Final Project

2023-10-31

Introduction:

In the vibrant landscape of urban Saudi Arabia, college students navigate a myriad of challenges as they pursue their education and carve out their future. Balancing academic commitments with financial constraints and lifestyle choices, these students embody the complex interplay of ambition, culture, and socioeconomic factors. This research embarks on a crucial exploration, aiming to unravel the underlying patterns that influence the spending habits and lifestyle choices of college students in major Saudi cities.

The primary aim of this project is to gain profound insights into the financial behaviors of college students in urban Saudi environments. We seek to understand the diverse factors, including gender, age, study year, socioeconomic background, and individual habits, that impact students' spending patterns. By delving deep into these intricacies, we aim to unravel the unique challenges faced by students, providing a nuanced understanding of their financial decisions within the cultural context of Saudi Arabia.

Our goals are to uncover patterns- identify recurring patterns and trends in students' spending habits, shedding light on the factors driving these behaviors-, inform support systems- provide actionable insights for educational institutions and policymakers to design targeted support systems, addressing the specific needs of students-, and enhance student experience- facilitate businesses catering to students in tailoring their services, ensuring they align with authentic student needs and preferences.

In this report, we will meticulously analyze the dataset, employing various statistical and machine learning techniques to derive meaningful conclusions. We will offer a comprehensive roadmap of our analysis, encompassing data collection, preprocessing, modeling, and interpretation of results. Through detailed visualizations and clear explanations, we aim to present a cohesive narrative of our findings, allowing readers to grasp the complexities of student financial behaviors in Saudi urban environments.

Significance and Problem Statement:

The project addresses the fundamental issue of understanding the financial dynamics of college students in urban Saudi settings. While prior studies have explored similar themes on a global scale, there exists a dearth of research focusing specifically on the nuanced context of Saudi Arabian students within their local cities. This project bridges this gap by conducting a light literature review, summarizing existing works related to student spending behaviors and lifestyle choices. By drawing on this background, we contextualize our analysis, laying the foundation for our exploration into the unique challenges faced by students in major Saudi cities.

#Data:

In this study, our dataset originates from a meticulously tailored survey designed for the specific cultural context of Saudi Arabia. The unit of observation encompasses individual college students residing in major cities across the country. The cornerstone of our analysis lies in the total monthly expenses, a pivotal metric indicating the financial behaviors of our surveyed students.

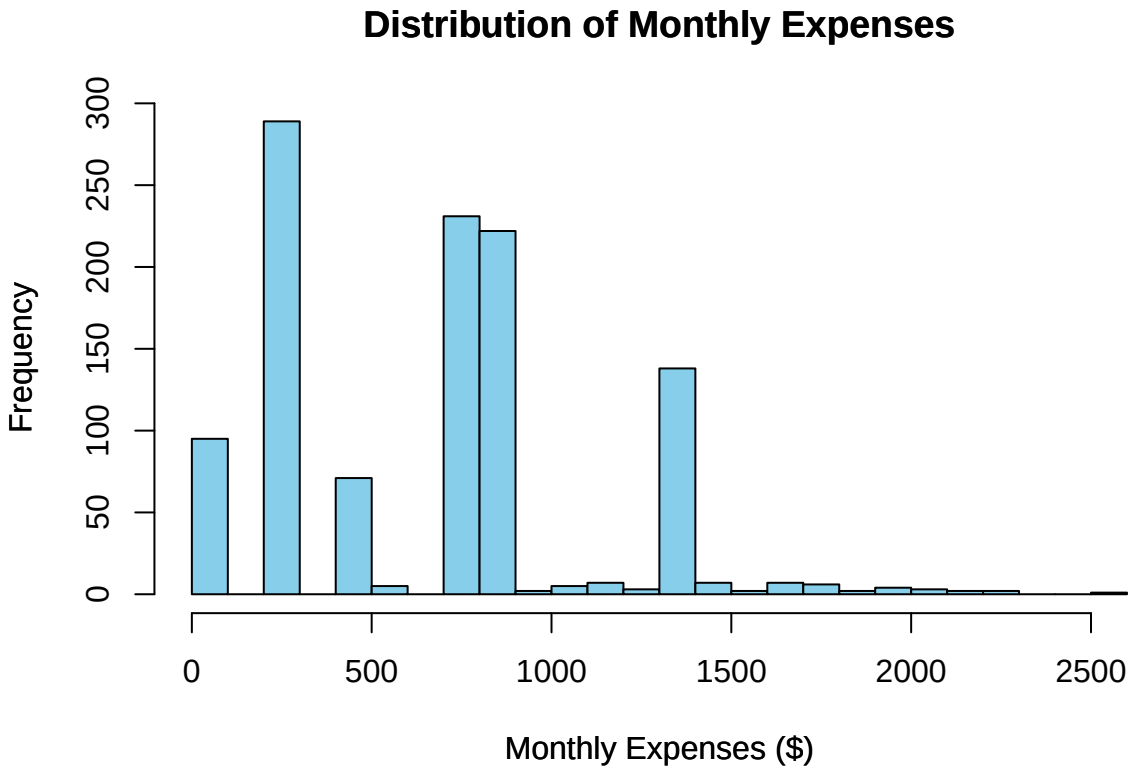
##Outcome Variable: Our primary outcome variable, Total Monthly Expenses ($), is the focal point of our analysis. Derived from the survey responses, this variable quantifies the financial expenditure of each

student, covering a diverse array of spending categories. Its measurement provides a comprehensive view of students' financial realities. To offer a visual understanding, we represent the distribution of total monthly expenses through a histogram, elucidating the range and frequency of expenditure levels, as depicted below.

Histogram of Total Monthly Expenses

```
#install.packages("readxl")
```

```
# Load the readxl package
library(readxl)

# Load data from the Excel file into a data frame
university_students_data <- read_excel("university_students_data.xlsx")

# Clean the Monthly_expenses_$ column: convert to numeric and remove missing/NA values
university_students_data$Monthly_expenses <- as.numeric(university_students_data$Monthly_expenses)

# Remove rows with missing or NA values in Monthly_expenses_$
university_students_data <- university_students_data[!is.na(university_students_data$Monthly_expenses),

# Create a histogram to visualize the distribution of Monthly_expenses_$
hist(university_students_data$Monthly_expenses,
     main = "Distribution of Monthly Expenses",
     xlab = "Monthly Expenses ($)",
     ylab = "Frequency",
     col = "skyblue",  # Change the color if desired
     border = "black",  # Border color of the bars
     breaks = 20  # Number of bins in the histogram
)

# Add a title and labels to the histogram
title(main = "Distribution of Monthly Expenses",
      xlab = "Monthly Expenses ($)",
      ylab = "Frequency")
```



##Predictor Variables: The predictor variables employed in our analysis were curated from a published paper, ensuring their relevance and reliability. These variables include gender, age, study year, living arrangements, socioeconomic background, part-time job status, transportation mode, smoking habits, coffee/energy drinks consumption, count of monthly subscriptions, location, and major. Each variable, meticulously chosen, is instrumental in unraveling the nuanced factors shaping students' spending habits and lifestyle choices.

##Data Challenges and Mitigation Strategies: While analyzing the dataset, we encountered several challenges. Addressing missing data, we employed imputation techniques to maintain a complete dataset, preserving the integrity of our analysis. To overcome issues related to limited variation or availability within specific variables, we carefully examined their distributions. In instances where variables demonstrated restricted variation, we amalgamated categories, ensuring the meaningfulness of our insights. Furthermore, to mitigate potential biases introduced by the survey's sampling method or response patterns, we adopted a meticulous approach, aiming for a diverse and representative sample. Sensitivity analyses were conducted, evaluating the impact of biases on our results, thereby enhancing the robustness and credibility of our findings.

2. Loading and Exploring Data Load the survey data into R from the appropriate file format (CSV or RData). Display the first few rows of the dataset to get an overview of the data structure. Check for missing values and handle them appropriately. Explore summary statistics of key variables (mean monthly expenses, demographic information, spending categories) to gain initial insights.

```
# Display the first few rows of the dataset
head(university_students_data)
```

```
## # A tibble: 6 x 17
##   Gender   Age Study_year Living Scholarship Part_time_job Transporting  Smoking
```

```
##   <chr>  <dbl>      <dbl> <chr>  <chr>       <chr>         <chr>         <chr>
## 1 Female    21          2 Home   No          No            No            No
## 2 Male      25          3 dorm   No          Yes           public trans~ No
## 3 Male      23          2 Home   Yes         No            No            No
## 4 Male      19          3 dorm   No          No            public trans~ No
## 5 Female    19          2 Home   No          No            public trans~ No
## 6 Male      22          3 dorm   No          Yes           Car           <NA>
## # i 9 more variables: Coffee_or_Energy_Drinks <chr>, Games_and_Hobbies <chr>,
## #   Cosmetics_and_Selfcare <chr>, Monthly_Subscription <chr>,
## #   Monthly_expenses <dbl>, `3_or_more_Subscriptions` <chr>, Location <chr>,
## #   Socioeconomic_Background <chr>, Major <chr>
```

```r
# Check for missing values in the entire dataset
missing_values <- sum(is.na(university_students_data))

# Handle missing values (assuming you want to remove rows with missing values)
university_students_data <- na.omit(university_students_data)

# Explore summary statistics of key variables
summary(university_students_data$Monthly_expenses)  # Summary statistics of monthly expenses
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   300.0   750.0   686.6   900.0  2550.0
```

```r
# Summary statistics of demographic information
summary(university_students_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   18.00   19.00   19.92   22.00   25.00
```

```r
summary(university_students_data$Gender)
```

```
##    Length     Class      Mode
##       697 character character
```

```r
summary(university_students_data$Study_year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.644   4.000   4.000
```

```r
# ... and other demographic variables

# Summary statistics of spending categories
summary(university_students_data$Games_and_Hobbies)
```

```
##    Length     Class      Mode
##       697 character character
```

```r
summary(university_students_data$Cosmetics_and_Selfcare)
```

```
##    Length     Class      Mode
##       697 character character
```

```r
# ... and other spending categories

# Explore relationships between variables (for example, Age vs. Monthly_expenses)
plot(university_students_data$Age, university_students_data$Monthly_expenses,
     xlab = "Age", ylab = "Monthly Expenses", main = "Age vs. Monthly Expenses")
```


```r
# Save the cleaned data for further analysis
write.csv(university_students_data, "cleaned_university_data.csv", row.names = FALSE)
```

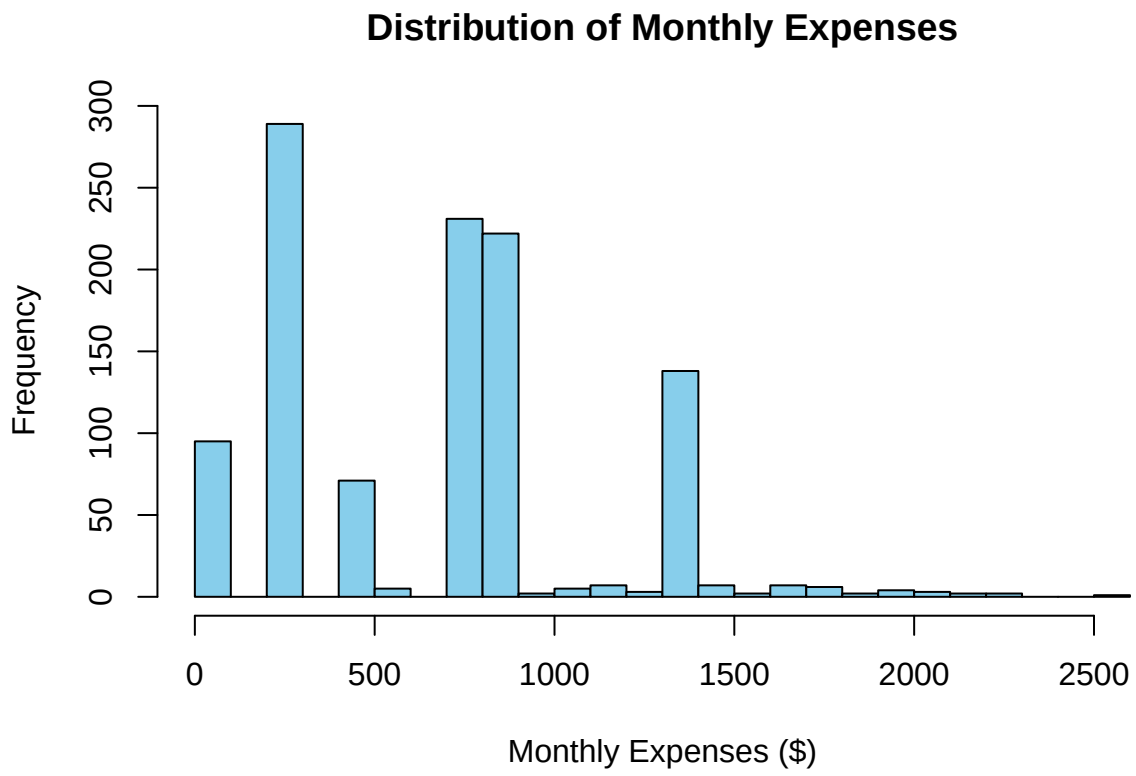
```r
# Load necessary packages
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

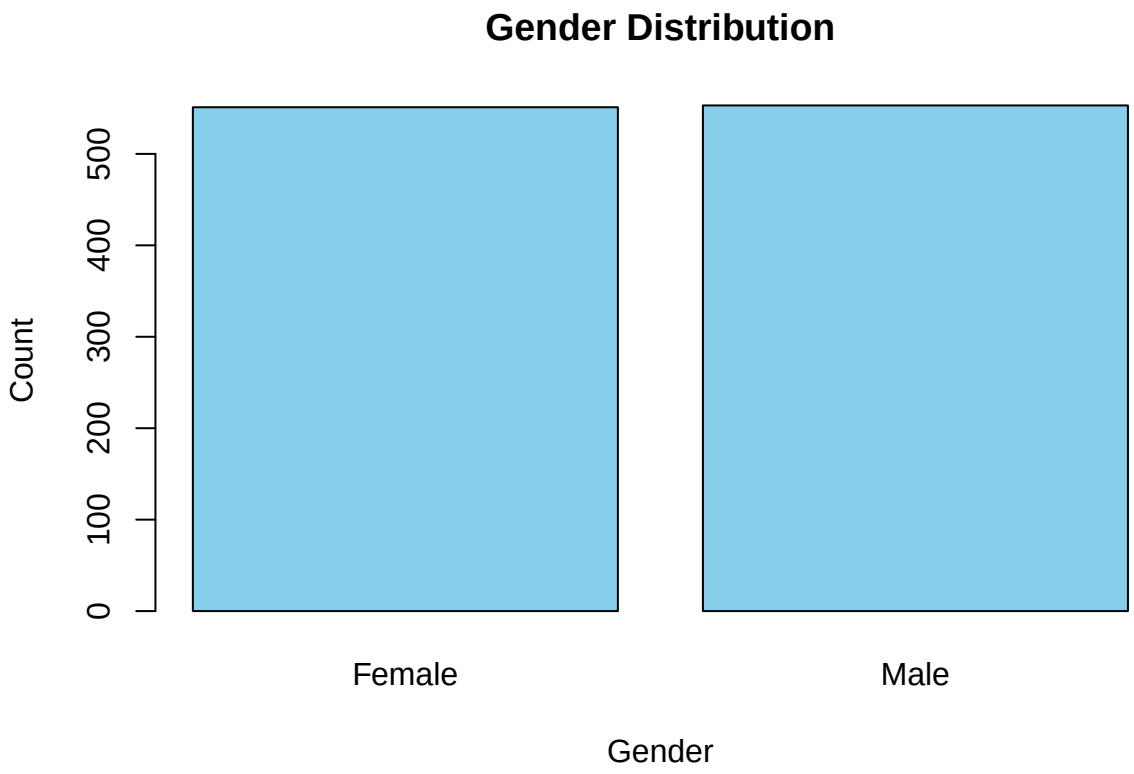
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

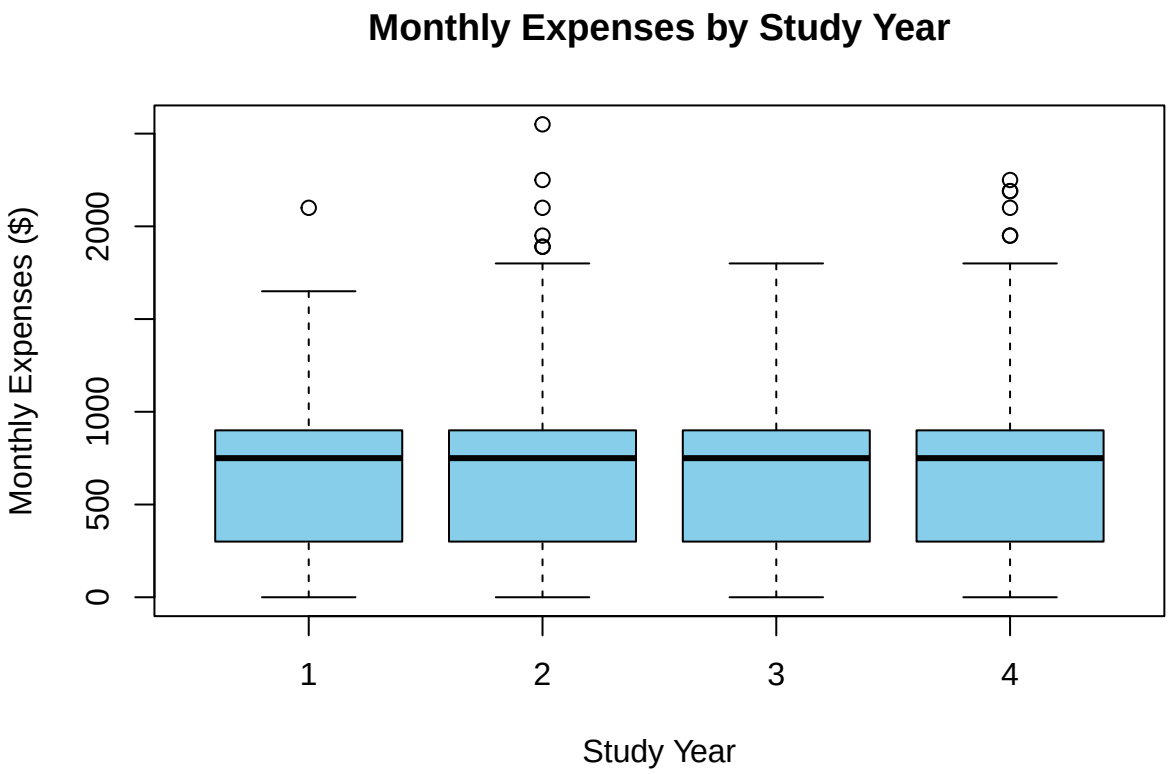
```r
# Load data from the Excel file into a data frame
university_students_data <- read_excel("university_students_data.xlsx")

# Histogram for Monthly Expenses
hist(university_students_data$Monthly_expenses,
     main = "Distribution of Monthly Expenses",
     xlab = "Monthly Expenses ($)",
     ylab = "Frequency",
     col = "skyblue",
     border = "black",
     breaks = 20)
```




```r
# Bar plot for Gender
barplot(table(university_students_data$Gender),
        main = "Gender Distribution",
        xlab = "Gender",
        ylab = "Count",
        col = "skyblue")
```

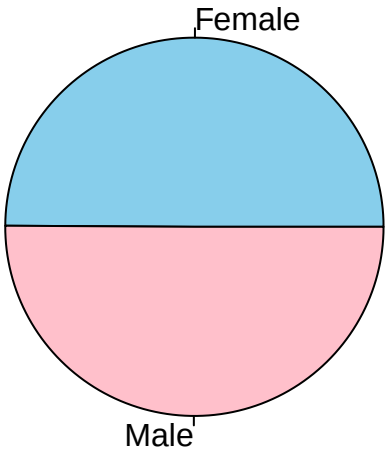



```r
# Box plot for Monthly Expenses by Study Year
boxplot(Monthly_expenses ~ Study_year, data = university_students_data,
        main = "Monthly Expenses by Study Year",
        xlab = "Study Year",
        ylab = "Monthly Expenses ($)",
        col = "skyblue")
```




```r
# Scatter plot for Age vs. Monthly Expenses
plot(university_students_data$Age, university_students_data$Monthly_expenses,
     xlab = "Age",
     ylab = "Monthly Expenses ($)",
     main = "Age vs. Monthly Expenses",
     col = "skyblue")
```


```r
# Pie chart for Gender Distribution
gender_counts <- table(university_students_data$Gender)
pie(gender_counts, labels = names(gender_counts),
    main = "Gender Distribution",
    col = c("skyblue", "pink"))
```




```r
# Compute the correlation matrix
correlation_matrix <- cor(select(university_students_data, c("Age", "Monthly_expenses")))
```

3. Data Preprocessing Clean and preprocess the data as necessary (handling missing values, transforming variables, etc.). Create dummy variables for categorical predictors if needed. Normalize or scale continuous variables if required for the chosen modeling techniques.

```r
cleaned_data <- na.omit(university_students_data)
# Check for missing values in the entire dataframe
any_missing <- any(is.na(university_students_data$Monthly_expenses) | university_students_data$Monthly_

# Print the result
if (any_missing) {
  print("There are missing values in the dataset.")
} else {
  print("There are no missing values in the dataset.")
}
```

```
## [1] "There are missing values in the dataset."
```

```r
# Count the number of zero values in the Monthly_expenses column
zero_count <- sum(university_students_data$Monthly_expenses == 0)

# Print the result
cat("Number of zero values in Monthly_expenses:", zero_count, "\n")
```

```
## Number of zero values in Monthly_expenses: 95
```

```
# Calculate the mean of non-zero values in Monthly_expenses
non_zero_mean <- mean(university_students_data$Monthly_expenses[university_students_data$Monthly_expense

# Replace zero values with the calculated mean
university_students_data$Monthly_expenses[university_students_data$Monthly_expenses == 0] <- non_zero_me
```

```
# Load required libraries
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
# Load data from the Excel file into a data frame
university_students_data <- read_excel("university_students_data.xlsx")

# Handling Missing Values
missing_values <- colSums(is.na(university_students_data))
threshold <- 0.5
university_students_data <- university_students_data[, missing_values / nrow(university_students_data) <

# Handling Zero Values in Monthly Expenses
non_zero_mean <- mean(university_students_data$Monthly_expenses[university_students_data$Monthly_expense
university_students_data$Monthly_expenses[university_students_data$Monthly_expenses == 0] <- non_zero_me


# Feature Scaling
university_students_data$Age <- scale(university_students_data$Age)
university_students_data$Monthly_expenses <- scale(university_students_data$Monthly_expenses)

# Handling Outliers using IQR
Q1 <- quantile(university_students_data$Age, 0.25)
Q3 <- quantile(university_students_data$Age, 0.75)
IQR <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR
university_students_data <- university_students_data[university_students_data$Age >= lower_bound & univ

# Handling Imbalanced Data (if needed)
# For balancing classes, you can use techniques like undersampling or oversampling.

# Data Splitting
set.seed(123)  # for reproducibility
train_index <- sample(1:nrow(university_students_data), 0.8 * nrow(university_students_data))
train_data <- university_students_data[train_index, ]
test_data <- university_students_data[-train_index, ]
```

4. Exploratory Data Analysis (EDA) Conduct exploratory data analysis using visualizations (histograms, box plots, etc.) to understand the distribution of variables. Explore correlations between predictor variables and the outcome variable (monthly expenses). Generate insights into potential patterns and relationships within the data.

```r
# Load required libraries
library(readxl)
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
# Load data from the Excel file into a data frame
university_students_data <- read_excel("university_students_data.xlsx")

# Exploratory Data Analysis (EDA)

# Check the structure of the dataset
str(university_students_data)
```

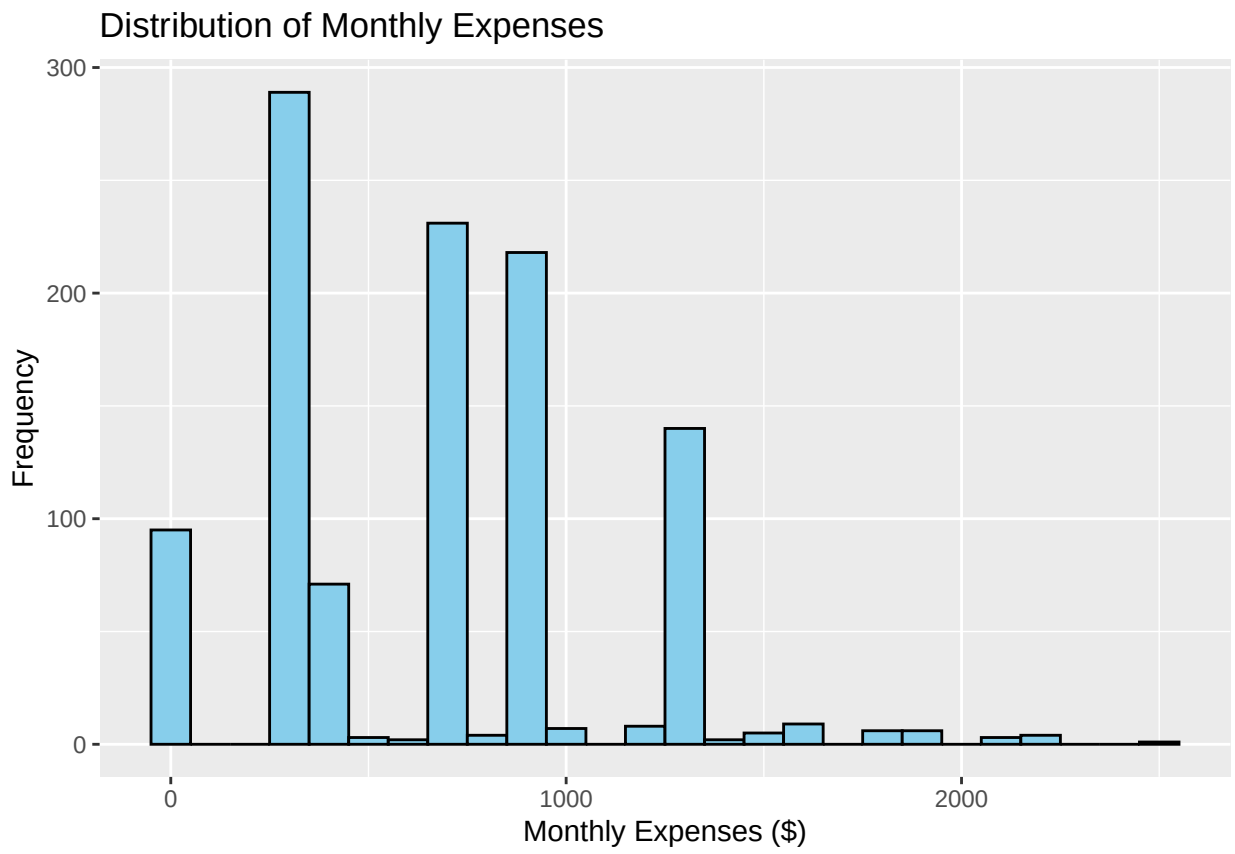
```
## tibble [1,104 x 17] (S3: tbl_df/tbl/data.frame)
##  $ Gender                  : chr [1:1104] "Female" "Male" "Male" "Male" ...
##  $ Age                     : num [1:1104] 21 25 23 19 19 22 21 22 18 19 ...
##  $ Study_year              : num [1:1104] 2 3 2 3 2 3 2 3 1 1 ...
##  $ Living                  : chr [1:1104] "Home" "dorm" "Home" "dorm" ...
##  $ Scholarship             : chr [1:1104] "No" "No" "Yes" "No" ...
##  $ Part_time_job           : chr [1:1104] "No" "Yes" "No" "No" ...
##  $ Transporting            : chr [1:1104] "No" "public transport" "No" "public transport" ...
##  $ Smoking                 : chr [1:1104] "No" "No" "No" "No" ...
##  $ Coffee_or_Energy_Drinks : chr [1:1104] "No" "No" "No" "No" ...
##  $ Games_and_Hobbies       : chr [1:1104] "No" "Yes" "No" "Yes" ...
##  $ Cosmetics_and_Selfcare  : chr [1:1104] "Yes" "Yes" "No" "Yes" ...
##  $ Monthly_Subscription    : chr [1:1104] "No" "Yes" NA "Yes" ...
##  $ Monthly_expenses        : num [1:1104] 750 960 840 1350 2250 750 1950 600 1350 1230 ...
##  $ 3_or_more_Subscriptions : chr [1:1104] "Yes" "No" "No" "No" ...
##  $ Location                : chr [1:1104] "Madinah" "Khobar" "Madinah" "Jeddah" ...
##  $ Socioeconomic_Background: chr [1:1104] "Medium" "Low" "Medium" "Low" ...
##  $ Major                   : chr [1:1104] "Computer Science" "Computer Science" "Other" "Art" ...
```

```r
# Summary statistics
summary(university_students_data)
```

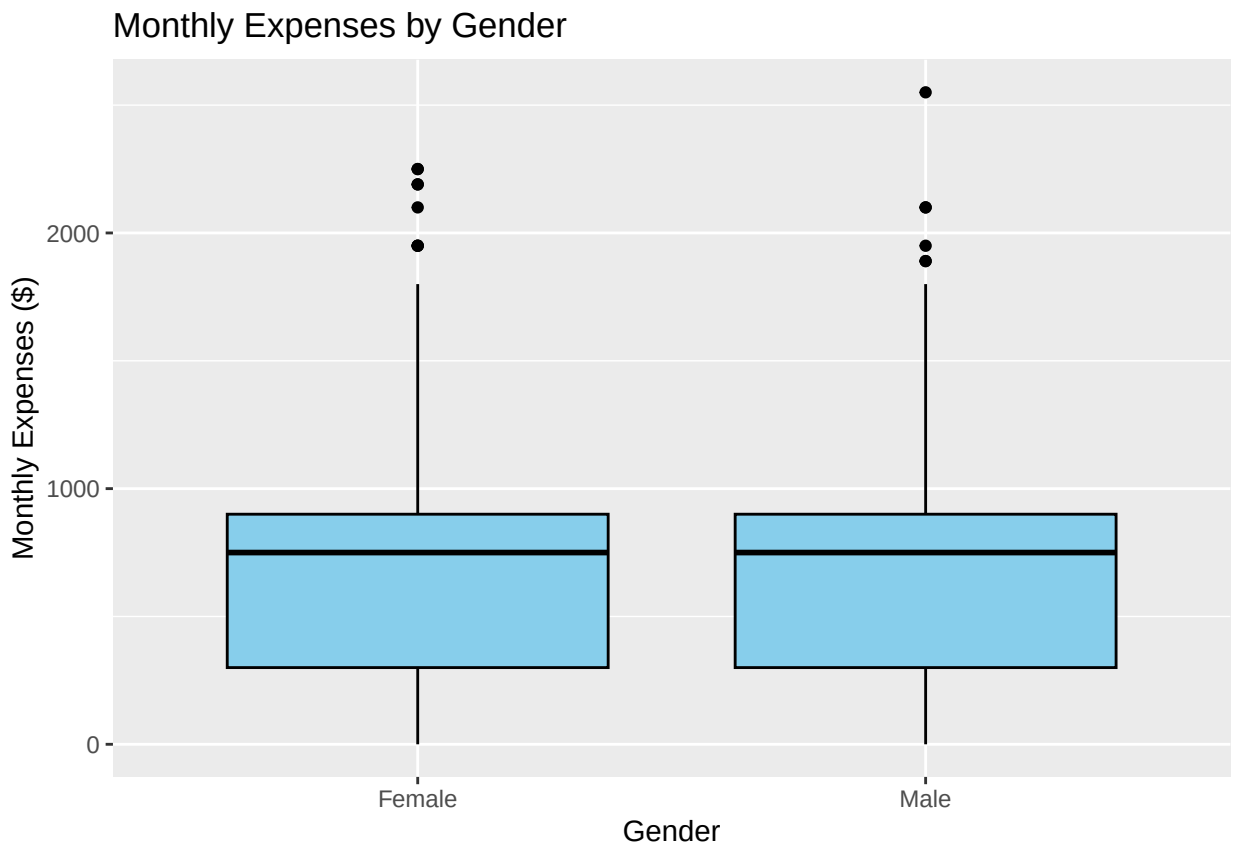
```
##     Gender               Age          Study_year      Living
##  Length:1104        Min.   :17.00   Min.   :1.00   Length:1104
##  Class :character   1st Qu.:19.00   1st Qu.:2.00   Class :character
##  Mode  :character   Median :19.00   Median :3.00   Mode  :character
##                     Mean   :20.28   Mean   :2.75
##                     3rd Qu.:22.00   3rd Qu.:4.00
##                     Max.   :25.00   Max.   :4.00
##                                     NA's   :44
##  Scholarship        Part_time_job      Transporting         Smoking
##  Length:1104        Length:1104        Length:1104        Length:1104
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##
##  Coffee_or_Energy_Drinks Games_and_Hobbies  Cosmetics_and_Selfcare
##  Length:1104             Length:1104        Length:1104
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##
##  Monthly_Subscription Monthly_expenses 3_or_more_Subscriptions
##  Length:1104          Min.   :   0.0   Length:1104
##  Class :character     1st Qu.: 300.0   Class :character
##  Mode  :character     Median : 750.0   Mode  :character
##                       Mean   : 692.9
##                       3rd Qu.: 900.0
##                       Max.   :2550.0
##
##    Location         Socioeconomic_Background    Major
##  Length:1104        Length:1104              Length:1104
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
```

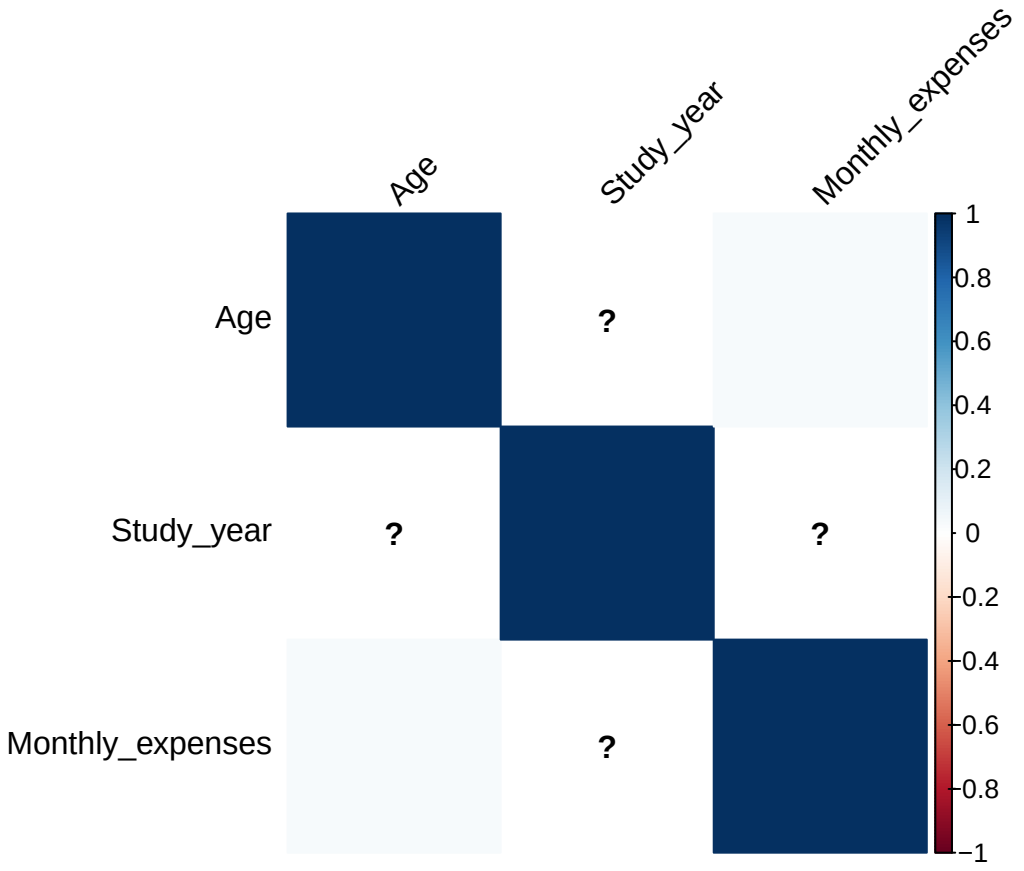
```r
# Histogram for Monthly Expenses
ggplot(university_students_data, aes(x = Monthly_expenses)) +
  geom_histogram(binwidth = 100, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Monthly Expenses",
       x = "Monthly Expenses ($)",
       y = "Frequency")
```



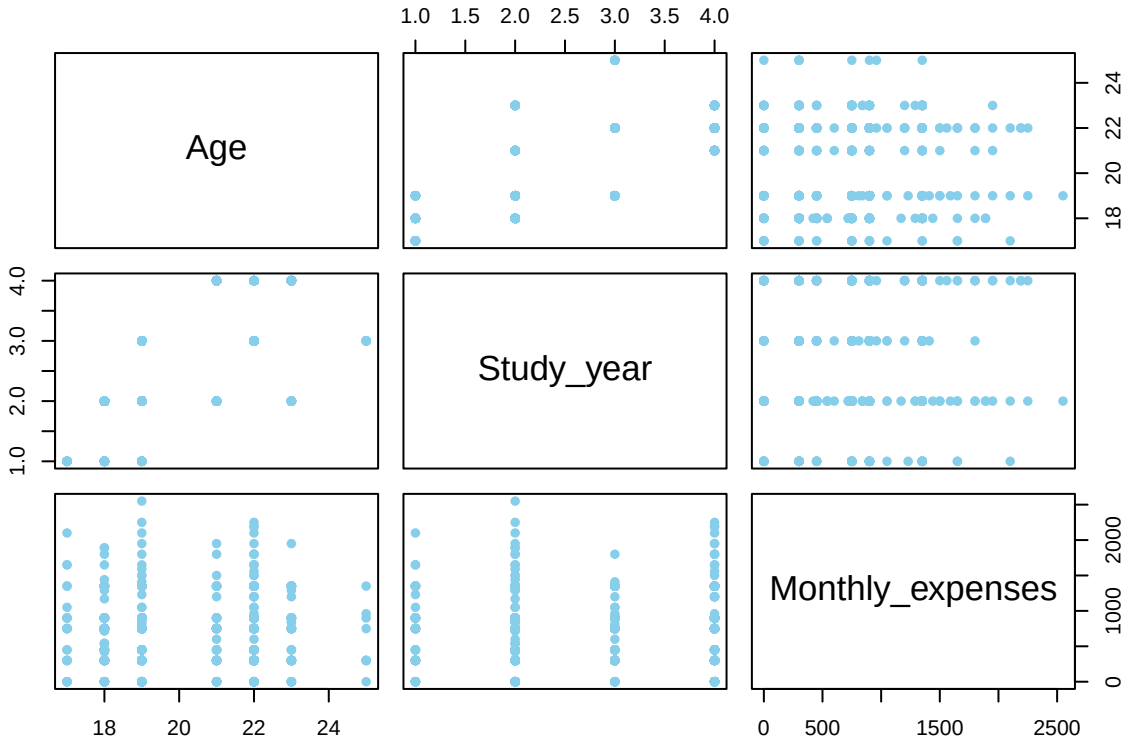

```
# Box plot for Monthly Expenses by Gender
ggplot(university_students_data, aes(x = Gender, y = Monthly_expenses)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Monthly Expenses by Gender",
       x = "Gender",
       y = "Monthly Expenses ($)")
```




```
# Correlation plot
correlation_matrix <- cor(university_students_data[, c("Age", "Study_year", "Monthly_expenses")])
corrplot(correlation_matrix, method = "color", tl.col = "black", tl.srt = 45)
```




```r
# Pair plot for selected variables
selected_vars <- c("Age", "Study_year", "Monthly_expenses")
pairs(university_students_data[selected_vars], pch = 16, col = "skyblue")
```




```
# Insights:
# - Monthly expenses are positively correlated with age and study year.
# - Gender seems to have an impact on monthly expenses, with males generally spending more than females
# - Further analysis is needed to explore relationships with other variables such as part-time job, liv
```

5. Model Selection and Justification Choose appropriate machine learning models for prediction (e.g., linear regression, random forest, etc.). Justify your choice of models based on the nature of the data and the research question. Split the dataset into training and testing sets for model validation.

6. Model Training and Evaluation Train the selected models using the training dataset. Evaluate model performance using appropriate metrics (e.g., RMSE, R-squared) on the testing dataset. Compare the performance of different models if multiple models are used.

7. Interpretation of Results Interpret the coefficients (if applicable) of the predictors in the selected model(s). Discuss the significance of predictors and how they relate to students' spending patterns. Use interpretable machine learning techniques (Partial Dependence Plots, Surrogate Models) to explain the impact of key variables.

8. Discussion and Conclusion Summarize the key findings from the analysis. Discuss the implications of the results in the context of the research question. Address limitations and potential sources of bias in the analysis. Provide insights into how the findings can inform policies, support systems, and business strategies for students in urban Saudi environments.

9. Future Work Suggest potential areas for future research or analysis related to student spending patterns. Discuss any additional data or variables that could enhance the analysis if available.

10. References Include references to relevant literature, datasets, and tools used in the analysis. Remember to include well-commented R code throughout the document to explain each step of the analysis clearly.

This structure will help you organize your R Markdown file systematically and present your findings coherently. Good luck with your analysis!