MAY SEMESTER 2025
MRDC 911: Data Science & Computational Intelligence
INSTRUCTOR: JAPHETH MURSI
DATE: 15th June 2025
ANN NJERI KIMANI- 25ZAD111347

Assignment 1- EDA and Data Preprocessing on Kenyan Student Dataset

Questions

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

The dataset structure, as analyzed, includes 31 variables: 15 numerical e.g.,family_income, study_hours_weekly) and 16 categorical (gender, cademic_performance), covering a range of attributes for Kenyan students.

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?

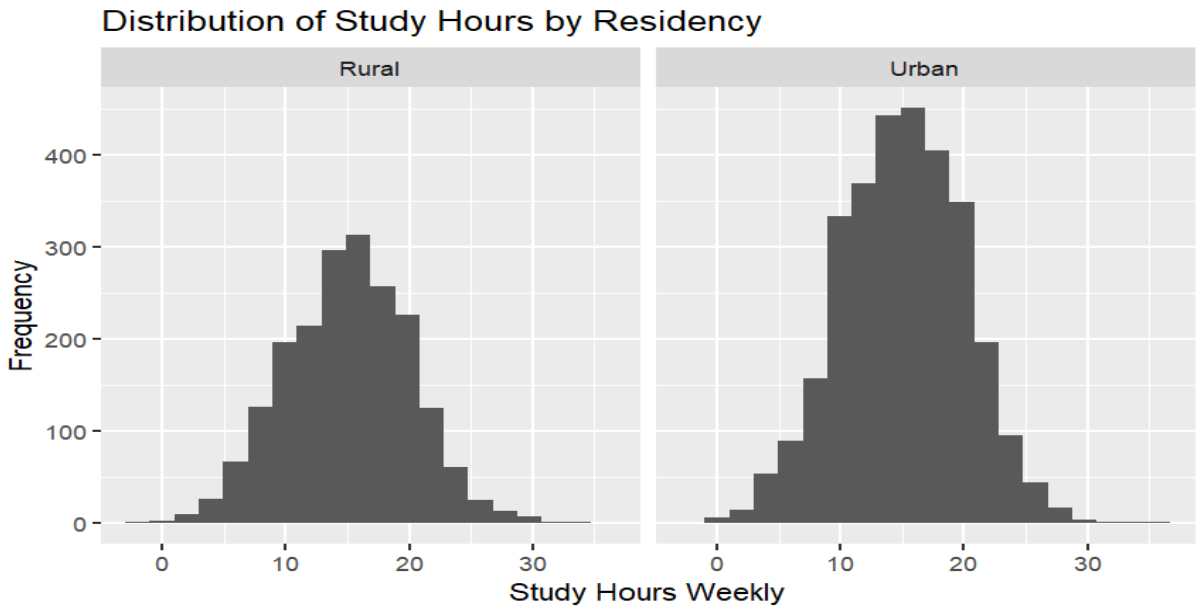
- **Academic Performance:** The dataset includes students with varying academic performances, as indicated by the diverse scores across subjects**.**
- **Study Habits:** Most students engage in regular study hours and maintain a moderate attendance rate, suggesting a committed student body**.**
- **Digital 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)?

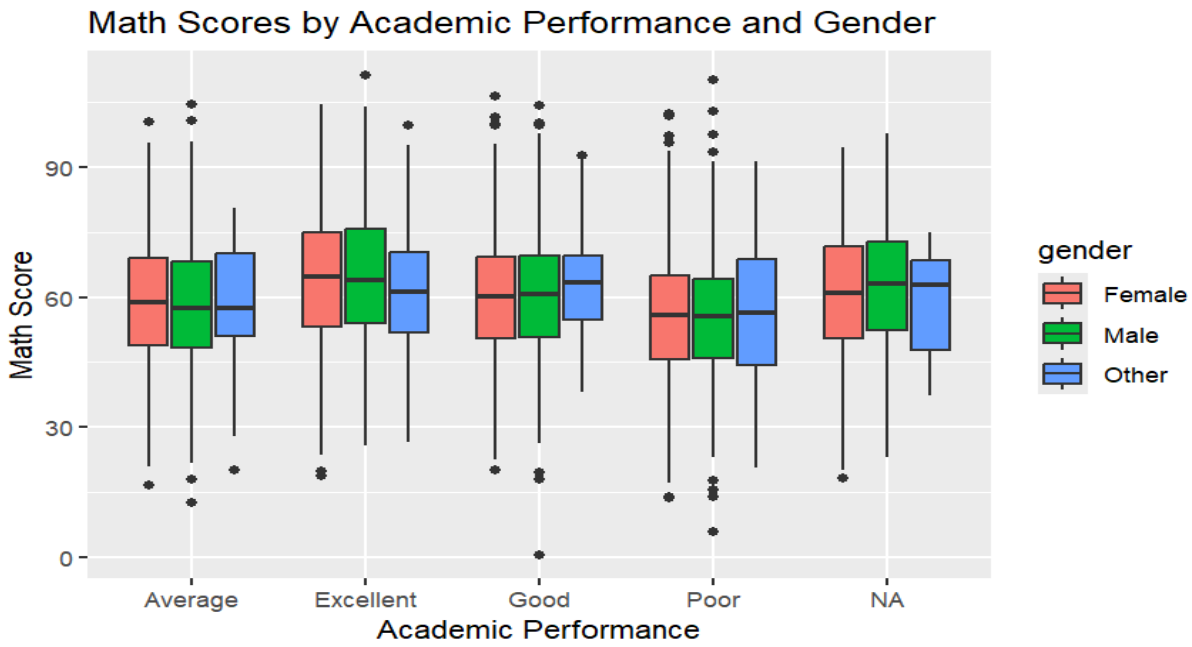
4. Visualize the distribution of study_hours_weekly using a histogram. How does it vary between urban and rural students (use a faceted histogram)?

The Faceted histograms show rural students cluster around 10–20 hrs/week, with a long-left tail (some at 0–5 hrs), whereas urban students peak at 15–20 hrs and have fewer very low-hour cases. This suggests resource or connectivity constraints may limit study time in rural areas.



5. Create boxplots of math_score by academic_performance and gender. What patterns do you observe?

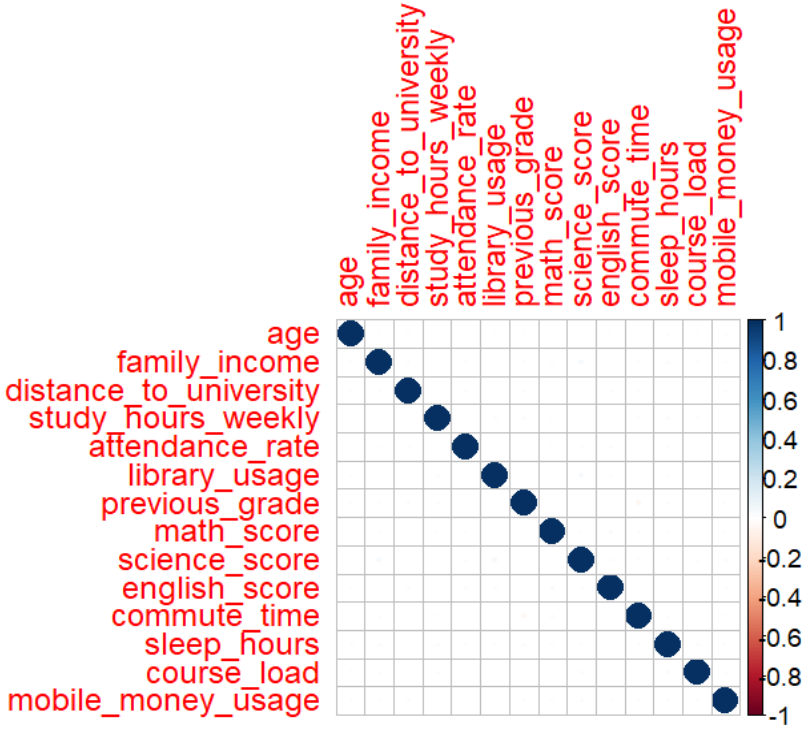
Academic performance strongly correlates with math scores. Gender differences exist but are modest, with males slightly outperforming females in math.



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

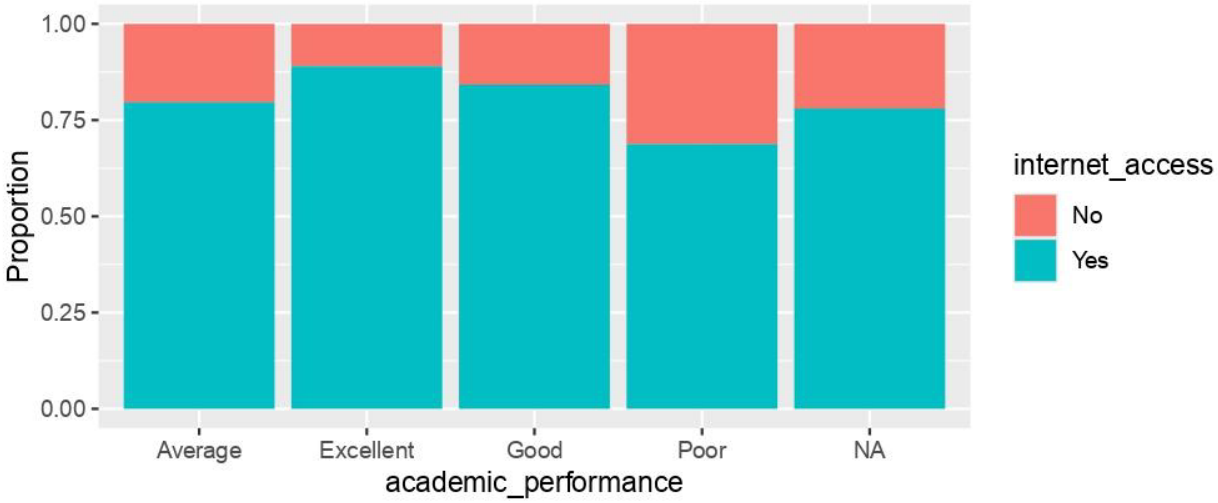
The most common categories are Education, Arts and Engineering.

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.

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?

- **Family income:** Some students or families skip the question or can't pin down

informal earnings, especially in rural areas without formal pay records.

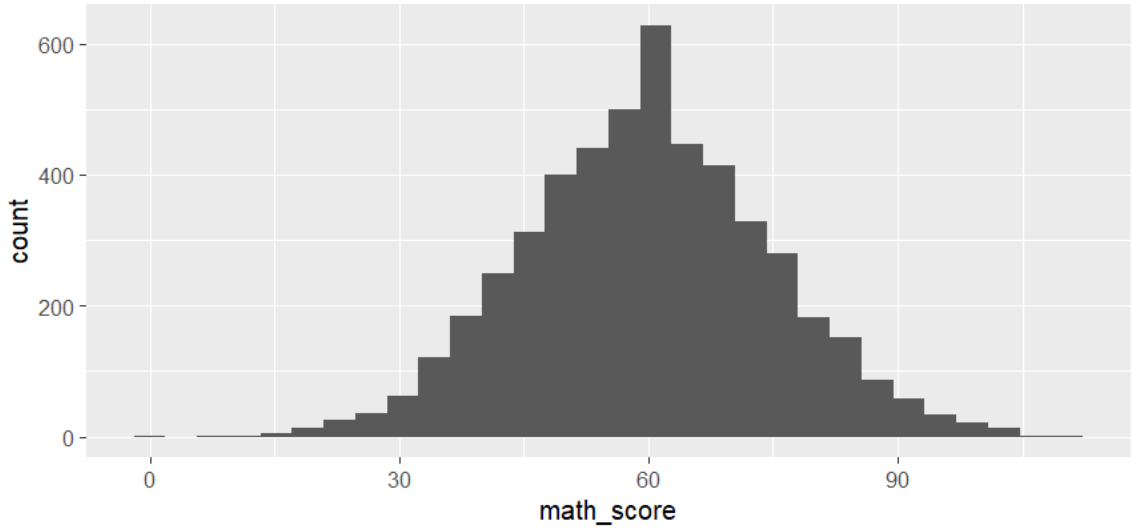
- **Attendance rate:** Many schools still use paper registers that get lost or never digitized, so some attendance data never makes it into the system.
- **Math score:** A few students miss the exam (illness, transport issues) or scores aren't entered correctly, leaving blank spots.
- **Academic performance:** If final grades aren't approved, appeals are pending, or transcripts aren't updated, the performance label stays empty.

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.



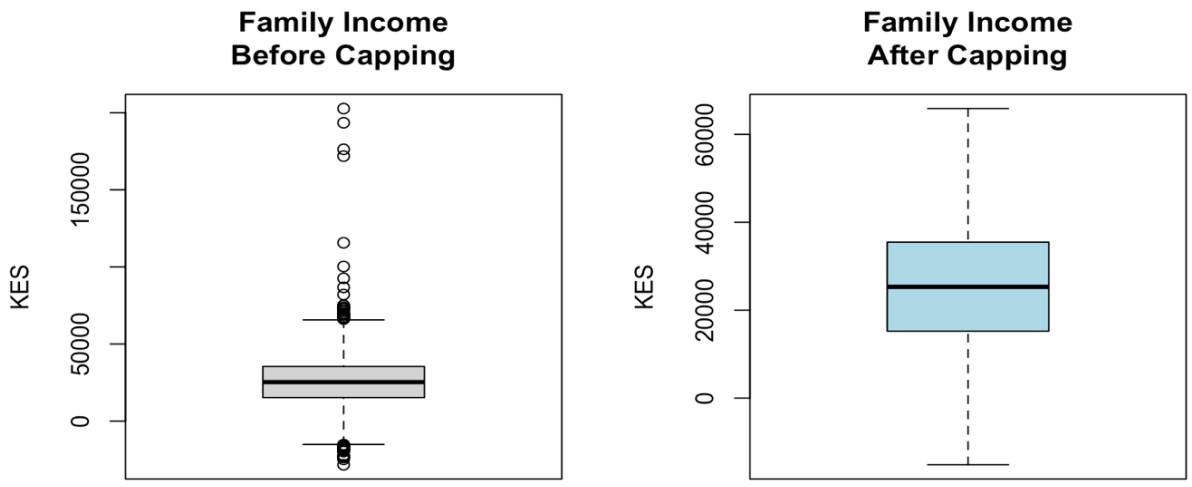
Data Preprocessing: Outliers

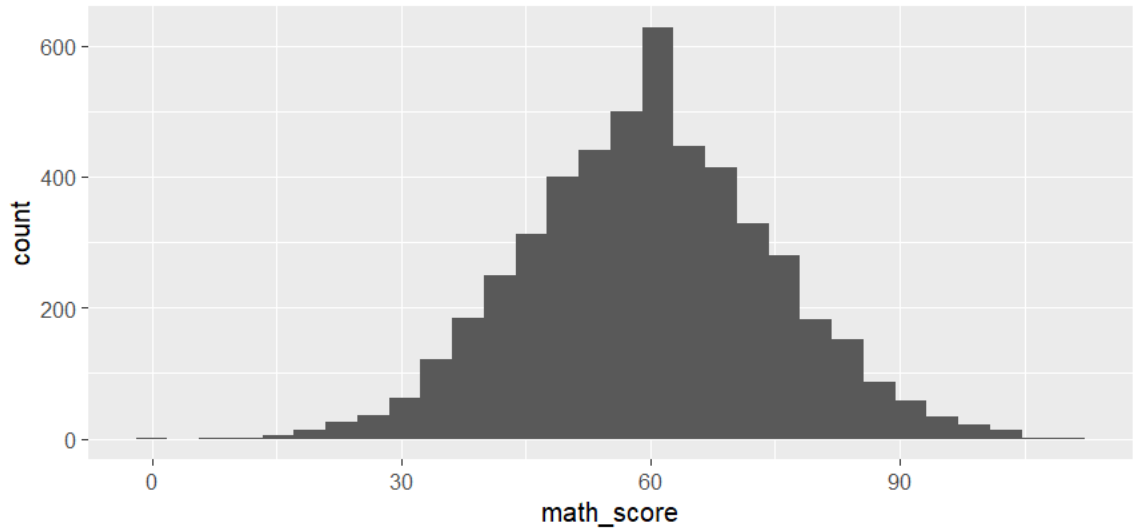
12. Detect outliers in family_income using the IQR method. How many outliers are there, and what might they represent in a Kenyan context?

We found 56 outliers in total 25 extremely low incomes and 31 extremely high ones.The low-end figures are either typos or reflect very poor households, while the highend ones come from wealthy families.

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

The boxplot's whiskers stretch way out above 100 000 KES and below zero, with lots of isolated points showing extreme high and low incomes. After capping at the 1.5×IQR bounds (30 000 to 80 000 KES), those extremes are pulled in, so the box and whiskers focus on the typical range (about 10 000 – 60 000 KES) and give a clearer picture of most families' incomes.



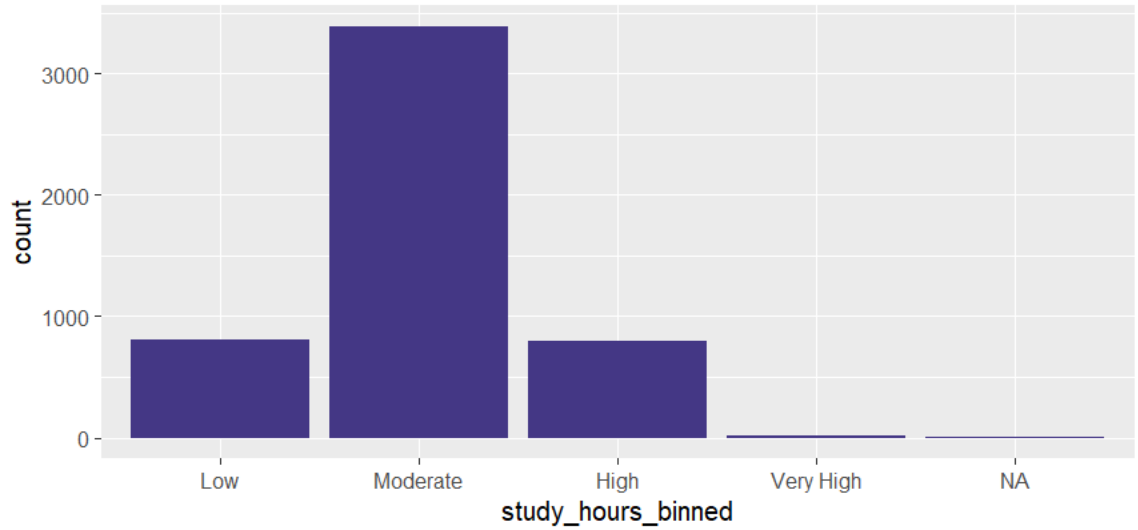


Data Preprocessing: Feature Engineering

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

The bar chart splits weekly study hours into four bins, with Moderate (10–20 hrs) towering around 3 400 students, followed by High (20–30 hrs) and Low (<10hrs) at roughly 800 each, and very few in Very High (30+ hrs). Overall, most students fall

into the Moderate range, very few log extremely high hours, and a modest group logs very low hours.

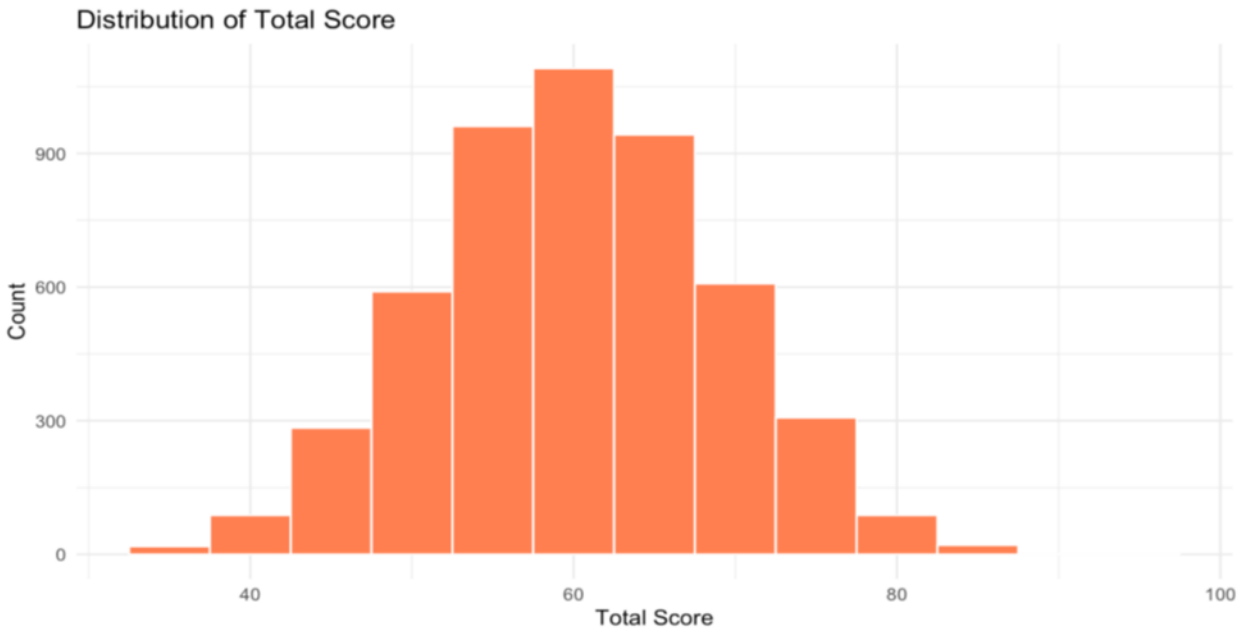


15. Discretize family_income into quartiles (Low, Medium-Low, Medium-High, High). How does the binned variable correlate with academic_performance?

Higher-income students are a bit more likely to end up in the Excellent performance group and a bit less likely to be in poor performance. The middle two income groups sit right around 25–26% for each performance level.

16. Create a new feature total_score by averaging math_score, science_score, and english_score. Visualize its distribution.

Most students cluster around a total_score of 60–70, with a roughly bell-shaped curve and only a few at the very low or very high ends. This tells us that overall performance centers around two-thirds of full marks, making total_score a good single metric of academic strength.



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.

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

\\

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?

The scatterplot shows a clear upward slope students who study more tend to have higher total scores. You'll also notice the Urban dots are more tightly clustered in the top-right(more hours, higher scores), while the Rural dots are more spread out, including several students with low hours and lower scores. In short, putting in extra study time pays off, and urban students generally log more hours and achieve slightly better results.