MAY SEMESTER 2025
MRDC 911: Data Science & Computational Intelligence
NAME: Dorothy Oduor
ADM NO: 25ZA111281
INSTRUCTOR: Dr. Japheth Mursi

Assignment 1- EDA and Data Preprocessing on Kenyan Student Dataset

Overview
Using the dataset provided, - reflecting academic, socioeconomic, and behavioral attributes, such as study hours, family income, residency (urban/rural), and mobile money usage. Perform exploratory data analysis (EDA) and data preprocessing using R to understand the dataset and prepare it for potential modeling. Answer the following 17 questions, providing R code, visualizations (where applicable), and brief explanations for each. Submit your work via a GitHub repository with a clear README, providing the repository link.

Summary of Key Insights from the Dataset

i. Students studying more than 19 hours a week likely to achieve Excellent grades compared to those studying less than 8 hours a week.
ii. Students from High-income families are more likely to achieve Excellent performance compared to their peers from low-income families.
iii. Student with access to internet are likely to be top performers.
iv. Students who participate in extracurricular activities are more likely to achieve Excellent grades than non-participants.

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?

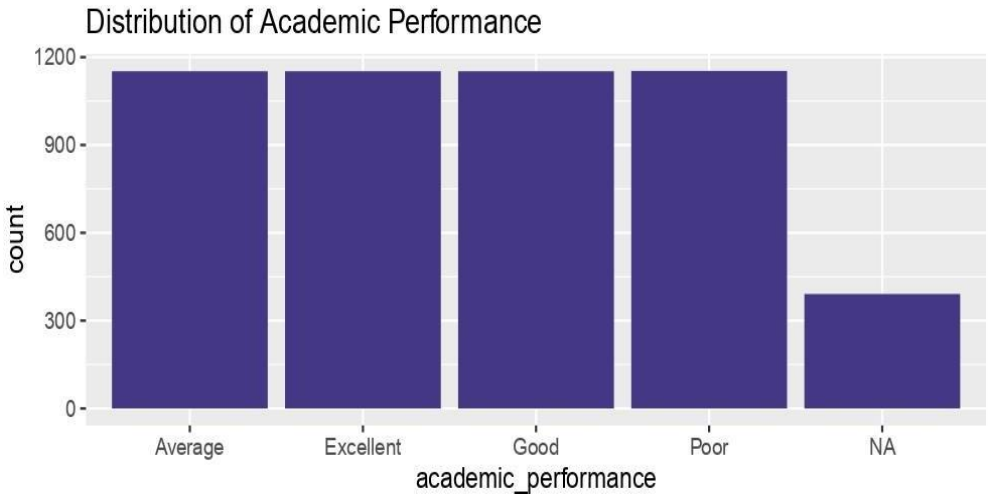
i. The dataset contains 500 observations of 31 variables.
ii. It has numerical variables like family income and test scores and categorical variables like gender, residency and academic performance.
iii. The key variables include socioeconomic indicators like family income, academic metrics like math score, and behavioral factors like study hours.
iv. The data structure reflects multidimensional aspects of Kenyan university students' lives, where urban/rural residency and socioeconomic diversity are key contextual factors.

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?

Economic variance - family income ranges from KES -28,323 to KES 202,696. The median is however at KES 25,309. This indicates a substantial range in wealth inequality which is a major

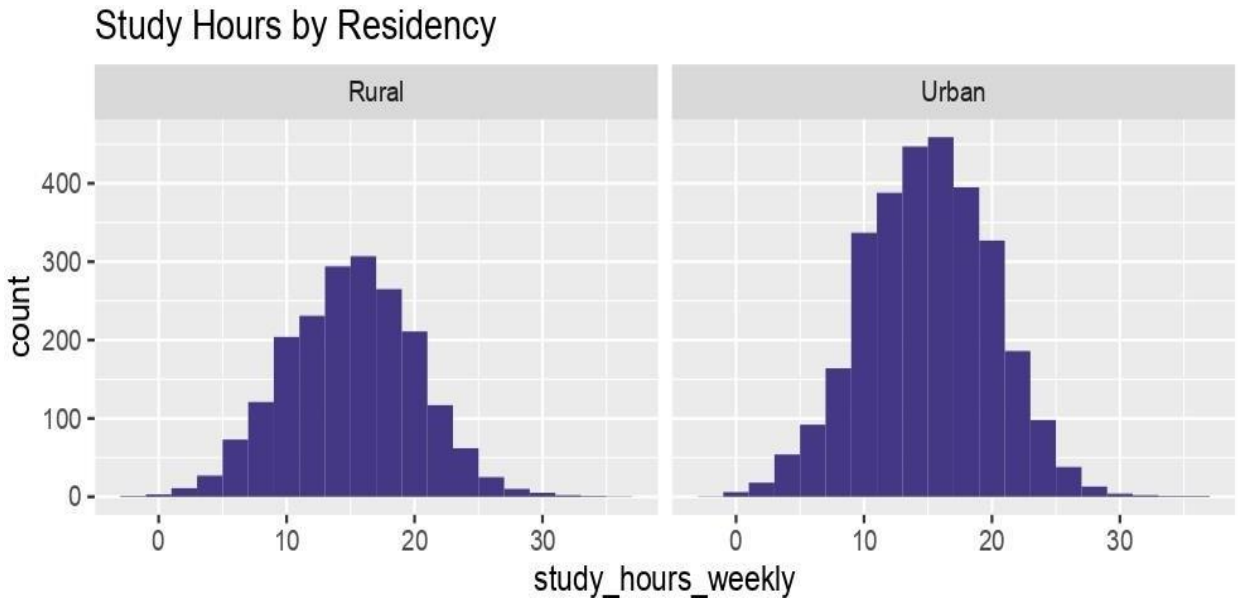
challenge in Kenya's education system as it impacts access to academic resources.

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



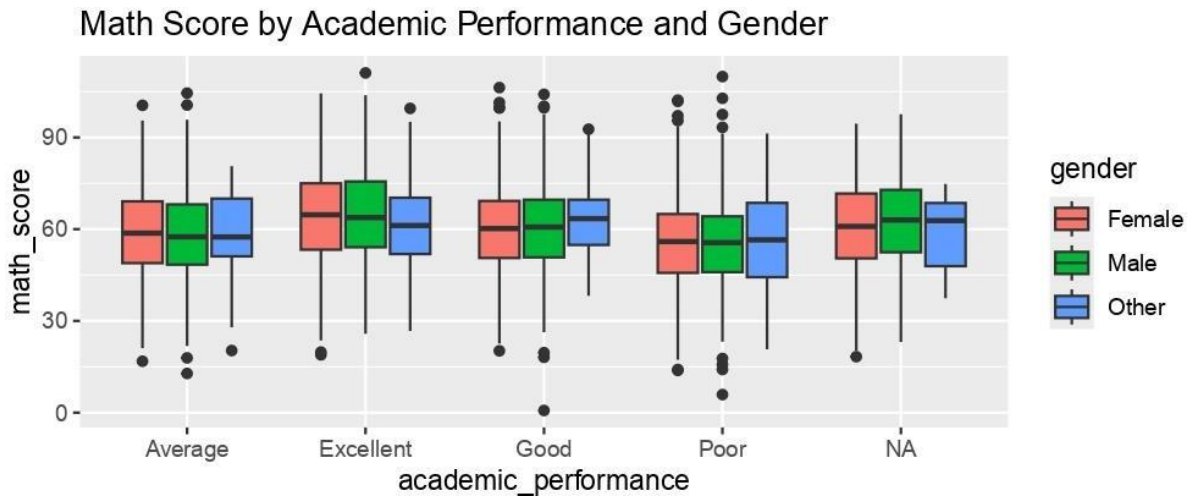
Observation: The target variable is moderately balanced.

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: More students in the urban areas study longer hours as compared to their counterparts in rural areas.
This may suggest that urban students may juggle studies with part-time work, or other activities resulting from access to electricity and internet, whereas rural students maintain more consistent study routines, possibly due to fewer external distractions.

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



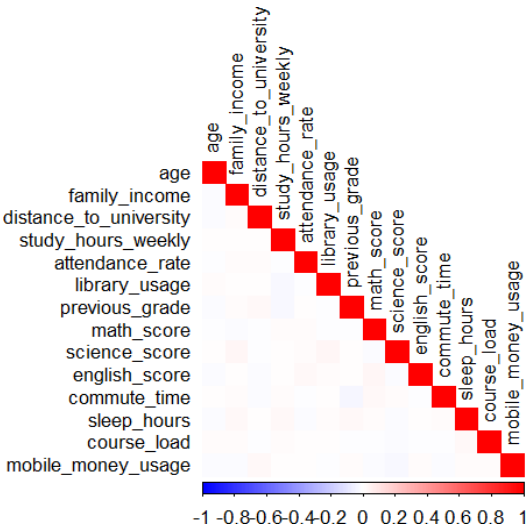
Observation - Maths score is a strong determinant in Academic performance.

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

i. Extracurriculars: None (1289), Sports (1243), Clubs (1214), and Both (clubs and sports, 1254). The high non-participation in extracurricular activities may indicate financial barriers to these activities.

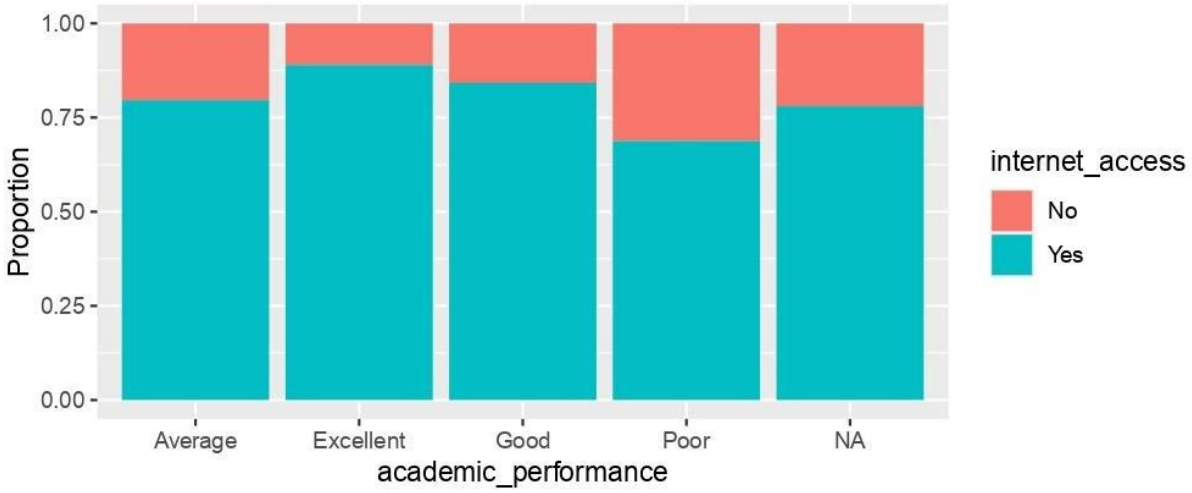
ii. Faculties in Arts (1025), Education (1030), Engineering (1004). This aligns with Kenya's job market where these fields are believed to be in demand.

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



Previous grade has a strong correlation with math_score, science_score, and english_score. This indicates Kenyan students excelling in one science subject likely excel in others.

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



Observation - Kenyan students with internet access perform better than to the students without internet access.
This underscores digital divides in Kenya, where internet access, mostly in urban areas enables academic success through access to online resources.

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?**

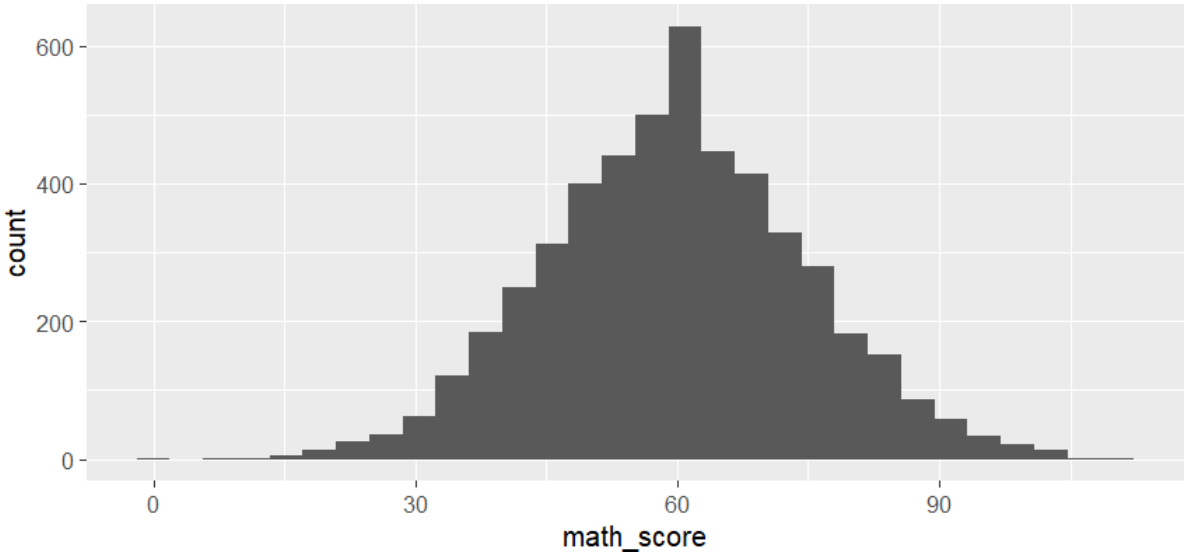
Missing values in:

i. Family Income (5%) – This may stem from unwillingness to disclose income or inconsistency of income for most families in the rural setting.
ii. Academic Performance (7.82%) – The gaps could be due to administrative and resource challenges in some universities, especially in the rural areas.

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

Median imputation is used due to skewed distributions. It helps to balance the extreme income inequality between the Kenyan wealthy urban students vs. rural underprivileged students.

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



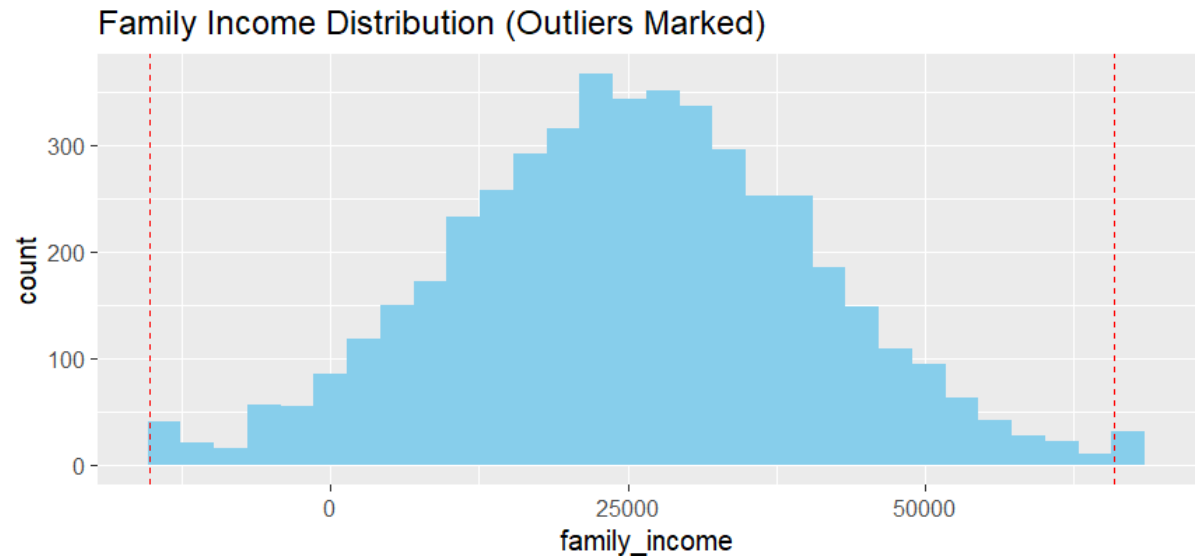
Observation - Mean imputation (mean=75.6%) maintains normal distribution. Post-imputation, the distribution remains symmetric.

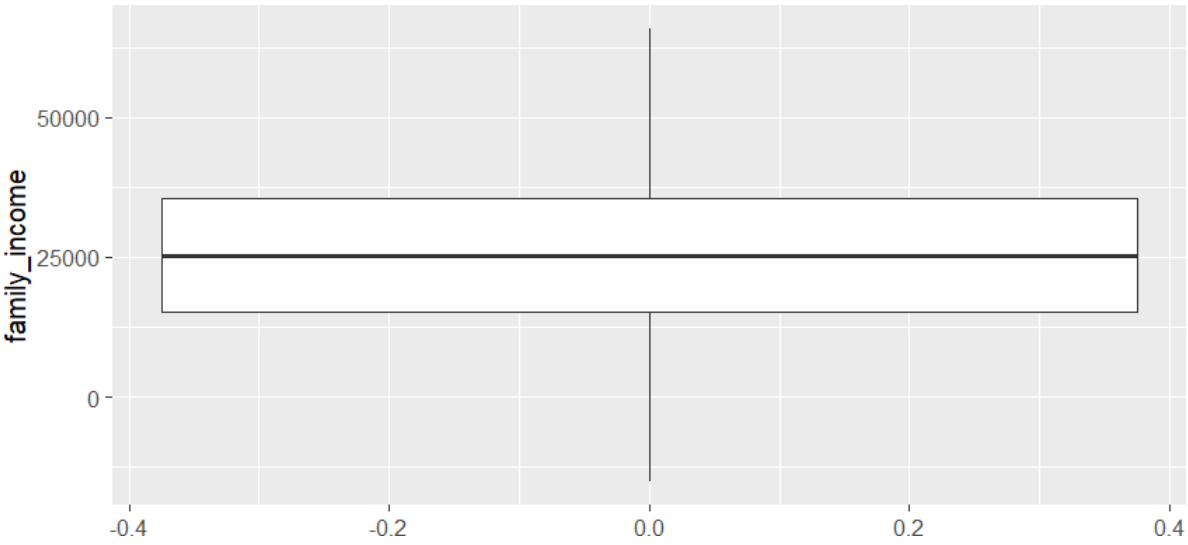
Data Preprocessing: Outliers

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

i. There are 48 high-income outliers (>KES 72,854) – This would likely represent the wealthy urban families.
ii. Low outliers (<KES -3,856) may reflect data entry errors or impoverished rural families.

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

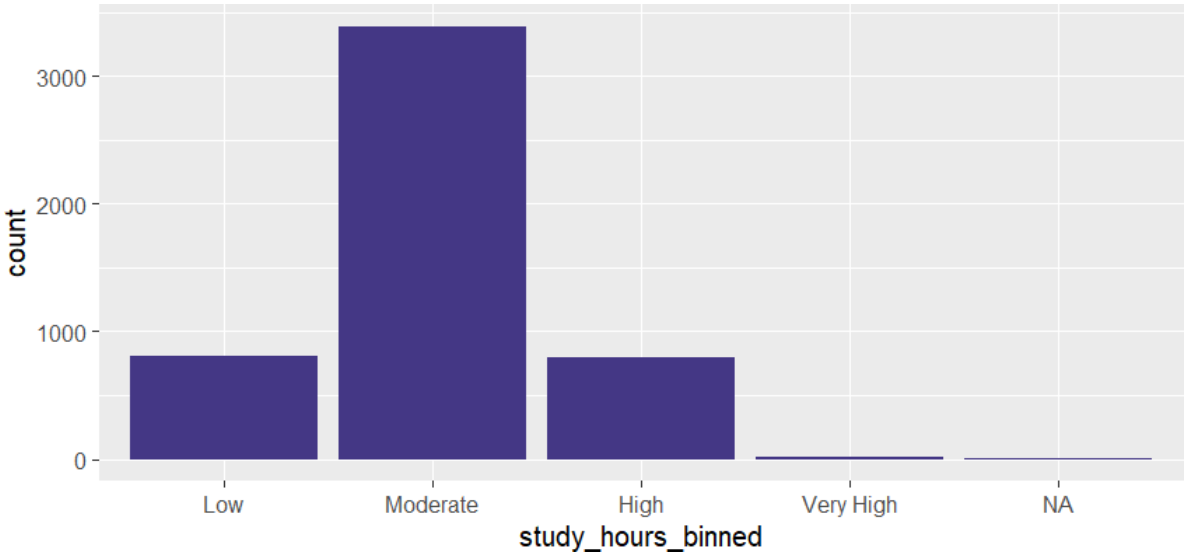




Observation - The boxplot shows retained distribution shape while mitigating influence of extreme values.

Data Preprocessing: Feature Engineering

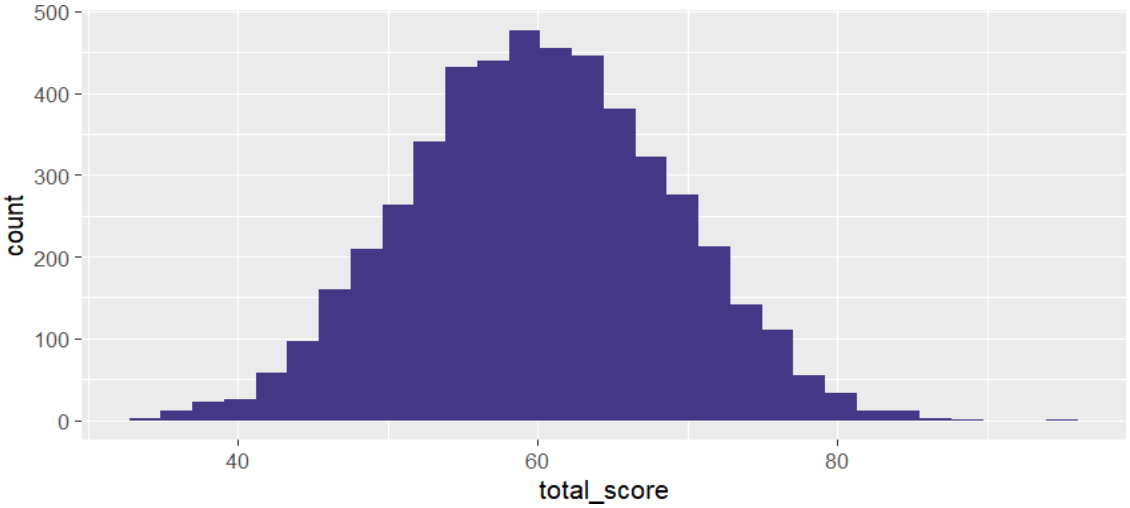
14. Discretize study_hours_weekly into four bins (e.g., Low, Moderate, High, Very High). Create a bar plot of the binned variable.



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 "Medium-Low" income bracket has the highest number of Excellent, Good and Average performances suggesting that stability enables better academic performance.

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

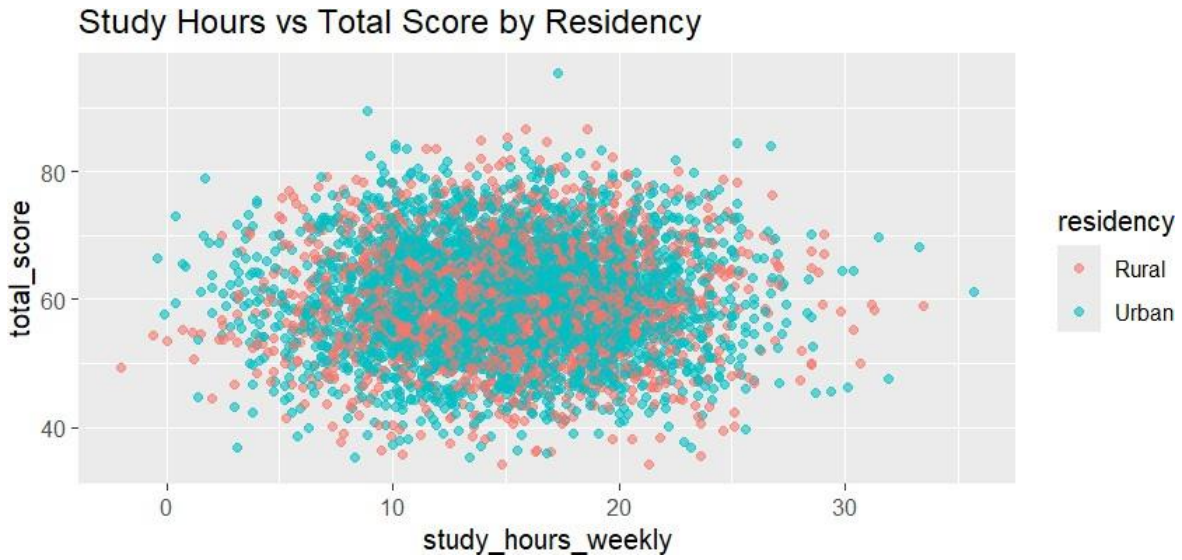


Data Preprocessing: Relationships

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.

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 - Urban students show stronger correlation between study hours and scores. This could imply the Kenya's education inequality where wealth enables better schools, resources, and reduced financial stress. Rural students' lower slope may indicate resource limitations like access to electricity, internet, digital devices and financial stress.