EDA on Student Performance Dataset

Name: 黃家福
Student ID: 41347051S

1. Objective(s)

The primary goal of this analysis is to explore the dataset containing student performance records and identify key factors influencing academic success.

Specifically, the analysis will:

- Examine distributions of student scores across different subjects.
- Compare performance between demographic groups (e.g., gender).
- Analyze correlations between features such as study time, attendance, and scores.
- Summarize findings that can inform potential interventions to improve student outcomes.

2. Description

The dataset includes records for 1,000 students, each with information about demographics, test preparation, and academic performance. The variables include:

- **Gender**: The gender of the student (`male` / `female`)
- **Race/ethnicity**: The student's racial or ethnic background
- **Parental level of education**: The highest level of education attained by the student's parent(s) or guardian(s)
- **Lunch**: Whether the student receives free or reduced-price lunch (`standard` / `free/reduced`)
- **Test preparation course**: Whether the student completed a test preparation course (`completed` / `none`)
- **Math score**: The student's score on a standardized mathematics test
- **Reading score**: The student's score on a standardized reading test
- **Writing score**: The student's score on a standardized writing test

It is worth noting that the race/ethnicity information has been anonymized into generalized group labels (e.g., "Group A", "Group B") rather than actual ethnic identifiers. This was likely done to prevent controversy or bias. While this protects privacy and avoids reinforcing stereotypes, it may also restrict the ability to study cultural or ethnic patterns in academic performance in more depth.

That said, this limitation does not significantly impact the exploratory data analysis overall, as the anonymized groupings still allow for meaningful comparisons and pattern detection across demographic

segments.

3. Pseudocode

- Import libraries for data handling and visualization.
- Load dataset from CSV into a dataframe.
- Check for missing values to ensure data integrity.
- Calculate average score from math, reading, and writing scores.
- Visualize score distributions and demographic comparisons using appropriate plots (e.g., violin, box, bar plots).
- Identify patterns and differences in scores across demographic and support-related factors.

4. Algorithms / Approaches

- **Data Quality Check:**
  Confirm the dataset contains no missing values to ensure completeness and reliability of the analysis.
- **Feature Engineering:**
  Create an average score by calculating the mean of math, reading, and writing scores to simplify performance comparisons.
- **Visualizing Score Distributions:**
  Use violin plots to display score distributions by subject and race/ethnicity, showing variability and density.
- **Group Comparisons:**
  Employ box plots, swarm plots, and boxen plots to compare average scores across gender, lunch type, and test preparation course status.
- **Statistical Summary for Parental Education:**
  Calculate mean scores and standard deviations to describe group differences and variability for parental education categories, as illustrated in the corresponding bar plot.
- **Demographic Proportion Visualization:**
  Use pie charts to illustrate the distribution of students across race/ethnicity, parental education, lunch type, and test preparation status.

5. Summary & Key Findings

- **Gender:**
  The student population is nearly evenly split between male (50.8%) and female (49.2%) students (see Figure 1.1a). Female students tend to outperform males slightly, with an average score about 3 points higher (see Figure 1.1b).

- **Race/Ethnicity:**
  Group C represents the largest ethnic category (32.3%) (see Figure 1.2a). Group E students achieve the highest average scores among the race/ethnicity groups (see Figure 1.2b).
- **Parental Level of Education:**
  Approximately 39.2% of students have parents who did not continue education beyond high school (see Figure 1.3a). Students whose parents attained some college education or higher score better on average. The bar plot (see Figure 1.3b) includes mean scores with error bars representing standard deviation, highlighting this positive correlation.
- **Learning Condition Observation:**
  About two-thirds of students fall into the categories of having standard lunch and not completing test preparation (see Figure 1.4a; Figure 1.5a), implying that many do not simultaneously benefit from both academic support and socioeconomic advantage.
- **Lunch Type:**
  Students receiving standard-priced lunch outperform those receiving free or reduced-price lunch, indicating socioeconomic factors may influence academic outcomes (see Figure 1.4b).
- **Test Preparation Course:**
  Completion of a test preparation course is linked to noticeably higher average scores compared to those who did not complete the course (see Figure 1.5b).

Figures

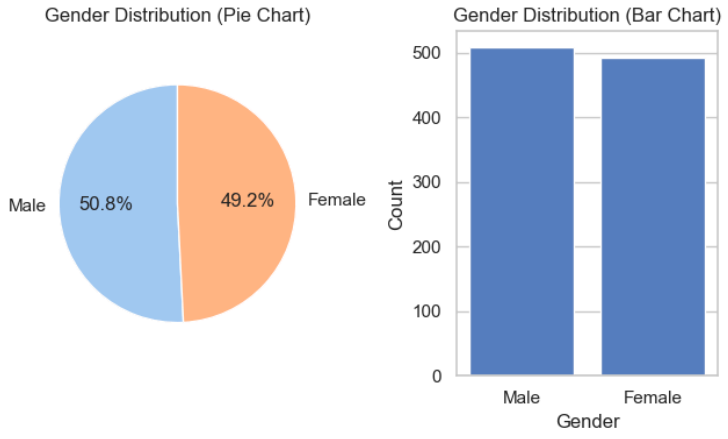

Figure 1.1a: Gender Distribution.

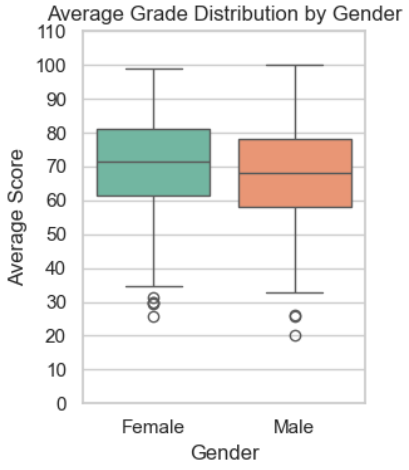

Figure 1.1b: Score Distribution by Gender.

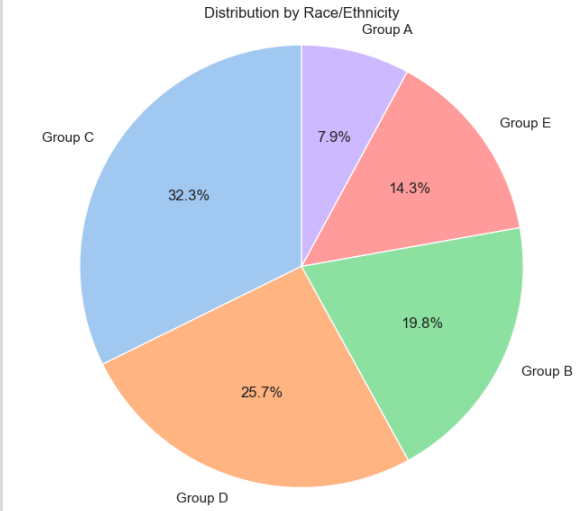

Figure 1.2a: Race/Ethnicity Distribution.

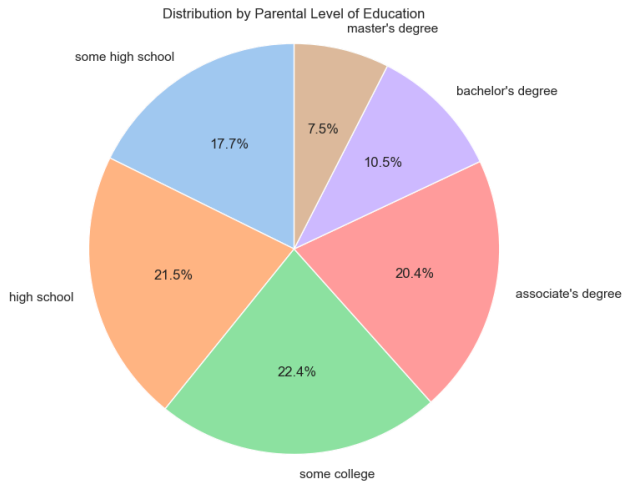

Figure 1.3a: Parental Level of Education Distribution.

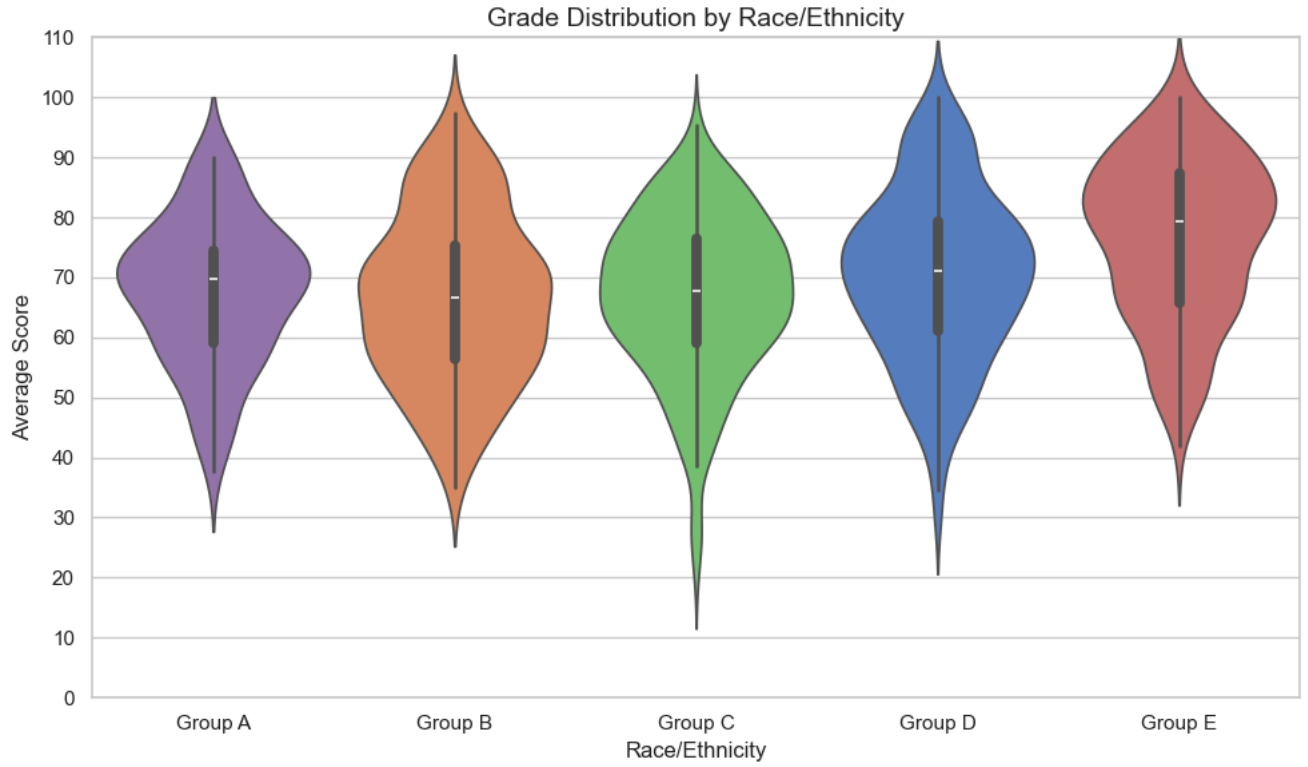

Figure 1.2b: Grade Distribution by Race/Ethnicity.

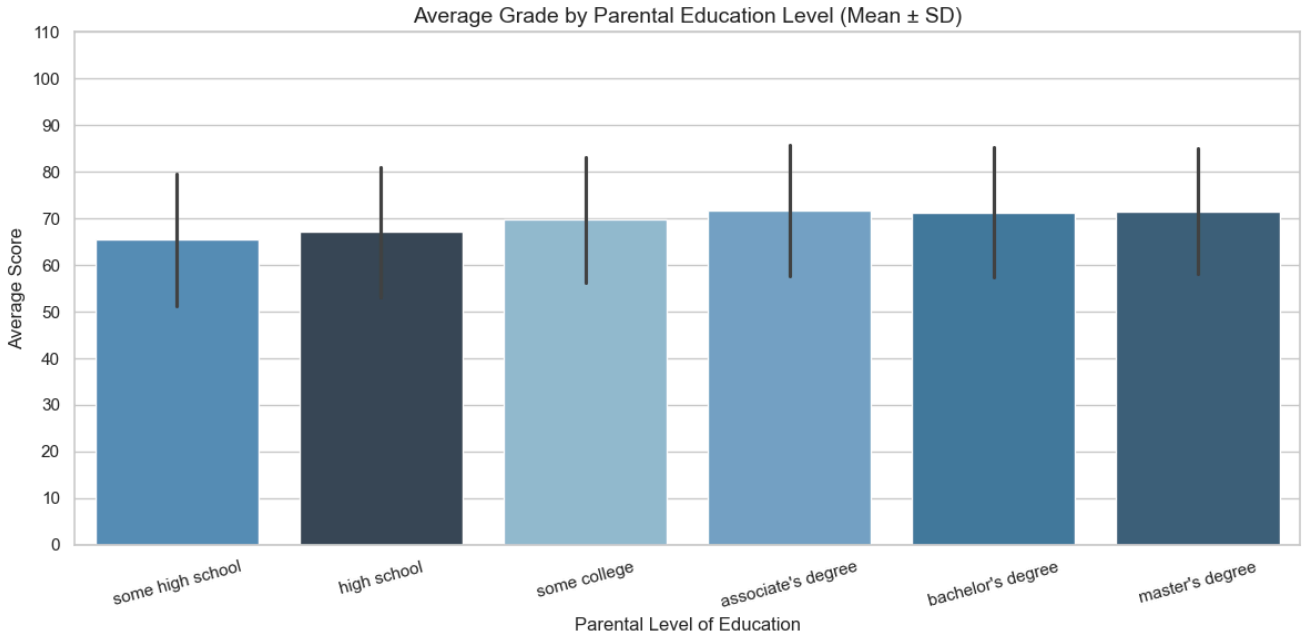

Figure 1.3b: Grade Distribution by Parental Level of Education.

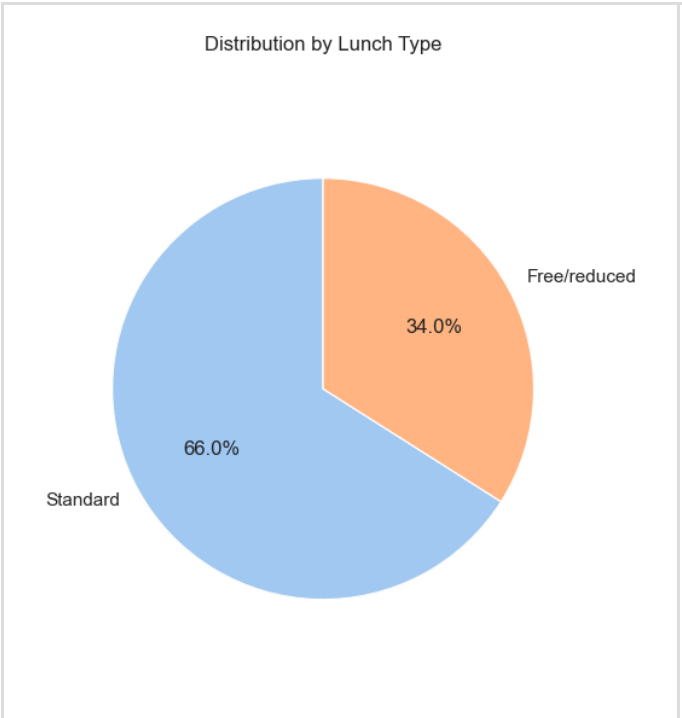

Figure 1.4a: Lunch Type Distribution.

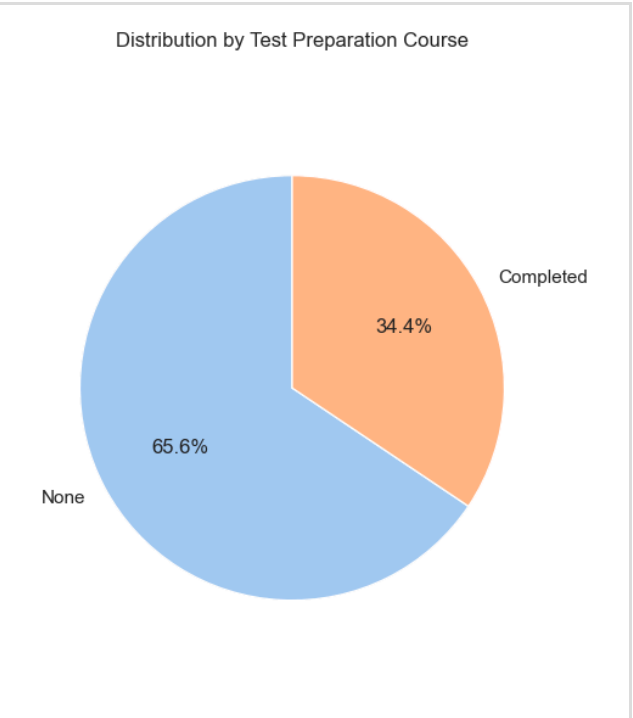

Figure 1.5a: Test Preparation Distribution.

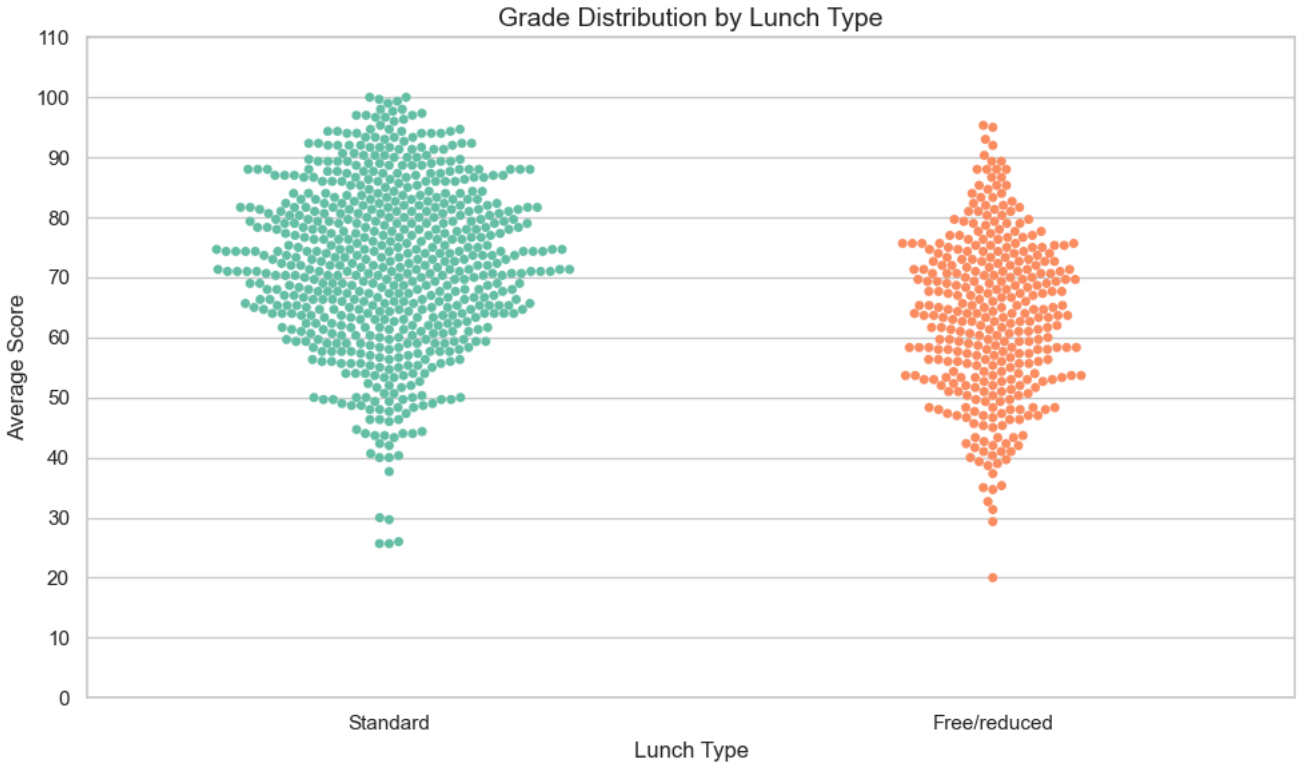

Figure 1.4b: Grade Distribution by Lunch Type.

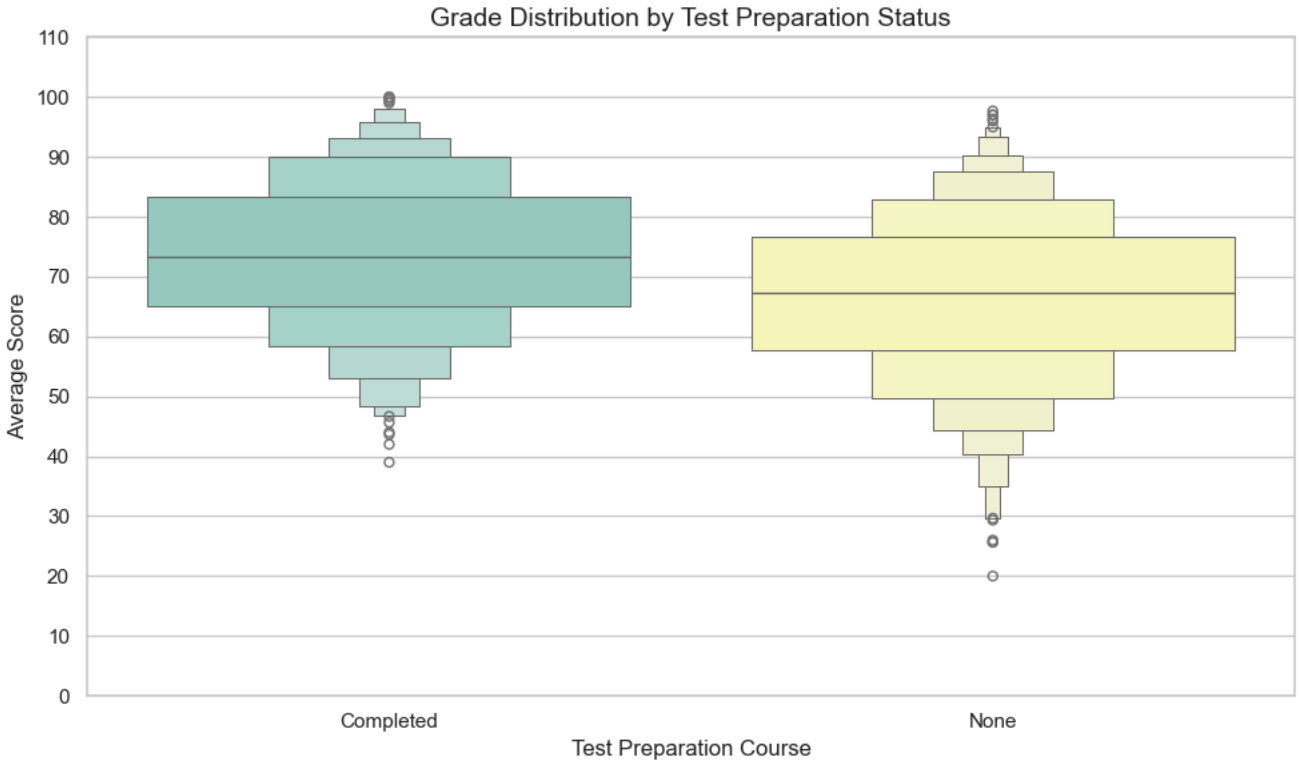

Figure 1.5b: Grade Distribution by Course Preparation.

6. Appendix / References

- **Dataset Source:**
  Student performance prediction on Kaggle
- **Project Repository:**
  GitHub – EDA on Student Performance