Name: Zalo Austine

ADM: 25ZAD111180

MRDC 911: Data Science & Computational Intelligence

Githublink:https://github.com/zaloAustine/Master-Data-Science-Exploratory-data

Assignment 1- EDA and Data Preprocessing on Kenyan Student Dataset

1. I loaded 5 000 student records and confirmed the dataset has 15 numeric and 16 categorical columns

   Numeric variables (15):

   student_id, age, family_income, distance_to_university, study_hours_weekly, attendance_rate, library_usage, previous_grade, math_score, science_score, english_score, commute_time, sleep_hours, course_load, mobile_money_usage

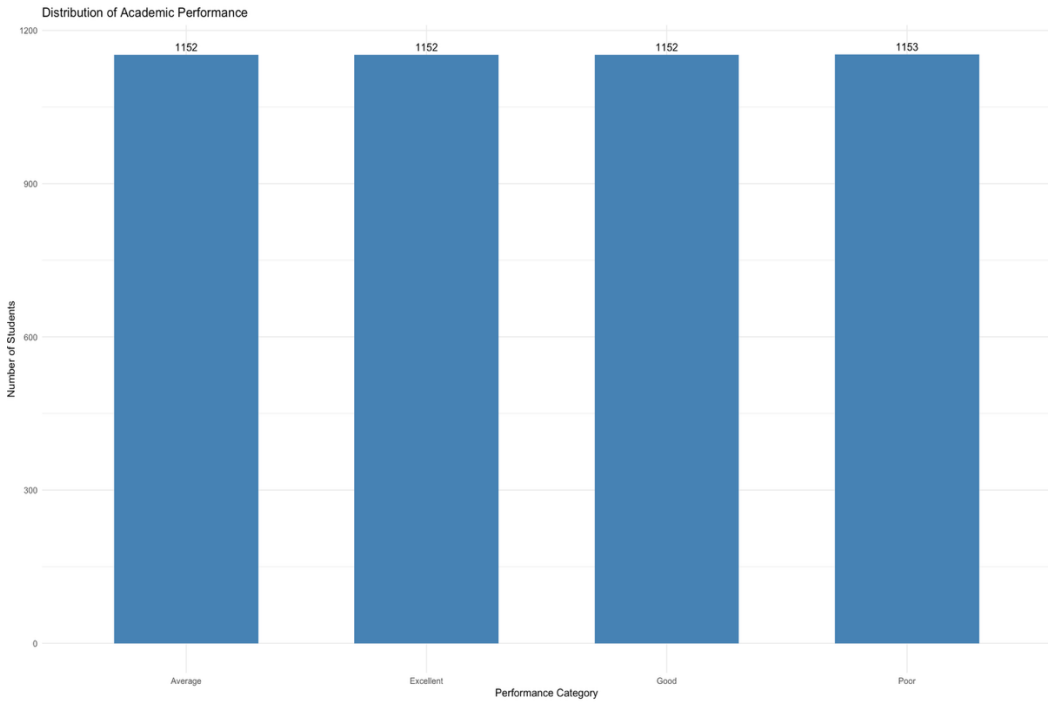
   Categorical variables (16):

   gender, residency, socioeconomic_status, parental_education, extracurricular_activities, internet_access, device_ownership, study_group_participation, scholarship_status, campus_housing, part_time_job, stress_level, faculty, year_of_study, health_status, academic_performance

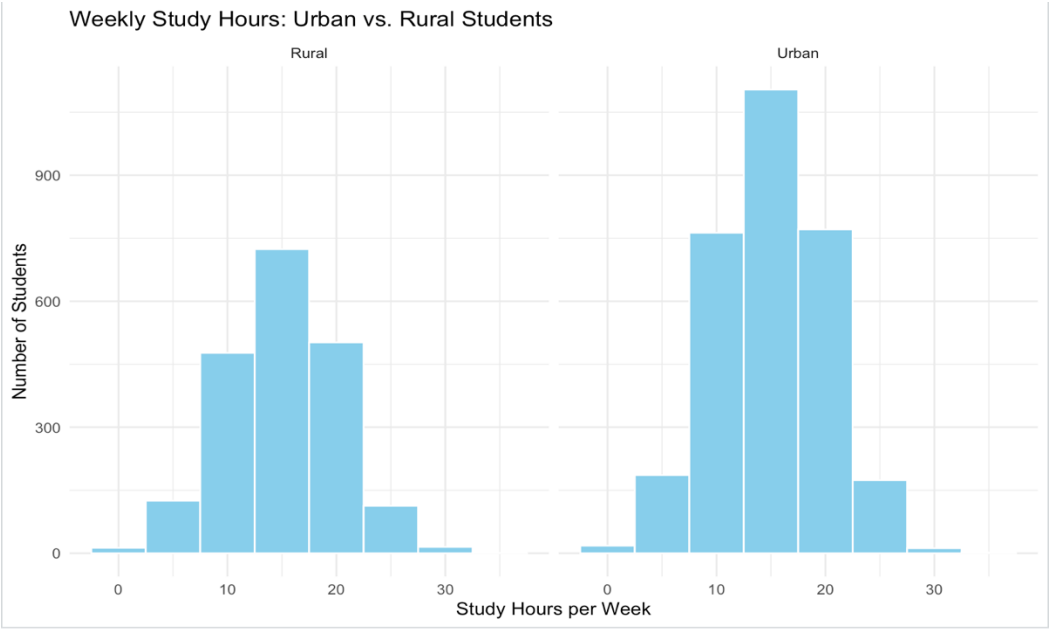
2. Numeric features such as family_income (mean = 25 448 KES, median = 25 309; SD = 16 039) and study_hours_weekly (mean/median = 15 hrs; SD = 5) show moderate central tendencies but wide spreads. The large difference between min (–28 323 KES) and max (202 696 KES) for income highlights outliers, while negative values in study hours point to data-entry errors.

	variable	count	missing	mean	median	min	max	sd
	student_id	5000	0	2500.50	2500.50	1.00	5000.00	1443.52
	age	5000	0	23.52	24.00	17.00	30.00	3.77
	family_income	4750	250	25447.84	25308.74	-28322.75	202696.20	16038.67
	distance_to_university	5000	0	49.70	49.40	0.00	100.00	28.83
	study_hours_weekly	5000	0	15.04	15.10	-2.00	35.70	5.03
	attendance_rate	4750	250	0.75	0.75	0.50	1.00	0.15
	library_usage	5000	0	5.02	5.00	-5.60	14.70	3.01
	previous_grade	5000	0	69.94	69.90	40.00	100.00	17.45
	math_score	4850	150	59.80	59.60	0.70	111.10	15.03
0	science_score	5000	0	60.30	60.40	9.30	129.20	15.00
1	english_score	5000	0	60.22	60.20	-4.80	110.40	15.07
2	commute_time	5000	0	29.82	29.80	-26.50	77.20	15.19
3	sleep_hours	5000	0	7.00	7.00	1.40	11.90	1.52
4	course_load	5000	0	15.05	15.00	3.10	26.40	3.02
5	mobile_money_usage	5000	0	2957.16	2955.62	-4368.45	12916.71	1990.83

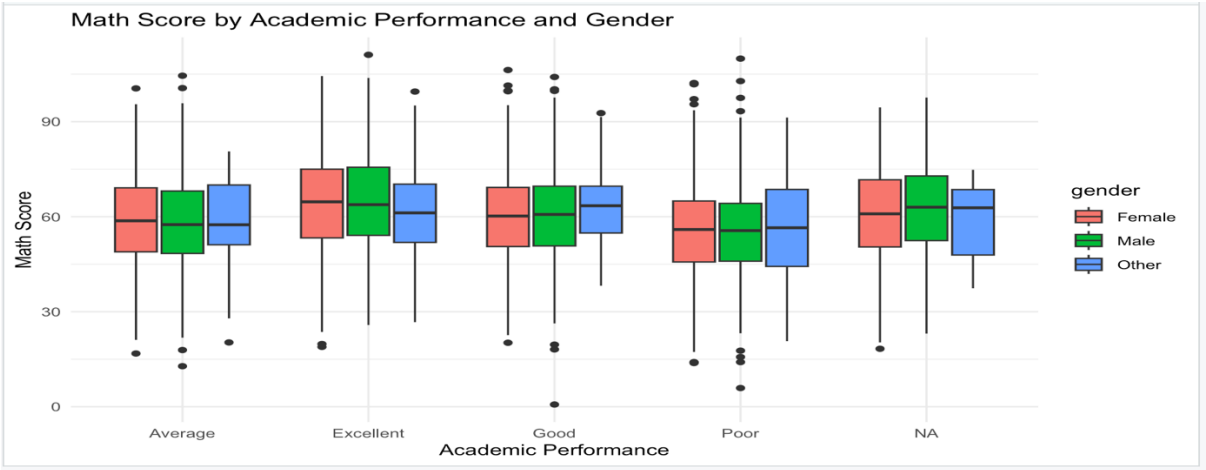
3. Yes each of the four categories has essentially the same count (1 152–1 153 students). With roughly 25% in each bucket, the target is very well balanced,



4. 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.



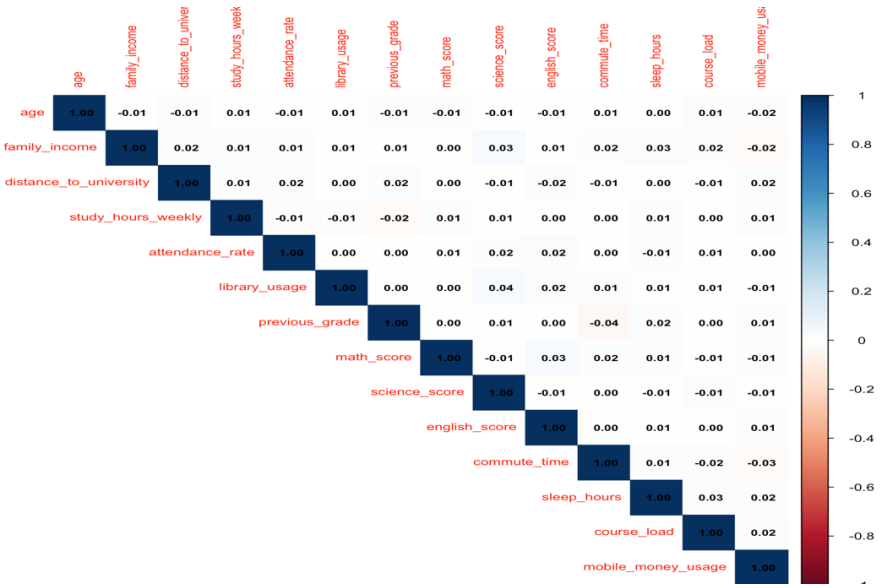
5. Math scores climb steadily from Poor to Excellent, matching the performance labels. Within each category, Male and female score almost the same showing gender has virtually no effect though Male performm slightly better. Students labeled Poor show a widespread and some very low scores, while Excellent students cluster tightly at the top.



6. The proportions for each extracurricular category are almost identical around 24–26% of students fall into each group: **None:** 25.8%. **Both (Sports + Clubs):** 25.1% **Sports only:** 24.9% **Clubs only:** 24.3%

   All five faculties recruit very similar shares of students each hovers around roughly one-fifth of the population but education being the Highest: **Education:** 20.6% **Arts:** 20.5% **Engineering:** 20.1%**Sciences:** 19.5% **Business:** 19.3%

7. Nothing really stands out every correlation is tiny. The largest is around +0.04 (library usage vs. science score), followed by about +0.03 for things like family income vs. math score or income vs. commute time. All other pairs sit near zero, so there are no meaningful linear relationships.



8. *data: data$internet_access and data$academic_performance*
   X-squared = 163.55, df = 3, p-value < 2.2e-16

   A χ^2 of 163.6 with p-value < 2.2e-16 means link between internet access and grades are far beyond random chance. Students without home internet show up in **Poor** much more than you'd expect, while those with internet are over-represented in **Excellent**.

9. **Family income (5%)**: Some students or families skip the question or can't pin down informal earnings, especially in rural areas without formal pay records.

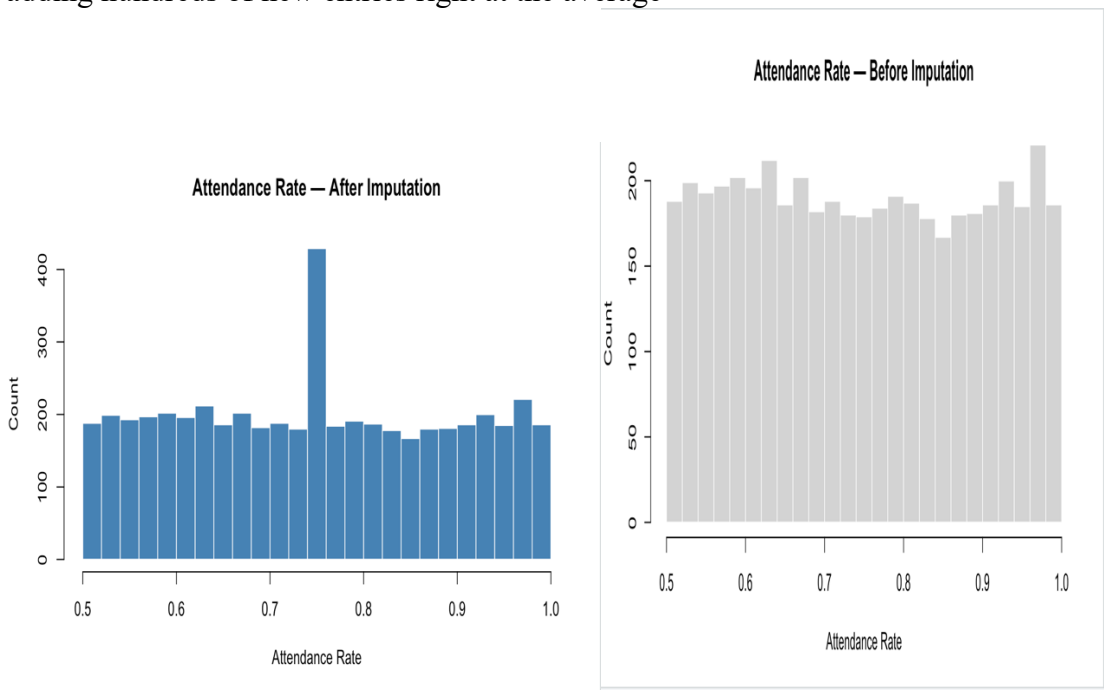
   Attendance rate (5%): Many schools still use paper registers that get lost or never digitized, so some attendance data never makes it into the system.

   Math score (3%): A few students miss the exam (illness, transport issues) or scores aren't entered correctly, leaving blank spots.

   Academic performance (7.8%): If final grades aren't approved, appeals are pending, or transcripts aren't updated, the performance label stays empty.

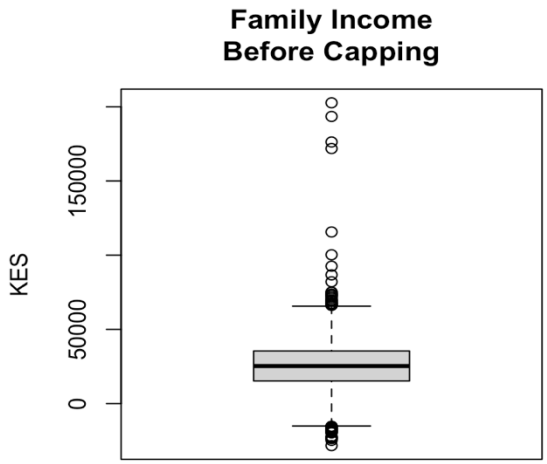
10. We will fill any gaps in **family_income** and **math_score** with that variable's **middle value** (its median).The median is the "middle" student's income or score, so it isn't skewed by a few very high or very low numbers meaning our replacement is more representative of a typical student.

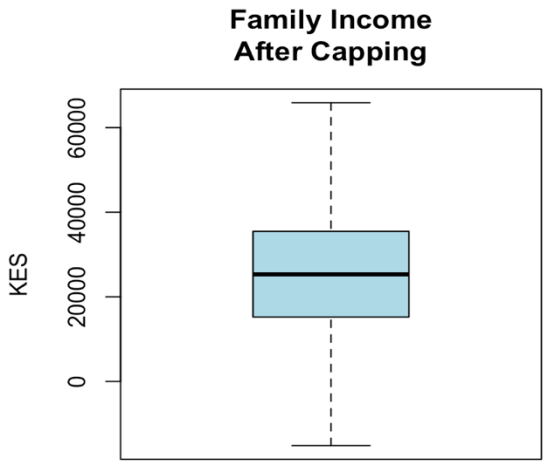
11. Before imputation, the attendance-rate histogram is a smooth and uniformly spread with no single value dominating After we fill in the Nulls with the mean you see a tall spike at that exact 75% mark, pushing the rest of the distribution unchanged but adding hundreds of new entries right at the average



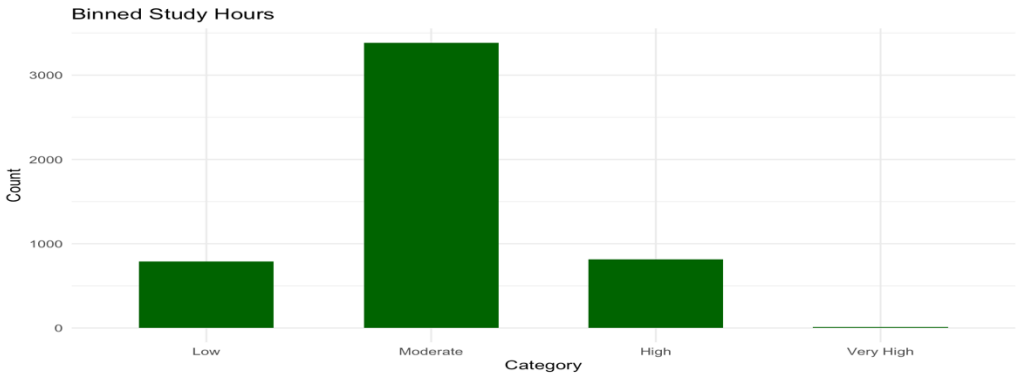
12. We found **56** outliers in total **25** extremely low incomes and **31** extremely high ones. The low-end figures are either typos or reflect very poor households, while the high-end ones come from wealthy families.

13. Before capping, 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.





14. 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 (<10 hrs)** 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.



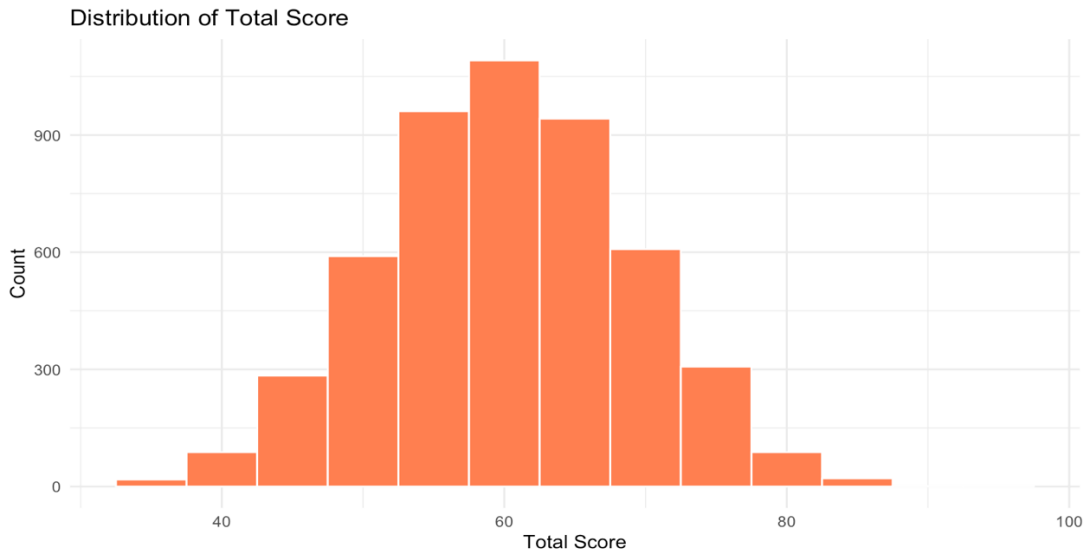
15. Higher-income students are a bit more likely to end up in the Excellent performance group and a bit less likely to be in poor performance. The middle two income groups sit right around 25–26% for each performance level.

```
> ct <- table(inc_bins, data$academic_performance)
> print(ct)

inc_bins       Average Excellent Good Poor
  Low              286       264  268  281
  Medium-Low       258       272  282  278
  Medium-High      268       259  280  278
  High             276       301  268  258
> print(round(prop.table(ct, 1), 2))

inc_bins       Average Excellent Good Poor
  Low             0.26      0.24 0.24 0.26
  Medium-Low      0.24      0.25 0.26 0.26
  Medium-High     0.25      0.24 0.26 0.26
  High            0.25      0.27 0.24 0.23
>
```

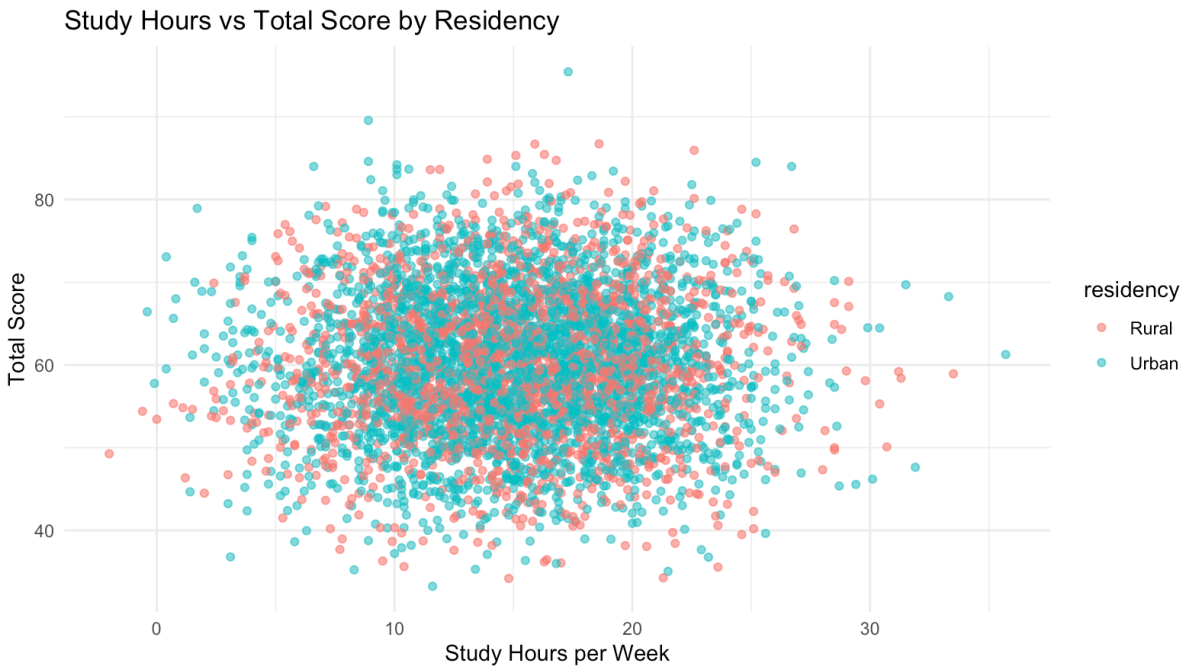
16. Most students cluster around a **total_score** of **60–70**, with a roughly bell-shaped curve and only a few at the very low or very high ends. This tells us that overall performance centers around two-thirds of full marks, making total_score a good single metric of academic strength.



17. You'll see that students who don't take part in any activities are over-represented in the Poor and Average groups, while those in **Sports only** or **Both (Sports + Clubs)** show higher counts in the Good and Excellent categories

	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. 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.



Summary

When it comes to study habits, urban students generally put in more consistent hours, often hitting around 15 to 20 hours a week, and they tend to score higher overall. In contrast, rural students show a broader range in their study times, with many putting in less than 10 hours and scoring lower. This difference likely due to better access to study resources, reliable electricity, and internet connectivity in urban areas.

Speaking of internet access, students who have it at home are much more likely to be among the Excellent performers, while those without it are often found in the Poor category. This highlights just how crucial digital access is for research, assignments, and studying it's a key factor in achieving academic success.

Lastly, there's a clear link between family income and academic performance. The highest-income quartile boasts a larger percentage of Excellent students (about 27%) compared to just 24% in the lowest quartile, and they have fewer “Poor” performers (around 23% versus 26%). In Kenya, where many people depend on irregular cash flows or informal jobs, having more financial resources means better access to study materials, transportation, and less economic stress, all of which contribute to improved academic outcomes.