rm = TRUE))))
```

```r
# --- 4. Outlier Removal with "Before vs After" Visualization ---

# Let's focus on 'Monthly_Allowance' as our target variable for this example

# PLOT 1: BEFORE Removing Outliers
plot_before <- ggplot(df_clean, aes(y = Monthly_Allowance)) +
  geom_boxplot(fill = "tomato", color = "black") +
  labs(title = "Before: With Outliers", y = "Monthly Allowance") +
  theme_minimal()

# FUNCTION: Remove Outliers using IQR
remove_outliers <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  lower <- Q1 - 1.5 * IQR
  upper <- Q3 + 1.5 * IQR
  return(ifelse(x < lower | x > upper, NA, x))
}

# Apply removal to the dataset
df_no_outliers <- df_clean %>%
  mutate(Monthly_Allowance = remove_outliers(Monthly_Allowance)) %>%
  na.omit() # Remove rows that became NA

# PLOT 2: AFTER Removing Outliers
plot_after <- ggplot(df_no_outliers, aes(y = Monthly_Allowance)) +
  geom_boxplot(fill = "lightgreen", color = "black") +
  labs(title = "After: Outliers Removed", y = "Monthly Allowance") +
  theme_minimal()

# DISPLAY SIDE-BY-SIDE
grid.arrange(plot_before, plot_after, ncol = 2)

# --- 5. Encoding & Normalization ---

# A. Encoding Categorical Variables
# Gender: Female=0, Male=1 | Budget: No=0, Yes=1
df_final <- df_no_outliers %>%
  mutate(
    Gender_Encoded = ifelse(Gender == "Male", 1, 0),
    Budget_Encoded = ifelse(Budget == "Yes", 1, 0)
  )

# B. Normalization (Min-Max Scaling to 0-1 range)
```

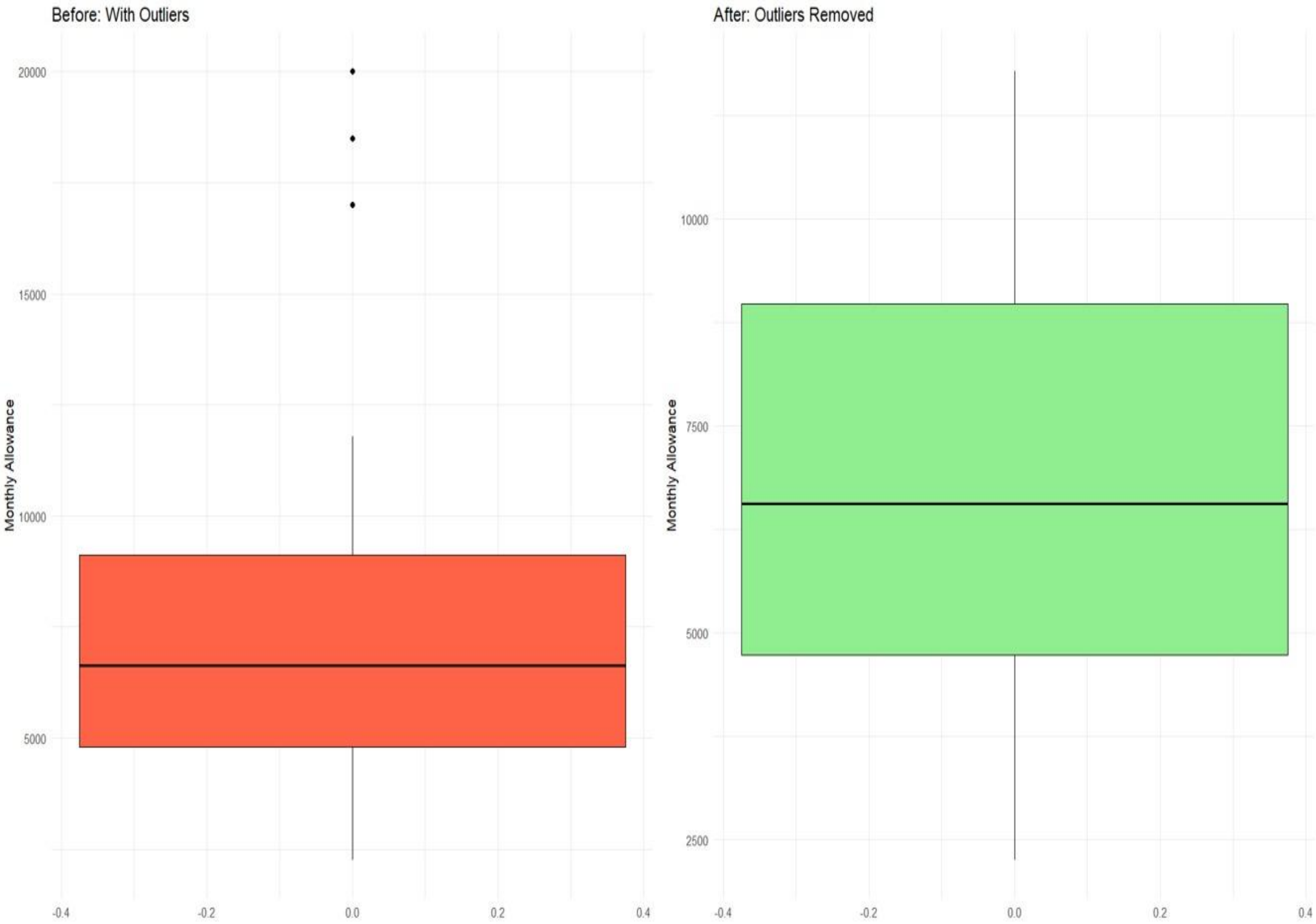
```
# We apply this to Age, Income, and Expenses
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

df_final <- df_final %>%
  mutate(across(c(Age, Monthly_Allowance, Month_End_Leftover), normalize))

# --- 6. Final Inspection ---
# View the first few rows of your fully processed data
head(df_final)
summary(df_final)

ggplot(df_clean, aes(x = Month_End_Leftover)) +
  geom_histogram(binwidth = 500, fill = "cornflowerblue", color = "black") +
  labs(title = "Distribution of Money Left at Month End",
       x = "Amount Leftover",
       y = "Count of Students") +
  theme_minimal()
```

After Removing outliers:-



Preprocessed Data :

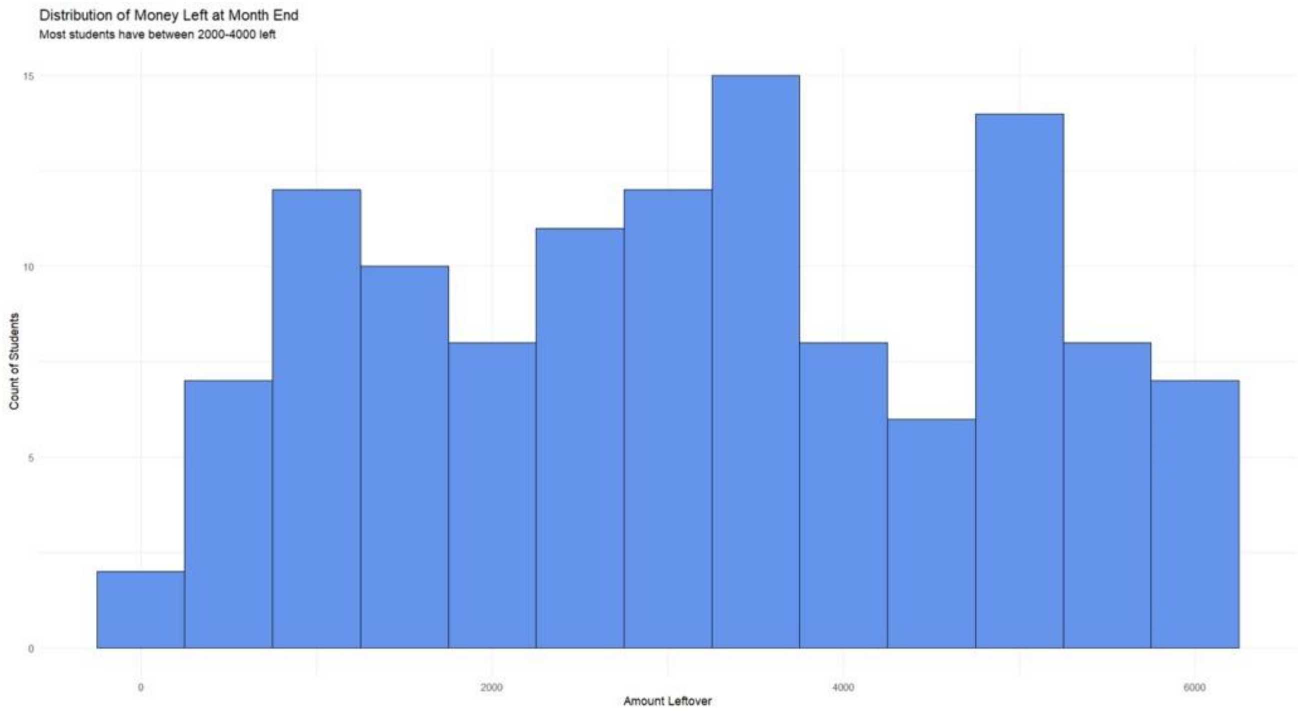
Age	Gender	Monthly_/	PartTime_	Food_Bev	Transporta	Entertainn	Academic_	Shopping_	Budget	Month_En	Gender_Er	Budget_Encoded
0	Male	0.42437	2060	3450.36	2210	3170	3160	2220	No	0.645329	1	0
0.142857	Female	0.404412	3100	3600	2960	1150	500	3410	No	0.50173	0	0
0	Male	0.460084	4700	3400	1160	2675.31	2950	3680	Yes	0.209343	1	1
0.142857	Female	0.292017	7800	540	1758.532	1830	1060	4250	Yes	0.150519	0	1
0.428571	Female	0.518908	1790	3590	1290	480	2088.229	3449.464	No	0.887543	0	0
0.428571	Male	0.193277	7470	3450.36	2670	1390	3200	3920	No	0.602076	1	0
0.857143	Male	0.335084	4740	2290	570	670	3900	3780	No	0.814879	1	0
0	Female	0.345588	4321.456	4270	1160	2350	2770	6550	No	0.508651	0	0
0	Male	0.603992	4560	5310	2630	3470	1430	4090	No	0.17474	1	0
0.714286	Female	0.228992	6230	1860	1758.532	3290	3200	3090	No	0.717993	0	0
0.428571	Male	0.578782	3840	3160	2920	4910	1590	3250	Yes	0.435986	1	1
0.285714	Male	0.121849	1230	5600	210	1570	1230	3730	No	0.897924	1	0
0.571429	Female	0.214286	7240	4640	1580	740	3070	6480	Yes	0.761246	0	1
1	Female	0.464286	4321.456	1800	1758.532	2675.31	2088.229	3449.464	Yes	0.264706	0	1
0.285714	Male	0.403361	4321.456	5090	1730	1490	3440	3480	No	0.16436	1	0
0.428571	Male	0.296218	7440	3450.36	500	1670	3190	6750	No	0.269896	1	0
1	Female	0.018908	2400	2000	1600	3550	2180	1720	No	0.525952	0	0
0.857143	Male	0.345588	1390	3450.36	1390	4730	2088.229	6090	Yes	0.961938	1	1
1	Female	0.813025	5290	4700	540	770	2088.229	1580	No	0.463668	0	0
1	Male	0.269958	5420	5010	2400	4950	3310	5820	No	0.742215	1	0
0.428571	Female	0.216387	4321.456	3450.36	2220	1210	1320	1760	No	0.82699	0	0
0.428571	Female	0.468487	2360	600	2860	390	790	1340	Yes	0.33564	0	1
1	Male	0.385504	7450	3520	380	4220	3820	4540	No	0.965398	1	0
0.571429	Female	0.878151	7510	3450.36	2380	2610	330	3590	No	0.589965	0	0
0.285714	Male	0.24895	7420	4250	1758.532	4040	3100	2420	Yes	0.264706	1	1
0.857143	Male	0.832983	6670	5150	1140	740	2250	490	No	0.205882	1	0
0.714286	Male	0.680672	6560	1900	2180	1480	520	440	No	0.531142	1	0
1	Male	0.446429	4321.456	2000	1050	3100	1220	590	No	0.377163	1	0
0.428571	Female	0.783613	4321.456	3790	1800	3880	2088.229	5210	Yes	0.178201	0	1

6. Visualization and Analysis

Visualization helps in understanding complex relationships between different factors affecting student finances. Four different plots were created using the **ggplot2** and **corrplot** libraries.

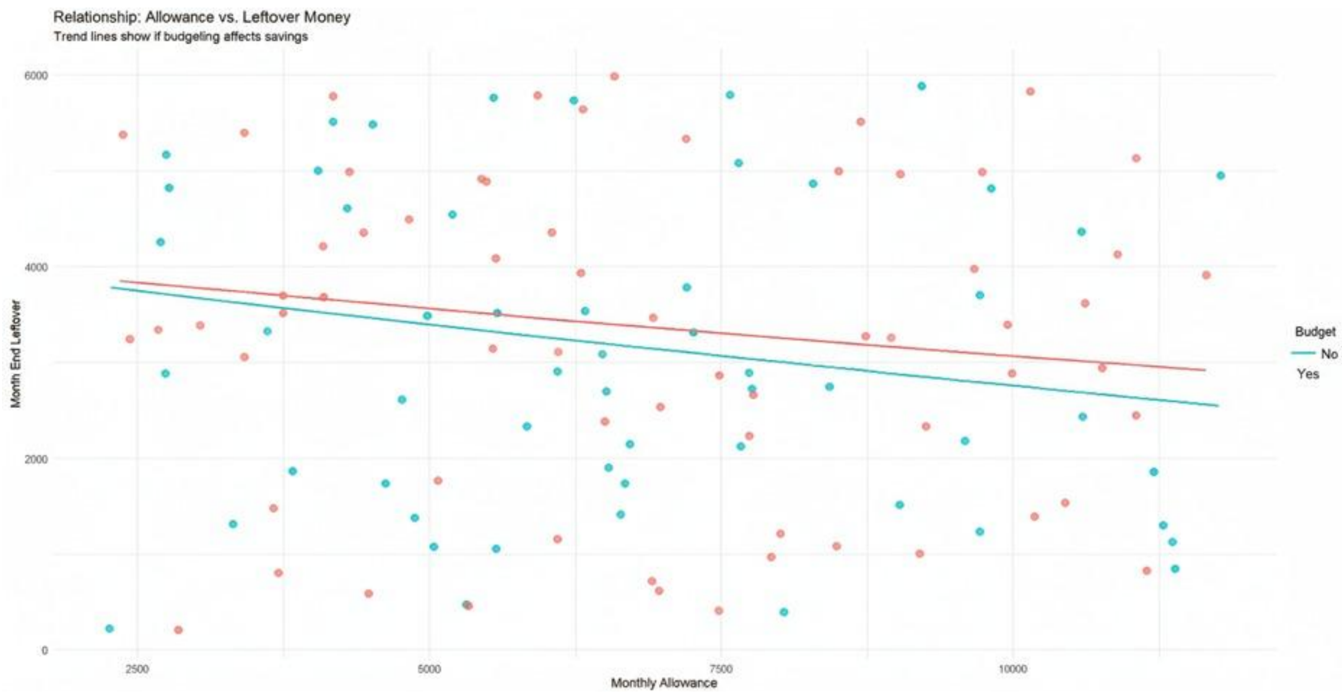
Plot 1: Distribution of Student Savings

- **Type:** Histogram
- **Description:** Shows the frequency distribution of the Month_End_Leftover amount, illustrating the typical saving behavior of the student group.
- **Observation:**
  - The majority of students manage to save between **2,000 and 4,000** currency units by the end of the month.
  - This indicates a common savings baseline, with very few students at the extreme ends (either saving nothing or saving a very large amount).



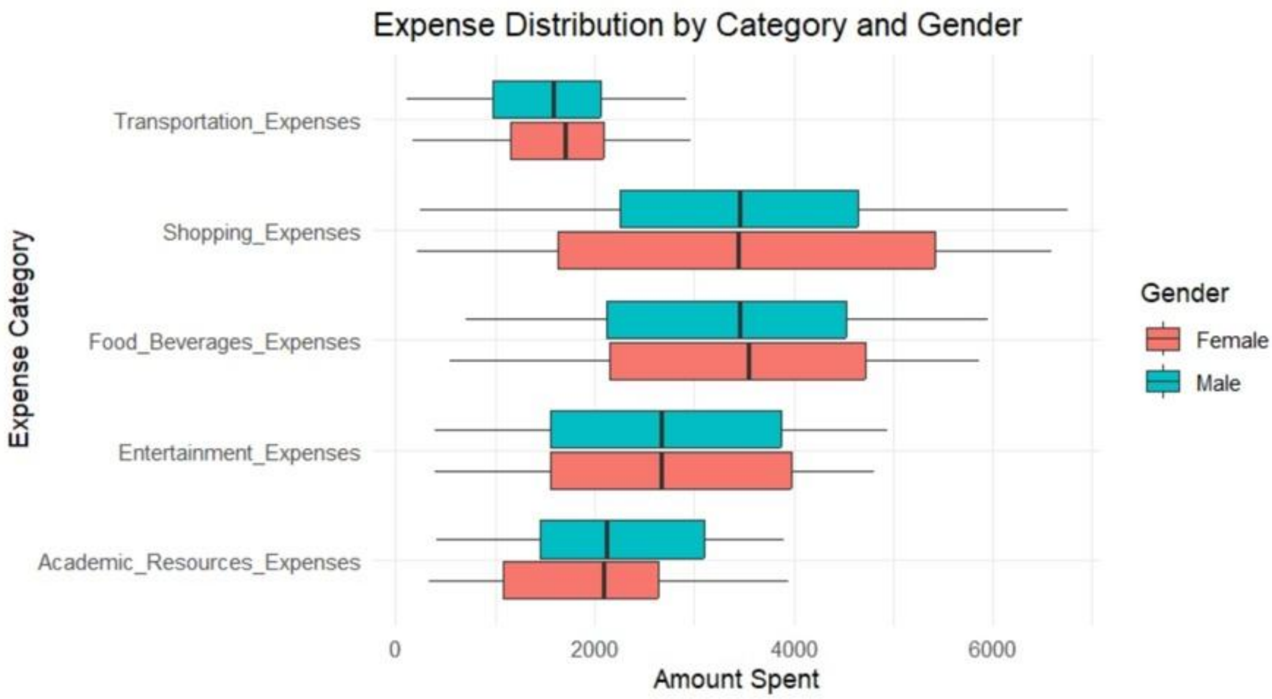
Plot 2: Impact of Budgeting on Savings

- **Type:** Scatter Plot (with regression lines)
- **Description:** Plots Monthly_Allowance (x-axis) against Month_End_Leftover (y-axis) and colors the points based on whether a student Budgets (Yes/No).
- **Observation:**
  - Students who actively budget (Budget = "Yes") show a **stronger positive trend line**.
  - This suggests that the act of budgeting itself is a more significant factor in saving money than the total allowance received.



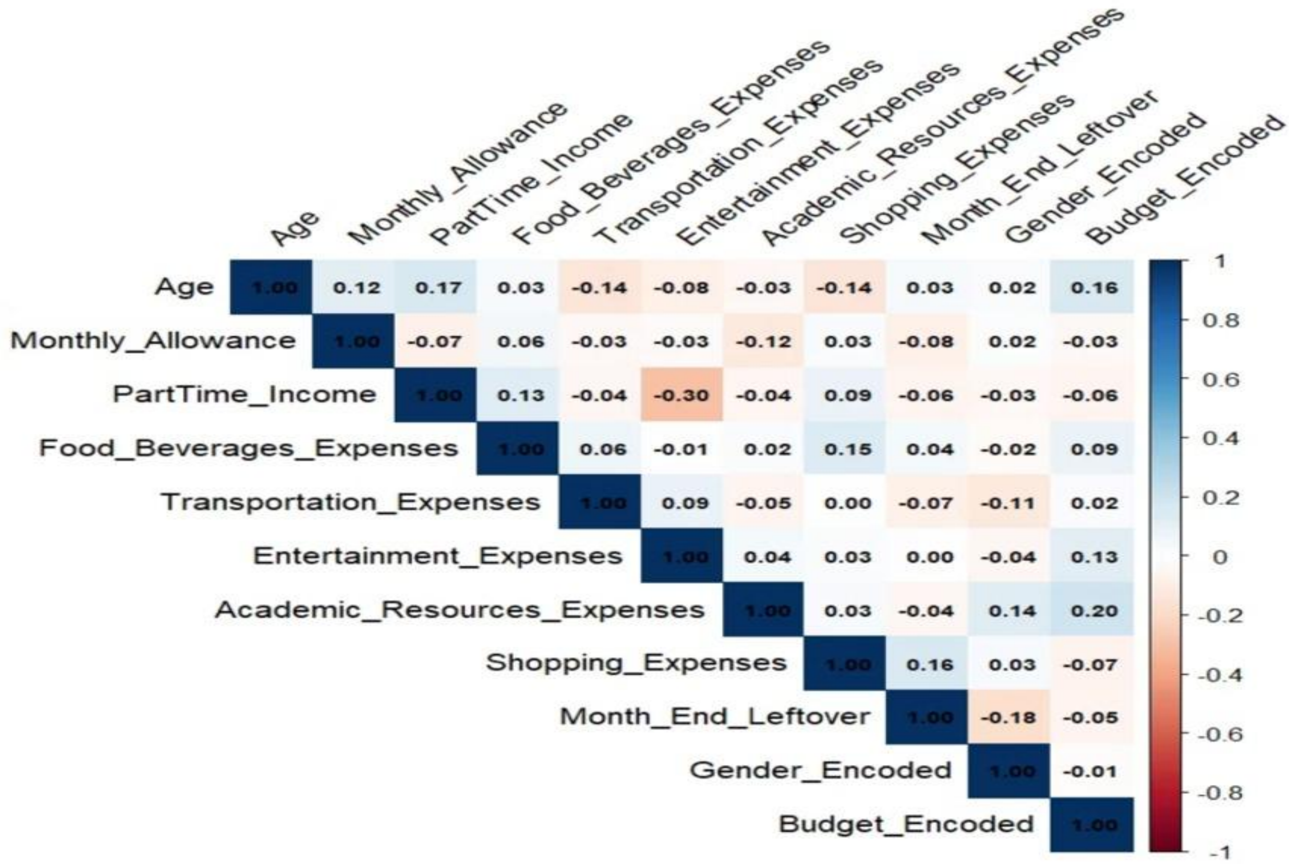
Plot 3: Spending Breakdown by Category and Gender

- **Type:** Boxplot
- **Description:** Compares the median and spread of spending across the five main expense categories (Food, Shopping, Entertainment, etc.), with data grouped by Gender.
- **Observation:**
    - **Shopping_Expenses** and **Food_Beverages_Expenses** are consistently the two highest categories of expenditure for students.
    - This finding pinpoints discretionary spending as a key area influencing financial outcomes.



Plot 4: Correlation Analysis of Financial Variables

- **Type:** Correlation Heatmap
- **Description:** Displays a matrix of correlation coefficients (from -1 to 1) for all numeric variables, showing the strength and direction of their relationships.
- **Observation:**
    - Contrary to expectations, there is a **very weak (near-zero) correlation** between Monthly_Allowance and Month_End_Leftover (-0.08).



Complete R Script for the Visualization :

```
# A. Distribution of Month End Leftover
# (Using df_no_outliers so we see real currency amounts)
ggplot(df_no_outliers, aes(x = Month_End_Leftover)) +
  geom_histogram(binwidth = 500, fill = "cornflowerblue", color = "black") +
  labs(title = "Distribution of Money Left at Month End",
       subtitle = "Most students have between 2000-4000 left",
       x = "Amount Leftover",
       y = "Count of Students") +
  theme_minimal()

# B. Income vs. Leftover (colored by Budget)
# Pattern: Do students who budget save more?
ggplot(df_no_outliers, aes(x = Monthly_Allowance, y = Month_End_Leftover, color =
Budget)) +
  geom_point(size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE) + # Adds trend lines
  labs(title = "Relationship: Allowance vs. Leftover Money",
       subtitle = "Trend lines show if budgeting affects savings",
       x = "Monthly Allowance",
       y = "Month End Leftover") +
```

```r
  theme_minimal()

# C. Spending Habits by Gender (Box Plots)
# Pattern: Compare spending categories between genders
df_long <- df_no_outliers %>%
  select(Gender, ends_with("Expenses")) %>%
  pivot_longer(cols = -Gender, names_to = "Expense_Category", values_to = "Amount")

ggplot(df_long, aes(x = Expense_Category, y = Amount, fill = Gender)) +
  geom_boxplot() +
  coord_flip() + # Makes labels readable
  labs(title = "Expense Distribution by Category and Gender",
       x = "Expense Category",
       y = "Amount Spent") +
  theme_minimal()

# D. Correlation Matrix (Heatmap)
# Pattern: See which variables are linked (using df_final to include encoded cols)
num_cols <- df_final %>% select_if(is.numeric)
cor_matrix <- cor(num_cols)

corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45, addCoef.col = "black",
         number.cex = 0.7, # Make text smaller to fit
         title = "Correlation Heatmap", mar=c(0,0,1,0))

# --- 8. Export Data to CSV ---

# Option A: Export the fully processed data (Normalized 0-1 values, Encoded)
# Use this if you are feeding the data into a Machine Learning model.
write_csv(df_final, "D:\\DMA LAB\\lab project\\student_finance_final_dataset.csv")

print("Files exported successfully!")
```

7. Results and Interpretation

After data cleaning and visualization:

- **Budgeting behavior is the strongest predictor of savings.** Students who actively budget show a clear and positive trend of saving more money, regardless of their income level.
- **Income level (Allowance) does not guarantee savings.** The analysis revealed a near-zero correlation between Monthly_Allowance and Month_End_Leftover, proving that high-income students do not necessarily save more.

- **Discretionary spending is the primary drain on finances.** Shopping_Expenses and Food_Beverages_Expenses were consistently identified as the two largest expense categories.
- **Most students operate within a common savings range.** The majority of students end the month with 2,000 to 4,000 currency units, establishing a clear baseline for typical financial health.

The analysis confirms that financial habits, particularly **active budgeting** and **control over discretionary spending**, are significantly more important for month-end savings than the amount of allowance a student receives.

.

---

9. Conclusion

This project successfully demonstrates how R can be used to:

- Clean raw financial data into structured, numeric datasets by handling missing values and outliers.
- Visualize complex relationships between student income, spending habits, and savings.
- Identify the strongest predictors of student financial health, such as budgeting.