FACULTY OF BUSINESS INFORMATICS

جامعة مصر للمعلوماتية
EGYPT UNIVERSITY OF INFORMATICS

Machine Learning Report

Maya Yakout (21-101019)

Christine Nagy (21-101058)

Jana Ahmed (21-101052)

Nouran Ahmed (21-101027)

Aya Ahmed (21-101008)

Submitted to: Dr. Anas Ismail

Course: BBA344 - Machine Learning for Descriptive & Predictive Analytics

23rd May, 2024

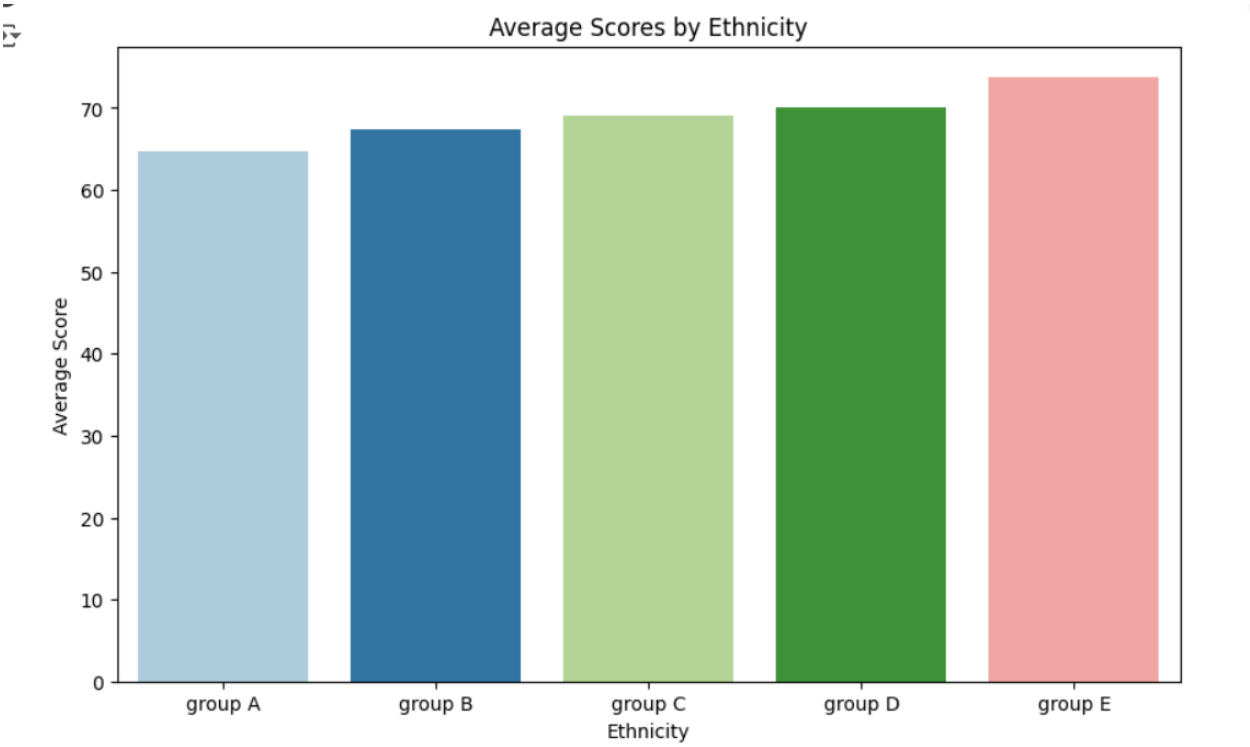
Contents

Introduction

The student study performance contains 1001 rows and 8 columns that have information on students and their academic performance. It has columns as gender, race_ethnicity, parental_level_of_education, lunch, test_preparation_course in math_score, reading score, and writing_score that will help us understand how these variables affect the students' scores in math, reading, and writing.

Question 1: Ethnicity and Achievement:

Are there any ethnic groups (A, B, C, D, E) that consistently score higher or lower across all subjects?



This bar chart shows the difference in average scores in three different subjects' math, reading and writing according to their ethnic groups.

Group A average score is around 63

Group B average score is around 68

Group C average score is around 70

Group D average score is around 70

Group E average score is around 72

We will use this function Data.groupby because we want all the ethnicity in the same group (for example, Group A) to be gathered and then all the scores (math, reading, writing) that is related to

the same ethnicity group, we will calculate its average score. We named this line in the code grouped_vars.

Then we used sns.barplot to help make the bar chart we want to visualize the data. So, we will use the grouped_vars that we made which will calculate the average of each score according to the group ethnicity to help us draw the bar plot by plotting each avearge score for each subject with the ethnic group.

Question 2: Test Prep Effectiveness:

Is there a significant improvement in scores for students who completed a test preparation course compared to those who didn't?

We made a function completed_avg to calculate the average score of students who completed the preparation course in math, reading and writing. Then we made another function named none_avg to calculate the average of students who did not complete any preparation course in math, reading and writing.

Then we calculated the difference between the students who completed the course and the students who didn't complete the course.

Then we printed the average scores for both:

Average Completed Preparation Course:

math_score 69.695531 reading_score 73.893855 writing_score 74.418994

Average for No Preparation Course:

math_score 64.077882 reading_score 66.534268 writing_score 64.504673

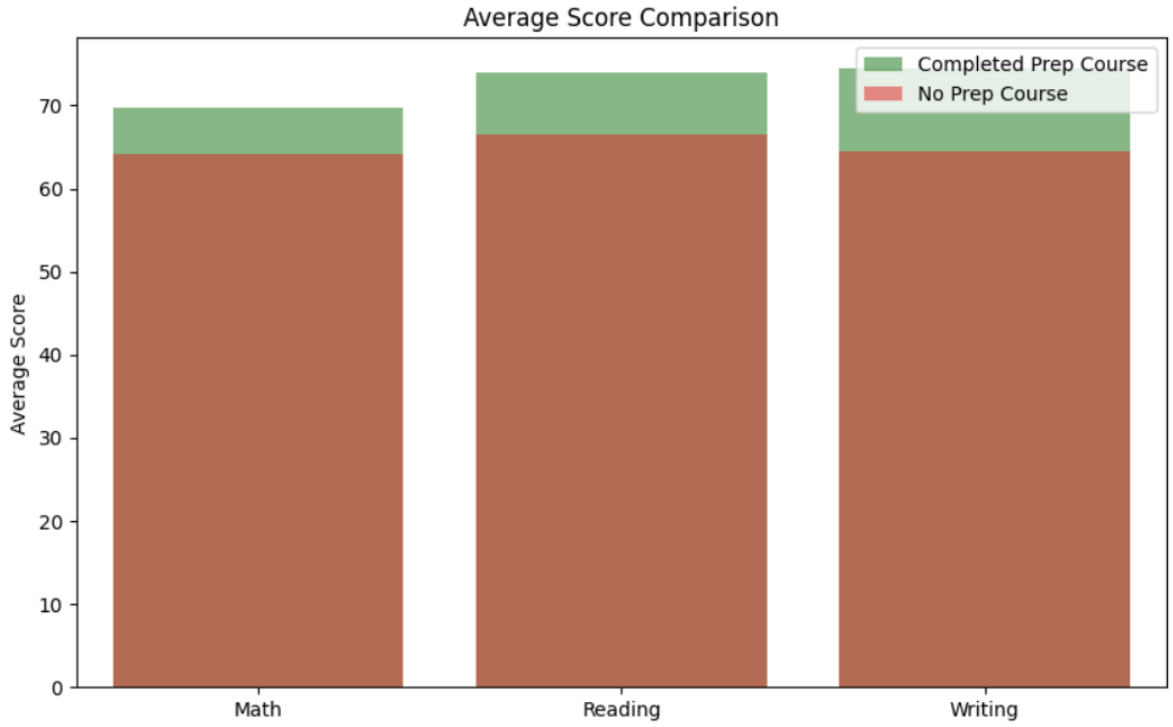
Difference in Average of scores:

math_score 5.617649 reading_score 7.359587 writing_score 9.914322

Then we used the function sns.barplot to visualize the average scores we got.

So, we will plot the outputs we got on a bar chart comparing the average scores in math, reading and writing between students who completed the test preparation course and those who didn’t complete the test preparation course.

The students that completed (green) and students that didn't complete (red).



Question 3: Gender Gap:

Is there a significant difference in average scores (math, reading, writing) between male and female students?

This graph shows average scores by gender (Male or Female) for math-score and Reading-score and Writing-score.

Math Scores: The average math score for males is higher than the average math score for females.

Reading Scores: The average reading score for females is higher than the average reading score for males.

Writing Scores: The average writing score for females is higher than the average writing score for males.

Gender	Math-score	Reading-score	Writing-score
Female	63.633205	72.608108	72.467181
Male	68.728216	65.473029	63.311203

Math-score for male = 68.728

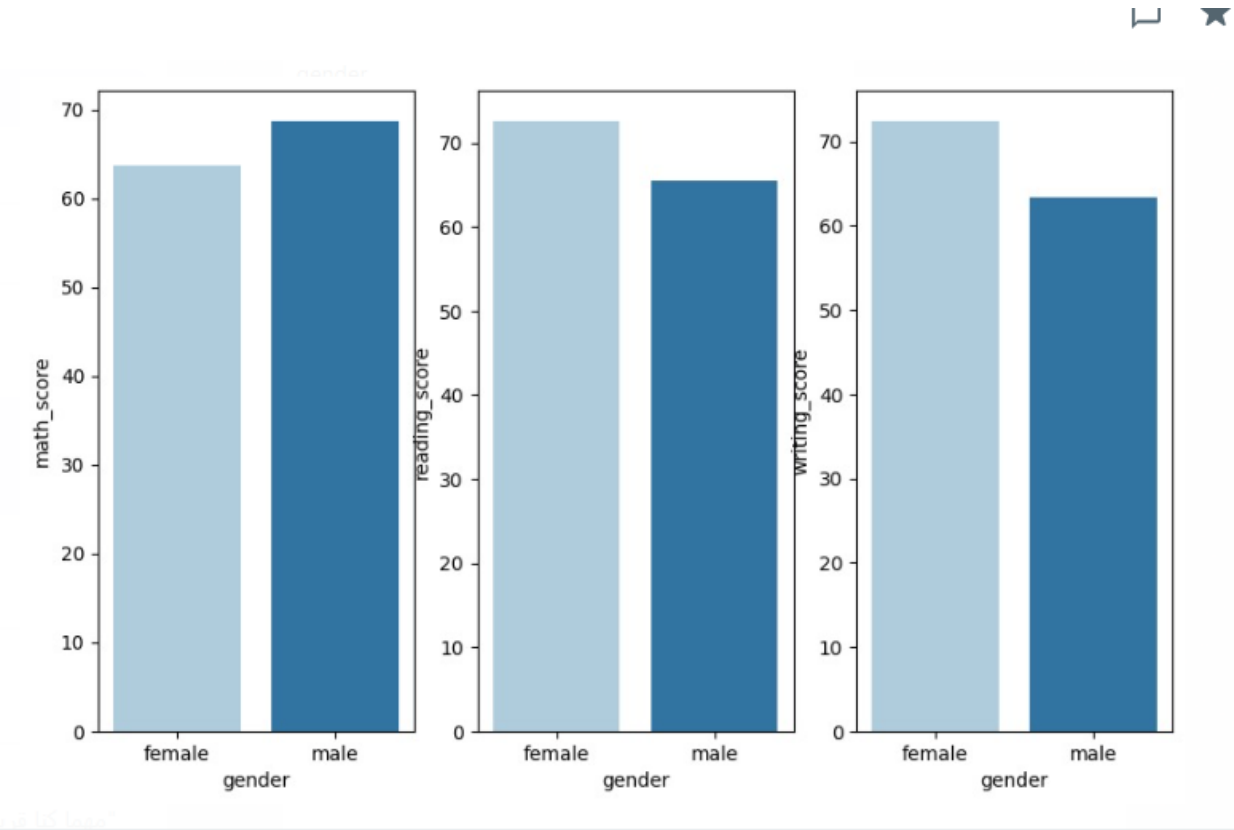
Math-score for female = 63.633

Reading-score for male = 65.473

Reading-score for female = 72.608

Writing –score for male = 63.311

Writing-score for female = 72.467



Question 4: Lunch and Test Performance

Do students who have standard lunch perform better than those who on free/reduced programs?

Yes, students who had standard lunch performed better than those on free/reduced program.

- Observing the box plot, we find that the median scores for ‘math’, ‘reading’, and ‘writing’ are higher for students with 'standard' lunch than those with 'free/reduced' lunch.
- Less variability and fewer outliers in the 'standard' lunch group, which indicates more consistent and better performance among these students.

- The coefficients for the linear regression models (positive values) indicate a positive relationship between having a standard lunch (coded as 1) and higher test scores.
- By observing the scatter plot, the three scores for ‘math’, ‘reading’, and ‘writing’ indicate good performance if the scatter plots show points closely aligned with the 45-degree line.
- Clustering around the line suggests that the models are reliable and that the predicted performance is consistent with the actual.
- The steps used:

a. Plotting box plots, function used: plt.figure(), plt.subplot(), sns.boxplot(), plt.title(), and plt.show(). This function creates boxplots to visualize the distribution of math, reading, and writing scores based on lunch type ('standard' or 'free/reduced').

b. Mapping Lunch Type to Numerical Values, Function used: ‘map()’. ‘standard’ was set as 1 and ‘free/reduced’ was set as 0 to do the regression analysis.

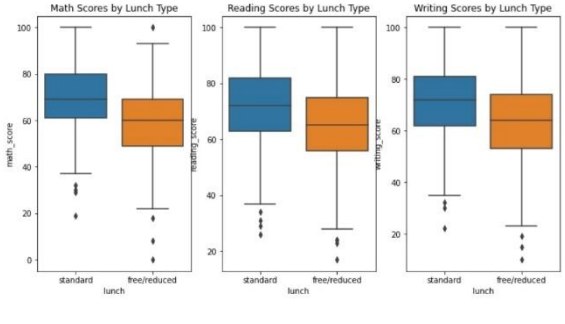
c. Splitting Data for Training and Testing, Function Used: train_test_split() from sklearn.model_selection. This code splits the data into training and testing sets for math, reading, and writing scores with a random state of 42.

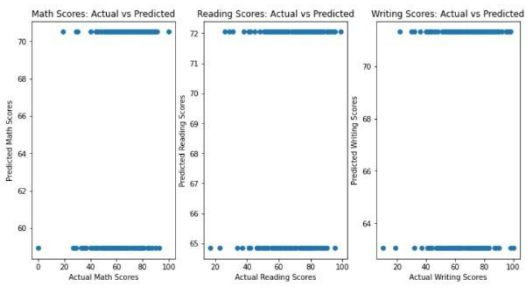
d. Training Linear Regression Models, Function Used: LinearRegression() and fit() from sklearn.linear_model. These functions train three separate linear regression models to predict math, reading, and writing scores based on the lunch type.

e. Making the predictions, Function Used: predict(). This function uses the trained models to predict the test set scores for math, reading, and writing.

f. Plotting Actual vs Predicted Scores, Function Used: plt.figure(), plt.subplot(), plt.scatter(), plt.xlabel(), plt.ylabel(), plt.title(), and plt.show(). These functions create scatter plots to visualize the relationship between actual and predicted scores for math, reading, and writing.

g. Printing Model Shapes and Coefficients, Function used: print(). This function prints the shapes of the training and testing sets and the coefficients of the linear regression models. The coefficients indicate the impact of lunch type on the test scores.



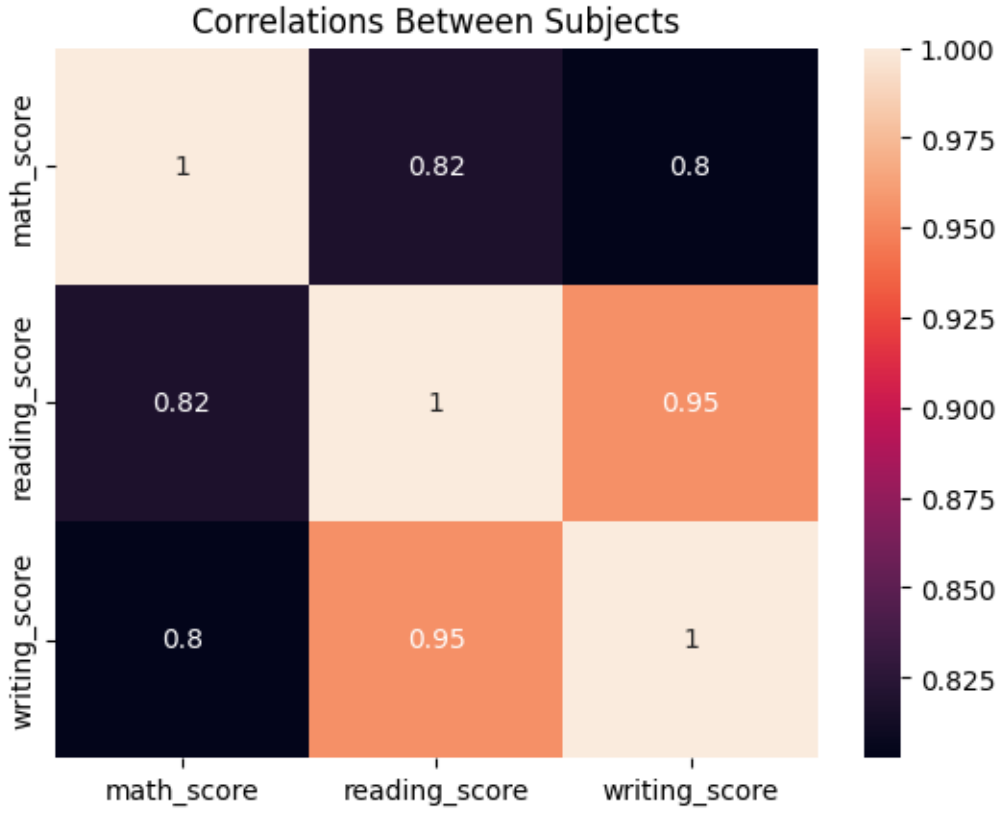



```
(800, 1) (200, 1) (800,) (200,)
Math Model Coefficient: [11.54937151]
Reading Model Coefficient: [7.1994257]
Writing Model Coefficient: [8.26324799]
```

Question 5: Subject Correlations:

Is there a correlation between performance in Math and Reading scores? Do students who excel in Writing also tend to score well in Math?

The graphic shows the correlation coefficients between the math, reading, and writing scores and stores.



Correlations between Subjects:

	math_score	reading_score	writing_score
math_score	1	0.81758	0.802642
reading_score	0.81758	1	0.954598
writing_score	0.802642	0.954598	1

- **Math Score:**
  - The correlation between math scores and themselves is 1.000000, indicating a perfect positive correlation.

    - The correlation between math scores and reading scores is 0.817580, indicating a strong positive correlation. This suggests that students who perform well in math tend to also perform well in reading.
    - The correlation between math scores and writing scores is 0.802642, indicating a strong positive correlation. This suggests that students who perform well in math tend to also perform well in writing.
- **Reading Score:**
    - The correlation between reading scores and math scores is 0.817580, consistent with the math-reading correlation.
    - The correlation between reading scores and themselves is 1.000000, indicating a perfect positive correlation.
    - The correlation between reading scores and writing scores is 0.954598, indicating a very strong positive correlation. This suggests that students who perform well in reading tend to also perform well in writing.
- **Writing Score:**
    - The correlation between writing scores and math scores is 0.802642, consistent with the math-writing correlation.
    - The correlation between writing scores and reading scores is 0.954598, consistent with the reading-writing correlation.
    - The correlation between writing scores and themselves is 1.000000, indicating a perfect positive correlation.

- The correlation analysis indicates strong positive relationships among math, reading, and writing scores. The relationship between reading and writing scores is particularly strong (0.954598), suggesting that students who excel in reading are likely to excel in writing too. Similarly, there are strong positive correlations between math and reading scores (0.817580) and between math and writing scores (0.802642), indicating that high performance in one subject generally correlates with high performance in the others.

- These findings can guide educational strategies by emphasizing the interconnected nature of these academic skills. Enhancing performance in one subject may positively impact performance in the other subjects.

Question 6: Parental Education Impact:

Do students with parents holding higher degrees (Master's, Bachelor's) outperform those with parents with lower educational attainment (High School, Associate's)?

1. Handling Categorical Data:

The code checks if the "parental_level_of_education" column is categorical data (type 'object'). If it is, it converts the categories (e.g., "High School Diploma", "bachelor's degree") into numerical labels using pd.factorize. This makes it suitable for linear regression.

2. Preparing Data for Modeling:

It extracts the "parental_level_of_education" data as a NumPy array and reshapes it for compatibility with the model. It uses parental education level for prediction.

3. Building Linear Regression Models:

The code creates three separate linear regression models (one for each subject: math, reading, writing). Each model is trained using the "parental_level_of_education" data (X) and the subject scores (math, reading, writing) from the data frame.

4. Predicting Scores for New Data:

It creates a new data point with specific parental education levels (e.g., levels 6 and 7, which could represent categories). This new data is reshaped to match the format expected by the model.

5. Making Predictions:

It uses the trained models (math_model, reading_model, writing_model) to predict scores for each subject based on the new parental education levels.

6. Printing Predicted Scores:

```
Predicted Math Scores:
Parental Education: 6, Predicted Math Score: 61.943287965724494
Parental Education: 7, Predicted Math Score: 60.69118526736281

Predicted Reading Scores:
Parental Education: 6, Predicted Reading Score: 65.05695426400011
Parental Education: 7, Predicted Reading Score: 63.81501958082286

Predicted Writing Scores:
Parental Education: 6, Predicted Writing Score: 62.314868363284035
Parental Education: 7, Predicted Writing Score: 60.581515407465254
```

Question 7: Statistics and Boxplot Analysis

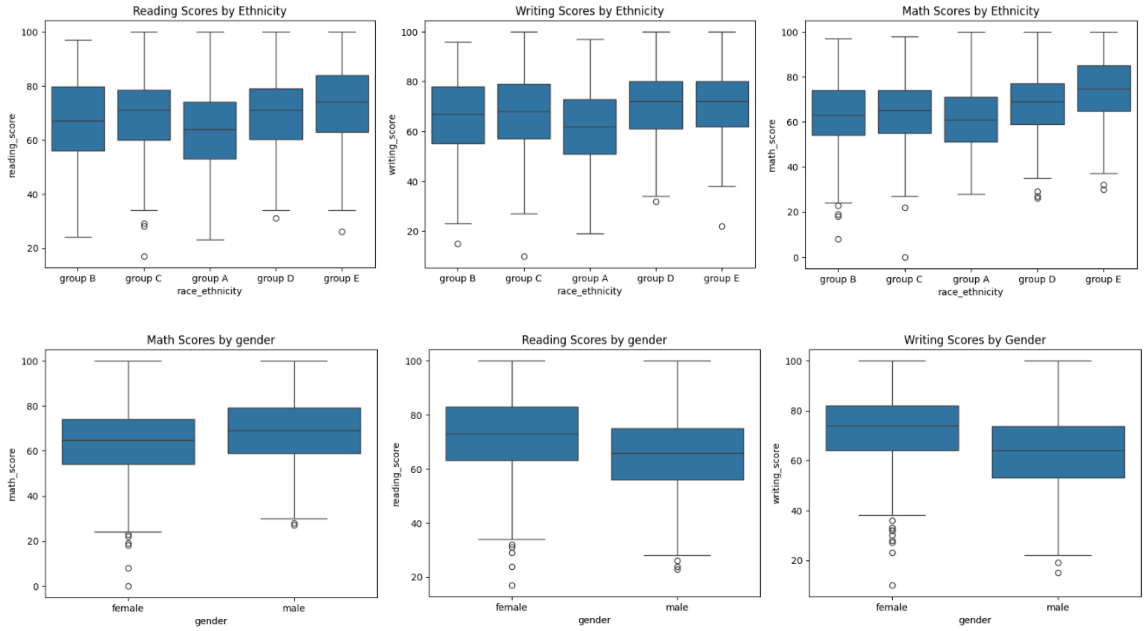
Across all demographics, which subject (Math, Reading, Writing) on average shows the highest and lowest scores?

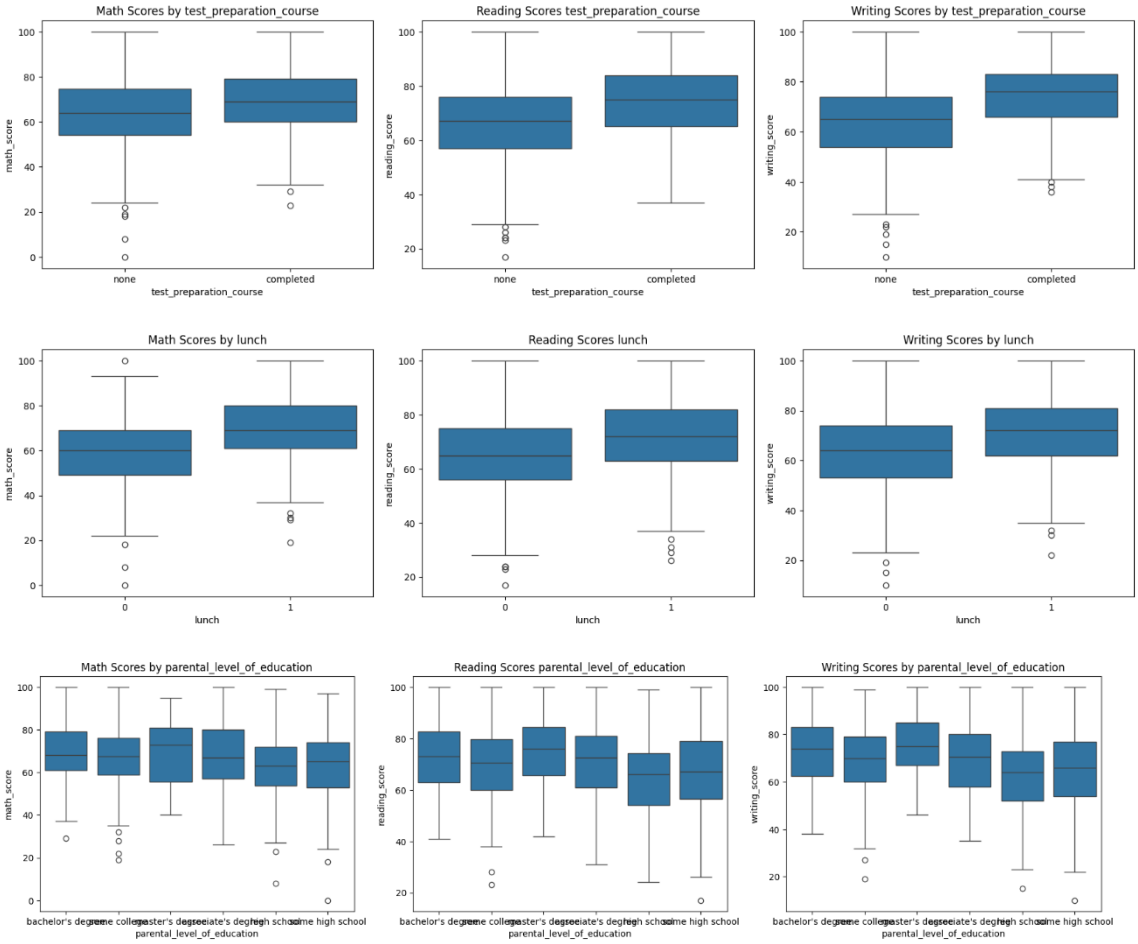
The code performs two main actions:

1. **Descriptive Statistics:** It calculates summary statistics (like mean, median, standard deviation) for each subject (math, reading, writing) to understand the overall distribution of scores. Which equal to:

	Descriptive Statistics by Subject		
	math_score	reading_score	writing_score
count	1000.00	1000.00	1000.00
mean	66.09	69.17	68.05
std	15.16	14.60	15.20
min	0.00	17.00	10.00
25%	57.00	59.00	57.75
50%	66.00	70.00	69.00
75%	77.00	79.00	79.00
max	100.00	100.00	100.00
Name	math_score	reading_score	writing_score
d.type:	float64	float64	float64

2. **Boxplot Analysis:** It creates boxplots to visualize how scores in each subject (math, reading, writing) are distributed across different student groups defined by categorical variables like ethnicity, gender, lunch type, parental education level, and test preparation course participation.




Math Scores by test_preparation_course
Reading Scores test_preparation_course
Writing Scores by test_preparation_course
math_score
reading_score
writing_score
none
completed
test_preparation_course
Math Scores by lunch
Reading Scores lunch
Writing Scores by lunch
0
1
lunch
Math Scores by parental_level_of_education
Reading Scores parental_level_of_education
Writing Scores by parental_level_of_education
bachelor's degree
high school
parental_level_of_education

References

Lab 7, Lab 8

Seaborn.boxplot#. seaborn.boxplot - seaborn 0.13.2 documentation. (n.d.). https://seaborn.pydata.org/generated/seaborn.boxplot.html

Learn. scikit. (n.d.). https://scikit-learn.org/stable/

Seaborn.heatmap#. seaborn.heatmap - seaborn 0.13.2 documentation. (n.d.). https://seaborn.pydata.org/generated/seaborn.heatmap.html

GeeksforGeeks. (2023, December 1). Pandas DataFrame Corr() method. https://www.geeksforgeeks.org/python-pandas-dataframe-corr/

Python statistics module. (n.d.). https://www.w3schools.com/python/module_statistics.asp