Pranavi Immani, Bhavya Vatsayavi, Allyson Wong

Team name: Spartan Analytics

CS22B-03

Professor Huynh-Westfall

May 3 2025

Table of Contents

Final Project Write-Up

1) Abstract

The purpose of this project is to analyze student academic performance using real-world data and identify patterns that may help educators understand student needs better. The dataset used for this project was found from a public Github repository system and is called "StudentsPerformance.csv", created by Rashida Nasrin. Our project used this dataset which contained the test scores of 1,000 students along with demographic information such as gender,

parental education level, lunch type, race/ethnicity, and participation in test preparation courses. Using these attributes, we built a reusable data analysis system in Python using Google Colab, using libraries like pandas, seaborn, and matplotlib to generate visualizations and insights.

2) Project Motivation & Dataset Description

One thing we've thought about when choosing to work on this project itself was how educators often lack certain intuitive tools to assess or understand when a student may need help. This also includes understanding how a student can fully thrive within subjects and get the full support they need. Usually, teachers may tend to realize but might not know how to initiate the assistance. As students ourselves, we wanted to try this to see if this could potentially assist and help teachers and professors. By analyzing the dataset, we aim to understand the effect of different factors on student outcomes and create visualizations to help educators identify and support at-risk students.

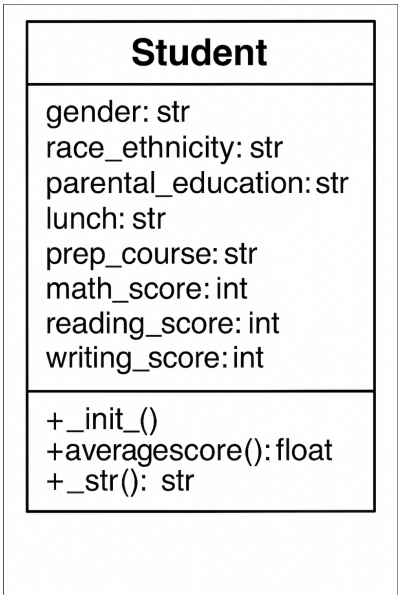
3) Research Questions

We formed questions to answer as our goal was to explore the relationship between these specific factors and use student performance to analyze it in Python on Colab. We formed these questions:

- ➢ Does gender influence academic performance across core subjects?
- ➢ How does parental education relate to student achievements?
- ➢ Do different racial/ethnic groups perform differently?
- ➢ What common traits do the top 10 students share?
- ➢ Does completing test prep significantly improve scores?

➢ Is there a strong correlation between math, reading and writing scores?

4) UML Design



Here is a UML of our Student class.

5) Object-oriented Programming Approach

In terms of OOP, we created a Student class which held individual student attributes and compute their average score.

```
# Creating Student Class

class Student:
    def __init__(self, gender, race_ethnicity, parental_education, test_prep, math_score, reading_score, writing_score):
        self.gender = gender
        self.race_ethnicity = race_ethnicity
        self.parental_education = parental_education
        self.test_prep = test_prep
        self.math_score = math_score
        self.reading_score = reading_score
        self.writing_score = writing_score

    def average_score(self):
        return (self.math_score + self.reading_score + self.writing_score) / 3
```

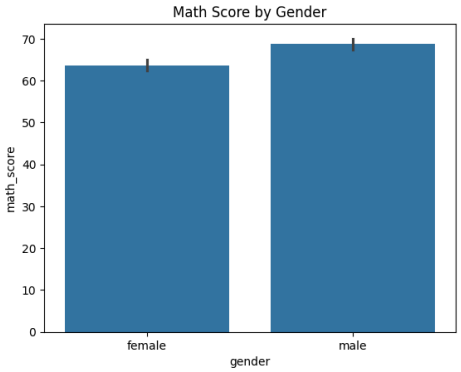
Each object stored a student's attribute (this includes gender, score, test prep, etc.) and calculated their average scores. This component helped make our code more structured and reusable. Furthermore, this project also provided us with a lot experience hands on in terms of data cleaning, visualization, and analytical thinking. We worked as a team to solve these problems

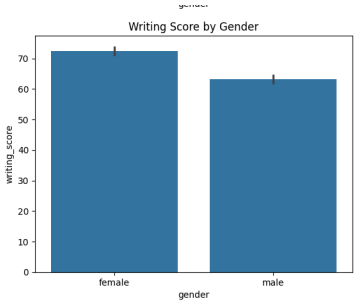
while applying both functional and OOP concepts. As mentioned earlier, we used Google Colab to implement our system in Python. Libraries such as pandas were used for our data cleaning and transformation, and seaborn/matplotlib for data visualization. Statistical tools were used, such as t-test and ANOVA, for significance testing. Our workflow was very organized, maintaining in this order:

- ➢ Column names
- ➢ Computing summary statistics
- ➢ Generating custom features (average + total score)
- ➢ Building the visualizations
- ➢ Applying hypothesis testing

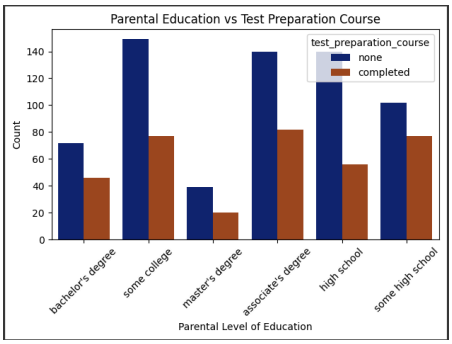
6) Data Visualization and Key Analysis

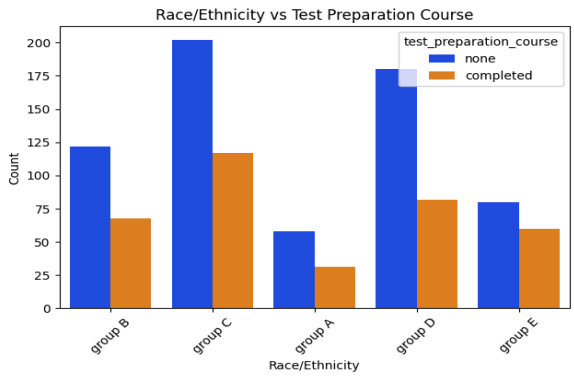
To begin with, for our analysis, we cleaned and normalized the dataset by using column names and confirming that no missing values were present. We created columns called average_score, whose purpose was to calculate the mean of each student's scores across math, reading, and writing. The new metric helped us summarize the overall performance and make good comparisons. Throughout our analysis, we used data visualization techniques to explore how performance was related to such demographics and educational background factors. Furthermore, we have found several important trends through analyzing this data. First, the gender appeared to influence performance. Female students consisted of scoring higher in subjects like reading and writing, while male students performed slightly better in subjects like math.

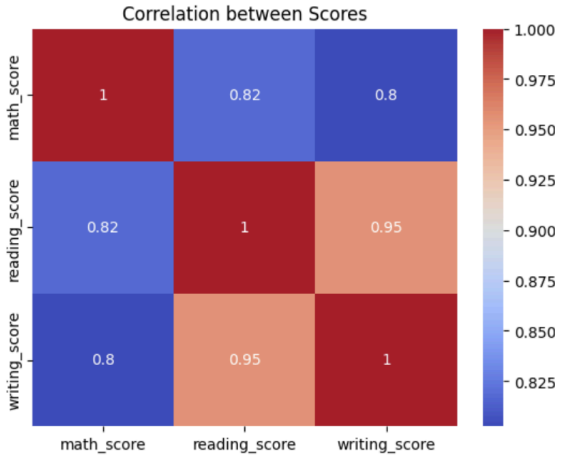




Participation in test prep courses also had a noticeable impact. For this, students who completed the course achieved higher average scores in all subjects. Additionally, parental education levels showed a positive correlation with student performance. Students whose parents held degrees (bachelors + master degrees) tend to score higher than those whose parents had completed up to high school/college without a degree. Not only that, but we also compared the race/ethnicity and rest prep courses to see which group has taken more prep courses.







We also observed the lunch types, which was a huge indicator for socioeconomic status, and how it was associated with their overall performance outcomes. Students receiving free/reduced price lunch scored much lower than their peers with a standard lunch/homemade lunch. This shows such external factors such as economic

background and how that might play a role into their academic success. Last but not least, our correlation analysis confirmed that the subjects math, reading, and writing scores are heavily interrelated. Students who excelled in those three subjects have seemed to perform well in all other subjects too.

7) Statistical Significance Testing

Overall, we expected to find strong differences in performance based on these demographics and academic performance factors. For testing our insights, we used t-test to compare gender based performance, and ANOVA to test variation between race/ethnic groups. We plotted bar graphs, boxplots, and heatmaps to visualize distribution and correlation (some images shown above are examples). These test helped validate that the patterns we analyzed were not indeed random. For example, the p-value for gender vs math score was really low, confirming a significant performance difference. Our testing strategy combined visual trends with a formal hypothesis test to ensure high accuracy.

```
from scipy.stats import ttest_ind

t_stat, p_value = ttest_ind(
    data[data['gender'] == 'male']['math_score'],
    data[data['gender'] == 'female']['math_score']
)
print(f"P-value: {p_value:.4f}")
```

```
P-value: 0.0000
```

```
[37] # Reading
ttest_ind(
    data[data['gender'] == 'male']['reading_score'],
    data[data['gender'] == 'female']['reading_score']
)

# Writing
ttest_ind(
    data[data['gender'] == 'male']['writing_score'],
    data[data['gender'] == 'female']['writing_score']
)
```

```
TtestResult(statistic=np.float64(-9.979557910004507), pvalue=np.float64(2.019877706867934e-22), df=np.float64(998.0))
```

8) Conclusion

In conclusion, the student performance analytics system we have built served as a valuable tool for being able to process and visualizing such educational data, while also identifying key performance indicators. The results could inform strategies for targeted interventions and improved support for students, especially if they are facing economic or academic challenges. This project could potentially be expanded by incorporating different but predictive modeling techniques by building a dashboard to display results in other formats that might be easier to see. Overall, we have strengthened our data analysis skills and deepened our understanding of how we can use this to analyze and compare datasets.

9) References

Data Source used: https://github.com/rashida048/Datasets/blob/master/StudentsPerformance.csv

Student Performance Analysis Notebook:

https://github.com/sharmaroshan/Students-Performance-Analytics/blob/master/Student_Performance.ipynb