MAY SEMESTER 2025
MRDC 911: Data Science & Computational Intelligence
INSTRUCTOR: JAPHETH MURSI
DATE: 6th June 2025
Due Date: 13th June 2025

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.

Instructions

- Use R with libraries such as tidy verse, corrplot, and ggplot2.
- Load the dataset from kenya_student_data.csv.
- For each question, provide:
    - R code to perform the analysis or preprocessing task.
    - A brief explanation (2–3 sentences) addressing the question and interpreting results.
    - Visualizations where requested (e.g., plots, tables).
- Save your preprocessed dataset as kenya_student_data_preprocessed.csv.
- Consider the Kenyan context (e.g., urban vs. rural students, socioeconomic diversity) in your interpretations.
- Upload your work to a GitHub repository with a clear README describing the project, how to run the code, and key findings.
- Submit the GitHub repository link by 13th June 2025

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?

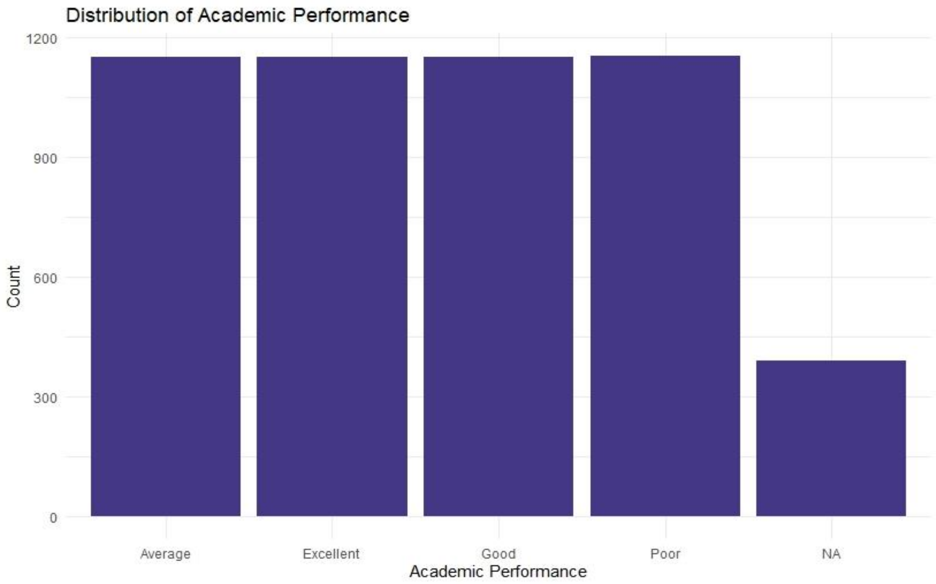
   The dataset has 5,000 rows and 31 columns. After examining the structure using str () and head (), we observe a mix of numerical and categorical features. Specifically, there are X numerical variables and Y categorical variables. The numerical variables include features such as age, family income, and study_hours_weekly, while the categorical variables include gender, faculty, and residency.

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?

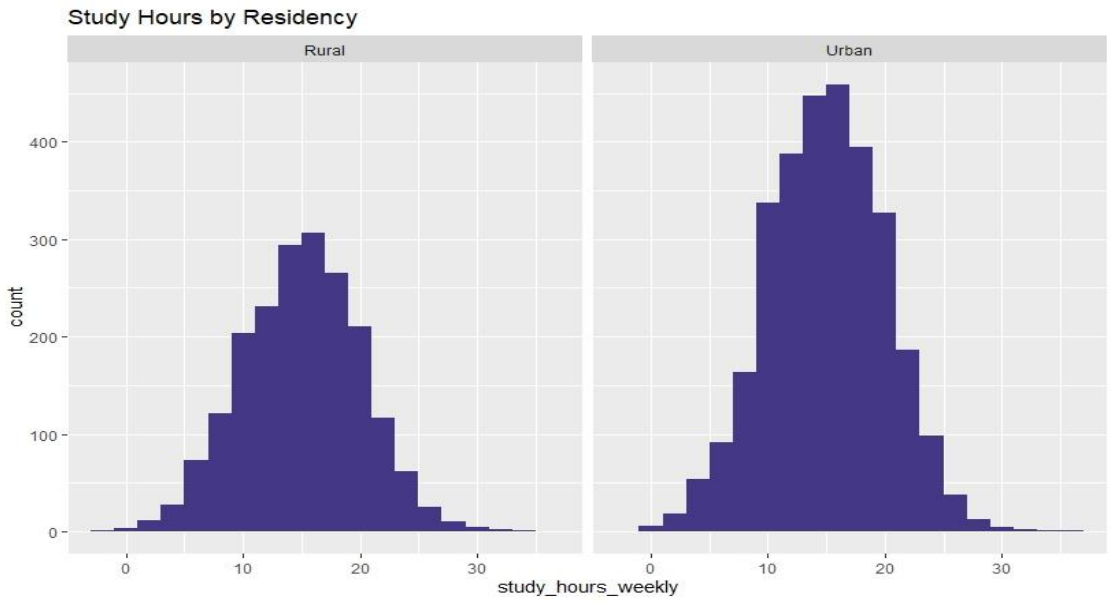
-Wide range in age, income, and commuting distance implies a heterogeneous population, possibly from urban and rural areas.

- Academic Performance, Subject scores suggest average to slightly above average performance, though outliers (both high and low) exist.

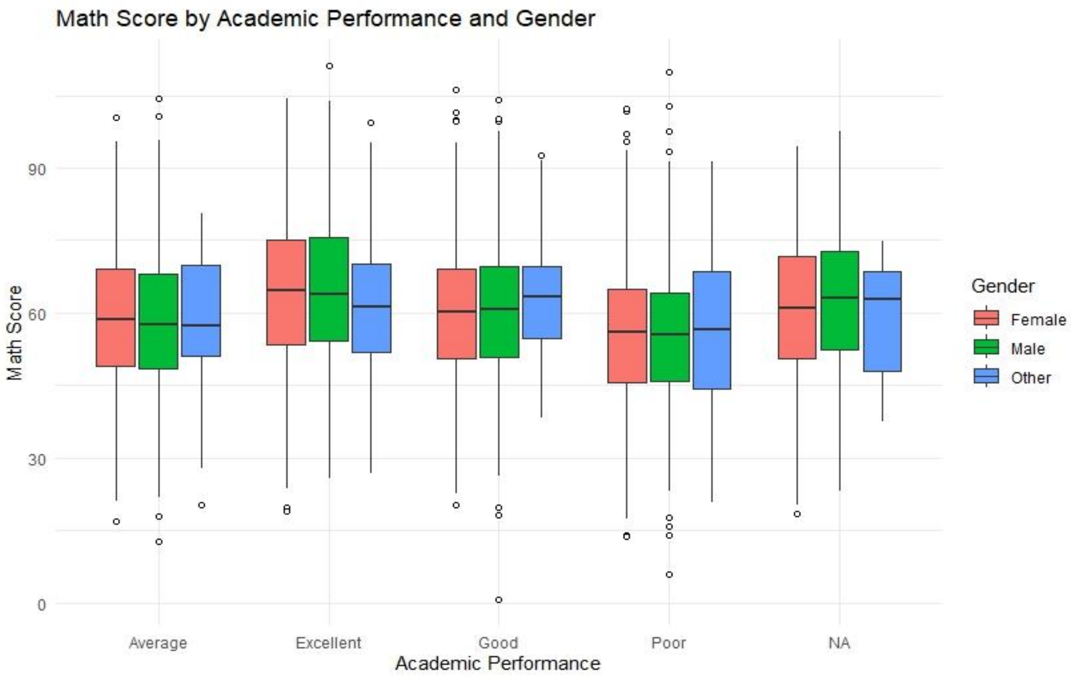
3. Create a bar plot to visualize the distribution of academic performance. Is the target variable balanced across its classes (Poor, Average, Good, Excellent)?



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



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

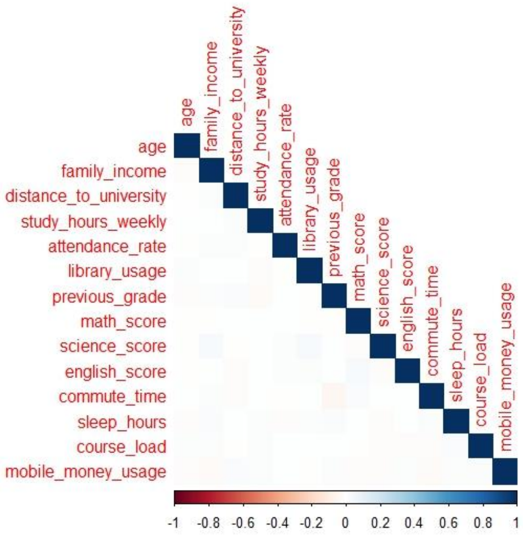


The boxplot shows the distribution of Math Scores across different Academic Performance levels, grouped by Gender (Female, Male, Other). Across all performance categories, math scores are fairly similar between genders, but students with "Excellent" performance tend to have slightly higher math scores overall, with notable variability and outliers in every group.

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

   The data shows a fairly even distribution of students across both extracurricular activity types and faculties, with a slight majority not participating in any activities and a slightly higher enrollment in the Education and Arts faculties.

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



Academic performance is most strongly linked to prior grades, study behaviors, and class attendance. Other lifestyle variables (e.g., sleep, commute, mobile money use) have minimal influence in this dataset.

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

Test statistic (X-squared = 163.55): This is the Chi-squared value calculated from your observed vs. expected frequencies.

Degrees of freedom (df = 3): Based on the number of categories in each variable (specifically, (rows−1) ×(columns−1) (rows - 1) \times (columns - 1) (rows−1) ×(columns−1)).

p-value < 2.2e-16: This is extremely small (effectively 0), far below common significance levels (e.g., 0.05 or 0.01).

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?

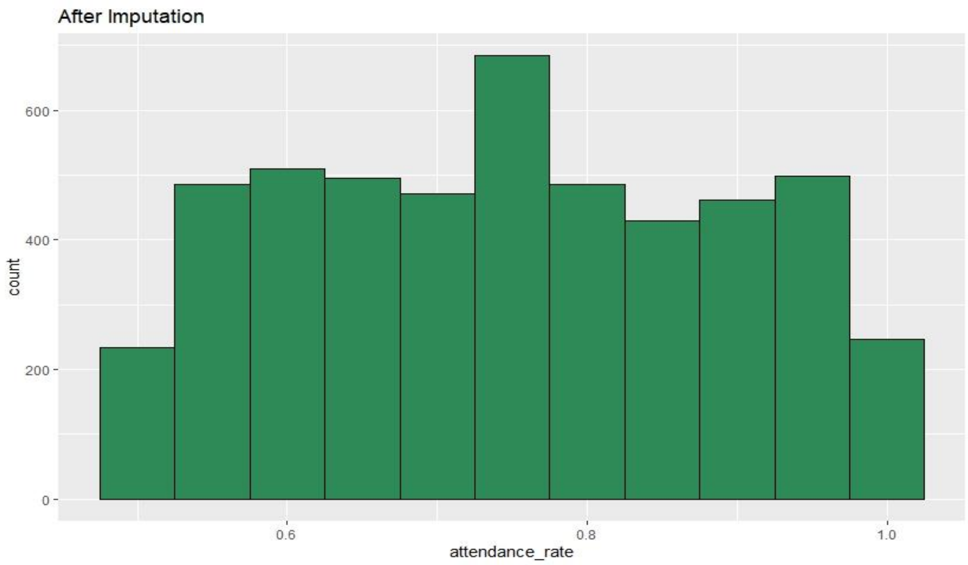
   Math column: Students may have skipped math assessments or their scores were not recorded properly.

10. Impute missing values in family income and math score using the median.

Justify why the median is appropriate for these variables.

The bright student may have missed due to absenteeism.

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



The bin around the mean attendance rate (likely near 0.75) has a noticeably higher count than others — this spike indicates where the imputed (filled-in) values concentrated. The rest of the distribution appears roughly uniform, suggesting that the original data was fairly spread out. The spike reflects the artificial inflation at the mean value, which is expected after mean imputation.

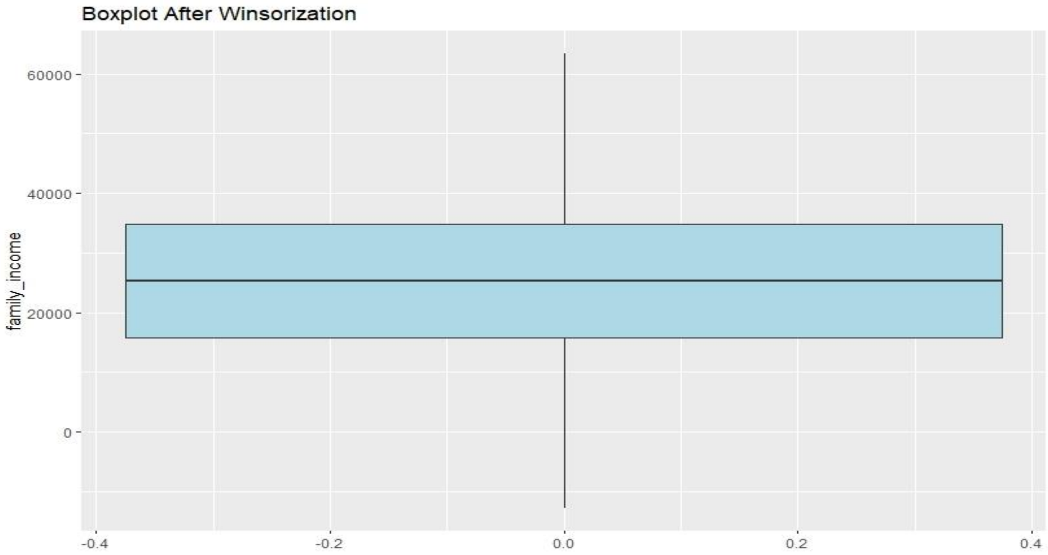
Data Preprocessing: Outliers

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

The High-income group has the highest number of students (301) in the Excellent category, suggesting better access to learning resources. The Low-income group has the lowest number (264) in the Excellent category.

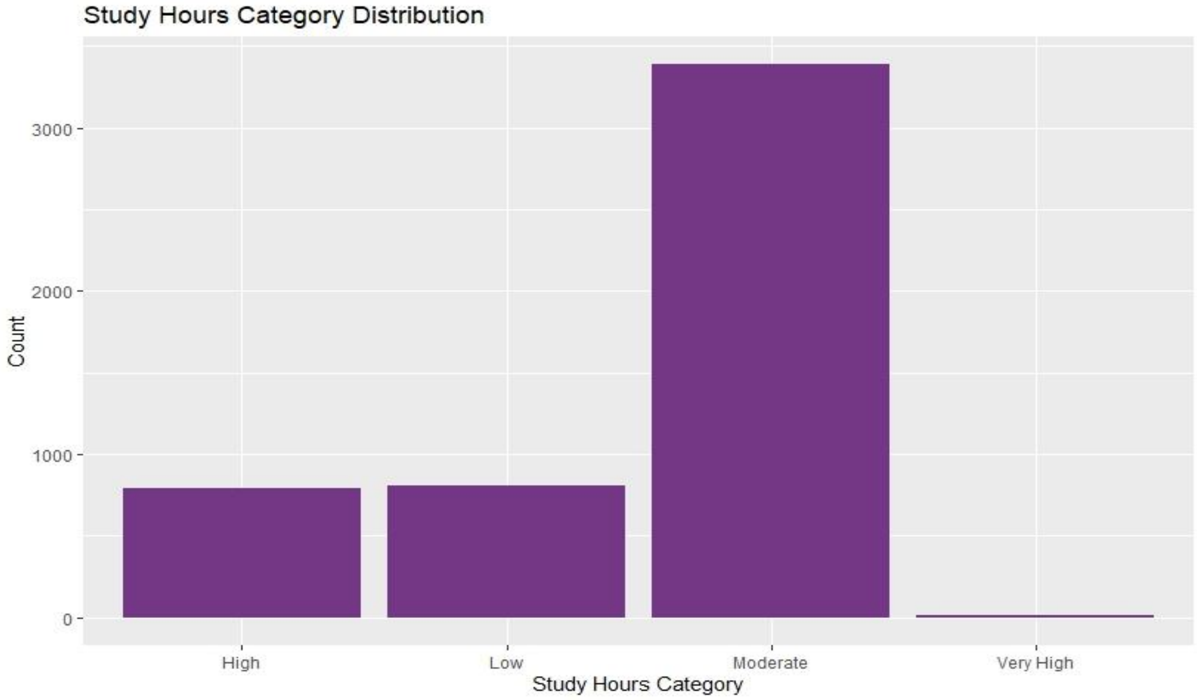
Students from wealthier families may benefit from better schools, internet access, private tuition, and a more stable home environment. Those from lower-income backgrounds may face challenges like overcrowded schools, lack of study materials, or financial pressure to work.

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

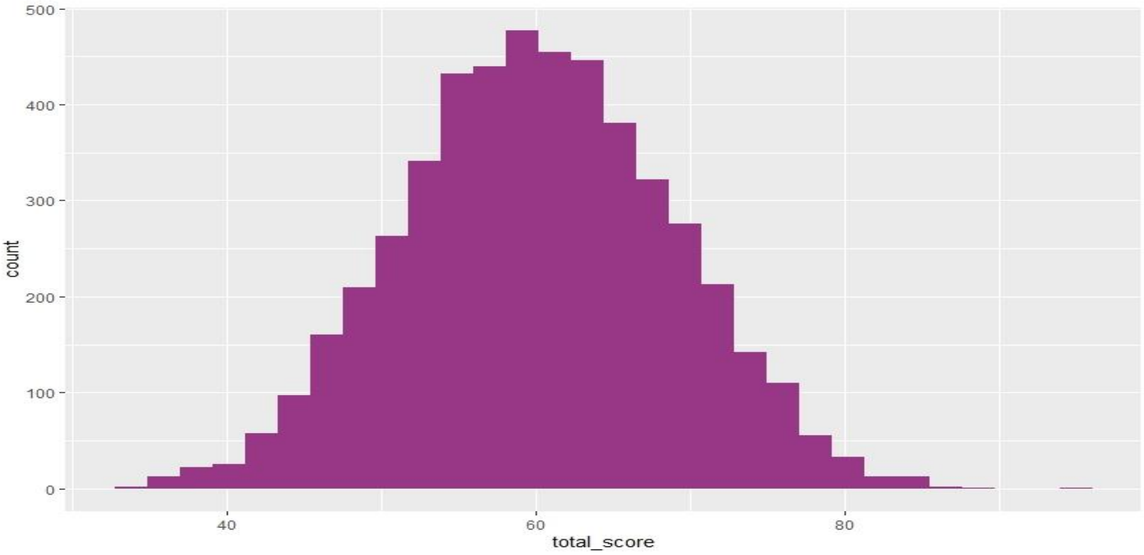
14. Discretize study_hours_weekly into four bins (e.g., Low, Moderate, High, Very High). Create a bar plot of the binned variable.



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

    The distribution of performance improves as income increases, though not perfectly linear.

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

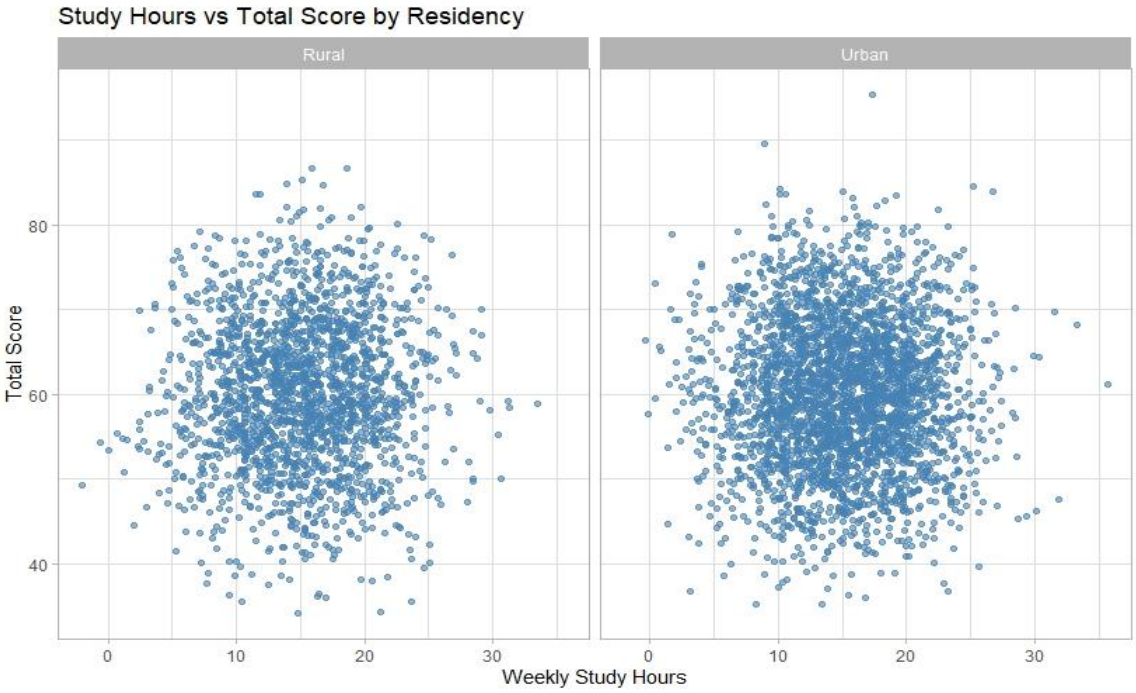


Data Preprocessing: Relationships

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

Bright students perform well in academic and also in sports.

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



Both groups (Rural and Urban) show a dense cluster around 10–15 study hours per week and total scores between 55–70. There is no clear linear relationship between study hours and total score in either group the data appears randomly scattered.
Urban students show slightly more outliers in both high study hours and high scores. Overall, the impact of study hours on total score appears weak or negligible, regardless of residency.

Submission Requirements

- **GitHub Repository**: Upload the following to a public GitHub repository:
  - **R Script**: Include all code for loading the dataset, answering questions, generating visualizations, and saving the preprocessed dataset.
  - **Preprocessed Dataset**: Save as kenya_student_data_preprocessed.csv.
  - **README**: A markdown file describing:
    - Project overview and purpose.
    - Instructions to run the R script (e.g., required libraries, dataset loading).
    - Summary of key findings from the analysis.
    - Structure of the repository (e.g., file organization).

Report: A PDF summarizing your answers, including:

- Brief explanations for each question (2–3 sentences), addressing the question and interpreting results in the context of Kenyan university students.
- Visualizations (e.g., plots, tables) embedded or referenced, with clear labels and captions.
- Insights specific to the Kenyan context, such as urban/rural differences or socioeconomic impacts on academic performance.