Dataset Analysis Report

Student Performance Dataset

1. Introduction

The Student Performance Dataset is used to analyze the academic performance of students based on various demographic, social, and educational factors. The objective of this analysis is to understand the structure of the dataset, identify different data types, assess data quality, and evaluate its suitability for machine learning tasks.

2. Dataset Overview

- **Number of Records:** 1000
- **Number of Features:** 8
- **Dataset Type:** Structured tabular data
- **Source:** Kaggle – Students Performance in Exams

The dataset contains information about students' gender, parental education, lunch type, test preparation course, and their scores in math, reading, and writing.

3. Feature Description & Data Types

Feature Name	Data Type
gender	Categorical
race/ethnicity	Categorical
parental level of education	Ordinal
lunch	Categorical
test preparation course	Binary
math score	Numerical
reading score	Numerical
writing score	Numerical

4. Target Variable

- **Target Variable:** math score
  The math score is chosen as the target variable for predicting student academic performance.
- **Input Features:**
  gender, race/ethnicity, parental level of education, lunch, test preparation course, reading score, writing score.

5. Data Quality Analysis

- The dataset contains **no missing values**, making it clean and ready for analysis.
- Categorical variables contain limited and well-defined categories.
- Numerical score columns range between 0 and 100.
- Some categorical features may require **encoding** before applying machine learning models.

6. Statistical Summary

Using df.describe(), the dataset shows:

- Mean, minimum, maximum, and standard deviation for math, reading, and writing scores.
- Balanced distribution of scores with moderate variation.
- No extreme outliers detected in numerical features.

7. Data Imbalance

- Gender distribution is fairly balanced.
- Test preparation course shows slight imbalance between students who completed and did not complete the course.
- Score distributions are reasonably spread, making the dataset suitable for supervised learning.

8. Machine Learning Suitability

The dataset is well-suited for:

- **Regression tasks** (predicting scores)
- **Classification tasks** (pass/fail prediction after transformation)

Minimal preprocessing is required:

- Encoding categorical variables
- Feature scaling (optional)

9. Conclusion

The Student Performance Dataset is clean, well-structured, and suitable for machine learning applications. It provides meaningful features that influence student performance and serves as an excellent dataset for understanding data types, exploratory data analysis, and predictive modeling.