Student Performances on Exams

Brandon Boyle, bboyle@bellarmine.edu

I. INTRODUCTION

The data set that I am working with looks at students performance over three exams taking into account outside factors such as lunch, test prep work, parents background, gender, and race. My goal for this project is to dial into the effects that lunch and test prep work have on students performance on exams. So whether a good lunch is actually beneficial or if test prep courses help improve exam scores.

II. BACKGROUND

The goal for this data set is to see if outside factors can have an effect on students performance on exams. Its looks at the scores from math, reading, and writing exams and also it looks at if that student had lunch, did the test prep, their parents background, and the students race and gender

III. EXPLORATORY ANALYSIS

This section will be similar to your exploratory analysis project. First, provide a summary of the data set similar to your first exploratory analysis: *e.g. this data set contains 398 samples with 7 columns with various data types*. In this summary, provide the data types of your columns (in a table) and then rather than providing tabular statistics and plots for each variable, provide only statistics and plots that seem unusual. For example, if one or two variables have significant missing values or the distribution of the variable is skewed or looks unusual note that. Provide the unusual statistics or plots in this section. Provide any other appropriate plots (e.g. correlation matrix, heatmaps, bar charts, etc.) that you deem necessary.

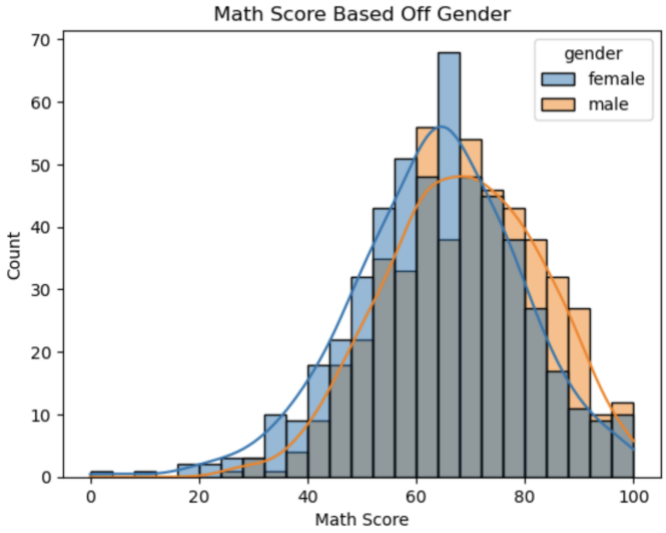
This data set had 8 columns and 1000 entries. All of which have no null values.

Table 1: Data Types

```
df.info()
```

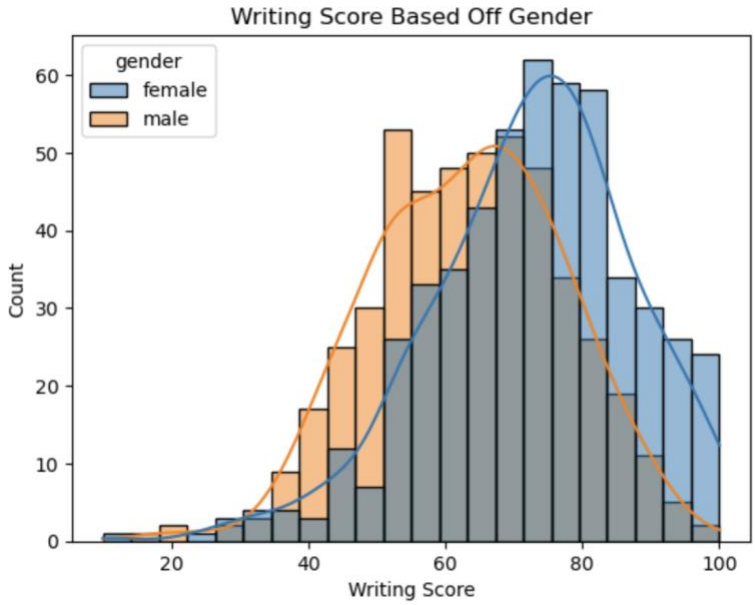
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64 
 6   reading score                1000 non-null   int64 
 7   writing score                1000 non-null   int64 
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

Graph 1 – Math Scores Based off Gender



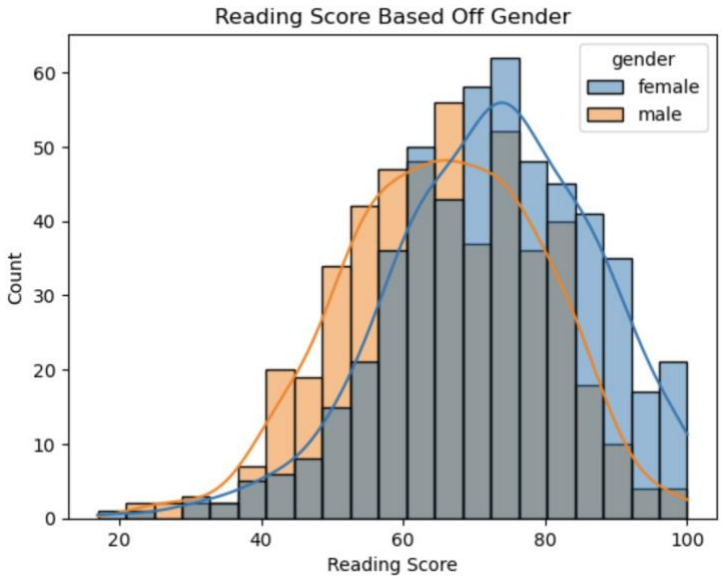
This graph shows that overall, man typically scored higher on the math exam then females. The graph is fairly symmetric with a couple outliers/a bigger tail on the left. The males bell curve is slightly shifted to the right compared to the females.

Graph 2 – Writing Score Based off Gender



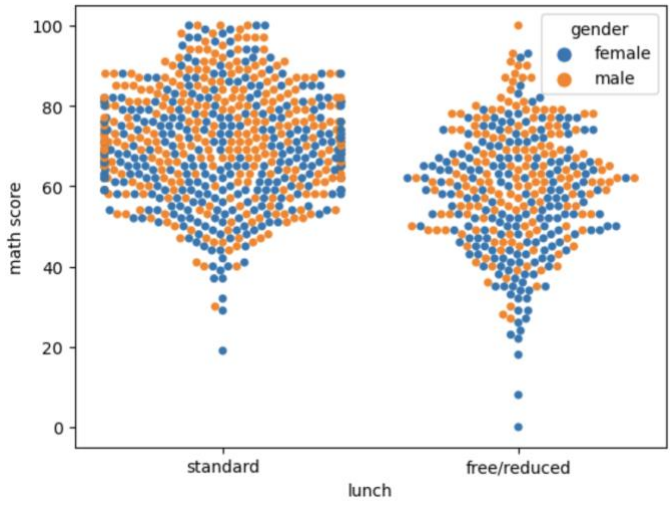
This graph shows us that taking everything into account, women tended to score higher on the writing exam with the majority of the data around an 80 while the mens majority is between 50 and 70. Both graphs are slightly left skewed.

Graph 3 – Reading Score Based Off Gender



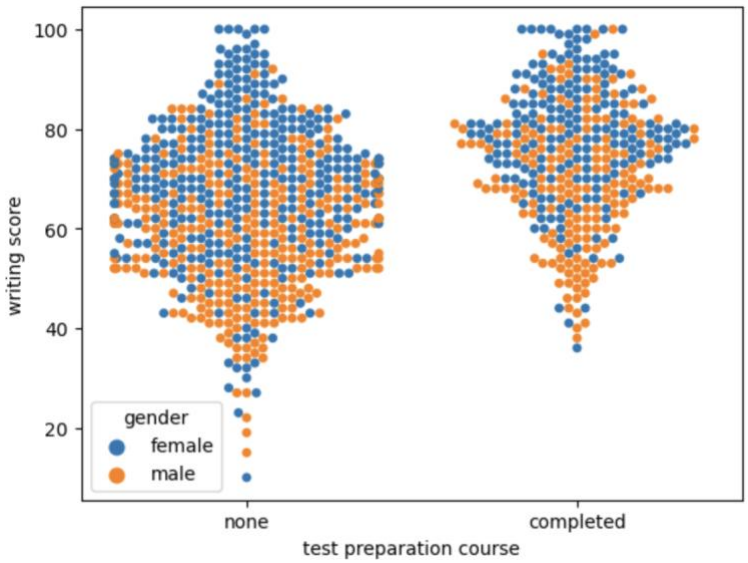
In this graph we see a similar trend to the writing scores with women tending to score higher on the reading score then men. The females curve is slightly skewed left while the males curve is just about symmetrical.

Graph 4 – Swarm plot of math test scores based off the two lunch options - color based on gender



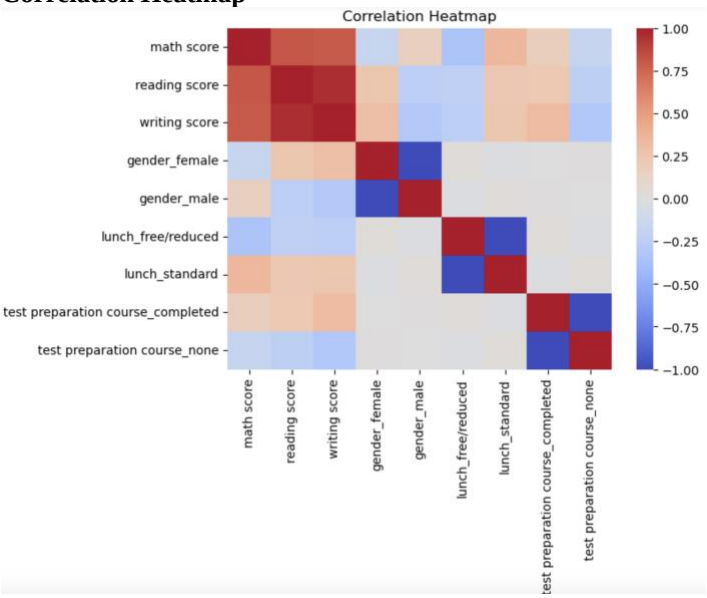
I wanted to get a look at the difference that lunch made to test scores and this graph about the math exam is a good visual representation of that. As we can see, more students had a standard lunch and because of that they tended to score higher on the exam. Gender is fairly random in their scores but based off the histogram from earlier, we know that males scored slightly higher on the math exam.

Graph 5 – Swarm Plot of writing scores based of test preparation – color based on gender



I found this graph really interesting. We can see that while most people did not do the test prep course, they tended to score lower on the writing exam. But we can also see more clearly that overall, test prep or not, females tended to score higher on this exam.

Correlation Heatmap



When I first made the heatmap, I included all the variable including race and parents level of education but when looking at the correlations, I found that those two variables had a very little and insignificant correlation so I just decided to drop them out of the final correlation heatmap. But from looking at the heatmap, we can see that there is a positive correlation between eating a standard lunch and getting a higher test score. There is also a positive correlation for completing the test prep and getting a better test score, while there are negative correlations between no lunch and test scores along with no test prep and test score. This backs up the argument that a good lunch and doing test prep is beneficial.

IV. METHODS

A. Data Preparation

This data set had no null values but after I looked at the correlations between all the variables, I decided to drop both race and parents level of education. They both had extremely low correlations compared to other variables and they both had a bunch of sub categories that were getting in the way.

B. Experimental Design

:

Table X: Experiment Parameters

Experiment Number	Parameters
	I did my experiments slightly different. I split the data with a test size of 20% but I did three different experiments with the test size of 20. All three had an x of ('gender_female', 'gender_male', 'lunch_free/reduced', 'lunch_standard', 'test preparation course_completed', 'test preparation course_none') but each had a different y as I wanted to look at all three test, math, writing, and reading.
1	20% split of data, compared math test score to the seven other variables
2	20% split of data, compared reading test score to the seven other variables
3	20% split of data, compared writing test score to the seven other variables

```
x=score[['gender_male','gender_female', 'lunch_free/reduced', 'lunch_standard',
       'test preparation course_completed', 'test preparation course_none']]
y=score[['math score']]
```

```
x1=score[['gender_male','gender_female', 'lunch_free/reduced', 'lunch_standard',
       'test preparation course_completed', 'test preparation course_none']]
y1=score[['reading score']]
```

```
x2=score[['gender_male','gender_female', 'lunch_free/reduced', 'lunch_standard',
       'test preparation course_completed', 'test preparation course_none']]
y2=score[['writing score']]
```

C. Tools Used

Anaconda Navigator 2.3.1, Jupyter Notebook 6.4.12, Python 3 for apple. In addition to Python I used the following libraries, numpy, pandas, matplotlib, and seaborn.

V. RESULTS

A. Classification Measures/ Accuracy measure

For Math Scores –

```
print(f"MSE: {mean_squared_error(y_test,y_pred):.2f}")
print(f"RMSE: {math.sqrt(mean_squared_error(y_test,y_pred)):.2f}")
print(f"R-square: {r2_score(y_test,y_pred):.2f}")
```

```
MSE: 185.76
RMSE: 13.63
R-square: 0.18
```

For Reading Scores –

```
print(f"MSE: {mean_squared_error(y1_test, y1_pred):.2f}")
print(f"RMSE: {math.sqrt(mean_squared_error(y1_test, y1_pred)):.2f}")
print(f"R-square: {r2_score(y1_test, y1_pred):.2f}")
```

```
MSE: 192.46
RMSE: 13.87
R-square: 0.12
```

For Writing Scores –

```
print(f"MSE: {mean_squared_error(y2_test, y2_pred):.2f}")
print(f"RMSE: {math.sqrt(mean_squared_error(y2_test, y2_pred)):.2f}")
print(f"R-square: {r2_score(y2_test, y2_pred):.2f}")
```

```
MSE: 192.63
RMSE: 13.88
R-square: 0.16
```

B. Discussion of Results

All of my models have pretty much the same RMSE and R-square. The biggest difference in the models is the MSE for the math scores. All the models are fairly similar as they all have the same x values and its only the y that is changing. I do think that the math scores MSE having the biggest difference is due to the fact that for math, the difference between the male and female scores is way smaller than the difference in genders for reading and writing scores.

C. Problems Encountered

I originally had a data set about shopping trends but as I was going through the process of making graphs and working on the MLR I realized that the data just wasn't sufficient enough to get solid results from so I switched my data set that I was using to this one about students exam scores. While it wasn't so much a problem, something that I had to deal with was the fact that I had three different subjects to look at for test scores so rather than making three models with different test sizes, I just decided to do three models that looked the three test individually.

D. Limitations of Implementation

There are many other factors that go into taking a test then just your lunch, a test prep course, and your parents schooling. Factors such as sleep, how much the student studied, if they were well focused during the exam, and many other little things can go into the grade these students get on their test.

E. Improvements/Future Work

I would have liked to look at a couple more variables such as sleep, time spent on studying, and any other distractions that could alter a students performance on a test.

VI. CONCLUSION

In the end I was able to pull some interesting trends from the data as well as answer the question I had originally wanted to answer. While men typically performed better on the math test, women tended to do better on both reading and writing test. I was also able to find that students who completed the test prep course as well as had a good lunch, they were the ones to do the best on the exams. But if you were able to either have a good lunch or complete the test prep course, you would have done much better than those who did neither. So for the most part the model did want it needed to and answered the questioned I had sought to find out.

REFERENCES

https://www.kaggle.com/datasets/spscientist/students-performance-in-exams
I used Kaggle for my dataset