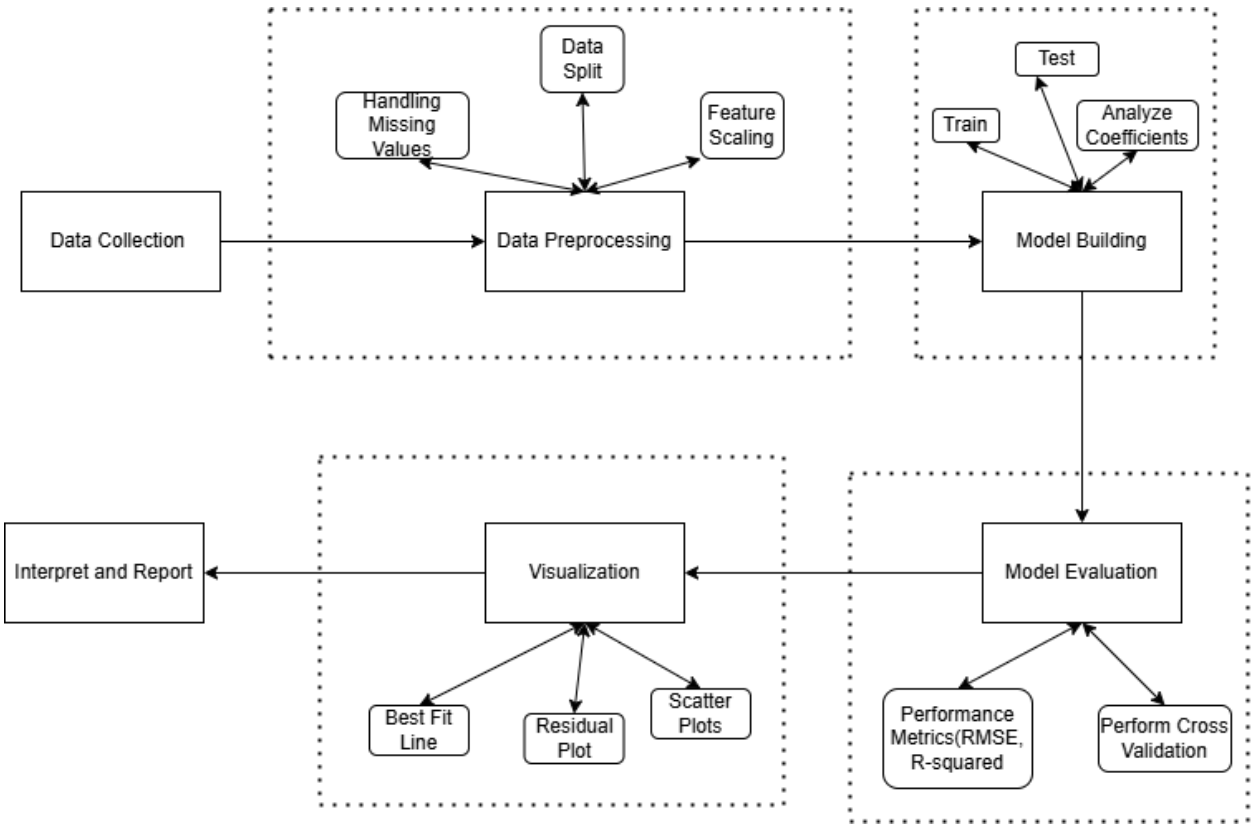
DSE 501 Term Project Proposal
Group 24 - Student Performance Analysis
Walker Mellon - 1219097204
Marie Rudasics - 1221746903
Raja Muhammad Shayan - 1232727147
Stavan Mehul Shah - 1233629139

Abstract
This study aims to understand the impact of demographic factors on student performance, focusing on math, reading, and writing scores. Using data on gender, race, parental education level, lunch program participation, and test preparation course completion, we analyze how socio-economic and demographic variables relate to academic outcomes. Statistical analysis allows us to identify trends and conclude how certain backgrounds influence performance in each subject. The insights gained can guide policymakers, educators, and parents in providing targeted support and resources, ultimately helping to reduce performance gaps and promote equitable educational opportunities for all students.

Methodology
This study attempts to explore the effect of demographic factors on students' achievement, interpreted as scores in mathematics, reading, and writing. Our dataset contains demographic variables: gender, race, parental education level, lunch program participation, and completion of test preparation courses that provide a foundation for examining socio-economic and demographic backgrounds relevant to academic outcomes. For such analysis, we will propose a methodology combining EDA with machine learning, selecting such models that balance interpretability with predictive accuracy.



Methodology Flowchart

Here, as we see in the above-given flowchart, we make use of linear regression as our main model for studying the relationship between demographic factors and students' performance in maths, reading, and writing scores. Linear regression is well-suited for the given analysis, as it provides a simple approach to identifying associations of independent demographic variables with outcomes on academic performance. At the same time, it enables us to interpret the extent to which each variable contributes to scores.

Methodology Steps:

Data Preprocessing:
We apply one-hot encoding to categorical demographic features (e.g., gender, race, parental education, lunch program, test preparation) to convert them into interpretable numerical inputs. Standard scaling is used to make feature magnitudes comparable, aiding model convergence and interpretability. The dataset is then split into training and testing sets (80:20) to build and evaluate the model's generalization on unseen data.

Model Building and Analysis of Student Performance:
We developed a linear regression model to examine the influence of demographic factors on students' performance scores. This model identifies the best-fit line that minimizes residuals, revealing general trends. Coefficient interpretation allows us to understand each factor's effect; for example, a positive coefficient for parental education implies higher performance in math with higher education levels. Significant factors are identified by low p-values, highlighting those that meaningfully affect scores.

Visualization:
Residual plots validate the model's linearity, while scatter plots, best-fit lines, and coefficient bar plots visually show the strength and direction of each factor's influence. Additionally, distribution plots (histograms and box plots) illustrate score disparities by demographic groups.

Model Evaluation:
We assess model performance using R-squared, MAE, and RMSE, complemented by cross-validation for robustness. This structured approach helps us draw actionable insights on how demographics affect academic outcomes, contributing valuable findings for education policy and targeted interventions. The dataset includes demographic factors like gender, race, parental education, lunch program participation, and test preparation, offering a basis to analyze the socio-economic and demographic background's role in academic performance.

Timeline
Checkpoint 1: **Due Nov. 5th**

- Within this checkpoint we will obtain and clean the dataset. This involves handling missing values, encoding categorical variables, and scaling features.

Checkpoint 2: **Due Nov 9th**

- For this checkpoint we will visualize our distributions, correlations, and summary statistics for both demographic and performance variables. This is where our exploratory data analysis takes place. We will also discover initial insights on potential relationships between demographics and scores.

Checkpoint3: **Due Nov. 12th**

- Within this checkpoint we will perform model selection and initial model training, such as linear regression tests and any alternative models (such as decision trees for comparison if needed). From there we will evaluate the model fit and refine accordingly using our training data results.

Checkpoint 4: **Due Nov. 16th**

- This checkpoint consists of model tuning and cross validation. The emphasis within this checkpoint is to optimize model parameters and perform cross-validation in order to ensure robustness of our chosen model. We will also perform statistical tests on coefficients and interpret their significance results. From there we will revisit our exploratory data analysis to validate our insights.

Checkpoint 5: **Due Nov. 19th**

- In our final checkpoint we will generate final visualizations such as residual or distribution plots. We will then compile our insights on demographic impacts and discuss the potential for intervention. Using this information, we will finalize our report and ensure that our visuals accurately represent the insights we gained from our exploratory data analysis.

Data Investigation

Summary Statistics for Student Math Scores

Mean, Median, Standard Deviation, Variance, and 5-Number Summary

math_mean	math_median	math_sd	math_variance	math_min	math_q1	math_q3	math_max
66.089	66	15.16308	229.919	0	57	77	100

Summary Statistics for Student Reading Scores

Mean, Median, Standard Deviation, Variance, and 5-Number Summary

reading_mean	reading_median	reading_sd	reading_variance	reading_min	reading_q1	reading_q3	reading_max
69.169	70	14.60019	213.1656	17	59	79	100

Summary Statistics for Student Writing Scores

Mean, Median, Standard Deviation, Variance, and 5-Number Summary

writing_mean	writing_median	writing_sd	writing_variance	writing_min	writing_q1	writing_q3	writing_max
68.054	69	15.19566	230.908	10	57.75	79	100

Fig 1. Summary Statistics by Subjects

The summary statistics for math, reading, and writing scores indicate consistent central tendencies, with mean scores around 66 to 69 across subjects. The standard deviations are similar, ranging from 14.6 to 15.2, suggesting comparable variability in performance. Each subject also has a minimum score of 0-17 and a maximum of 100, showing a wide range in student performance across all three areas.

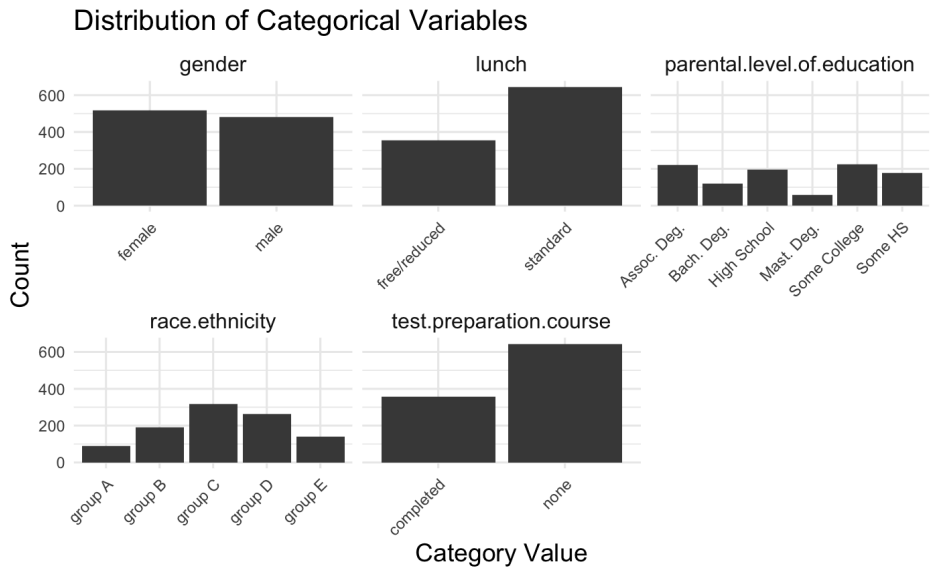


Fig 2. Bar Charts of Categorical Variables

These bar charts illustrate how categorical variables are distributed within the data set. This could provide insight into what demographic variables could drive student failures or successes within these three subjects. Imbalances of these variables could also provide explanations for potential sampling bias.

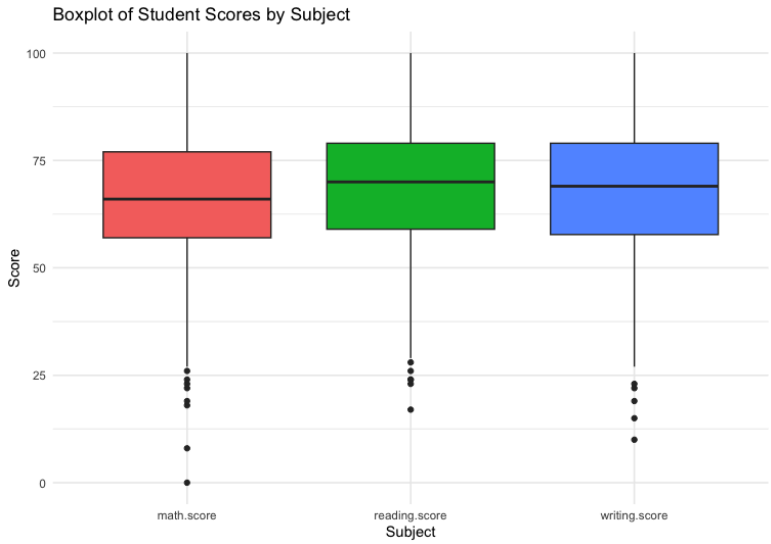


Fig 3. Boxplot of Student Scores by Subject

This boxplot shows the score distributions for math, reading, and writing. The median scores for each subject are relatively high, centered around 70–75. The range and spread are also similar, with some outliers at the lower end of the score spectrum. This suggests consistent performance across subjects, with a few students struggling significantly in each area.