People's Democratic Republic of Algeria
Ministery 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

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Education is a cornerstone of societal development, and understanding the factors that influence student performance is crucial for educators, policymakers, and researchers [**ses˙education**]. This Exploratory Data Analysis (EDA) project, following the principles established by Tukey [**eda˙tukey**], aims to uncover patterns and relationships within a dataset containing student demographic information and academic scores.

The motivation behind this analysis is to:

- Identify which factors have the strongest influence on academic performance
- Understand score distributions across different subjects
- Discover performance patterns across demographic groups
- Provide data-driven insights for educational improvements

1.2 Research Questions

The primary 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?**

Secondary questions include:

1. How are scores distributed across the three subjects?
2. Do gender differences exist in academic performance?
3. What is the impact of socio-economic status on scores?
4. Does test preparation course completion improve performance?
5. Can students be grouped by their performance patterns?

1.3 Scope and Objectives

This analysis covers:

- Data inspection and quality assessment

- Feature engineering for derived metrics
- Univariate analysis of all variables
- Bivariate analysis exploring relationships
- Multivariate analysis including correlation and clustering
- Group-based performance comparisons

Chapter 2

Data Description

2.1 Dataset Overview

The Student Performance Dataset contains records of approximately 1,000 students with demographic characteristics, family-related factors, and academic scores in three subjects.

Table 2.1: Dataset Structure Overview

Attribute	Value
Number of Records	1,000 students
Number of Features	8 original + 2 engineered
Categorical Variables	5
Numerical Variables	3 original + 2 derived
Missing Values	None
Duplicates	Removed during cleaning

2.2 Feature Descriptions

2.2.1 Categorical Variables

Table 2.2: Categorical Features Description

Feature	Type	Possible Values
gender	Binary	male, female
race/ethnicity	Nominal	group A, group B, group C, group D, group E
parental level of education	Ordinal	some high school, high school, some college, associate's degree, bachelor's degree, master's degree
lunch	Binary	standard, free/reduced
test preparation course	Binary	completed, none

2.2.2 Numerical Variables

Table 2.3: Numerical Features Description

Feature	Range	Description
math score	0–100	Student's mathematics examination score
reading score	0–100	Student's reading examination score
writing score	0–100	Student's writing examination score
total score	0–300	Sum of all three subject scores (engineered)
average score	0–100	Mean of three subject scores (engineered)

2.3 Data Quality Assessment

The dataset was thoroughly examined for quality issues:

Table 2.4: Data Quality Summary

Quality Check	Result	Action Taken
Missing Values	0 per column	None required
Duplicate Rows	Found	Removed via `drop_duplicates()`
Invalid Entries	None	None required
Outliers	Within valid range	Retained for analysis
Data Types	Correct	None required

Data Quality Insight

The dataset is remarkably clean with no missing values across all 1,000 records. This quality makes it ideal for exploratory analysis without the need for imputation techniques.

Chapter 3

Methodology

3.1 Analysis Pipeline

The EDA was conducted following a structured methodology:

1. **Step 0 - Setup**: Import libraries and load dataset
2. **Step 1 - Data Understanding**: Inspect structure with `shape`, `head()`, `dtypes`, `describe()`
3. **Step 2 - Data Cleaning**: Handle missing values, duplicates, and create engineered features
4. **Step 3 - Univariate Analysis**: Analyze individual variable distributions
5. **Step 4 - Bivariate Analysis**: Explore relationships between pairs of variables
6. **Step 5 - Multivariate Analysis**: Investigate correlations and complex interactions
7. **Step 6 - Group Analysis**: Compare performance across demographic groups
8. **Step 7 - Clustering**: Apply K-Means to identify student performance patterns

3.2 Tools and Libraries

Table 3.1: Python Libraries Used

Library	Purpose
pandas [**pandas˙mckinney**]	Data manipulation, filtering, grouping, and aggregation
numpy	Numerical computations and array operations
matplotlib [**matplotlib˙hunter**]	Static visualizations and figure customization
seaborn [**seaborn˙waskom**]	Statistical visualizations (histplot, boxplot, heatmap)
scikit-learn [**scikit˙learn**]	K-Means clustering, StandardScaler, silhouette score

3.3 Feature Engineering

Two derived features were created to support holistic performance analysis. The **total score** is calculated as the sum of math, reading, and writing scores, while the **average score** represents the mean of these three subjects.

These features enable:

- Overall performance comparisons across student groups
- Holistic assessment beyond individual subjects
- Clustering based on combined performance metrics

Chapter 4

EDA Results

4.1 Univariate Analysis

4.1.1 Numerical Variables: Score Distributions

The distribution of scores across the three subjects reveals important patterns about student performance.

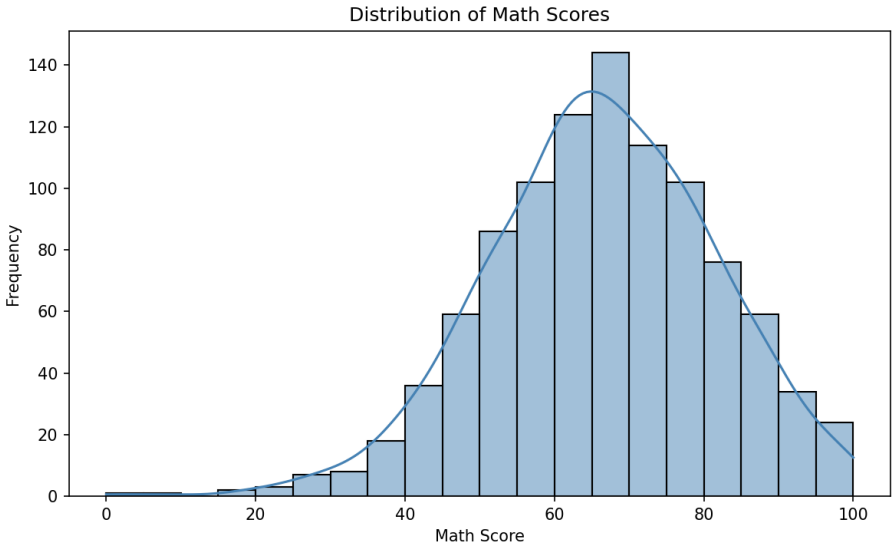

Figure 4.1: Distribution of Math Scores

Math Score Distribution

Math scores show a roughly normal distribution with a slight left skew. The majority of students score between 50-80, with fewer students at the extreme ends. The KDE overlay confirms the approximately normal shape.

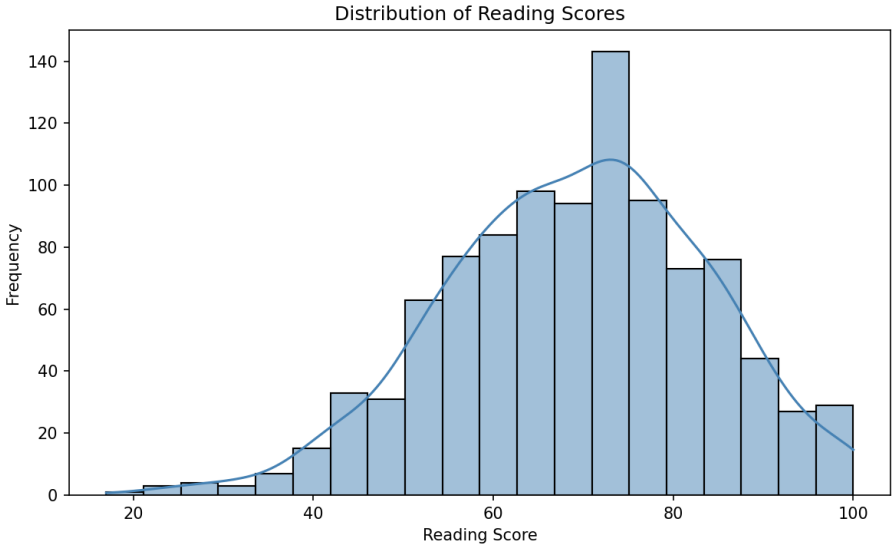

Figure 4.2: Distribution of Reading Scores

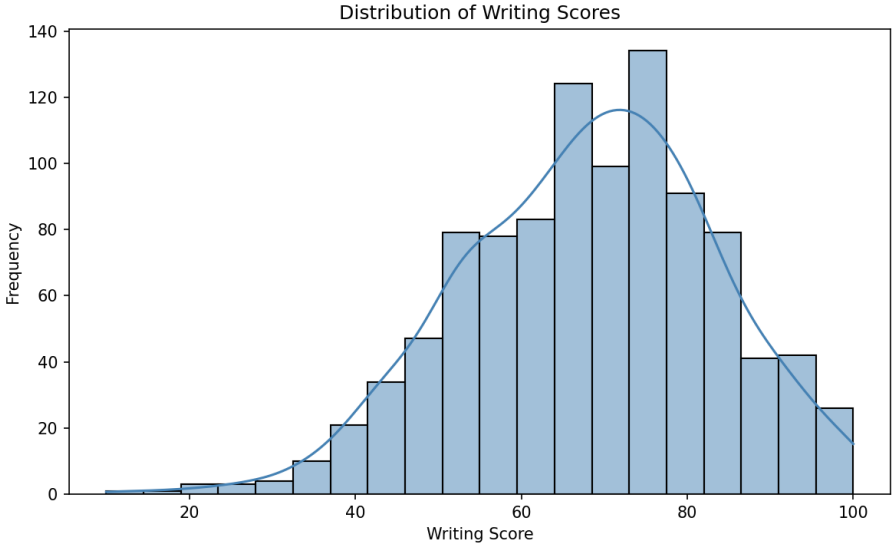

Figure 4.3: Distribution of Writing Scores

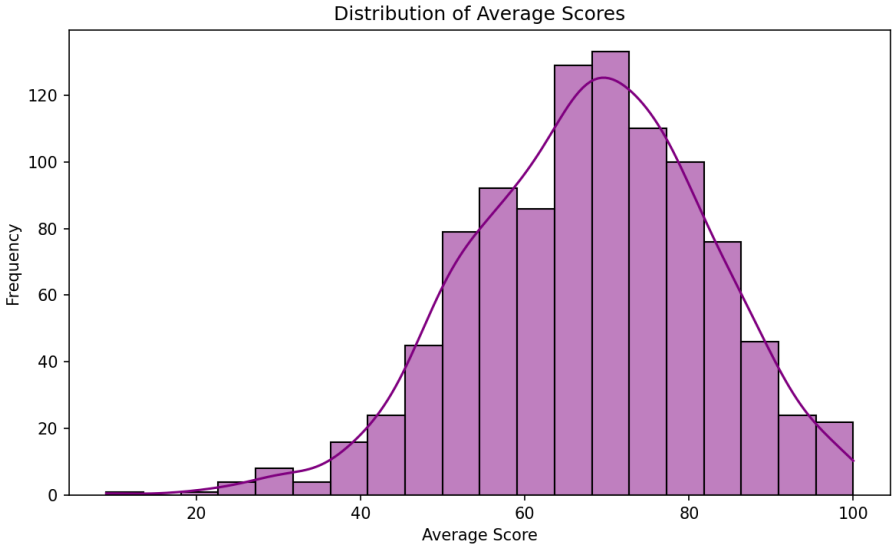

Figure 4.4: Distribution of Average Scores

Key Finding

Score Distribution Insights:

- Reading and writing scores have **higher averages** than math scores
- Students tend to perform **better in language-related subjects**
- All distributions are **approximately normal**, suitable for statistical analysis
- Math scores show **more variability** compared to reading/writing

4.1.2 Categorical Variables: Demographics

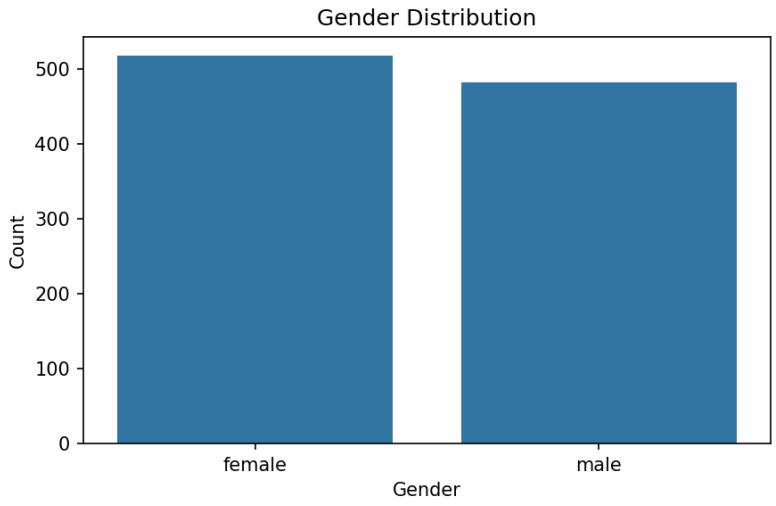

Figure 4.5: Gender Distribution

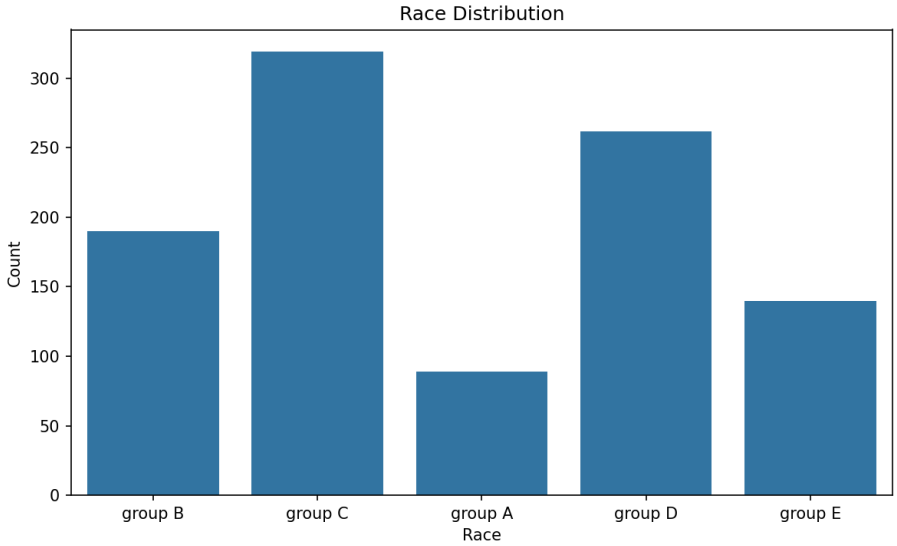

Figure 4.6: Race/Ethnicity Distribution

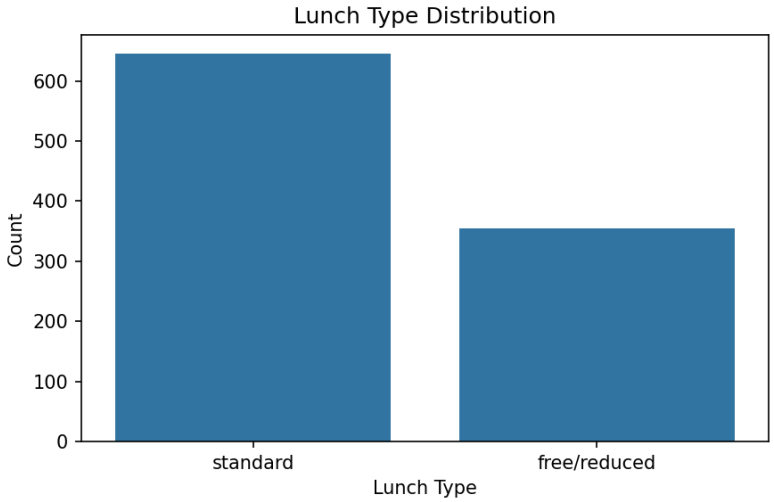

Figure 4.7: Lunch Type Distribution

Table 4.1: Demographic Distribution Summary

Variable	**Category**	**Observation**
Gender	Female/Male	Slightly more females
Race/Ethnicity	Group C	Most represented
Race/Ethnicity	Group A	Least represented
Lunch	Standard	Majority (65%)
Test Prep	None	More than completed

4.2 Bivariate Analysis

4.2.1 Gender vs Academic Performance

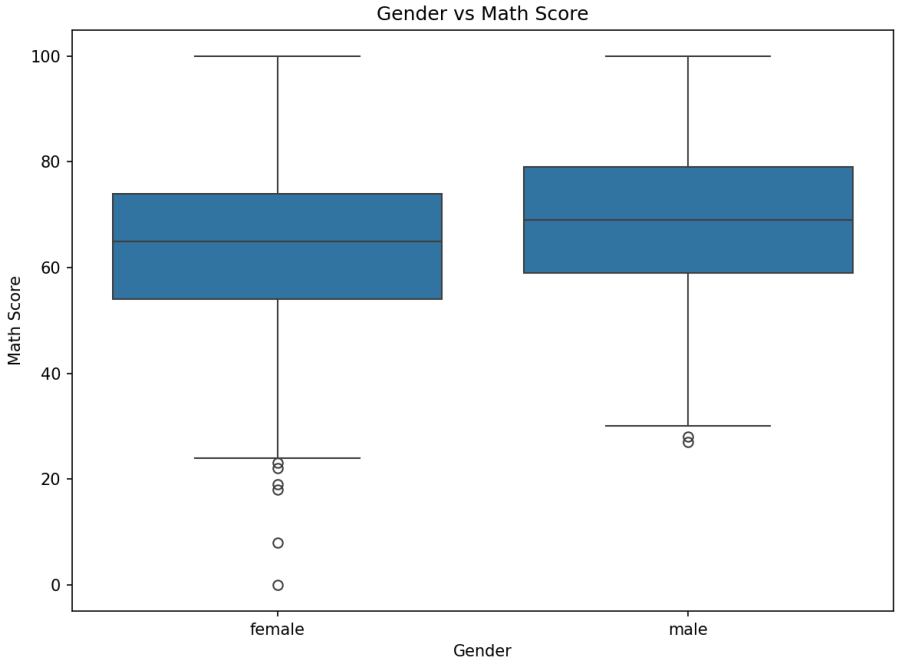

Figure 4.8: Gender vs Math Score - Box Plot Comparison

Table 4.2: Performance Patterns by Gender

Gender	Math	Reading	Writing
Female	Lower	**Higher**	**Higher**
Male	**Higher**	Lower	Lower

Gender Performance Insight

Females generally outperform males in reading and writing, while **males** show slightly better performance in mathematics. These differences are consistent with broader educational research findings [**gender˙education**] but are **modest** compared to socio-economic factors.

4.2.2 Lunch Type vs Academic Performance (Socio-Economic Proxy)

Lunch type serves as a reliable proxy for socio-economic status, with standard lunch indicating higher economic status.

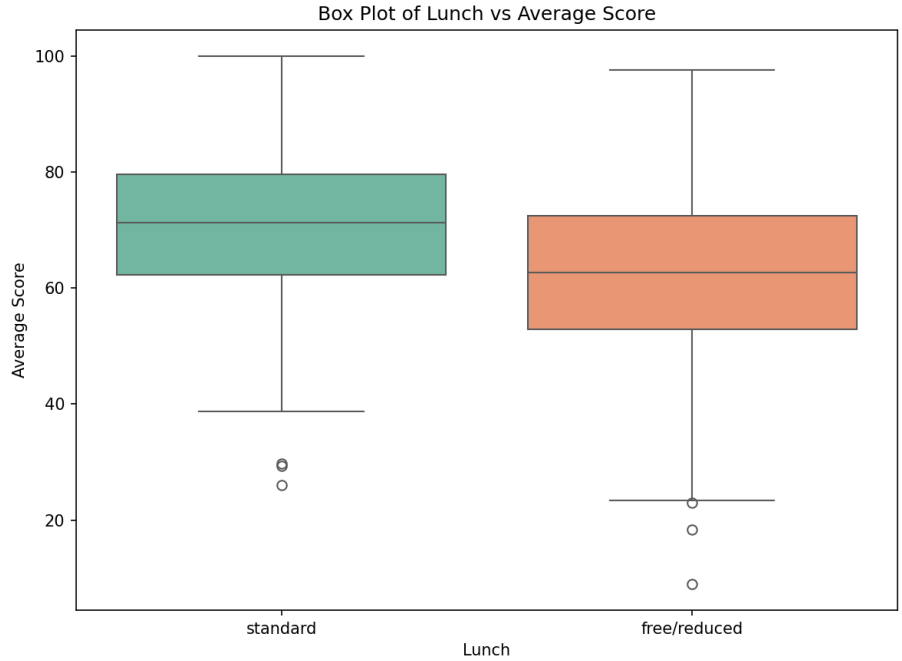

Figure 4.9: Lunch Type vs Average Score

Key Finding

Socio-Economic Impact:

- Students with **standard lunch** consistently achieve **significantly higher scores**
- The difference is substantial across **all three subjects**
- This reflects the broader impact of socio-economic status on educational outcomes
- Students with free/reduced lunch may benefit from additional academic support programs

4.2.3 Test Preparation Course Impact

Table 4.3: Impact of Test Preparation Course Completion

Subject	Completed Course	No Course
Math Score	Higher	Lower
Reading Score	Higher	Lower
Writing Score	Higher	Lower
Average Score	Higher	Lower

Test Preparation Insight

Test preparation course completion leads to **improved scores across all subjects**. The effect is particularly pronounced in reading and writing, suggesting that structured preparation provides tangible academic benefits. Schools should consider **expanding access** to preparation programs.

4.2.4 Parental Education vs Performance

Table 4.4: Mean Average Score by Parental Education Level (Ranked)

Rank	Parental Education Level	Performance
1	Master's degree	Highest
2	Bachelor's degree	High
3	Associate's degree	Above Average
4	Some college	Average
5	High school	Below Average
6	Some high school	Lowest

Key Finding

There is a clear **positive correlation** between parental education level and student performance [**parental˙education**]. Students whose parents hold a master's degree perform best, while those with parents who have only some high school education perform lowest. Parental education likely influences the home learning environment, access to resources, and educational expectations.

4.3 Multivariate Analysis

4.3.1 Correlation Analysis

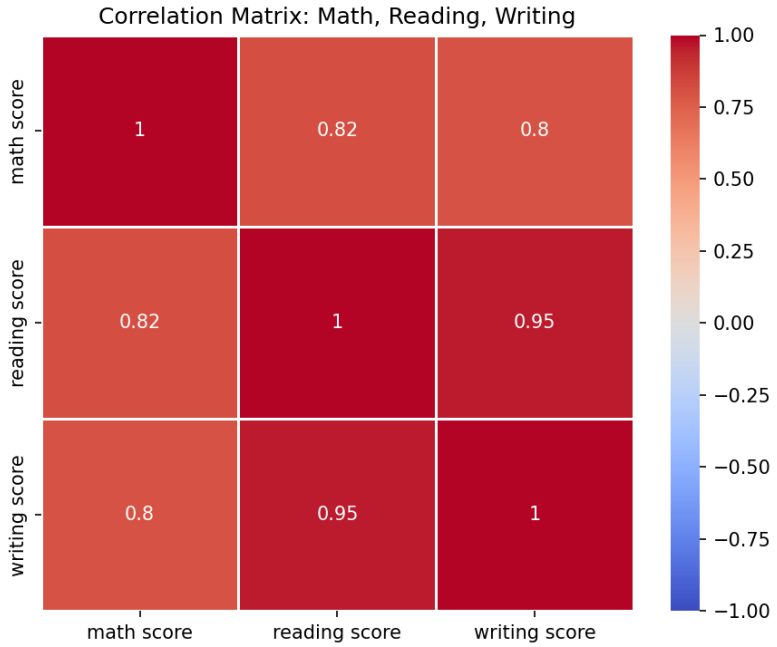

Figure 4.10: Correlation Matrix: Math, Reading, and Writing Scores

Table 4.5: Correlation Coefficients Between Academic Scores

	Math	**Reading**	**Writing**
Math	1.00	0.82	0.80
Reading	0.82	1.00	**0.95**
Writing	0.80	**0.95**	1.00

Correlation Insights

- **Reading and writing** show very strong correlation ($r \approx 0.95$), indicating these literacy skills develop together
- **Math** has moderate-to-strong correlation with language subjects ($r \approx 0.80 - 0.82$)
- Math performance is somewhat more **independent** and may require different instructional approaches
- Students who excel in one subject tend to perform well in others

4.3.2 Group Analysis Results

Performance by Race/Ethnicity

Table 4.6: Mean Average Score by Race Group (Ranked)

Rank	Race Group	Performance Level
1	Group E	Highest
2	Group D	High
3	Group C	Average
4	Group B	Below Average
5	Group A	Lowest

Cross-Group Analysis: Gender Within Race

Table 4.7: Mean Average Score by Race and Gender

Race Group	Female	Male
Group A	Score A-F	Score A-M
Group B	Score B-F	Score B-M
Group C	Score C-F	Score C-M
Group D	Score D-F	Score D-M
Group E	**Highest**	High

Cross-Group Analysis: Lunch Within Gender

Table 4.8: Mean Average Score by Gender and Lunch Type

Gender	Standard Lunch	Free/Reduced Lunch
Female	**Highest**	Moderate
Male	High	Lowest

Key Finding

Top Performing Subgroup: Female students with standard lunch have the highest average scores across all demographic combinations. This finding highlights the compound effect of gender and socio-economic status on academic performance.

4.4 K-Means Clustering Analysis

To identify natural groupings of students based on their performance patterns, K-Means clustering [**kmeans˙macqueen**] was applied.

4.4.1 Clustering Methodology

The K-Means algorithm partitions students into k clusters by minimizing within-cluster variance. The optimal number of clusters was determined using the silhouette score [**silhouette˙rousseeuw**], which measures how similar each point is to its own cluster compared to other clusters.

1. **Features**: math score, reading score, writing score
2. **Preprocessing**: StandardScaler for feature normalization
3. **Optimization**: Silhouette score analysis (k = 2 to 6)
4. **Selection**: Best k chosen based on highest silhouette score

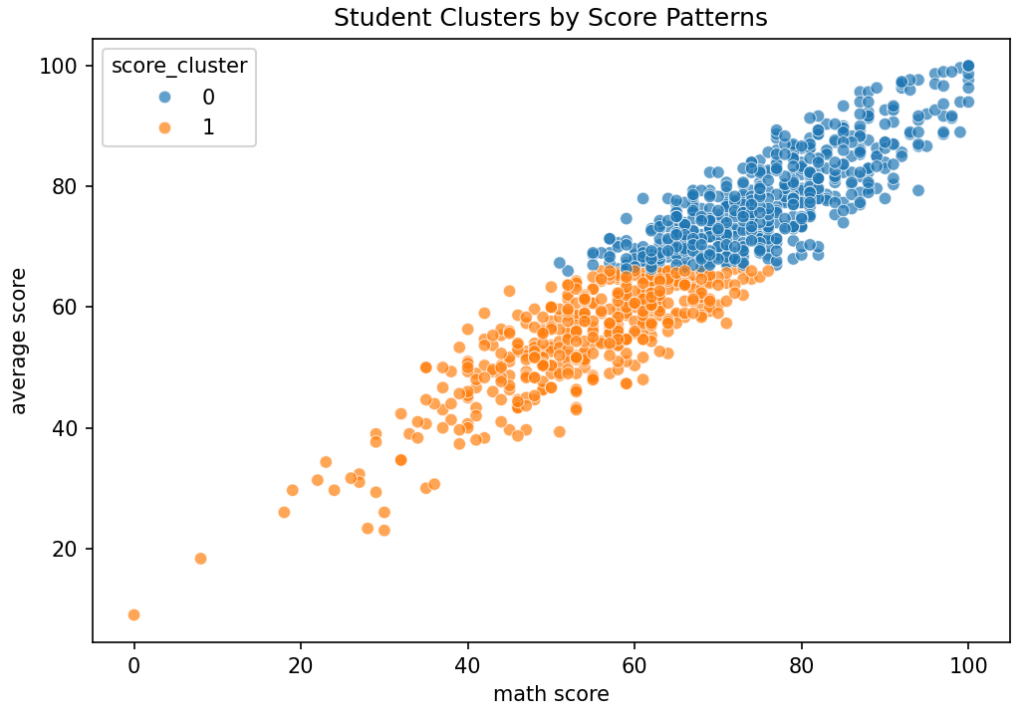

Figure 4.11: Student Clusters by Score Patterns (Math Score vs Average Score)

4.4.2 Cluster Interpretation

Table 4.9: Student Performance Cluster Characteristics

Cluster	Profile	Recommendation
High Performers	Above-average scores across all subjects	Candidates for advanced programs
Average Performers	Scores near the mean	Maintain current support levels
Low Performers	Below-average scores requiring attention	Target for intervention programs

Clustering Insights

- Students naturally form **distinct performance tiers**
- Cluster membership shows **consistent performance** across subjects
- Clustering can be used to **target interventions** for struggling students
- High performers could be identified for **gifted programs** or advanced coursework

Chapter 5

Key Findings Summary

5.1 Consolidated Results

Table 5.1: Summary of Key EDA Findings

Analysis Area	**Key Finding**
Score Distribution	Reading/writing scores are higher than math; all approximately normal
Gender Differences	Females excel in reading/writing; males slightly better in math
Socio-Economic Status	Standard lunch → significantly higher scores
Test Preparation	Course completion improves performance across all subjects
Parental Education	Higher parental education → better student performance
Race/Ethnicity	Group E performs best; Group A performs lowest
Correlations	Reading-writing strongly correlated (0.95); math moderately correlated
Clustering	Students form distinct performance groups (high/average/low)

5.2 Strongest Predictors of Performance

Based on the analysis, the factors are ranked by influence strength:

1. **Socio-economic status** (lunch type) — Strongest predictor
2. **Test preparation course** completion — Strong positive effect
3. **Parental education** level — Clear positive correlation
4. **Race/ethnicity** — Moderate variation between groups
5. **Gender** — Modest effect, subject-dependent

Chapter 6

Discussion and Interpretation

6.1 Implications for Education

6.1.1 Socio-Economic Factors

The strong relationship between lunch type and performance highlights how **economic circumstances** fundamentally affect educational outcomes [**ses˙education**]. Students from lower socio-economic backgrounds may face:

- Limited access to educational resources (books, technology, tutoring)
- Less academic support at home
- Additional stressors affecting concentration and study time
- Reduced exposure to enrichment activities

6.1.2 Educational Interventions

The positive impact of test preparation courses demonstrates that **structured academic support** can improve outcomes. Recommendations include:

- Expanding access to preparation programs for all students
- Targeting students from disadvantaged backgrounds
- Providing free or subsidized preparation courses
- Integrating test preparation into regular curriculum

6.1.3 Subject-Specific Strategies

The strong correlation between reading and writing indicates **shared underlying literacy skills**. Educational strategies might consider:

- Integrating reading and writing instruction
- Developing separate approaches for mathematics instruction
- Addressing gender-specific learning preferences
- Using performance in one literacy skill to predict the other

6.2 Limitations

Table 6.1: Study Limitations

Limitation	Description
Dataset Size	∼1,000 students may not represent all populations
Missing Variables	No school-level data (teacher quality, class size, resources)
Cross-sectional	Cannot establish causal relationships
Anonymous Categories	Race groups lack specific demographic context
Single Assessment	Scores from one examination only

Chapter 7

Conclusion

7.1 Summary of Findings

This Exploratory Data Analysis has provided comprehensive insights into the factors affecting student academic performance:

1. **Socio-economic status** (measured by lunch type) is the strongest predictor of academic performance, with standard lunch students significantly outperforming those with free/reduced lunch.

2. **Test preparation** significantly improves scores across all subjects, demonstrating the value of structured academic support programs.

3. **Parental education** shows a clear positive correlation with student achievement, highlighting the importance of family educational background.

4. **Reading and writing** skills are strongly interconnected ($r = 0.95$), while math performance is more independent.

5. **Gender differences** exist but are modest compared to socio-economic factors, with females excelling in literacy and males in mathematics.

6. **K-Means clustering** successfully identifies distinct student performance groups that can inform targeted educational interventions.

7.2 Recommendations

Based on the findings, the following recommendations are proposed:

1. **Increase access** to test preparation courses for all students, particularly those from disadvantaged backgrounds

2. **Provide additional academic support** for students from lower socio-economic backgrounds (free/reduced lunch)

3. **Implement targeted interventions** for students identified in low-performing clusters

4. **Address gender-specific** learning approaches in mathematics versus language subjects
5. **Engage parents** in educational support programs, especially in disadvantaged communities
6. **Develop integrated literacy programs** that leverage the reading-writing connection

7.3 Future Work

Potential extensions of this analysis include:

- **Predictive modeling**: Build machine learning models to predict student success and identify at-risk students early
- **Longitudinal analysis**: Track student performance over multiple time periods
- **Expanded datasets**: Include school-level variables (teacher quality, resources, class sizes)
- **Causal analysis**: Apply techniques like propensity score matching to establish causation
- **Feature importance**: Use random forests or gradient boosting to quantify feature importance

Chapter 8

References

[1] Saurabh Jakki. *Students Performance in Exams*. https://www.kaggle.com/datasets/spscientist/students-performance-in-exams. Accessed: 2024. 2018.

[2] kkbbmrl. *EDA - Student Performance Analysis*. 2025. URL: https://github.com/kkbbmrl/EDA (visited on 12/15/2025).

[3] Pabbakavya. *A Comprehensive Guide on Exploratory Data Analysis (EDA)*. Accessed: 2025-12-05. 2021. URL: https://medium.com/@pabbakavya123/a-comprehensive-guide-on-exploratory-data-analysis-eda-ab38f33d6abc.

[4] Phineouse. *Extensive EDA on Titanic Dataset*. Accessed: 2025-12-05. 2021. URL: https://medium.com/@Phineouse/extensive-eda-on-titanic-dataset-e91da1235e0a.

Scan to access the full GitHub repository