People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Mohamed El Bachir El Ibrahimi University of Borj Bou Arréridj
Faculty of Mathematics and Computer Science
Department of Computer Science


**Report**

Exploratory Data Analysis (EDA) Project

Specialty: M1 TIC

# THEME

# EDA On Students Performance Dataset

*Presented by:*

Zekhnine Bachir

Souici Mouhamed

*In front of the jury composed of:*

**President:** Dr. Laifa Meriem

**Examiner:** Dr. Khouadia Youcef

**2025/2026**

# Contents

# Chapter 1

# Introduction

The purpose of this report is to conduct an Exploratory Data Analysis (EDA) on the Student Performance Dataset in order to understand the factors that influence students' academic outcomes. The dataset contains demographic, socio-economic, and academic variables, which makes it suitable for analyzing patterns related to educational achievement.

The research question guiding this study is:

> **Which factors demographic, socio-economic, or academic have the strongest influence on student performance across mathematics, reading, and writing?**

This analysis aims to explore score distributions, relationships between features, and performance differences across student groups.

## 1.1 Data Description

The dataset includes approximately 1,000 students (rows), each described by several categorical and numerical variables. It contains demographic characteristics, family-related factors, and three academic subject scores.

## Main Features

- **gender**: male or female.
- **race/ethnicity**: Groups A–E (anonymous socio-cultural categories).
- **parental level of education**: highest education level (high school, associate, bachelor, master, etc.).

- **lunch**: standard or free/reduced.
- **test preparation course**: completed or none.
- **math score**, **reading score**, **writing score**: numerical values from 0 to 100.

## Data Quality and Cleaning

The dataset contains no missing values or invalid entries. Duplicates were checked and removed when present. Categorical variables were standardized (lowercase formatting), and numerical variables were validated. Two engineered features were created:

- **total_score**: sum of all three subjects.
- **average_score**: mean of the three subjects.

The dataset is clean, consistent, and ready for analysis.

## 1.2 Methodology

The EDA was structured into several steps:

## 1. Data Inspection

The dataset was loaded using Python and inspected via `head()`, `info()`, and summary statistics.

## 2. Data Cleaning

Missing values, duplicates, and formatting issues were checked. No major cleaning was required beyond standardizing categories.

## 3. Feature Engineering

New variables (`total_score` and `average_score`) were created to support deeper analysis.

## 4. Visual Exploration

Multiple visualizations were generated:

- Histograms for score distributions.
- Boxplots comparing performance across groups (gender, lunch, test preparation).
- Barplots for parental education.
- Correlation heatmap.

## 5. Statistical Analysis

Correlation coefficients were computed to quantify relationships between variables.

Sampling was considered but not necessary due to the small dataset size.

## 1.3 EDA Results

This section presents key findings from the visual and statistical analysis.

## Score Distributions

Reading and writing scores show higher averages than math scores. Histograms reveal that students tend to perform better in language-related subjects.

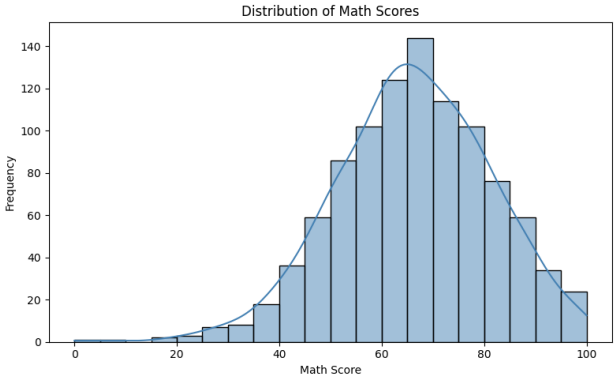


Figure 1.1: Distribution of Math Scores

# Group Comparisons

**Gender:** Females generally perform better in reading and writing, while males perform slightly better in math.

**Lunch Type:** Students with standard lunch achieve significantly higher scores, reflecting socio-economic effects.

**Test Preparation:** Students who completed the preparation course outperform those who did not across all subjects.

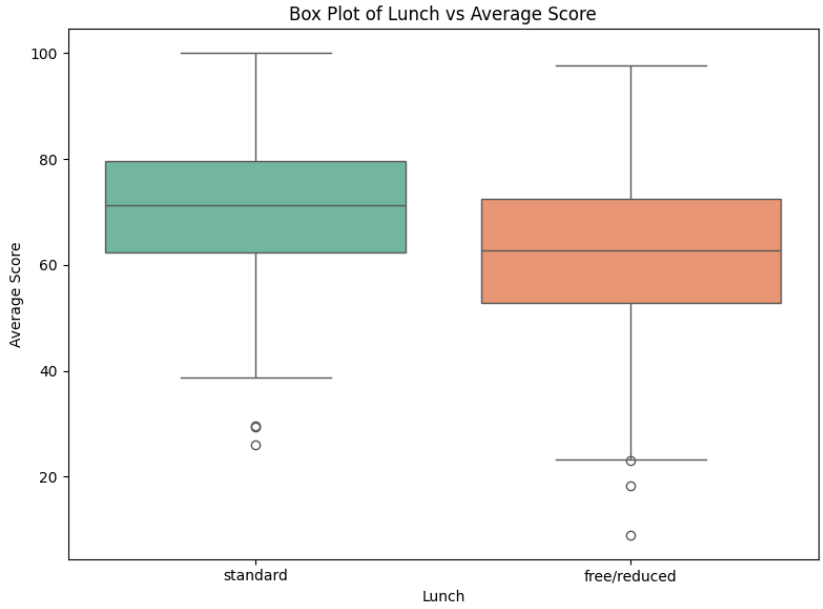


Figure 1.2: Lunch Type vs Average Score

# Correlation Analysis

Reading and writing scores are strongly correlated, while math shows moderate correlation with both. This suggests that literacy skills develop together, whereas math performance is more independent.

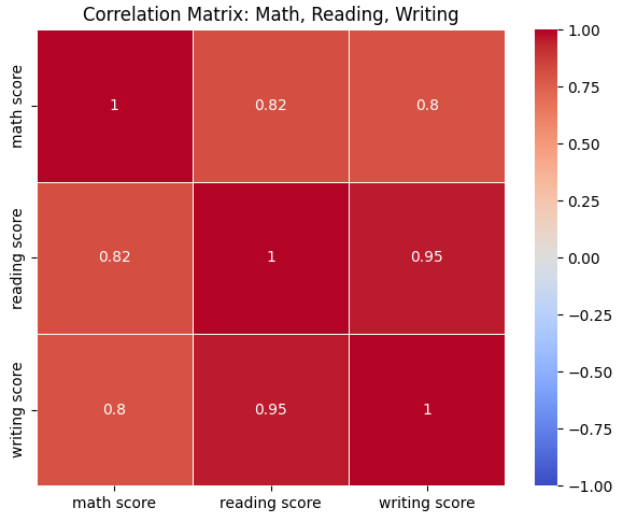


Figure 1.3: Correlation Heatmap of Academic Scores

## 1.4 Interpretation and Discussion

The analysis indicates that socio-economic and educational background are strong predictors of performance. Lunch type representing economic status shows a substantial effect on scores. Completion of a test preparation course also leads to higher performance, suggesting the importance of academic support.

Parental education influences results, with students whose parents have higher degrees performing better on average. Gender differences exist but are modest compared to socio-economic variables.

Overall, performance is shaped by a combination of academic, personal, and family-related factors.

## 1.5 Conclusion

This EDA concludes that student performance is influenced by multiple variables, with socio-economic status, test preparation, and parental education playing the most significant roles. Reading and writing scores are strongly related, indicating shared skill development. Math performance is more varied and appears to depend on different abilities.

The dataset is limited by its lack of school-level variables (teacher quality, resources, attendance). Future extensions could involve predictive modeling or analyzing larger educational datasets.

# References

[1] Pabbakavya. (2021) A comprehensive guide on exploratory data analysis (eda). Accessed: 2025-12-05. [Online]. Available: https://medium.com/@pabbakavya123/a-comprehensive-guide-on-exploratory-data-analysis-eda-ab38f33d6abc

[2] Phineouse. (2021) Extensive eda on titanic dataset. Accessed: 2025-12-05. [Online]. Available: https://medium.com/@Phineouse/extensive-eda-on-titanic-dataset-e91da1235e0a

[1, 2]

Scan to access the full GitHub repository