title: “Assignment 1: EDA and Data Preprocessing” output: pdf_document: toc: true number_sections: true keep_tex: false latex_engine: xelatex includes: in_header: preamble.tex highlight: tango fig_caption: true code_folding: hide echo: false —

```
knitr::opts_chunk$set(warning = FALSE)
```

Install and load required libraries

```
#install.packages(c("tidyverse","corrplot", "ggplot2","dplyr", "tidyr"))

#Load the libraries
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.2     ✓ tibble    3.3.0
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(ggplot2)
library(dplyr)
library(tidyr)
```

Exploratory Data Analysis (EDA)

```
#Load the dataset
data <- read_csv("kenya_student_data.csv")
```

```
## Rows: 5000 Columns: 31
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (16): gender, residency, socioeconomic_status, parental_education, extra...
## dbl (15): student_id, age, family_income, distance_to_university, study_hour...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#display data structure
cat("\n Structure of the data \n")
```

```
##
##  Structure of the data
```

```
str(data)
```

```
## spc_tbl_ [5,000 × 31] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ student_id               : num [1:5000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ age                      : num [1:5000] 18 20 27 28 27 21 18 21 20 22 ...
##  $ gender                   : chr [1:5000] "Male" "Female" "Male" "Female" ...
##  $ residency                : chr [1:5000] "Urban" "Rural" "Urban" "Urban" ...
##  $ socioeconomic_status     : chr [1:5000] "Middle" "High" "Low" "Middle" ...
##  $ parental_education       : chr [1:5000] "Tertiary" "Secondary" "Secondary" "Tertiary" ...
##  $ family_income            : num [1:5000] 15223 19615 12709 28037 NA ...
##  $ distance_to_university   : num [1:5000] 72.9 98.8 54 17.6 57.7 23.5 60.5 81.7 50.6 62 ...
##  $ study_hours_weekly       : num [1:5000] 7 10.3 14.6 4.8 16.9 18.3 14.5 13.6 13.7 11.7 ...
##  $ attendance_rate          : num [1:5000] 0.86 0.718 0.804 0.702 NA ...
##  $ library_usage            : num [1:5000] 1.4 1.8 3.6 3.7 1.1 1.6 3.6 7.5 4.1 6.4 ...
##  $ extracurricular_activities: chr [1:5000] "None" "Both" "Both" "Clubs" ...
##  $ internet_access          : chr [1:5000] "Yes" "Yes" "Yes" "Yes" ...
##  $ device_ownership         : chr [1:5000] "Both" "Both" "Laptop" "Laptop" ...
##  $ previous_grade           : num [1:5000] 52.2 63.4 71.1 66.9 41.2 91.1 61.9 57.7 47.2 58.2
...
##  $ math_score               : num [1:5000] 85.2 50.9 51.9 54.5 71.2 68.1 55.3 55.9 59.9 52.7
...
##  $ science_score            : num [1:5000] 44.8 70.3 58.1 47.5 72.7 57.5 56.4 67.9 56.3 61.3
...
##  $ english_score            : num [1:5000] 35.7 65.5 65.8 39.4 70.4 69.5 78.4 43.3 57.8 40.8
...
##  $ study_group_participation : chr [1:5000] "Yes" "No" "Yes" "Yes" ...
##  $ scholarship_status       : chr [1:5000] "None" "Partial" "Partial" "Partial" ...
##  $ campus_housing           : chr [1:5000] "On-Campus" "Off-Campus" "Off-Campus" "Off-Campu
s" ...
##  $ part_time_job            : chr [1:5000] "Yes" "Yes" "Yes" "No" ...
##  $ commute_time             : num [1:5000] 37.7 34.4 59.2 35.2 35.3 3.5 69 22.8 47.4 24.8
...
##  $ sleep_hours              : num [1:5000] 5.5 8.6 8.5 7.9 6.5 8.5 8.7 5.2 7.9 6.1 ...
##  $ stress_level             : chr [1:5000] "Low" "Moderate" "Moderate" "Moderate" ...
##  $ course_load              : num [1:5000] 16 18.6 10.2 19.1 15.3 15.1 15.6 15.2 13.5 18.7
...
##  $ faculty                  : chr [1:5000] "Sciences" "Business" "Education" "Education" ...
##  $ year_of_study            : chr [1:5000] "2nd" "1st" "4th" "1st" ...
##  $ mobile_money_usage       : num [1:5000] 4710 3179 1186 5742 5252 ...
##  $ health_status            : chr [1:5000] "Good" "Good" "Fair" "Fair" ...
##  $ academic_performance     : chr [1:5000] "Average" "Average" "Good" "Poor" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   student_id = col_double(),
##   ..   age = col_double(),
##   ..   gender = col_character(),
##   ..   residency = col_character(),
##   ..   socioeconomic_status = col_character(),
##   ..   parental_education = col_character(),
##   ..   family_income = col_double(),
##   ..   distance_to_university = col_double(),
##   ..   study_hours_weekly = col_double(),
##   ..   attendance_rate = col_double(),
##   ..   library_usage = col_double(),
```

```
##   ..   extracurricular_activities = col_character(),
##   ..   internet_access = col_character(),
##   ..   device_ownership = col_character(),
##   ..   previous_grade = col_double(),
##   ..   math_score = col_double(),
##   ..   science_score = col_double(),
##   ..   english_score = col_double(),
##   ..   study_group_participation = col_character(),
##   ..   scholarship_status = col_character(),
##   ..   campus_housing = col_character(),
##   ..   part_time_job = col_character(),
##   ..   commute_time = col_double(),
##   ..   sleep_hours = col_double(),
##   ..   stress_level = col_character(),
##   ..   course_load = col_double(),
##   ..   faculty = col_character(),
##   ..   year_of_study = col_character(),
##   ..   mobile_money_usage = col_double(),
##   ..   health_status = col_character(),
##   ..   academic_performance = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# display first few rows
cat("\n First 6 rows of the data \n")
```

```
##
##  First 6 rows of the data
```

```
head(data)
```

```
## # A tibble: 6 × 31
##   student_id   age gender residency socioeconomic_status parental_education
##        <dbl> <dbl> <chr>  <chr>     <chr>                <chr>
## 1          1    18 Male   Urban     Middle               Tertiary
## 2          2    20 Female Rural     High                 Secondary
## 3          3    27 Male   Urban     Low                  Secondary
## 4          4    28 Female Urban     Middle               Tertiary
## 5          5    27 Other  Urban     High                 Primary
## 6          6    21 Male   Urban     Middle               None
## # ℹ 25 more variables: family_income <dbl>, distance_to_university <dbl>,
## #   study_hours_weekly <dbl>, attendance_rate <dbl>, library_usage <dbl>,
## #   extracurricular_activities <chr>, internet_access <chr>,
## #   device_ownership <chr>, previous_grade <dbl>, math_score <dbl>,
## #   science_score <dbl>, english_score <dbl>, study_group_participation <chr>,
## #   scholarship_status <chr>, campus_housing <chr>, part_time_job <chr>,
## #   commute_time <dbl>, sleep_hours <dbl>, stress_level <chr>, …
```

```
#Count how many columns are numeric vs categorical
numeric_count <- sum(sapply (data, class) == "numeric")

categorical_count <- sum(sapply (data, class) == "character")

# Output the counts
cat("📊 Numeric columns:", numeric_count, "\n")
```

```
## 📊 Numeric columns: 15
```

```
cat("🔤 Categorical columns:", categorical_count, "\n")
```

```
## 🔤 Categorical columns: 16
```

Q1.How many numerical and categorical variables are there?

📊 Numeric columns: 15 🔤 Categorical columns: 16

```
#Summary Stats for Numerical Variables
data |> select(where(is.numeric)) |>  summary()
```

```
##    student_id        age        family_income    distance_to_university
##  Min.   :   1   Min.   :17.00   Min.   :-28323   Min.   :  0.0
##  1st Qu.:1251   1st Qu.:20.00   1st Qu.: 15216   1st Qu.: 25.1
##  Median :2500   Median :24.00   Median : 25309   Median : 49.4
##  Mean   :2500   Mean   :23.52   Mean   : 25448   Mean   : 49.7
##  3rd Qu.:3750   3rd Qu.:27.00   3rd Qu.: 35477   3rd Qu.: 74.8
##  Max.   :5000   Max.   :30.00   Max.   :202696   Max.   :100.0
##                                 NA's   :250
##  study_hours_weekly attendance_rate  library_usage    previous_grade
##  Min.   :-2.00      Min.   :0.5000   Min.   :-5.600   Min.   : 40.00
##  1st Qu.:11.50      1st Qu.:0.6212   1st Qu.: 3.000   1st Qu.: 54.70
##  Median :15.10      Median :0.7452   Median : 5.000   Median : 69.90
##  Mean   :15.04      Mean   :0.7482   Mean   : 5.025   Mean   : 69.94
##  3rd Qu.:18.60      3rd Qu.:0.8772   3rd Qu.: 7.100   3rd Qu.: 85.10
##  Max.   :35.70      Max.   :1.0000   Max.   :14.700   Max.   :100.00
##                     NA's   :250
##    math_score    science_score   english_score     commute_time
##  Min.   :  0.7   Min.   :  9.3   Min.   : -4.80   Min.   :-26.50
##  1st Qu.: 49.7   1st Qu.: 50.1   1st Qu.: 50.10   1st Qu.: 19.20
##  Median : 59.6   Median : 60.4   Median : 60.20   Median : 29.80
##  Mean   : 59.8   Mean   : 60.3   Mean   : 60.22   Mean   : 29.82
##  3rd Qu.: 69.9   3rd Qu.: 70.1   3rd Qu.: 70.30   3rd Qu.: 40.40
##  Max.   :111.1   Max.   :129.2   Max.   :110.40   Max.   : 77.20
##  NA's   :150
##   sleep_hours    course_load    mobile_money_usage
##  Min.   : 1.4   Min.   : 3.10   Min.   :-4368
##  1st Qu.: 5.9   1st Qu.:13.10   1st Qu.: 1613
##  Median : 7.0   Median :15.00   Median : 2956
##  Mean   : 7.0   Mean   :15.05   Mean   : 2957
##  3rd Qu.: 8.1   3rd Qu.:17.00   3rd Qu.: 4276
##  Max.   :11.9   Max.   :26.40   Max.   :12917
##
```

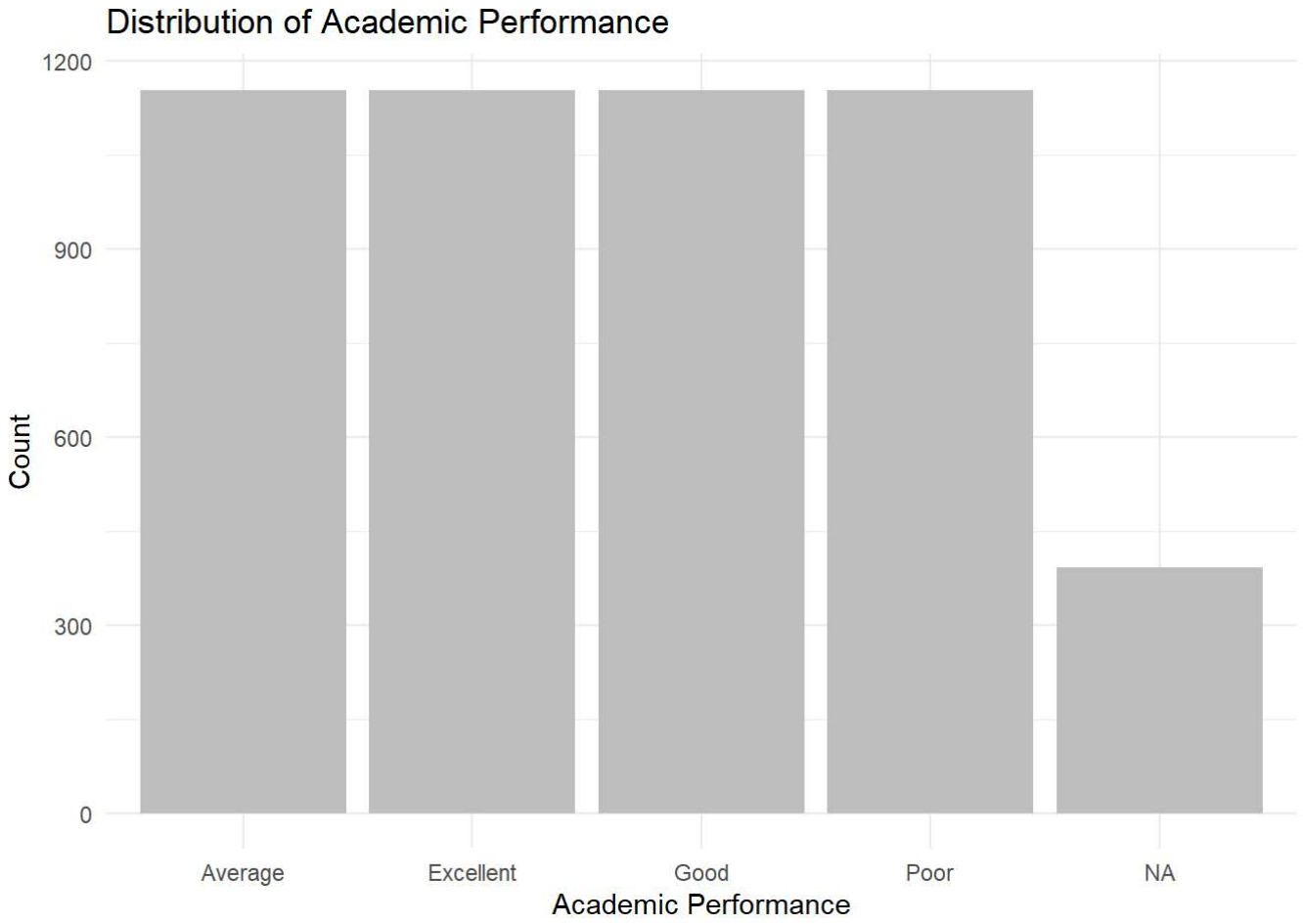
Q2. What insights do these provide about the data?

These summary statistics provide an overview of the dataset's key characteristics.The central tendency; mean and median show where most data points cluster while the minimum, maximum and quartiles show potential anomalies and help identify outliers.

```
# View class distribution - class counts
table(data$academic_performance)
```

```
##
##   Average Excellent      Good      Poor
##      1152      1152      1152      1153
```

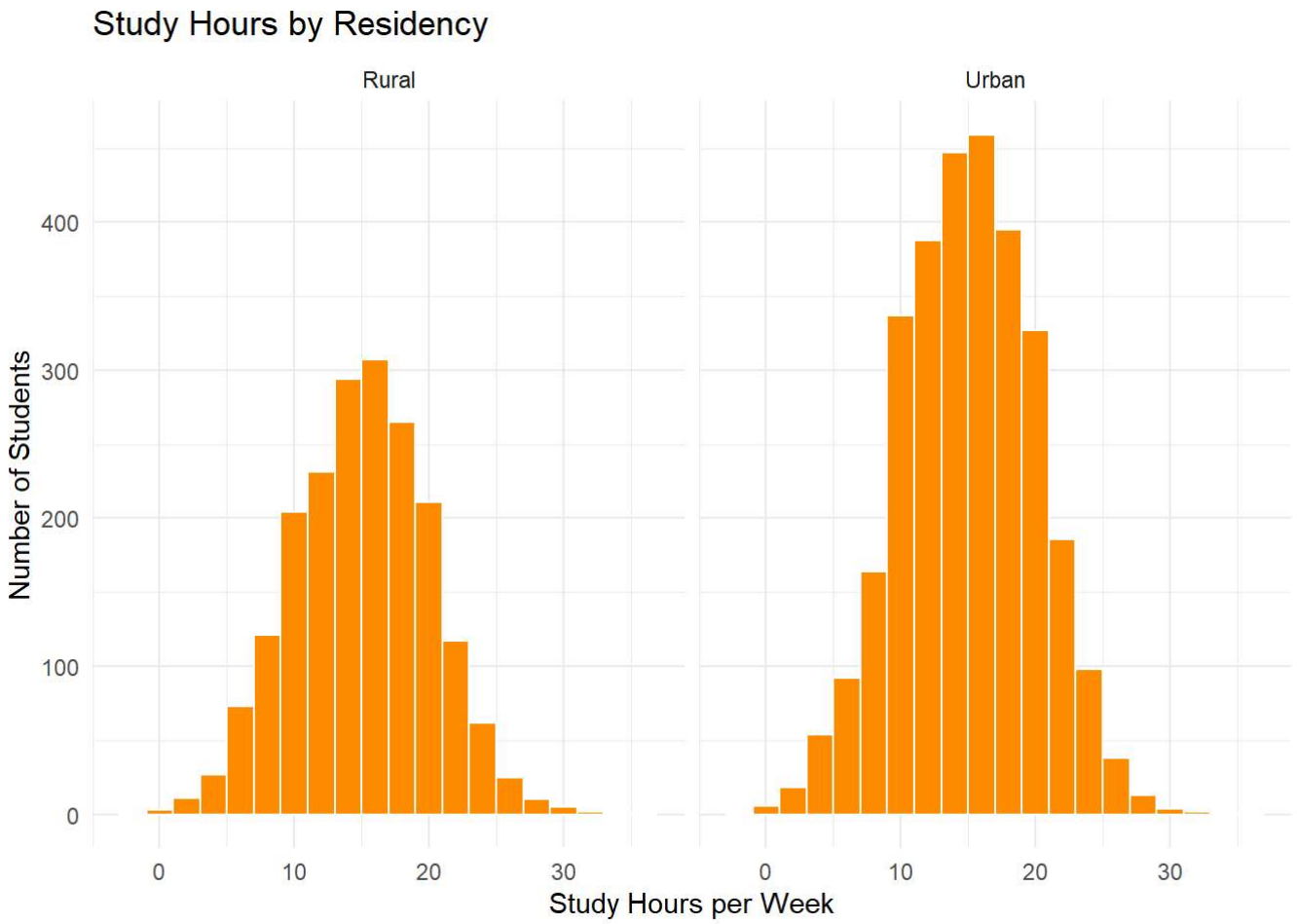
```
#Plot bargraph
data |> ggplot(aes(academic_performance)) +
        geom_bar(fill = "grey") +
        labs(title = "Distribution of Academic Performance",
            x = "Academic Performance",
            y = "Count") +
         theme_minimal()
```



Q3. Is the target variable balanced across its classes?

Yes, the bar heights are roughly equal suggesting that the variables are balanced across its classes.

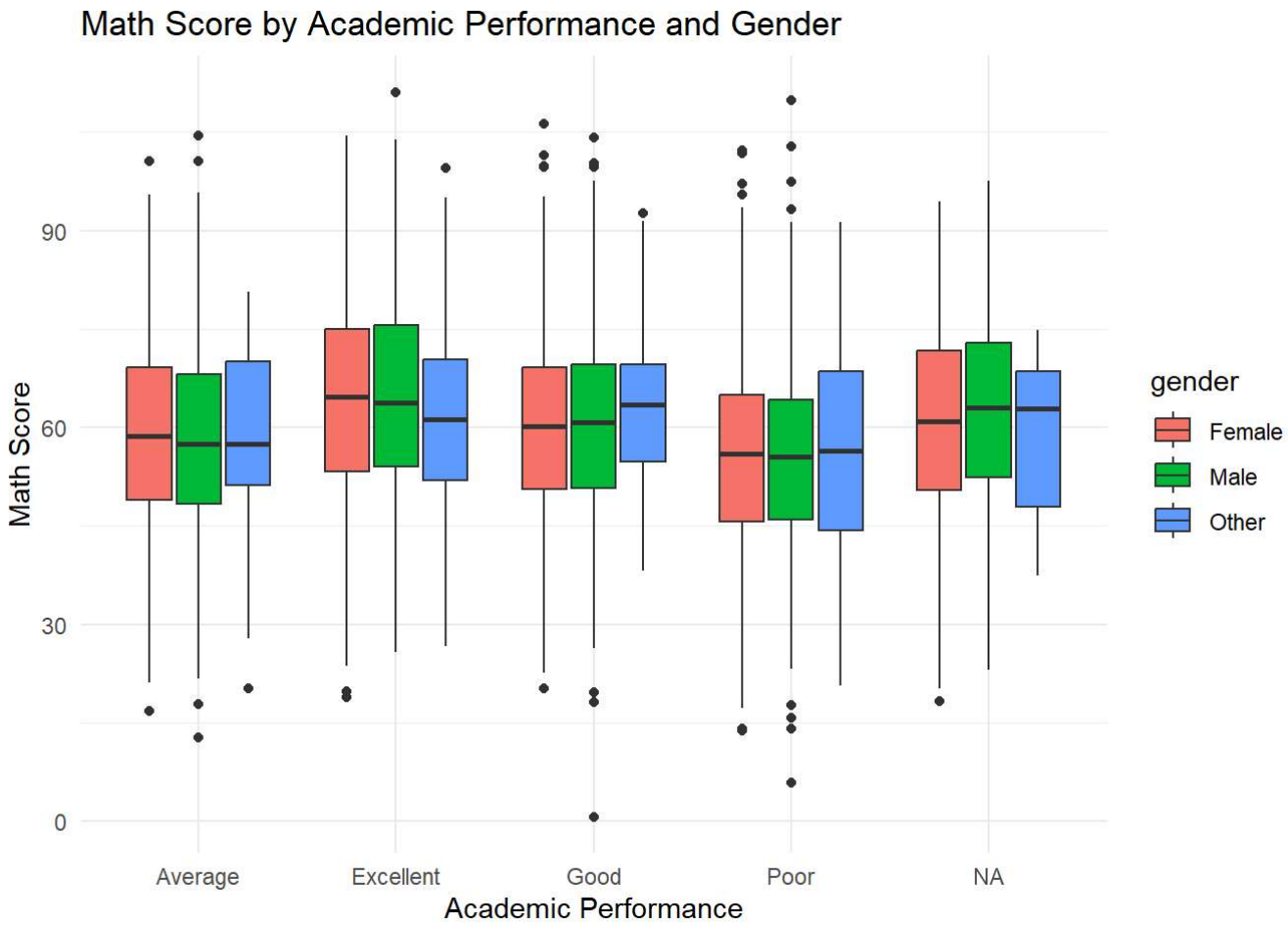
```
#Histogram of study hours by residency
data |> ggplot(aes(study_hours_weekly)) +
        geom_histogram (binwidth = 2, fill = "darkorange", color = "white") +
        facet_wrap(~residency) +
        labs(title = "Study Hours by Residency",
             x = "Study Hours per Week",
            y = "Number of Students") +
        theme_minimal()
```



Q4. How does it vary between urban and rural students?

Both urban and rural histograms look fairly symmetric, there is no skewness. Meaning that most students, regardless of residency, cluster around similar study hours despite differences in numbers. In the Kenyan context, this could suggest similarity in academic effort, access to learning materials or national standardization of study expectations. The greater number of urban students could simply reflect urban population density or sampling design.

```
#Boxplots of math_score by performance and gender
data |> ggplot(aes(x = academic_performance, y = math_score)) +
        geom_boxplot(aes(fill = gender)) +
        labs (title= "Math Score by Academic Performance and Gender",
                x = "Academic Performance",
                y = "Math Score") +
        theme_minimal()
```



Q5. Boxplots of math_score by performance and gender. What patterns do you observe?

The Median math score across performance categories seem to increase from “Poor” to “Excellent” showing males and females who are excellent in academic performance have high median Math scores. Male and female median Math scores in all academic performance categories seem to be the same. There are outlier math scores in all categories of the academic performance. ➞ Academic performance gradient: A clear increase in math scores from Poor → Excellent validates the reliability of academic_performance categorization. ➞ Variability: Large spread among Average or Poor students may reflect inequalities — some students excelling despite low overall category due to factors like school resources or individual effort.

```
# Proportions for extracurricular_activities
data |> count(extracurricular_activities) |>
        mutate(prop = n/ sum(n) *100)
```

```
## # A tibble: 4 × 3
##   extracurricular_activities     n  prop
##   <chr>                      <int> <dbl>
## 1 Both                        1254  25.1
## 2 Clubs                       1214  24.3
## 3 None                        1289  25.8
## 4 Sports                      1243  24.9
```

```
# Proportions for faculty
data |> count(faculty) |>
        mutate(prop = n/ sum(n) *100)
```

```
## # A tibble: 5 × 3
##   faculty         n  prop
##   <chr>       <int> <dbl>
## 1 Arts         1025  20.5
## 2 Business      967  19.3
## 3 Education    1030  20.6
## 4 Engineering  1004  20.1
## 5 Sciences      974  19.5
```

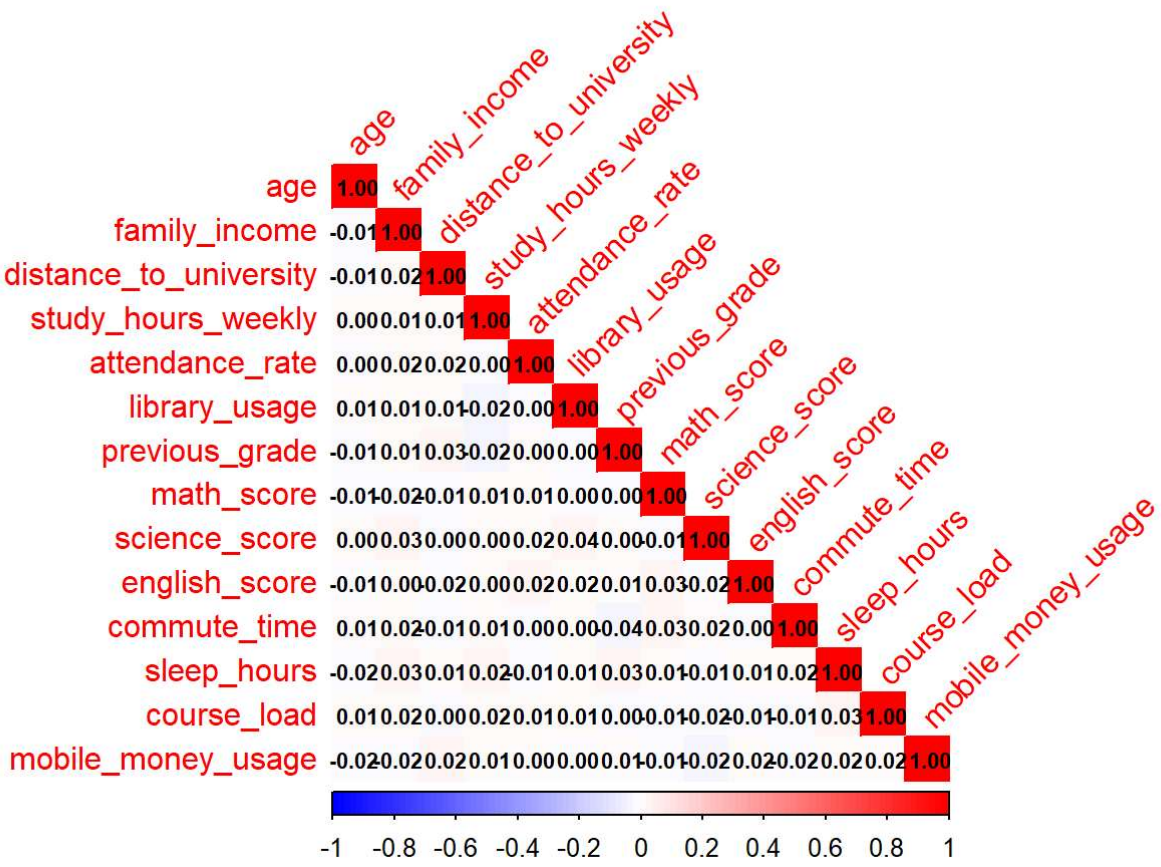
Q6. Proportions in extracurricular_activities and faculty. Which categories are most common?

The most common category in extracurricular activities is ‘None’ and the most common faculty is ‘Education’.

```
# Select the numeric variables, excluding 'student_id'
num_data <- data |>
    select(where(is.numeric), -student_id)

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

# Visualize correlation matrix using a heatmap
corrplot(corr_matrix, method = "color", type = "lower",
         tl.col = "red", tl.srt = 45, addCoef.col = "black",
         number.cex = 0.7,  col = colorRampPalette(c("blue", "white", "red"))(200))
```



Q7. Correlation Matrix (exclude student_id). Which pairs have the strongest correlations?

Age and family_income, distance_to_university, previous_grade, or math_score all have a strong negative correlation (-0.01). Mobile_money_usage and math_score also have a strong negative correlation of -0.01 while Science_score and library_usage is the closest to 1 with a positive correlation of 0.04.

```
# Create a contingency table
cont_table <- table(data$internet_access, data$academic_performance)

# Run chi-squared test
chisq.test(cont_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 163.55, df = 3, p-value < 2.2e-16
```

Q8. Chi-squared Test (internet_access vs academic_performance). Is internet_access associated with academic_performance?

The p-value 2.2e-16, is less than 0.05 indicating there is statistically significant association between internet_access and academic_performance

Data Preprocessing: Missing Values

```
# Calculate number and percentage of missing values per column
missing_summary <- sapply(data, function(x) sum(is.na(x)))
missing_percent <- round(missing_summary / nrow(data) *100, 2)

# Combine and display
missing_report <- data.frame( Column = names(missing_summary),
                              Missing_Count = missing_summary,
                              Missing_Percent = missing_percent)

# Show only columns with missing data
missing_report[missing_report$Missing_Count > 0, ]
```

```
##                                  Column Missing_Count Missing_Percent
## family_income                family_income           250            5.00
## attendance_rate           attendance_rate           250            5.00
## math_score                     math_score           150            3.00
## academic_performance academic_performance           391            7.82
```

Q9. Identify Missing Values. Why might these variables have missing data in a Kenyan context?

Family Income (5%) -This is sensitive financial information and students may not know or may avoid sharing. → Often, discussing income in Kenya is culturally sensitive, especially in lower-income or rural households.

Attendance Rate (5%) -This could be due to poor record-keeping or informal school systems. → Many schools (especially in rural areas) lack digitized records; paper-based systems often have gaps. → This could also reflect inconsistent attendance tracking.

Math Score (3%) -Could be ungraded students, recent transfers, or exam absences. → Students who missed exams due to illness, unpaid fees, or transfer cases.

Academic Performance (7.82%) -May be missing if performance is yet to be evaluated or withheld. → In Kenya, this could be missing if students are newly admitted, assessments are pending, or results are withheld due to fees.

```
# Impute family_income with median
median_income <- median(data$family_income, na.rm = TRUE)
data$family_income[is.na(data$family_income)] <- median_income

# Impute math_score with median
median_math <- median(data$math_score, na.rm = TRUE)
data$math_score[is.na(data$math_score)] <- median_math
```

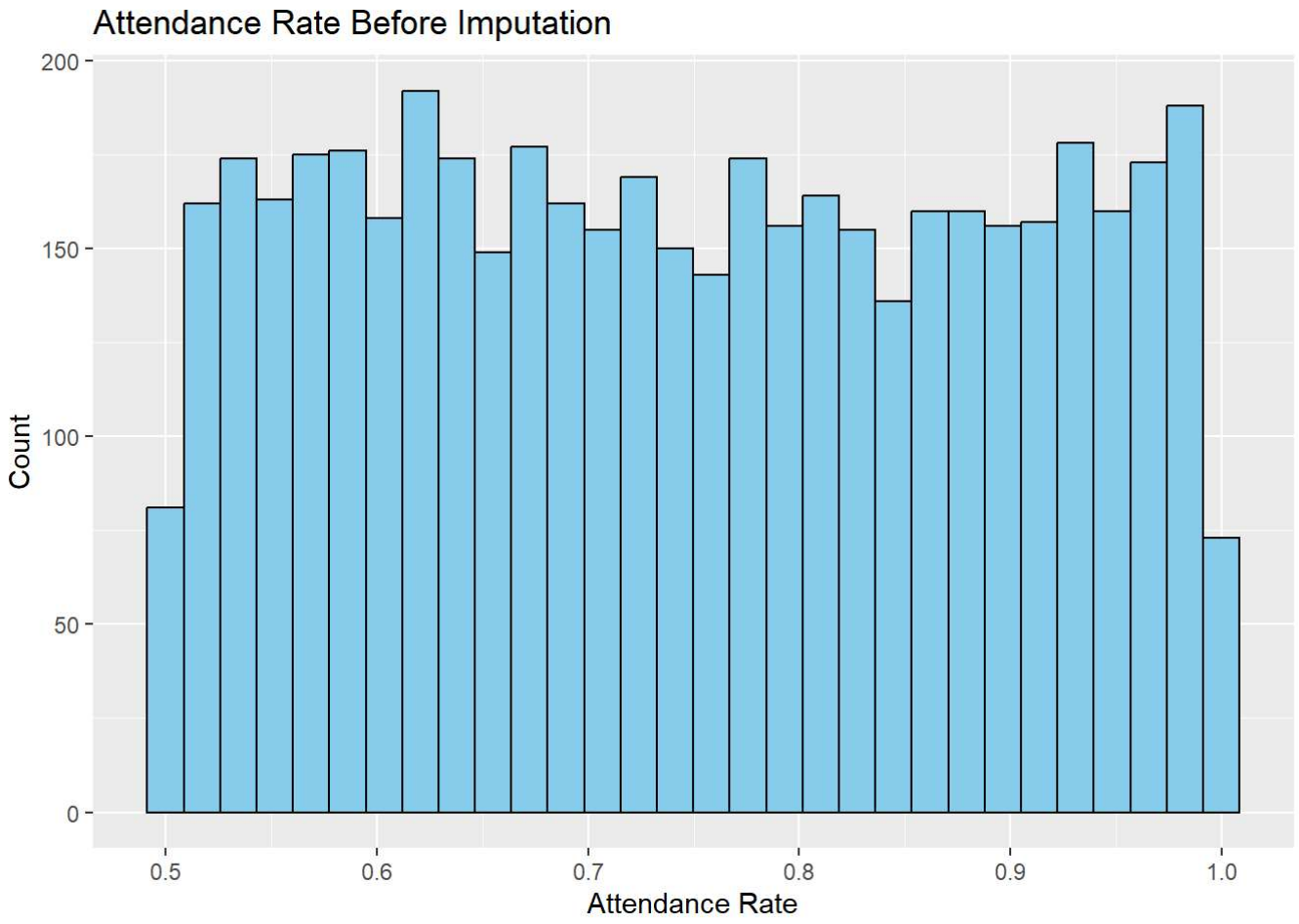
Q10: Impute family_income and math_score with median. Why Use the Median?

✅ WMedian is robust to outliers: Income and exam scores can have extreme values (e.g., very high earners or top scorers). The median is not affected by these outliers, making it more robust than the mean.

✅ Skewed distributions: Both family_income and math_score are often right-skewed (especially income). The median better represents the center of such distributions.

```
# Histogram before imputation
data |> ggplot(aes(attendance_rate)) +
      geom_histogram(fill = "skyblue", color = "black", na.rm = TRUE) +
      labs(title = "Attendance Rate Before Imputation", x = "Attendance Rate", y = "Count")
```

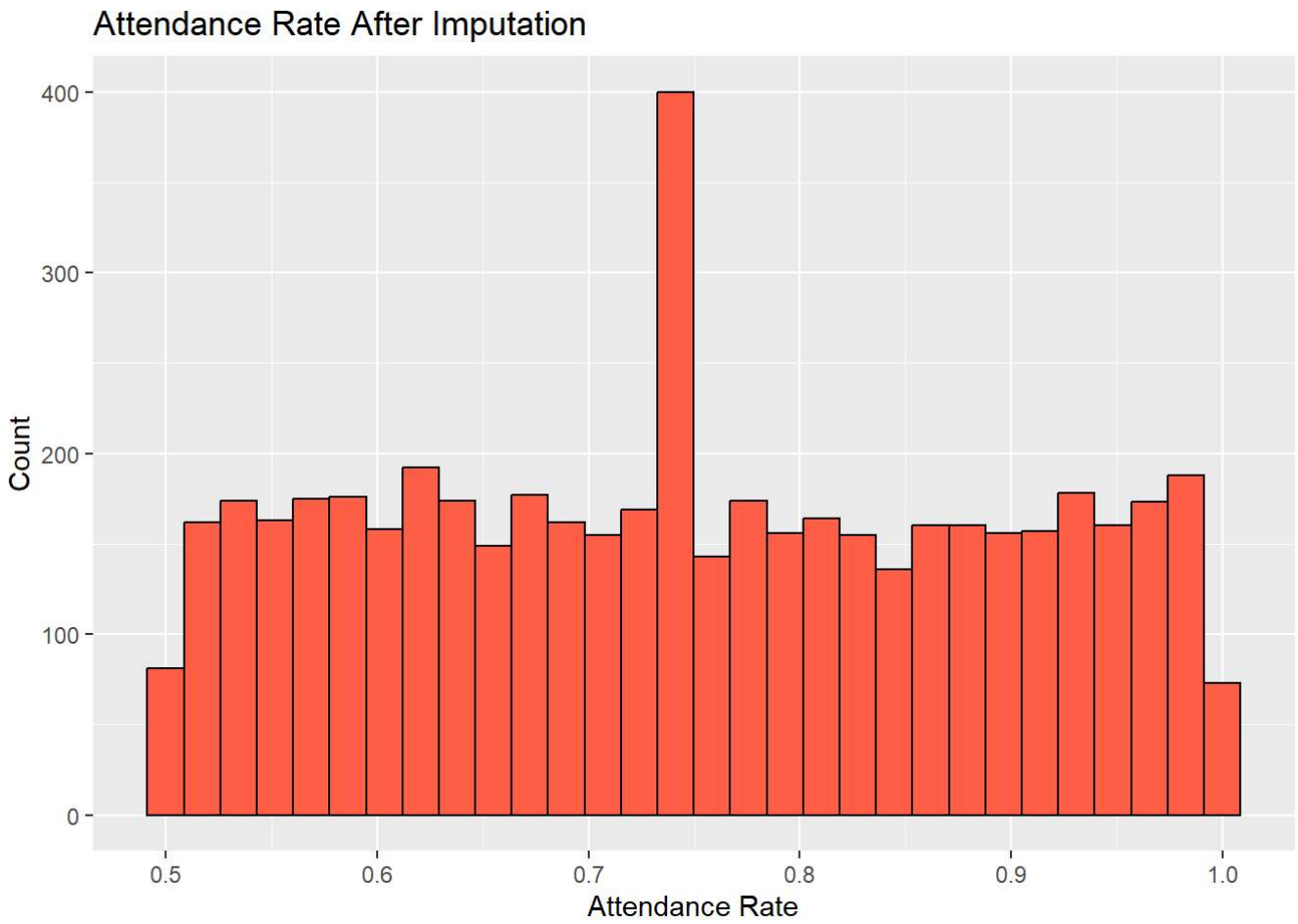
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Impute missing values with the mean
data$attendance_rate[is.na(data$attendance_rate)] <- mean(data$attendance_rate, na.rm = TRUE)

# Histogram after imputation
data |> ggplot(aes(attendance_rate)) +
        geom_histogram(fill = "tomato", color = "black") +
        labs(title = "Attendance Rate After Imputation", x = "Attendance Rate", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Q11. Impute attendance_rate with mean. Compare the distributions before and after imputation using histograms.

📝 Attendance_rate is a continuous variable and its distribution is typically symmetric. The mean imputation retains the central tendency and is acceptable when the missingness is random and small in proportion.

Statistical Impact Visual inspection shows no difference after imputation since the effect on the distribution is minor. The missingness is small (~5%).

Data Preprocessing: Outliers

```
# Compute Q1, Q3, and IQR
Q1 <- quantile(data$family_income, 0.25, na.rm = TRUE)
Q3 <- quantile(data$family_income, 0.75, na.rm = TRUE)
IQR<- Q3 - Q1

# Define outlier thresholds
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Detect outliers
outliers <- length(data$family_income[data$family_income < lower_bound | data$family_income > up
per_bound])

# Print results
cat("Number of outliers in family income:", outliers, "\n")
```

```
## Number of outliers in family income: 79
```

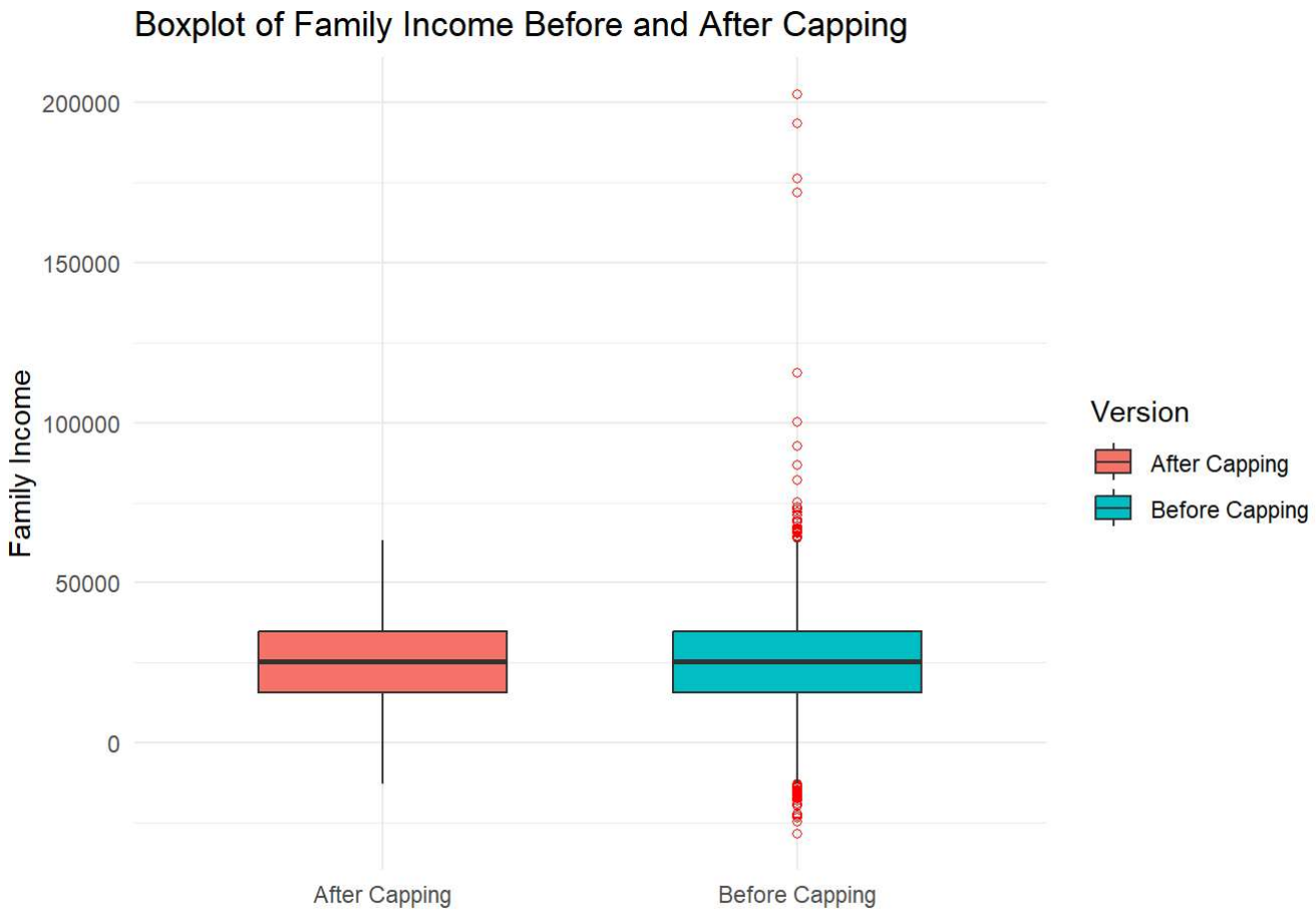
Q12. Detect Outliers (IQR method). What might these outliers represent in a Kenyan context?

🧠 Outliers in family_income may represent: 1. High-income families in affluent areas (e.g., Nairobi suburbs, upper-class). 2. Low outliers may reflect students from marginalized or rural communities. 3. Could indicate data entry errors (e.g., extra zero added or wrong currency). 4. Socioeconomic inequality in Kenya, so such income variation is realistic and informative.

```
# Cap values
data$family_income_capped <- data$family_income
data$family_income_capped[data$family_income > upper_bound] <- upper_bound
data$family_income_capped[data$family_income < lower_bound] <- lower_bound

# Prepare data for plotting
income_df <- data |>
  select(family_income, family_income_capped) |>
  pivot_longer(cols = everything(), names_to = "Version", values_to = "Income") |>
  mutate(Version = recode(Version,
                          "family_income" = "Before Capping",
                          "family_income_capped" = "After Capping"))

# Plot boxplots
ggplot(income_df, aes(x = Version, y = Income, fill = Version)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 1, width = 0.6) +
  labs(title = "Boxplot of Family Income Before and After Capping",
       x = "", y = "Family Income") +
  theme_minimal()
```



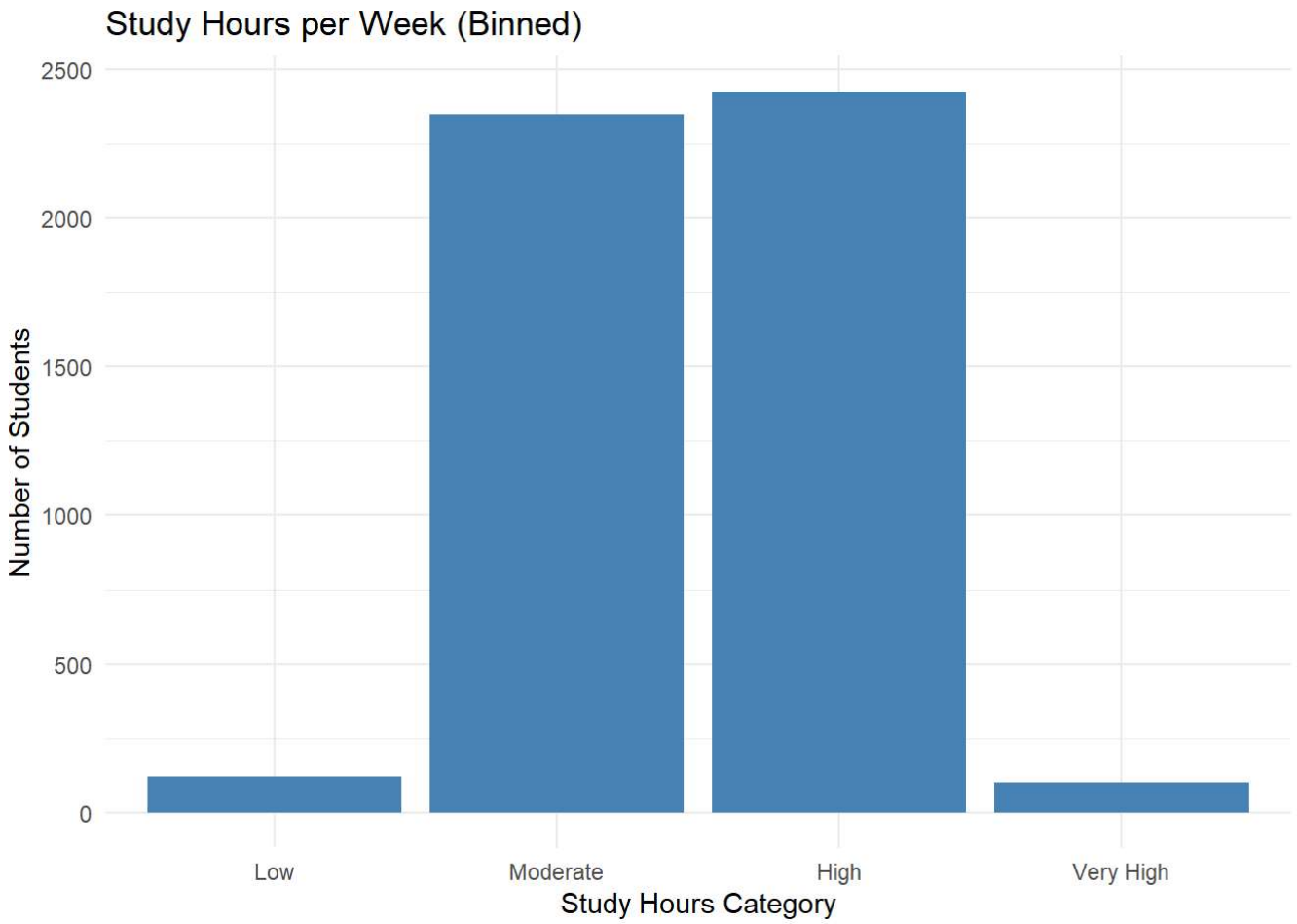
Q13. Cap family_income Outliers and Visualize the distribution before and after capping using boxplots.

🧠 Interpretation Before Capping there are extreme values (outliers) far from the box. After Capping, outliers are pulled within the upper/lower bounds hence reducing skewness and extreme influence. This is useful for robust statistical modeling.

Data Preprocessing: Feature Engineering

```
# use the cut() function to categorize the continuous variable; discretize into 4 bins ***
data$study_hours_binned <- cut(data$study_hours_weekly,
                     breaks = c(-Inf, 5, 15, 25, Inf), # Define fixed breakpoints
                     labels = c("Low", "Moderate", "High", "Very High"), right = TRUE)

# Bar plot of the binned variable
data |> ggplot(aes(study_hours_binned)) +
        geom_bar(fill = "steelblue") +
        #facet_wrap(~academic_performance) +
        labs(title = "Study Hours per Week (Binned)",
             x = "Study Hours Category",
            y = "Number of Students") +
       theme_minimal()
```



Q14.Discretize study_hours_weekly into bins and create a barplot

The barplot shows the distribution of study hours weekly after discretizing. Majority of the study hours are high, i.e, 15 to 24 hours.

```
# Discretize into quartiles with meaningful labels
data <- data |>
  mutate(family_income_binned = cut(family_income,
    breaks = quantile(family_income, probs = c(0, 0.25, 0.5, 0.75, 1), na.rm = TRUE),
    include.lowest = TRUE,
    labels = c("Low", "Medium-Low", "Medium-High", "High")))

# correlation with academic_performance using contingency table (counts)
cont_table <- table(data$family_income_binned, data$academic_performance)
cont_table
```

```
##
##                Average Excellent Good Poor
##   Low              300       281  279  298
##   Medium-Low       308       311  325  319
##   Medium-High      247       249  264  269
##   High             297       311  284  267
```

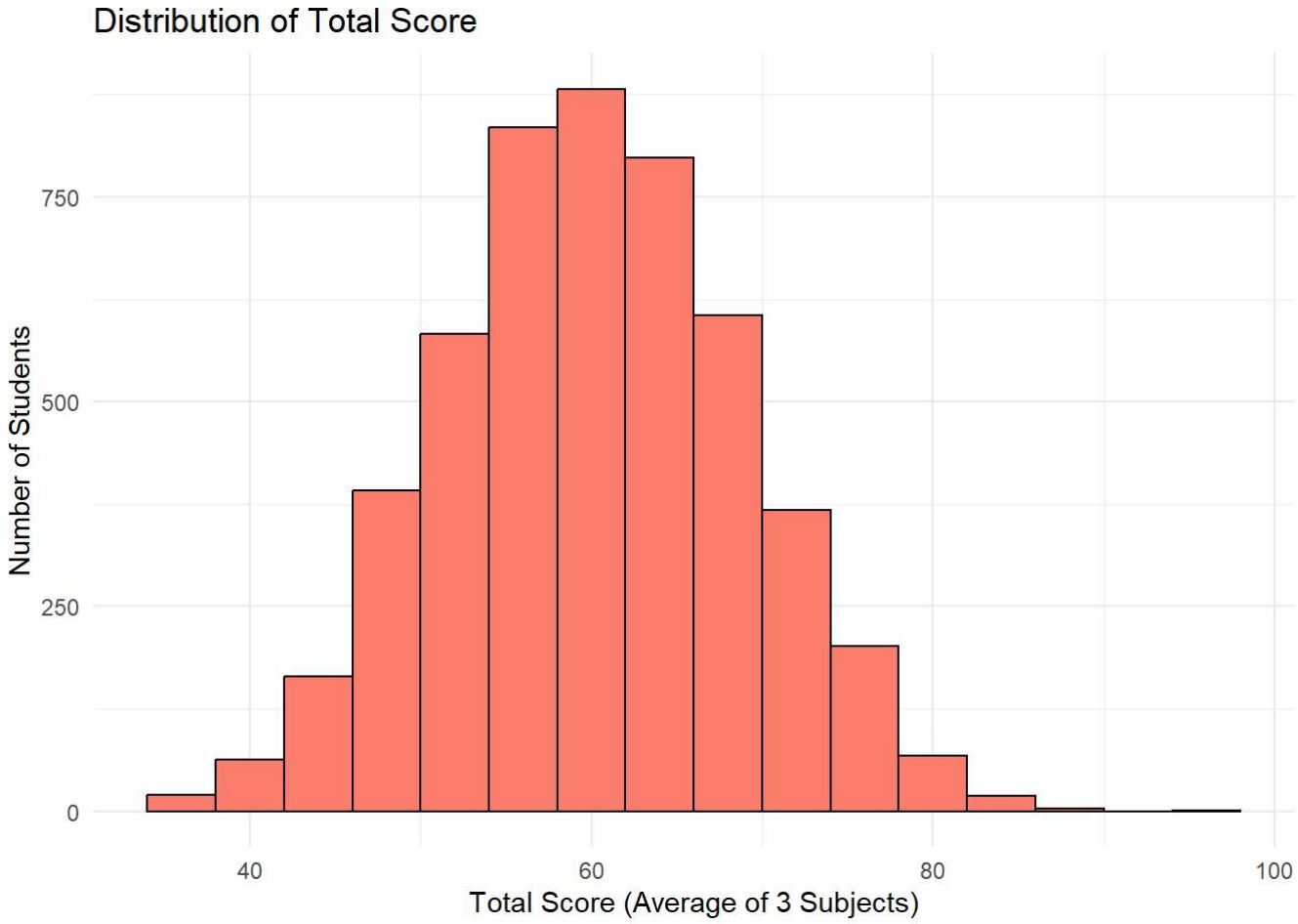
Q15. Discretize family_income into quartiles. How does the binned variable correlate with academic_performance?

🔍 Interpretation: High income quartile has a greater proportion of students with Excellent academic performance while the low income quartile has a great proportion of students with Average academic performance.

🔍 Possible Explanations in the Kenyan context: Some students from lower-income households may receive scholarships or support that helps them perform equally well. This can be confirmed with the scholarship_status variable.

```
# Create total_score by averaging the three subject scores
data$total_score <- rowMeans(data[, c("math_score", "science_score", "english_score")], na.rm =
TRUE)

# Plot the distribution of total_score
data |> ggplot(aes(x = total_score)) +
        geom_histogram(binwidth = 4, fill = "salmon", color = "black") +
      labs(title = "Distribution of Total Score",
       x = "Total Score (Average of 3 Subjects)",
       y = "Number of Students") +
      theme_minimal()
```



Q16. Create average score and visualize its distribution

🔍 The distribution of total_score is symmetric meaning that the mean and median are close to each other and there aren't extreme values pulling the data to one side.

Data Preprocessing: Relationships

```
#Contingency table for extracurricular_activities vs. academic_performance
activity_perf <- table(data$extracurricular_activities, data$academic_performance)
activity_perf
```

```
##
##          Average Excellent Good Poor
##   Both       302       285  282  297
##   Clubs      280       271  282  277
##   None       288       290  311  306
##   Sports     282       306  277  273
```

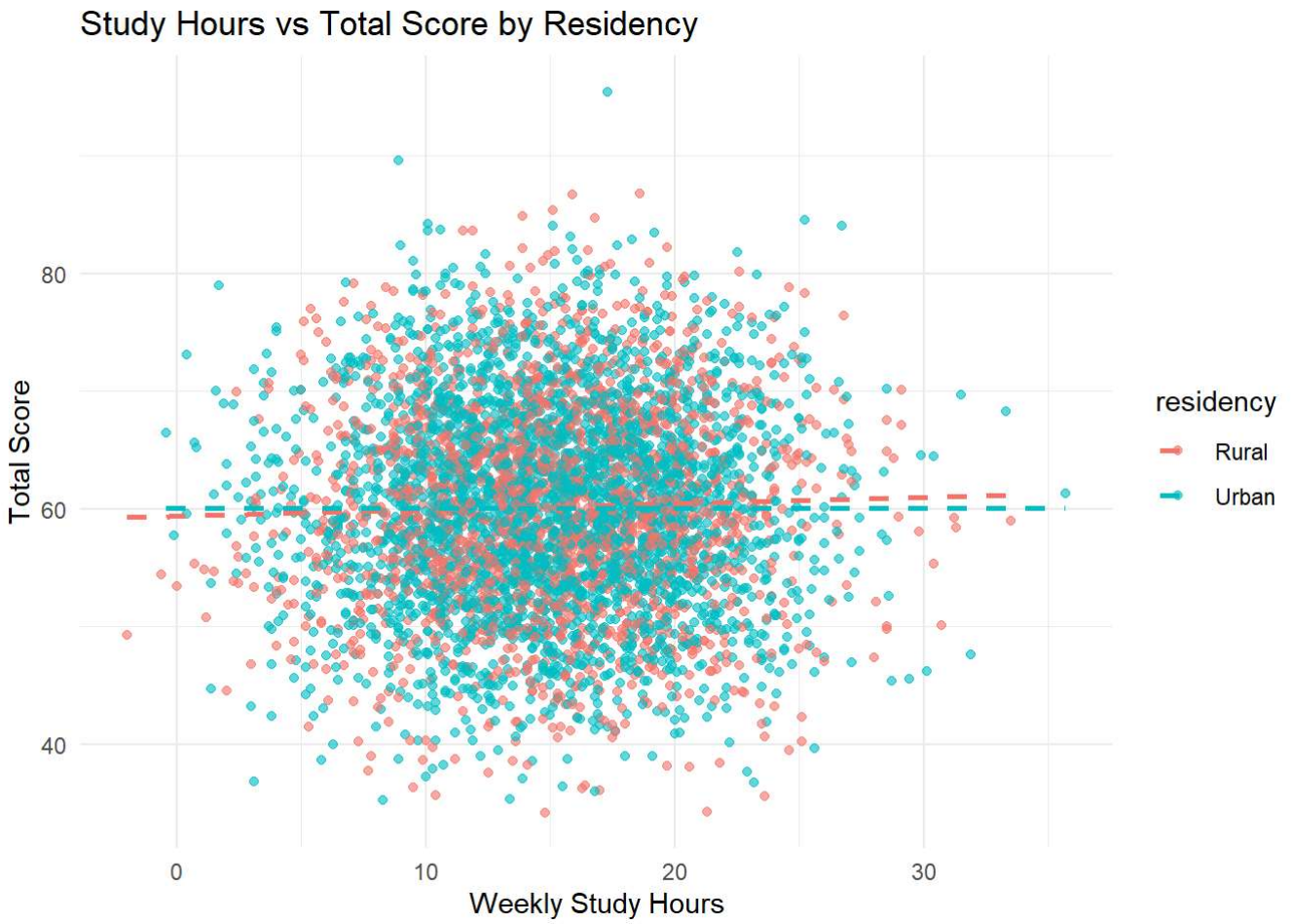
Q17. Contingency table for extracurricular_activities vs. academic_performance. What patterns suggest about student involvement?

🧠 Observations Sports group has the highest number of students with Excellent performance. None group, students in no extracurricular activity, has slightly more students with Poor and Good performance compared to other categories. Both sports and clubs has higher counts in Average followed by “Poor”.

This potentially suggests participation in extracurriculars does not have a strong or significant influence on academic performance in this dataset. There might be other confounding factors (e.g., socio-economic status, library usage, device ownership) that play a larger role.

```
#scatter plot between study_hours_weekly and total_score colored by residency
data |> ggplot(aes(study_hours_weekly, total_score, color = residency)) +
        geom_point(alpha = 0.6) +
        geom_smooth(method = "lm", se = FALSE, linetype = "dashed") +
        labs(title = "Study Hours vs Total Score by Residency",
              x = "Weekly Study Hours",
              y = "Total Score") +
        theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Q18. Visualize the relationship between study_hours_weekly and total_score colored by residency. What trends do you observe?

The regression line (from lm(total_score ~ study_hours_weekly)) cuts through the middle because it's a single global trend across all students, regardless of residency.

Urban and Rural points appear similarly spread – this suggests that the relationship between study_hours_weekly and total_score is not strongly different between Urban and Rural students or there might be very little overall correlation between study hours and total score, regardless of residency.

```
#Saving preprocessed dataset
write_csv(data, "kenya_student_data_preprocessed.csv")
```