**Data Analysis Process Summary**

1. **Importing Libraries:**

* Essential libraries (pandas, sklearn, matplotlib, seaborn) were used for data manipulation, analysis, and visualization.

1. **Loading the Dataset:**

* The dataset was loaded using pandas and includes the following columns: gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, and writing score.

1. **Exploratory Data Analysis (EDA):**

* **Data Structure:** 1000 rows and 8 columns.
* **Column Names:** The dataset contains both categorical variables (e.g., gender, race/ethnicity, parental level of education, lunch, test preparation course) and numerical variables (e.g., math score, reading score, writing score).
* **Data Types:** Categorical columns as objects, numerical scores as int64 or float64.
* **Unique Values Identified:**
  + **Gender:** 'female', 'male'
  + **Race/Ethnicity:** 'group A', 'group B', 'group C', 'group D', 'group E'
  + **Parental Level of Education:** 'associate's degree', 'bachelor's degree', 'high school', 'master's degree', 'some college', 'some high school'
* **Summary Statistics:**
  + Math score: mean = 66.13, std = 14.98, min = 0, max = 100.
  + Reading score: mean = 69.17, std = 14.60, min = 17, max = 100.
  + Writing score: mean = 68.05, std = 15.20, min = 10, max = 100.
* **Missing Data:** 39 missing values in the math score column.
* **Duplicate Data:** No duplicate rows were found in the dataset.

1. **Data Cleaning:**

* **Handling Missing Values:** Replaced missing values in math score with the mean using SimpleImputer.
* **Dropping Columns:** Removed unnecessary columns like lunch and test preparation course.
* **Label Encoding:**
  + Applied LabelEncoder to encode categorical columns into numerical values.
  + Columns encoded: gender, race/ethnicity, and parental level of education.

**Key Outputs:**

* **Shape of the Data:** 1000 rows and 6 columns after dropping columns.
* **Missing Values:** All missing values were addressed.