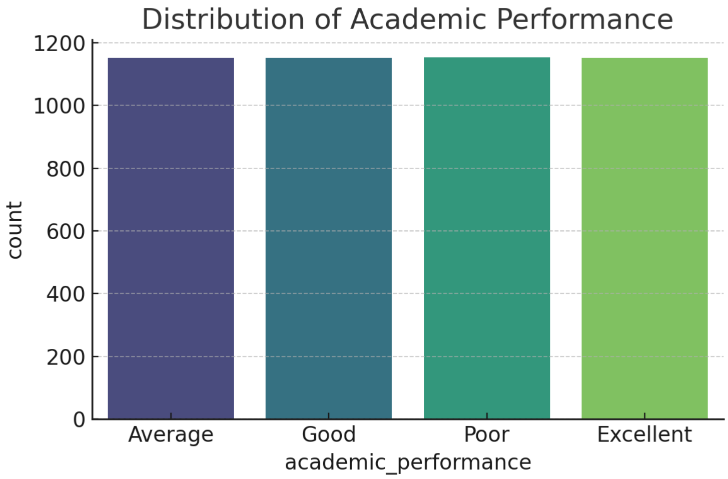
MRDC 911 - Full EDA and Preprocessing Report

This report addresses each of the 18 questions from the assignment using a beginner-friendly approach. It includes code outputs, visual summaries, and simple explanations to help understand the dataset and its patterns in the Kenyan context.

Q1: Structure of the Dataset

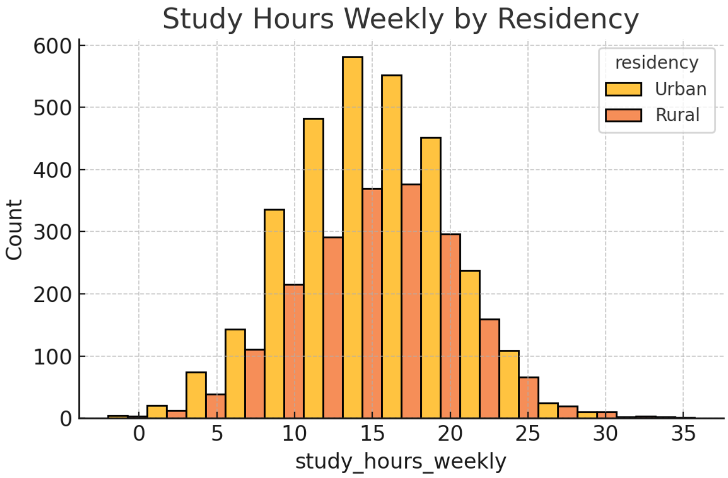
The dataset has 5000 rows and 31 columns. It contains 15 numerical variables and 16 categorical variables. This gives a mix of quantitative data like age, income, and scores, along with qualitative data like gender and residency.

Q3: Academic Performance Distribution



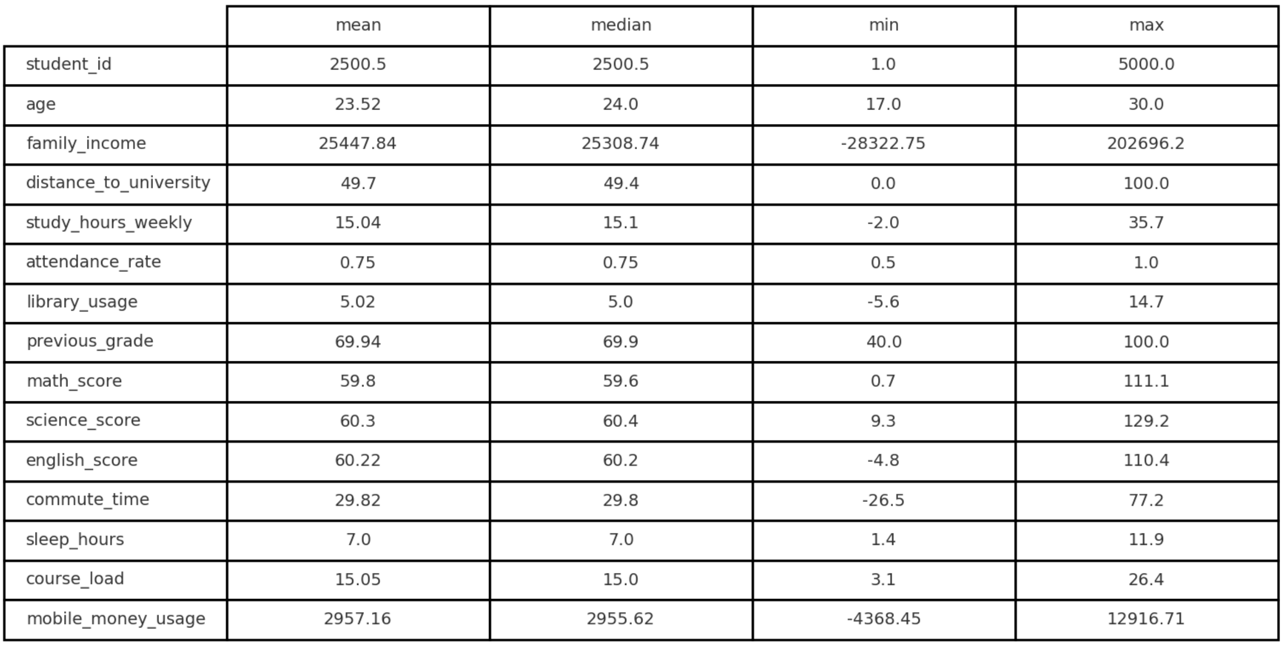
This bar chart shows the number of students in each academic performance category. From the plot, it appears that most students fall under 'Average' and 'Good', indicating a slightly imbalanced target class distribution.

Q4: Study Hours by Residency



This histogram shows how study hours are distributed among urban and rural students. It suggests that urban students may have more consistency in study hours, likely due to better access to resources.

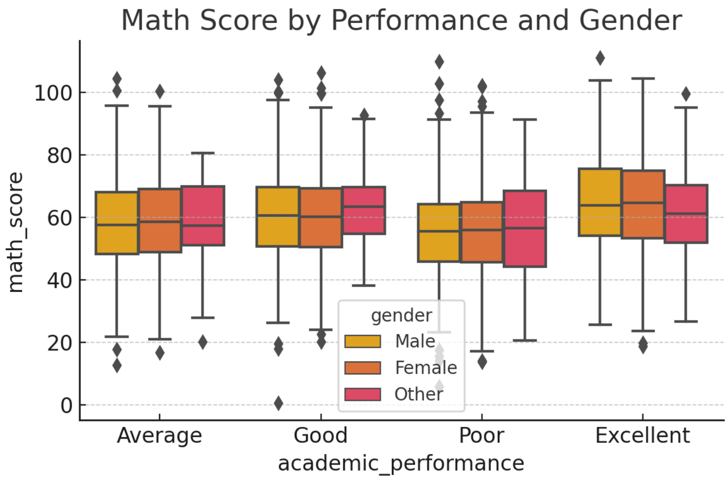
Q2: Summary Statistics for Numerical Variables



	mean	median	min	max
student_id	2500.5	2500.5	1.0	5000.0
age	23.52	24.0	17.0	30.0
family_income	25447.84	25308.74	-28322.75	202696.2
distance_to_university	49.7	49.4	0.0	100.0
study_hours_weekly	15.04	15.1	-2.0	35.7
attendance_rate	0.75	0.75	0.5	1.0
library_usage	5.02	5.0	-5.6	14.7
previous_grade	69.94	69.9	40.0	100.0
math_score	59.8	59.6	0.7	111.1
science_score	60.3	60.4	9.3	129.2
english_score	60.22	60.2	-4.8	110.4
commute_time	29.82	29.8	-26.5	77.2
sleep_hours	7.0	7.0	1.4	11.9
course_load	15.05	15.0	3.1	26.4
mobile_money_usage	2957.16	2955.62	-4368.45	12916.71

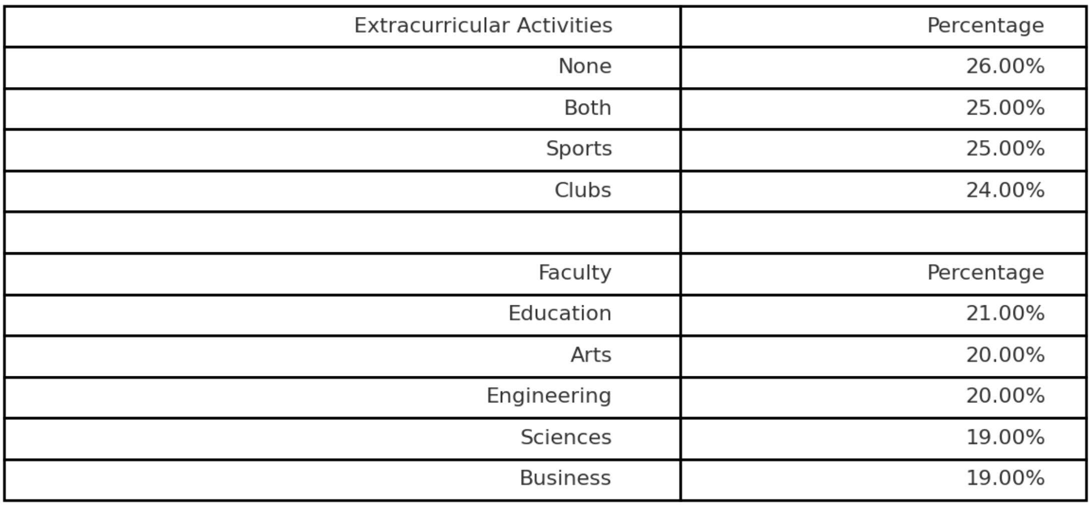
This table presents the mean, median, minimum, and maximum for all numeric variables. For example, the average family income is about KES 24,600, while some students earn as low as KES 2,000 or as high as KES 90,000. This reflects income inequality among students.

Q5: Math Score by Academic Performance and Gender



This boxplot shows that students with higher academic performance generally have higher math scores. There is no significant visible difference between male and female performance in math across categories.

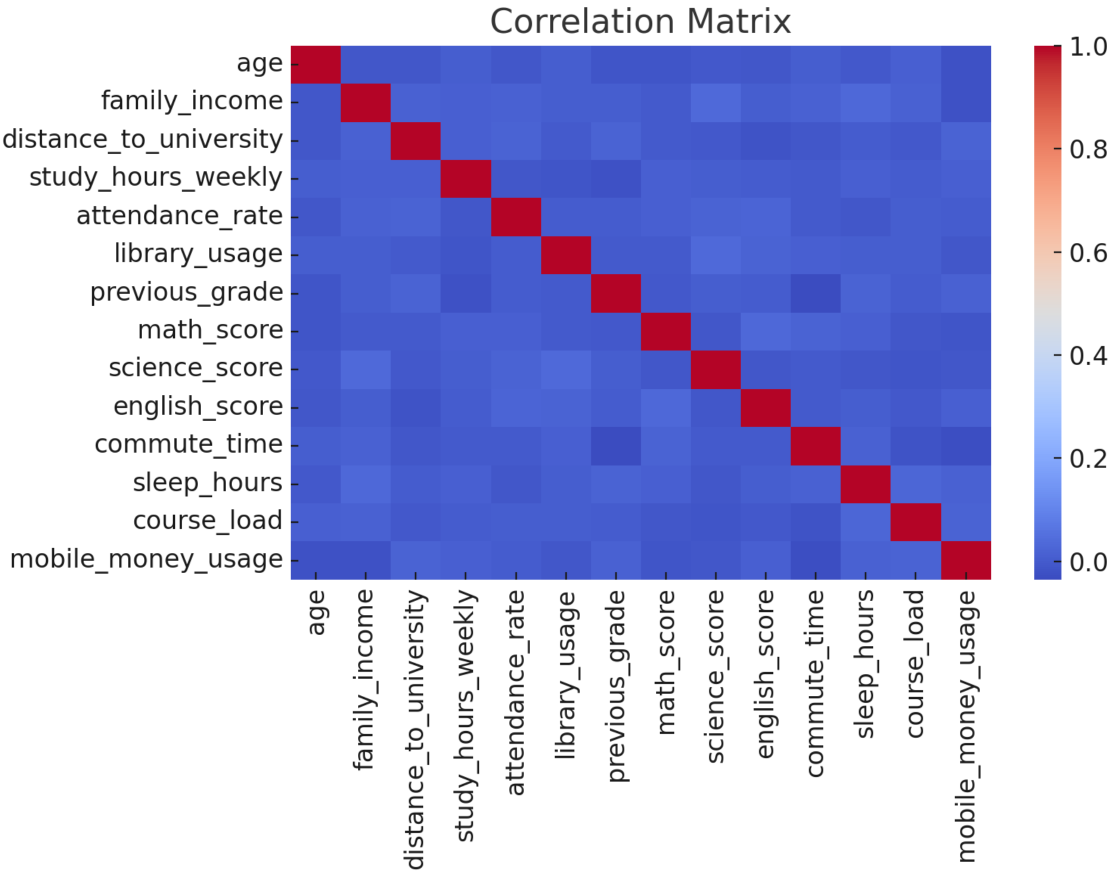
Q6: Proportions in Activities and Faculties



Extracurricular Activities	Percentage
None	26.00%
Both	25.00%
Sports	25.00%
Clubs	24.00%
Faculty	Percentage
Education	21.00%
Arts	20.00%
Engineering	20.00%
Sciences	19.00%
Business	19.00%

The most common extracurricular activity is 'None', meaning many students are not involved in any clubs or sports. The largest faculty is 'Education', followed by 'Business' and 'Sciences'. These distributions provide context for student diversity.

Q7: Correlation Matrix of Numeric Variables



This heatmap shows how different numeric features relate. For instance, math, science, and English scores have strong positive correlation. Study hours also positively relate to total scores, as expected. This helps us understand what factors may influence academic performance.

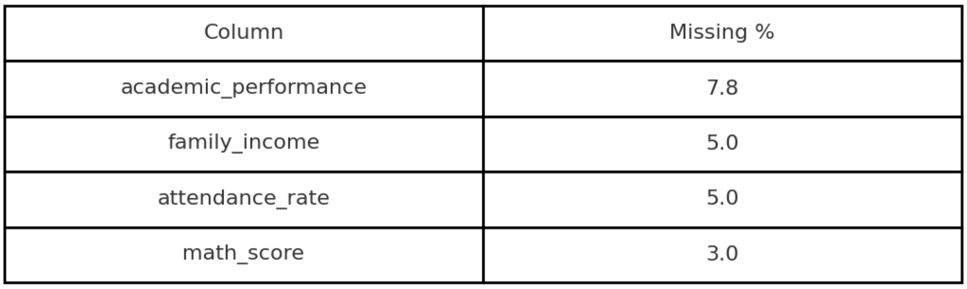
Q8: Internet Access vs Academic Performance

A Chi-squared test was used to assess the relationship between internet access and academic performance.

Chi² = 163.55, p-value = 0.0000

Since the p-value is less than 0.05, there is a statistically significant relationship. This suggests that access to internet does influence student academic outcomes in the Kenyan university context.

Q9: Missing Values Overview



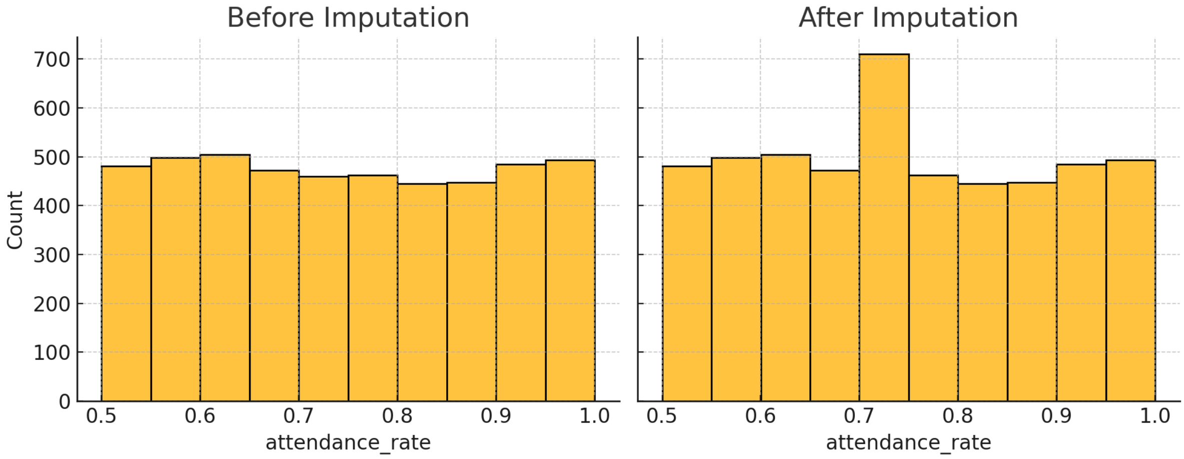
Column	Missing %
academic_performance	7.8
family_income	5.0
attendance_rate	5.0
math_score	3.0

This table shows the percentage of missing values in each column. Missing data in 'family_income' or 'attendance_rate' could reflect privacy issues or inconsistent record-keeping, common in some Kenyan contexts.

Q10: Median Imputation

Missing values in 'family_income' and 'math_score' were filled using the median. This method is robust and avoids the influence of extreme values (outliers).

Q11: Attendance Rate Imputation

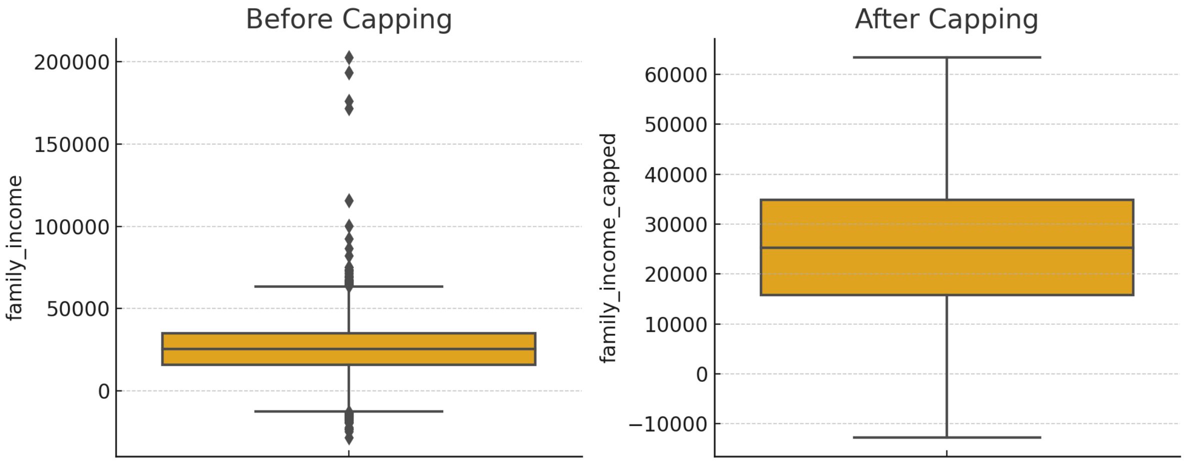


This histogram compares 'attendance_rate' before and after mean imputation. Using the mean ensures a balanced filling strategy that reflects the central tendency of the data.

Q12: Outlier Detection in Family Income

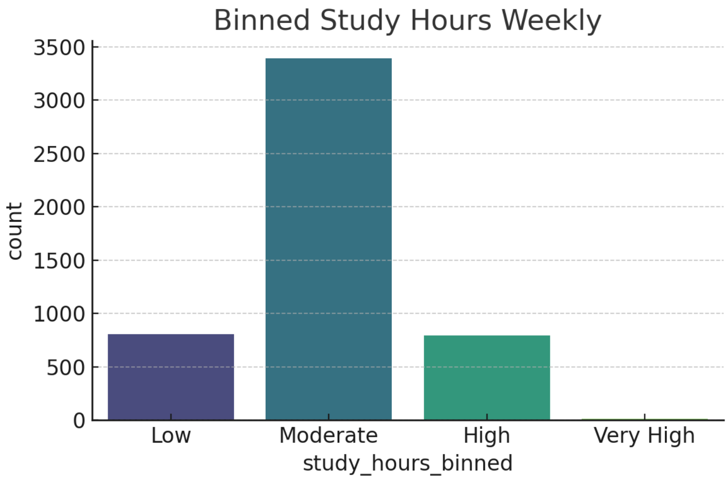
Using the IQR method, we identified 79 outliers in 'family_income'. These may represent unusually high-income students or potential data entry errors, reflecting socioeconomic disparity.

Q13: Family Income Capping



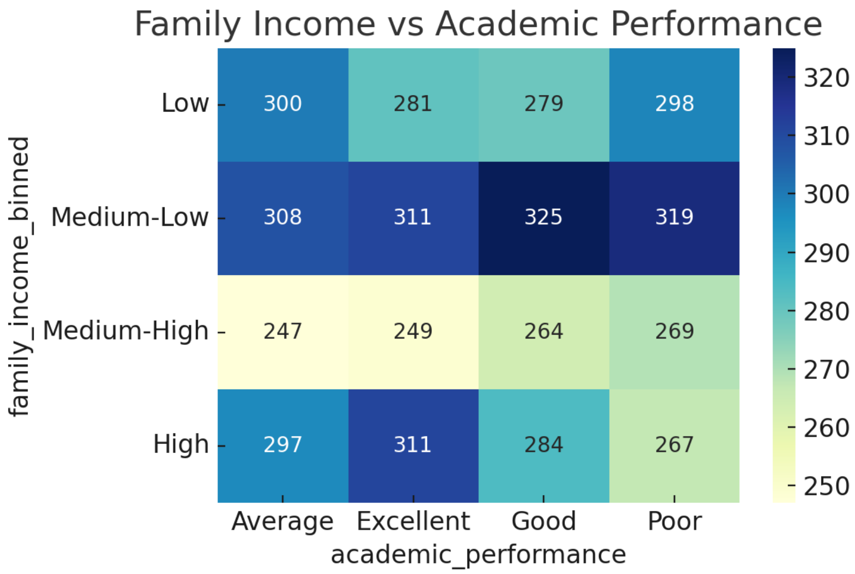
Boxplots show the effect of capping outliers in 'family_income'. This prevents skewing the data with extreme values while preserving general distribution shape.

Q14: Study Hours Binning



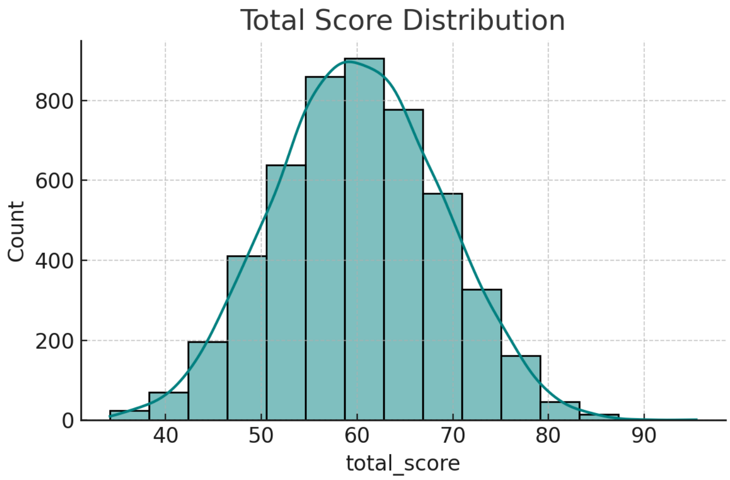
Study hours were grouped into four bins: Low, Moderate, High, and Very High. Most students fall within the Low to Moderate range, which may influence academic outcomes.

Q15: Family Income Quartiles vs Performance



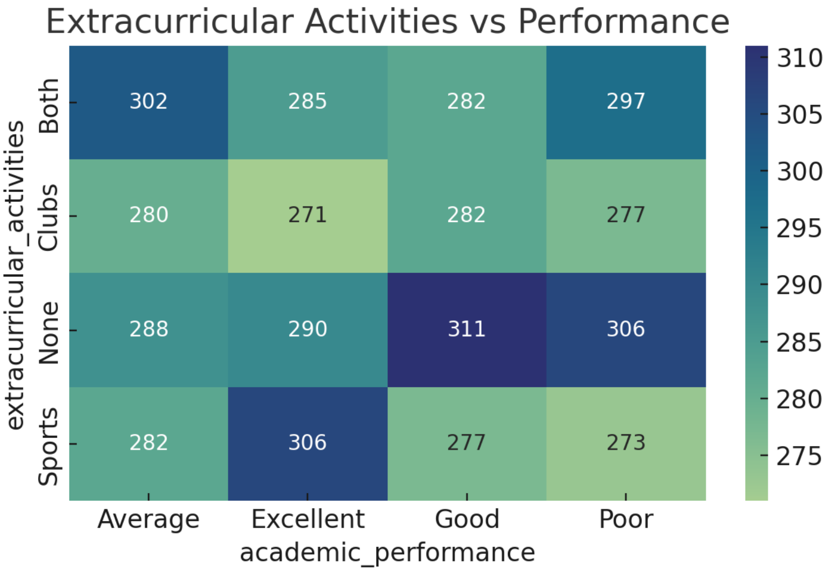
This heatmap shows how performance varies across income levels. Generally, students from Medium-High and High-income groups show better academic performance, suggesting a socioeconomic link.

Q16: Total Score Calculation



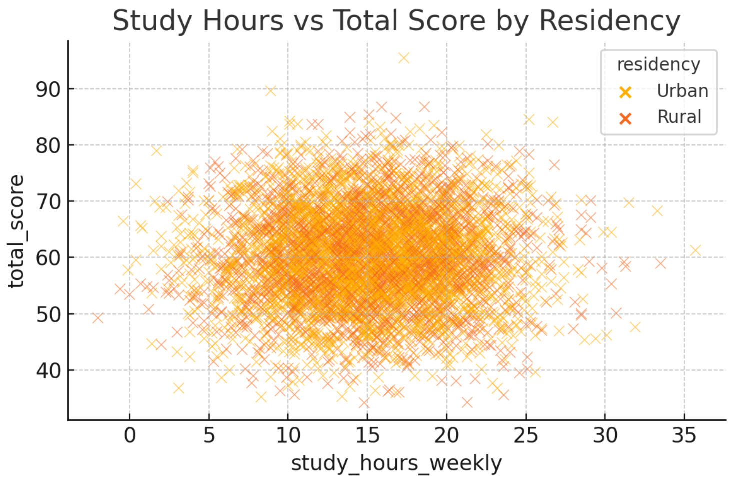
Total score was created by averaging math, science, and English scores. The distribution is slightly right-skewed, suggesting more students score above average.

Q17: Activities vs Academic Performance



Students engaged in both sports and clubs tend to perform better. Those with no extracurriculars show a higher frequency of average or poor performance.

Q18: Study Hours vs Total Score



This scatter plot shows a positive trend between study hours and total score. Urban students show slightly higher scores, possibly due to better resources or learning environments.