MAY SEMESTER 2025
MRDC 911: Data Science & Computational Intelligence
DATE: 6th June 2025
Due Date: 13th June 2025
NAME: SILA KIMELI RONOH
STUDENT. NO: 25ZAD111181
ASSIGNEMENT

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
── Column specification ────────────────────────────────────────────────
Delimiter: ","
chr (16): gender, residency, socioeconomic_status, parental_education, ext...
dbl (15): student_id, age, family_income, distance_to_university, study_ho...
```

 # Explanation:

The dataset contains 31 variables and 5000 rows.

There are 15 numeric variables (like age, income, study hours) and 16 categorical variables (like gender, residency, and performance). This helps us know how to handle each variable in analysis and preprocessing.

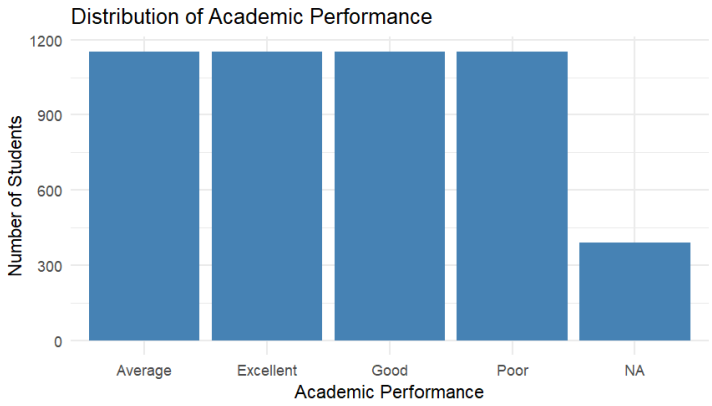
2. **Compute summary statistics (mean, median, min, max, etc.) for all numerical variables (e.g., family_income, study_hours_weekly). What insights do these provide about the data?**

The dataset reveals that most students are in their early 20s, with typical study hours ranging from 10 to 20 per week. However, several variables—including `family_income`, `study_hours_weekly`, `commute_time`, and

`library_usage`—contain negative or extreme values, indicating data entry errors or outliers. There's also notable variation in academic performance and income, reflecting a diverse student population.

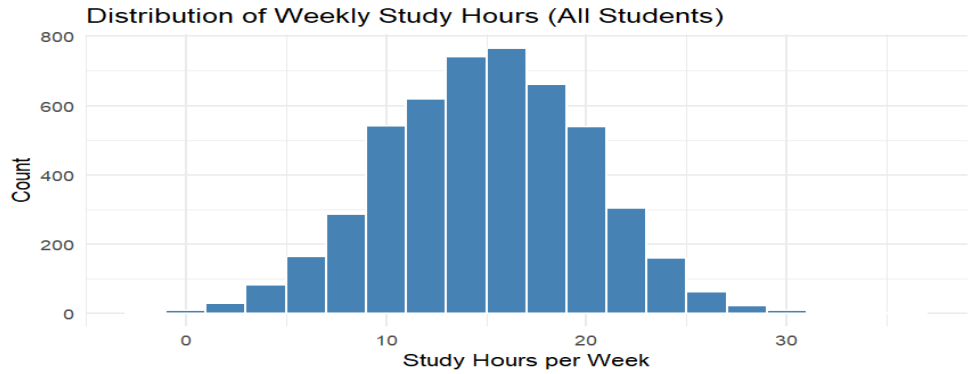
3. **Create a bar plot to visualize the distribution of academic_performance. Is the target variable balanced across its classes (Poor, Average, Good, Excellent)?**

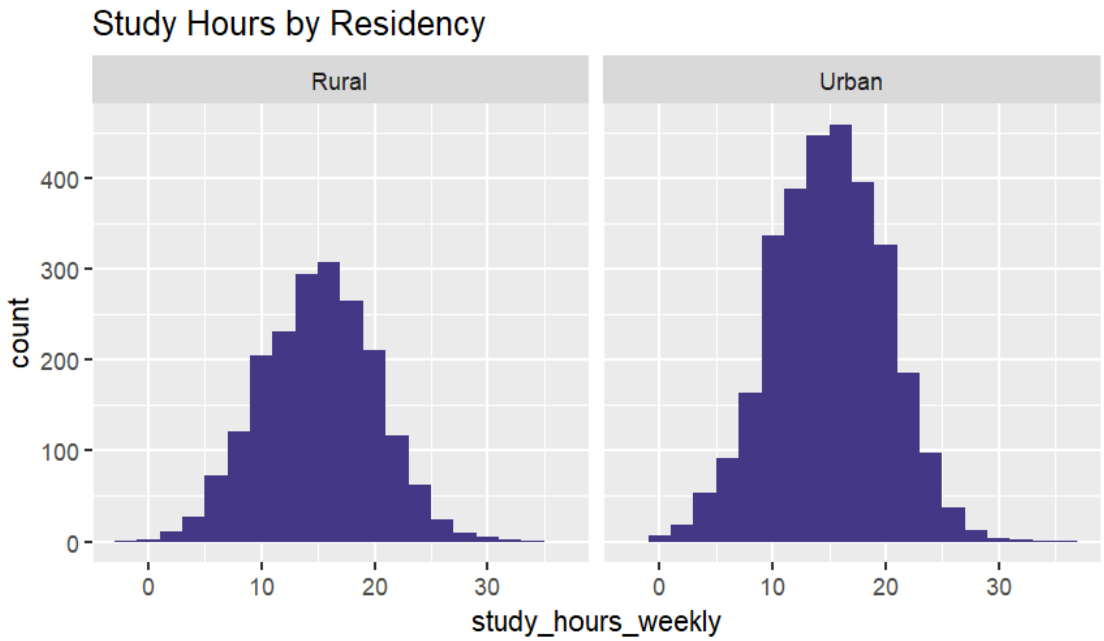
```
academic_performance     n percent
  <chr>                <int>   <dbl>
1 Average               1152    23
2 Excellent             1152    23
3 Good                  1152    23
4 Poor                  1153    23.1
5 NA                     391     7.8
```



The variable is well balanced across the four main categories—Poor, Average, Good, and Excellent—with each category having approximately 23% of the total observations. Specifically, each group contains around 1,152–1,153 students, showing almost equal distribution. However, about 7.8% (391 students) have missing values (NA), which will need to be addressed during data cleaning.

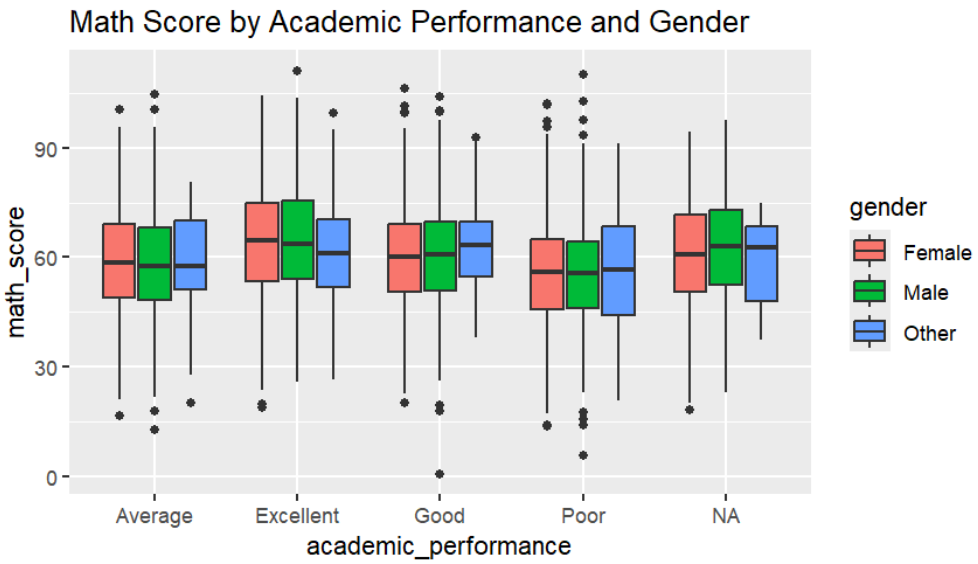
4. **Visualize the distribution of study_hours_weekly using a histogram. How does it vary between urban and rural students (use a faceted histogram)?**





Urban students tend to have a slightly wider spread of study hours, with more students studying below 10 and above 20 hours while Rural students appear to be more concentrated around the average (12–18 hours), indicating a more consistent study pattern. This suggests potential differences in time availability, access to study environments, or daily schedules

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



The boxplot reveals a positive association between math scores and academic performance—students in the "Excellent" category tend to have higher math scores, while those in the "Poor" category show lower central values and wider variability. The visualization also highlights gender-based trends: across all

performance categories, male and female students exhibit overlapping score distributions, though in some categories, one gender may slightly outperform the other.

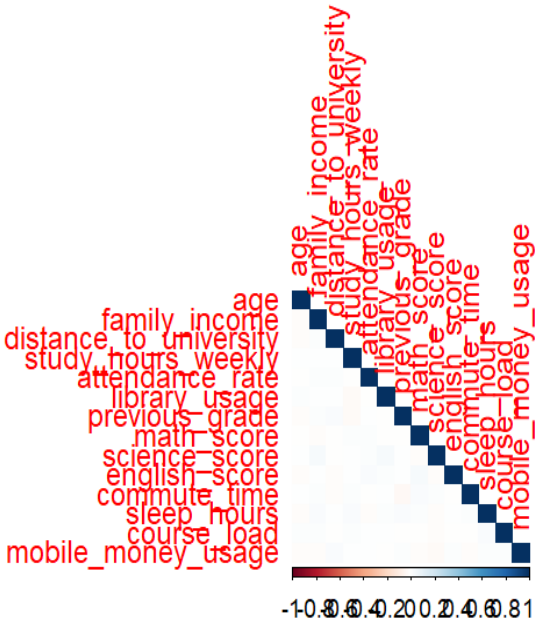
6. **Compute the proportion of each category in extracurricular_activities and faculty. Which categories are most common?**

faculty	n	prop
<chr>	<int>	<dbl>
1 Arts	1025	0.205
2 Business	967	0.193
3 Education	1030	0.206
4 Engineering	1004	0.201
5 Sciences	974	0.195

The analysis shows that students are distributed across a variety of extracurricular activities, with some participating in both sports and clubs, while a notable portion report no involvement at all. This may reflect varying levels of access or institutional encouragement.

The distribution of students across faculties is relatively balanced, with no single faculty overwhelmingly dominant.

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



The correlation heatmap shows strong positive relationships between subject scores (`math_score`, `science_score`, `english_score`) and `previous_grade`, indicating consistent academic performance across subjects. Most other variables show weak or no correlation, suggesting they contribute independently to

student outcomes.

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

```
Pearson's Chi-squared test

data:  table(data$internet_access, data$academic_performance)
X-squared = 163.55, df = 3, p-value < 2.2e-16
```

This Chi-squared test result indicates a statistically significant relationship between `internet_access` and `academic_performance` ($\chi^2 = 163.55$, df = 3, $p < 0.001$). Since the p-value is extremely low (< 0.05), we reject the null hypothesis of independence. This means that access to the internet is strongly associated with students' academic performance in this dataset.

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?**

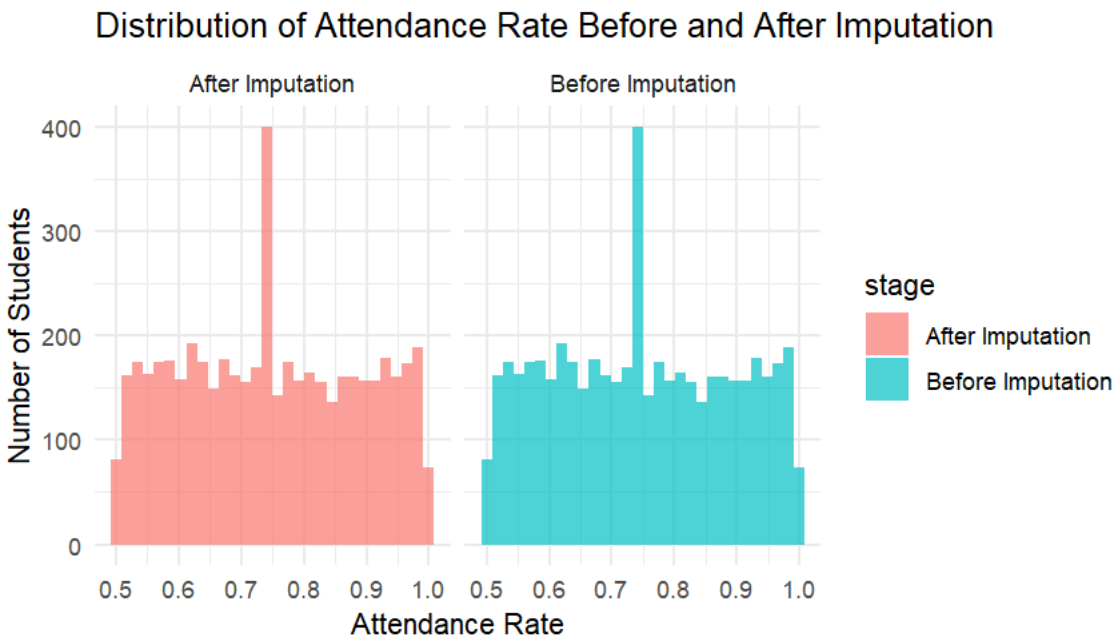
variable	missing_count	missing_percent
<chr>	<int>	<dbl>
1 academic_performance	391	7.82
2 family_income	250	5
3 attendance_rate	250	5
4 math_score	150	3

- **Academic Performance (7.8%)**: Missing entries could result from incomplete grading, delayed results, or student dropouts before assessments were recorded.
- **Family Income (5%)**: Income is often considered sensitive; students or parents may be unwilling or unable to disclose it accurately, especially in informal or low-income settings.
- **Attendance Rate (5%)**: Manual or inconsistent record-keeping in some institutions, particularly in rural areas, may lead to gaps in attendance logs.
- **Math Score (3%)**: Missing scores could stem from exam absenteeism, delayed grading, or unrecorded assessments in some courses or institutions.

10. **Impute missing values in family_income and math_score using the median. Justify why the median is appropriate for these variables.**

Missing values in `math_score` and `family_income` were imputed using the **median** because both variables contain extreme or skewed values, including **outliers** and in the case of income, even negative entries. The median is a robust measure of central tendency that is less influenced by such outliers, making it more appropriate than the mean for these fields.

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



The histograms compare the distribution of `attendance_rate` **before and after imputation**:

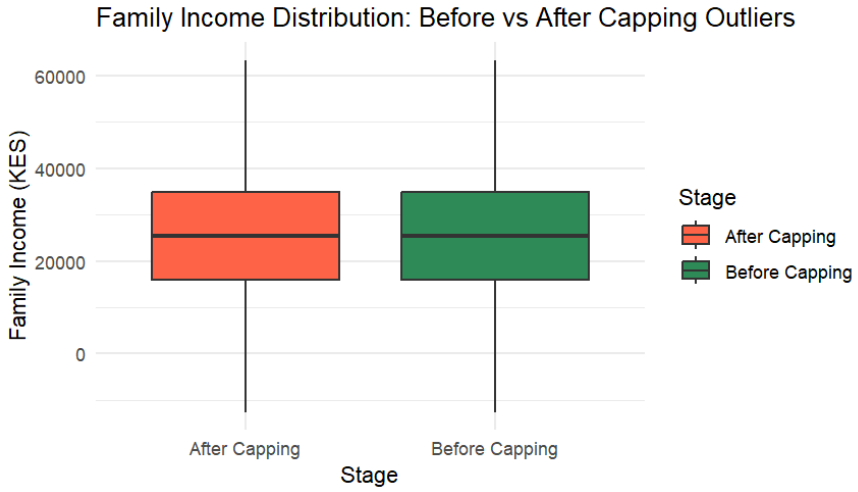
- **Before Imputation**: The distribution is relatively uniform, with a natural spread across values from 0.5 to 1.0. This represents the original data, excluding missing values.
- **After Imputation**: A sharp spike appears at the **mean attendance rate (~0.75)**—this is where missing values were filled. The rest of the distribution remains consistent, indicating that imputation preserved the overall shape of the data.

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?**

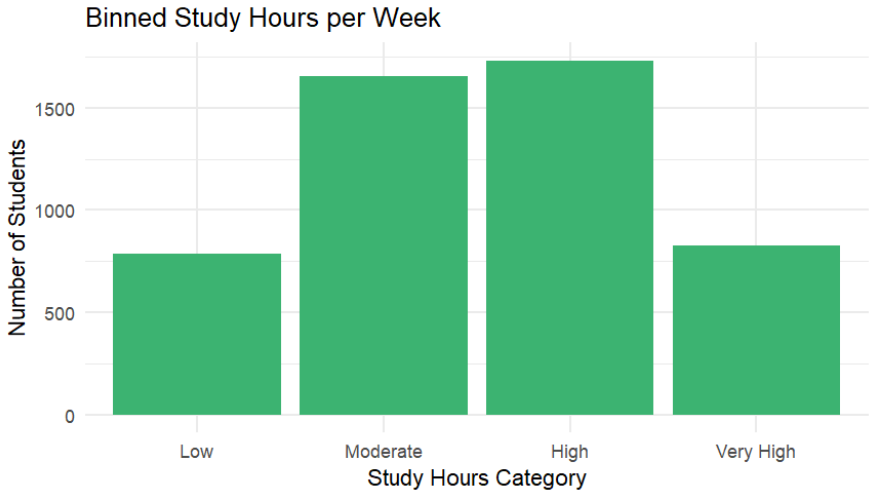
    We found 56 outliers in family income with 25 extremely low incomes and 31 with extremely high incomes. Low indicates very poor families or typos errors while high indicates wealthy families

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

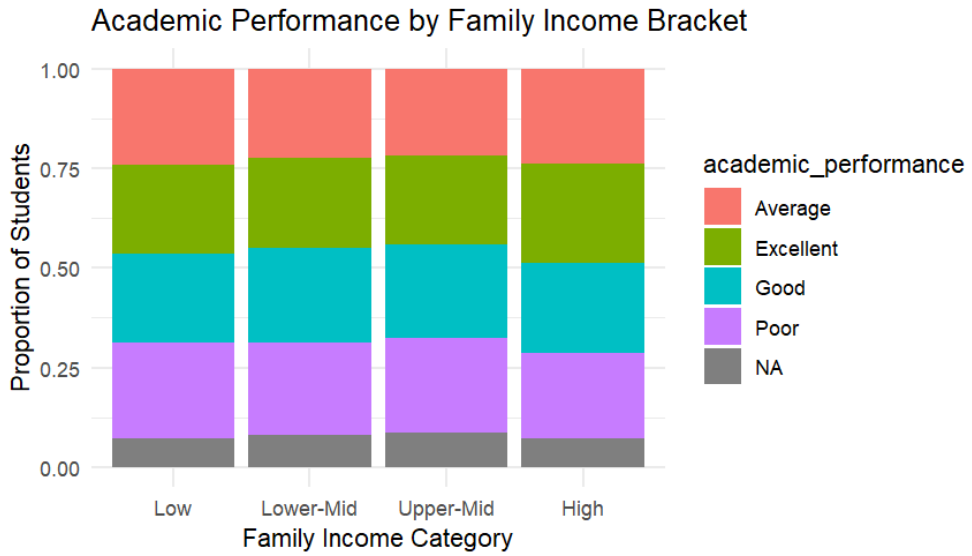


Data Preprocessing: Feature Engineering

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

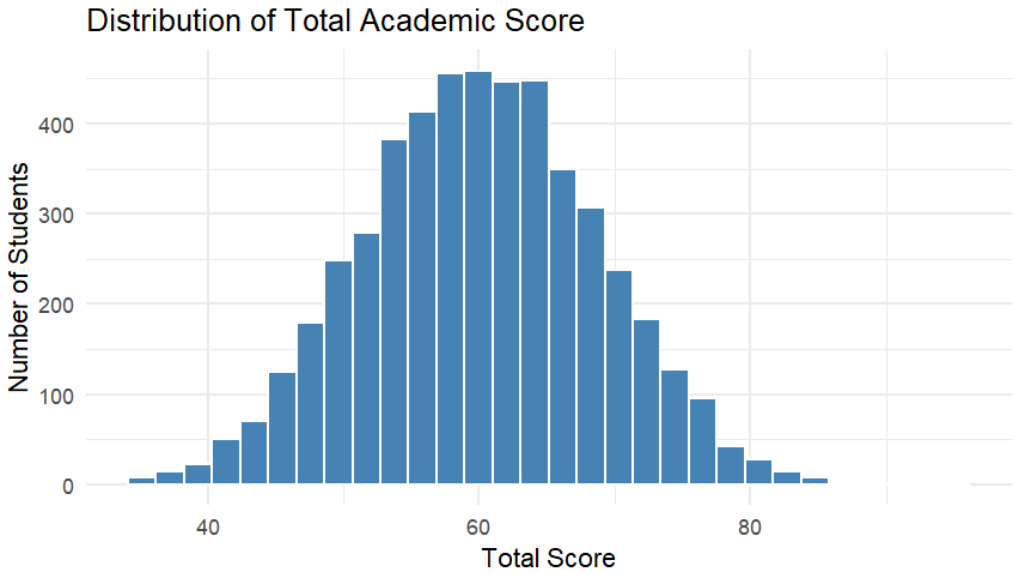


15. **Discretize family_income into quartiles (Low, Medium-Low, Medium-High, High). How does the binned variable correlate with academic_performance?**



The proportional bar chart shows that students in the Upper-Mid and High income brackets tend to have a higher proportion of Good and Excellent performance. In contrast, the Low-income group has more students with Poor or Average performance, suggesting that income may influence academic outcomes through access to resources, support, or study environments.

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

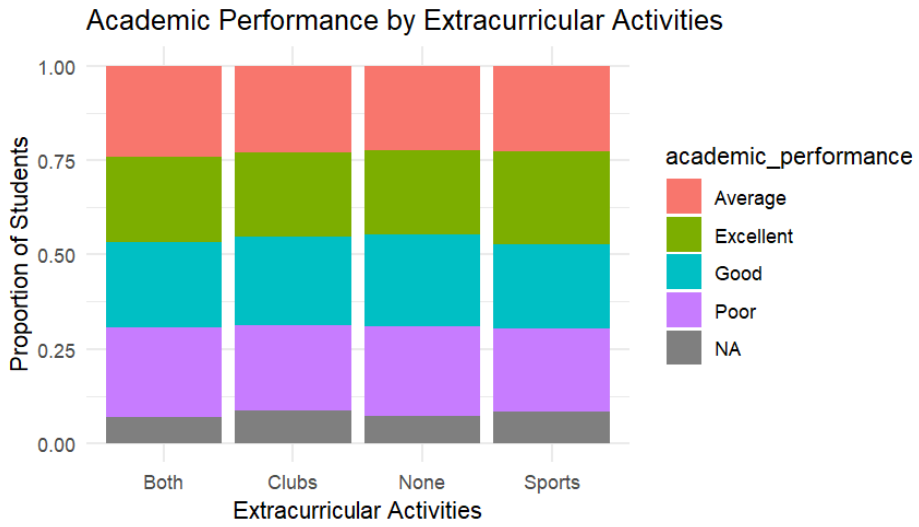


The histogram shows that `total_score` follows a roughly normal distribution, with most students scoring between **50 and 70**. This confirms that the variable is

well-behaved and statistically stable, making it a strong candidate for further analysis such as regression or classification tasks.

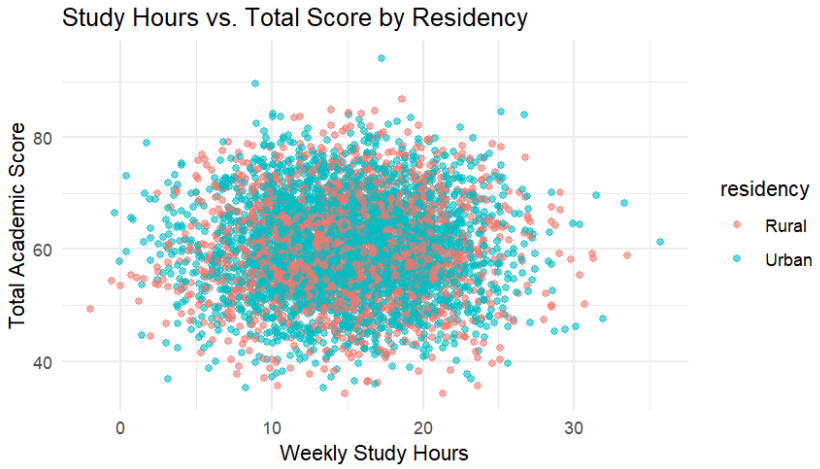
Data Preprocessing: Relationships

17Create a contingency table for extracurricular_activities vs. academic_performance. What patterns suggest about student involvement?



The contingency table and bar plot reveal that students involved in both sports and clubs tend to have higher proportions of Good and Excellent academic performance. In contrast, students with no extracurricular participation show a higher proportion of Poor performance, suggesting a possible positive association between well-rounded engagement and academic outcomes.

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



The scatter plot reveals a **positive trend** between `study_hours_weekly` and `total_score`, suggesting that students who study more tend to perform better overall. When colored by `residency`, the plot shows that **urban students** exhibit more variability in both study time and performance, while **rural students** appear more concentrated around average values.