MAY SEMESTER 2025
MRDC 911: Data Science & Computational Intelligence
INSTRUCTOR: JAPHETH MURSI
DATE: 12th June 2025
Esther Njoroge
Adm No. 25ZA111316

Assignment 1- EDA and Data Preprocessing on Kenyan Student Dataset

Questions And Answers.

Exploratory Data Analysis (EDA)

1. Load the dataset and display its structure (e.g., column names, data types, first few rows). How many numerical and categorical variables are there?

   Rows: 5000
   Columns: 31

2. Compute summary statistics (mean, median, min, max, etc.) for all numerical variables (e.g., family income, study_hours_weekly). What insights do these provide about the data?

   Academic Performance: The dataset includes students with varying academic performances, as indicated by the diverse scores across subjects.

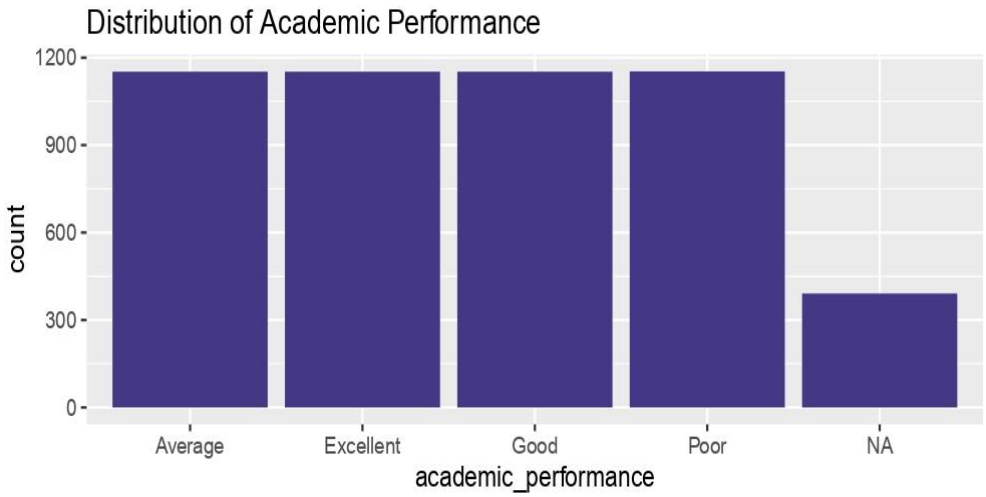
   Study Habits: Most students engage in regular study hours and maintain a moderate attendance rate, suggesting a committed student body.

   Digital Access: Most students have access to the internet and own digital devices, facilitating online learning and communication.

   Lifestyle Factors: Students exhibit a range of lifestyle choices, including varying commute times, sleep hours, and stress levels, which may influence academic performance.

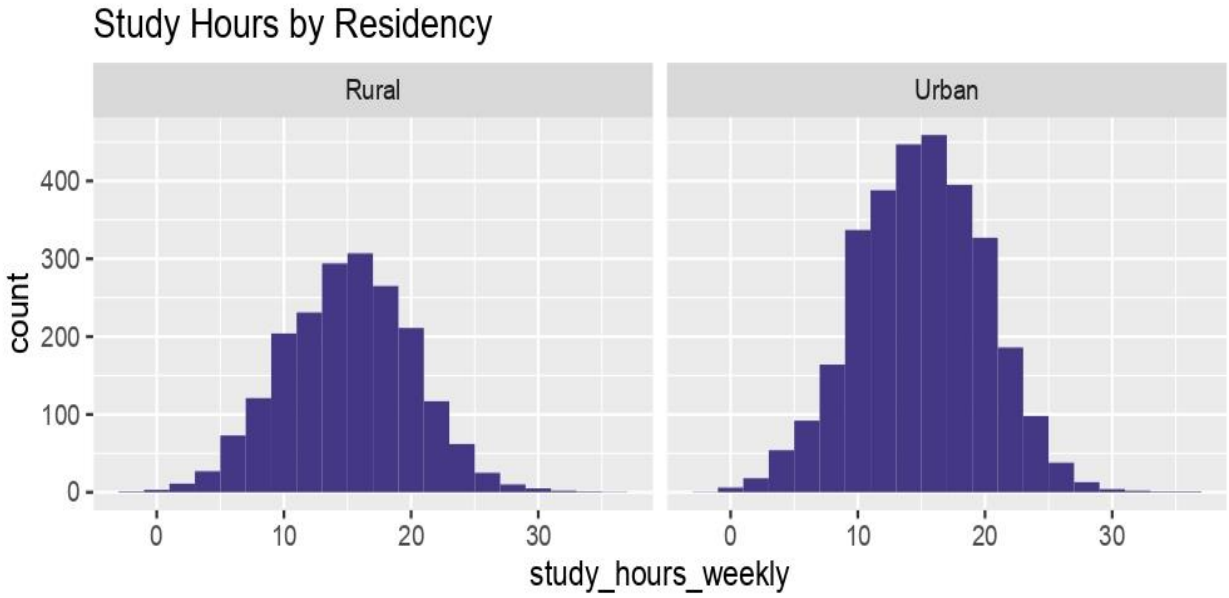
   Data Quality: Some variables contain negative values, which may indicate missing or erroneous data entries.

3. Create a bar plot to visualize the distribution of academic performance. Is the target variable balanced across its classes (Poor, Average, Good, Excellent)?



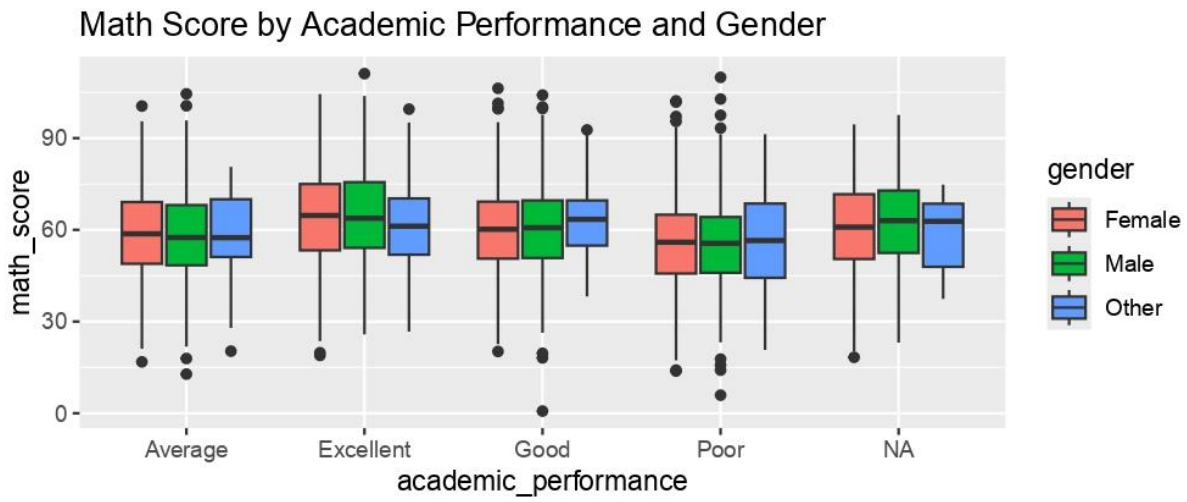
Observation: The variable of academic performance is balanced across all classes.

4. Visualize the distribution of study_hours_weekly using a histogram. How does it vary between urban and rural students (use a faceted histogram)?



Observation: The urban students have more study hours compared to the rural students.

5. Create boxplots of math score by academic performance and gender. What patterns do you observe?

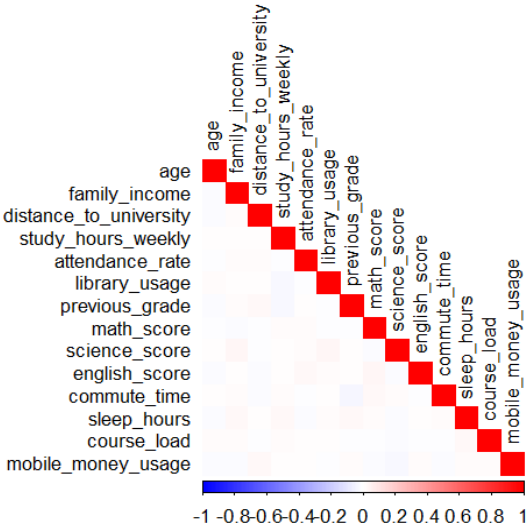


Observation: Academic performance strongly correlates with math scores. Gender differences exist but are modest, with males slightly outperforming females in math.

6. Compute the proportion of each category in extracurricular activities and faculty. Which categories are most common?

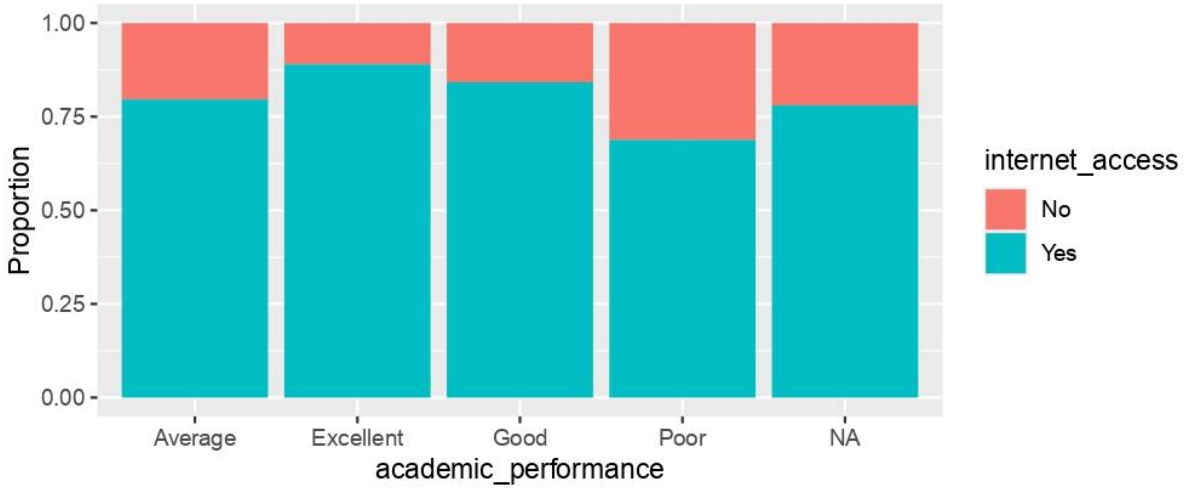
The most common categories are Education, Arts and Engineering.

7. Create a correlation matrix for numerical variables (excluding student_id) and visualize it using a heatmap. Which pairs have the strongest correlations?



Observation: The categories with the most correlation are: age, family income, distance to university, study hours, attendance rate and library usage.

8. Use a statistical test (e.g., chi-squared) to check if internet access is associated with academic performance. Interpret the results.



Observation: Those with internet access appear to be performing slightly better compared to those without internet access.

Data Preprocessing: Missing Values

9. Identify columns with missing values and report their percentages. Why might these variables have missing data in a Kenyan context?

Missing Variables in family income may be because of families not wanting to disclose how much they earn. Also in most Kenyan homes, the income is irregular especially for families who live in the rural areas and rely on farming or casual jobs.

Attendance rate: you may have students missing classes, also most of the attendance registers are manual, which is prone to errors or even getting misplaced. Teachers, strikes, especially in public schools can affect the attendance rate.

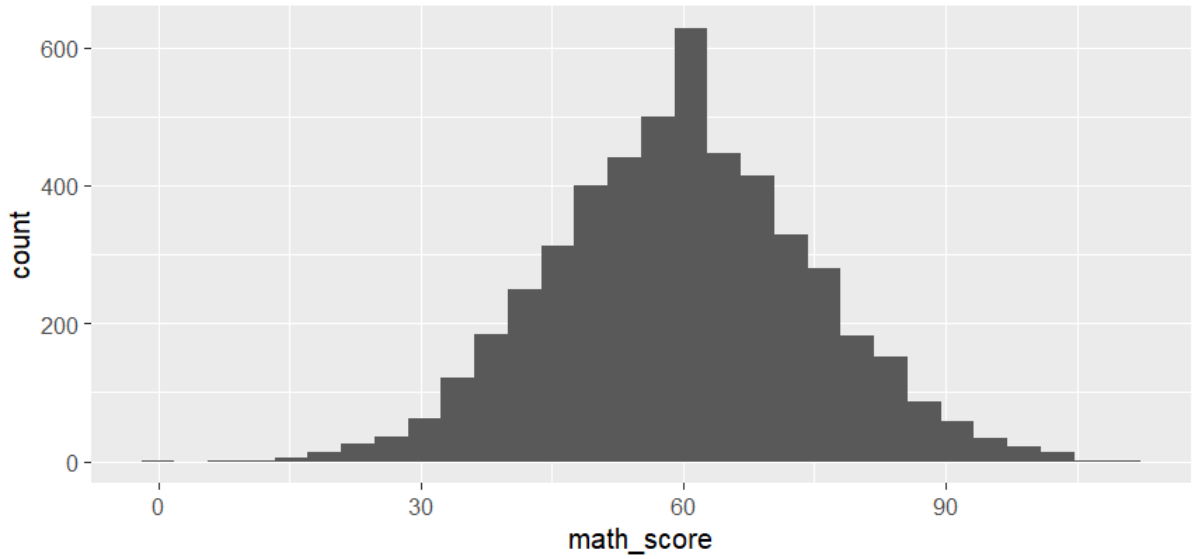
Math Score: You may have students with missing marks as the grading and marking system in most schools in Kenya are manual. Also, absenteeism due to factors such as school fees may affect the math score.

Academic performance: This may be because of lack of proper infrastructure, especially in rural areas where the students may lack materials like books and pens. The teachers may also be overburdened due to the teacher student ratio, where 1 teacher is expected to teach a whole stream.

10. Impute missing values in family income and math score using the median. Justify why the median is appropriate for these variables.

The median is appropriate since the missingness is random and not systemic. We may have better performing students missing scores due to absenteeism. Also, outliers in the dataset will not affect the medians position.

11. Impute missing values in attendance rate using the mean. Compare the distributions before and after imputation using histograms.



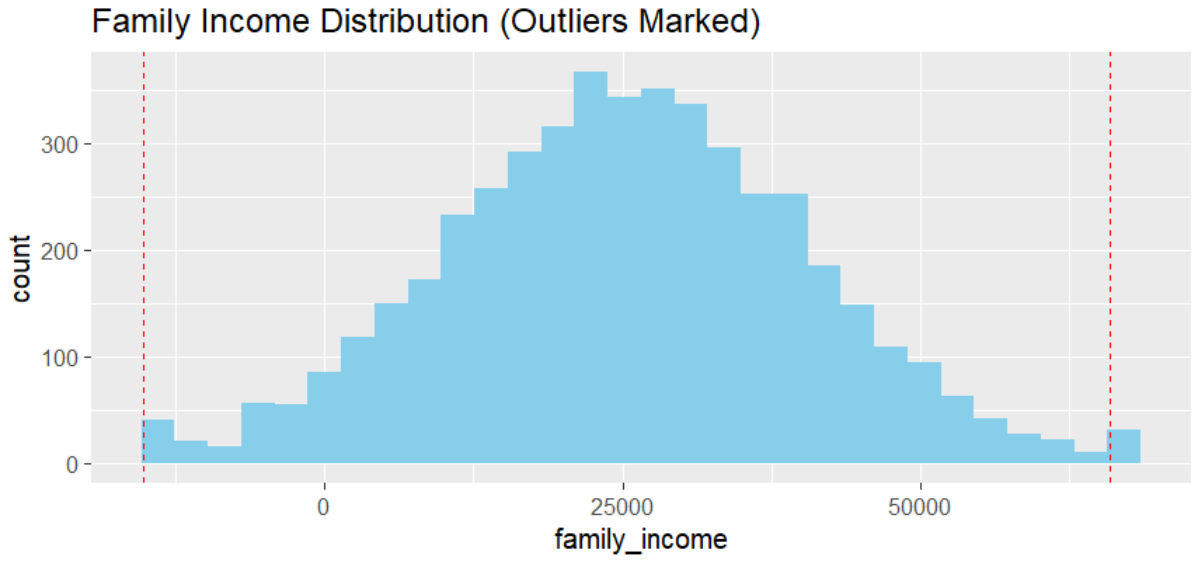
Observation: Attendance rates are typically normally distributed, making mean appropriate. Histograms verify if imputation preserves distribution shape.

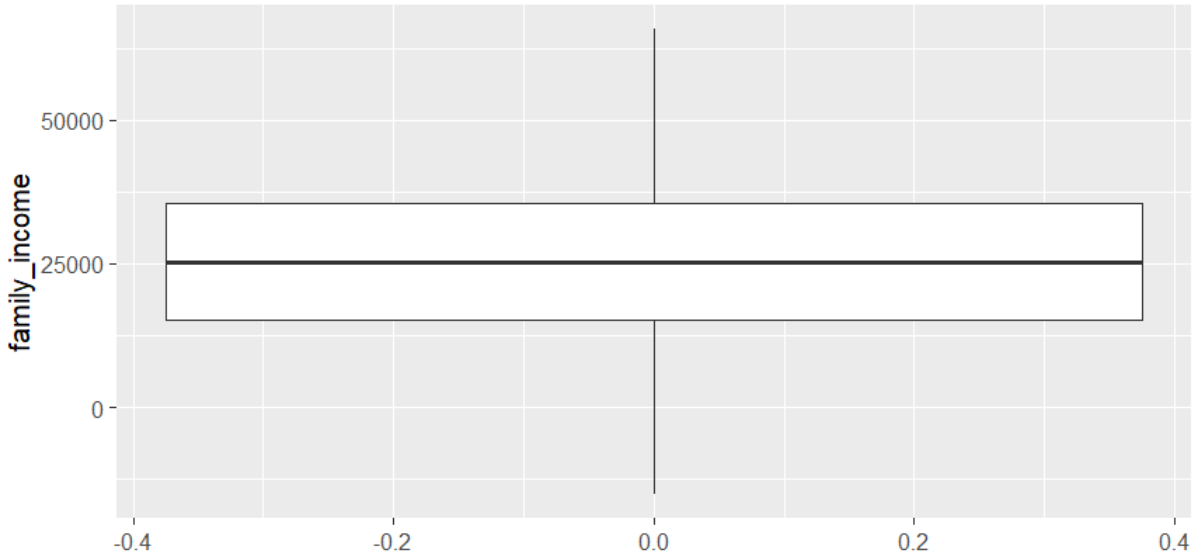
Data Preprocessing: Outliers

12. Detect outliers in family income using the IQR method. How many outliers are there, and what might they represent in a Kenyan context?

There are 56 outliers in family income, this could be because the families have irregular income may be from businesses or casual jobs. The outliers may also be because of families not wanting to share how much they earn.

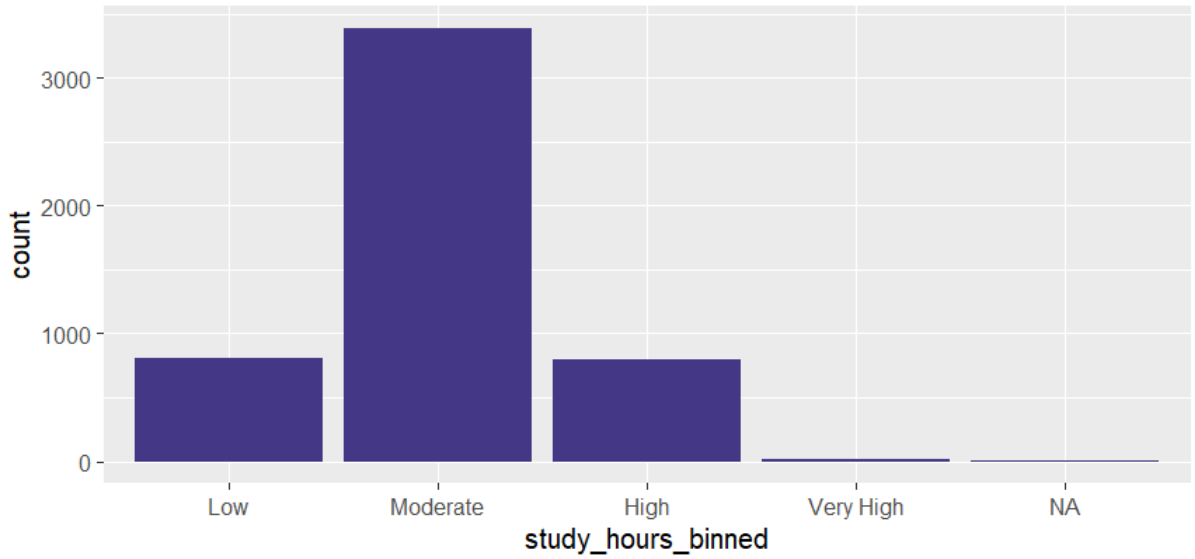
13. Cap outliers in family income at the 1.5*IQR bounds. Visualize the distribution before and after capping using boxplots.





Data Preprocessing: Feature Engineering

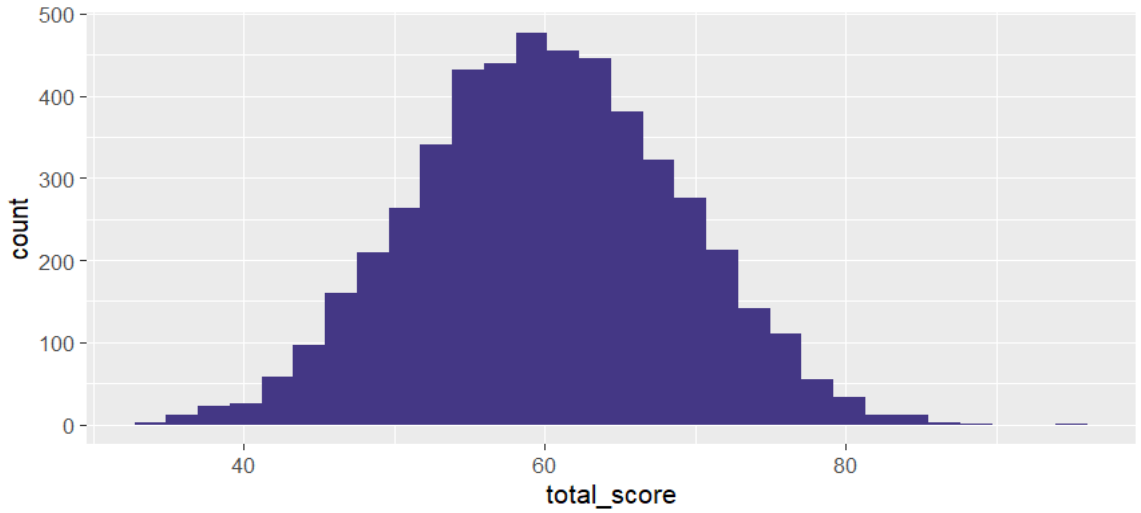
14. Discretize study_hours_weekly into four bins (e.g., Low, Moderate, High, Very High). Create a bar plot of the binned variable.



15. Discretize family income into quartiles (Low, Medium-Low, Medium-High, High). How does the binned variable correlate with academic performance?

Academic Performance Distribution: The "Excellent" category has the highest count in the "High" income bracket (301 students), suggesting a positive correlation between higher family income and better academic performance.
Consistency Across Brackets: The distribution of students across the "Good" and "Poor" categories remains relatively consistent across different income brackets, indicating that while higher income may be associated with better performance, it doesn't guarantee it.

16. Create a new feature total score by averaging math score, science score, and English score. Visualize its distribution.

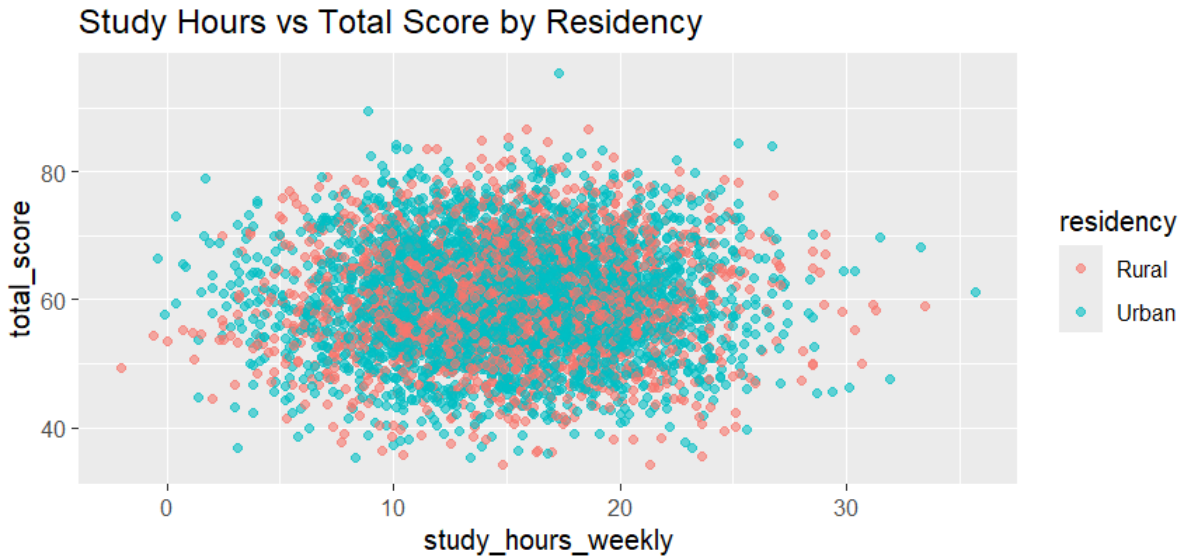


Data Preprocessing: Relationships

17. Create a contingency table for extracurricular activities vs. academic performance. What patterns suggest about student involvement?

Positive Impact of Sports: The higher number of "Excellent" students in the sports category suggests that involvement in sports may have a positive effect on academic performance.
Potential Overload with Both: Students participating in both clubs and sports may experience an overload, leading to a distribution skewed towards "Average" and "Poor" performance.
Non-Participation Advantage: The higher number of "Excellent" and "Good" students in the "None" category could indicate that not engaging in extracurricular activities allows students to focus more on academics.

18. Visualize the relationship between study_hours_weekly and total score (from Q16) using a scatter plot, colored by residency. What trends do you observe?



Observation: There appears to be a positive correlation between study hours and total score across all residency groups. However, the strength of this relationship varies by residency status, with urban residents showing a clearer upward trend.