Stutern Challenge

August 7, 2019

1 Abstratct

Measuring the performance of student is important for our educational system. The student performance analysis intends to show how each students perform by comparing the students score to different variables. The univariate analysis done in this project describes each variable independently while the multivariate analysis is done to find relationships between one or more variales. In order to improve the performsance of tstudents we have to know why and how they are being affected.

2 Introduction

In this project, we are going to analyze students performance in order to determine why they are doing well or bad, and what steps can be taken to improve on the current performance on the students.

3 Data Preparation

We first import all the models we will use for this project, then we read the data.After that we get basic info about the dataset which shows us that we have 1000 rows and 8 coulmns, 3 integer columns which are the scores of the students and 5 string columns. We also check if there are any missing values in the data set and we find none.

```
In [1]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
        from scipy.stats import kurtosis, skew, pearsonr

In [2]: data = pd.read_csv('StudentsPerformance (5).csv')
        data.head()

Out[2]:    gender race/ethnicity parental level of education         lunch  \
        0  female        group B           bachelor's degree      standard
        1  female        group C                some college      standard
        2  female        group B             master's degree      standard
        3    male        group A          associate's degree  free/reduced
```

```
4    male        group C                some college      standard

  test preparation course  math score  reading score  writing score
0                    none          72             72             74
1               completed          69             90             88
2                    none          90             95             93
3                    none          47             57             44
4                    none          76             78             75
```

```
In [3]: #get various information about the data set
        data.info()
```

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

```
In [4]: #check for null values
        data.isnull().sum()
```

```
Out[4]: gender                         0
        race/ethnicity                 0
        parental level of education    0
        lunch                          0
        test preparation course        0
        math score                     0
        reading score                  0
        writing score                  0
        dtype: int64
```

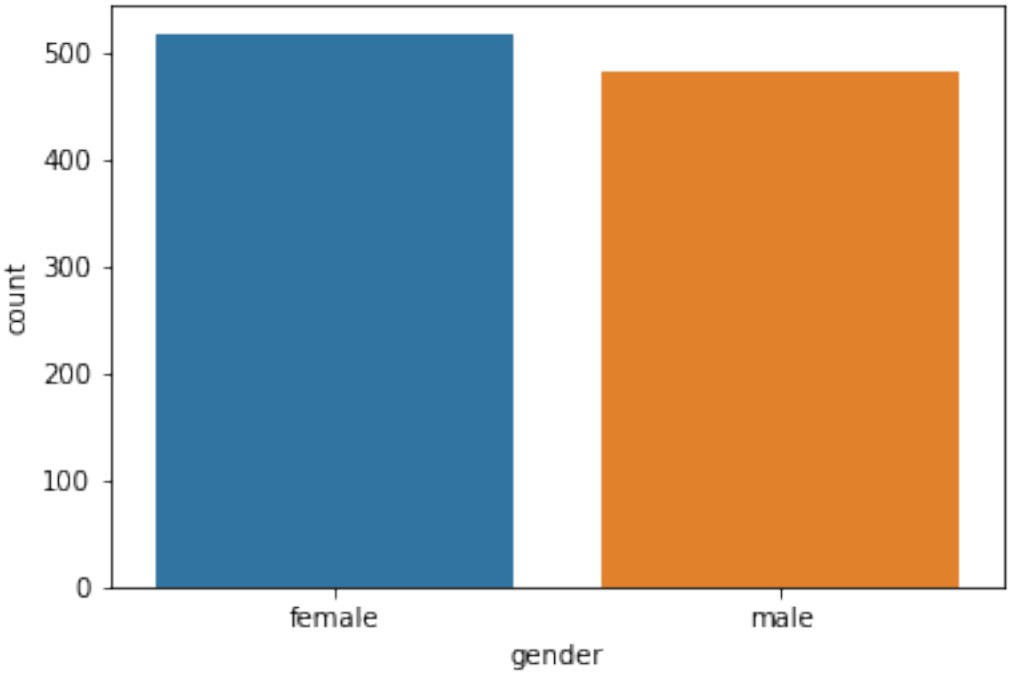
4 Univariate Analysis

From the below plot we see that the female gender is the more dominant gender than the male gender

```
In [5]: #Gender
        print(data['gender'].value_counts())
        sns.countplot(data['gender'])
```

```
female    518
male      482
Name: gender, dtype: int64
```

```
Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x7f9375f1d2b0>
```

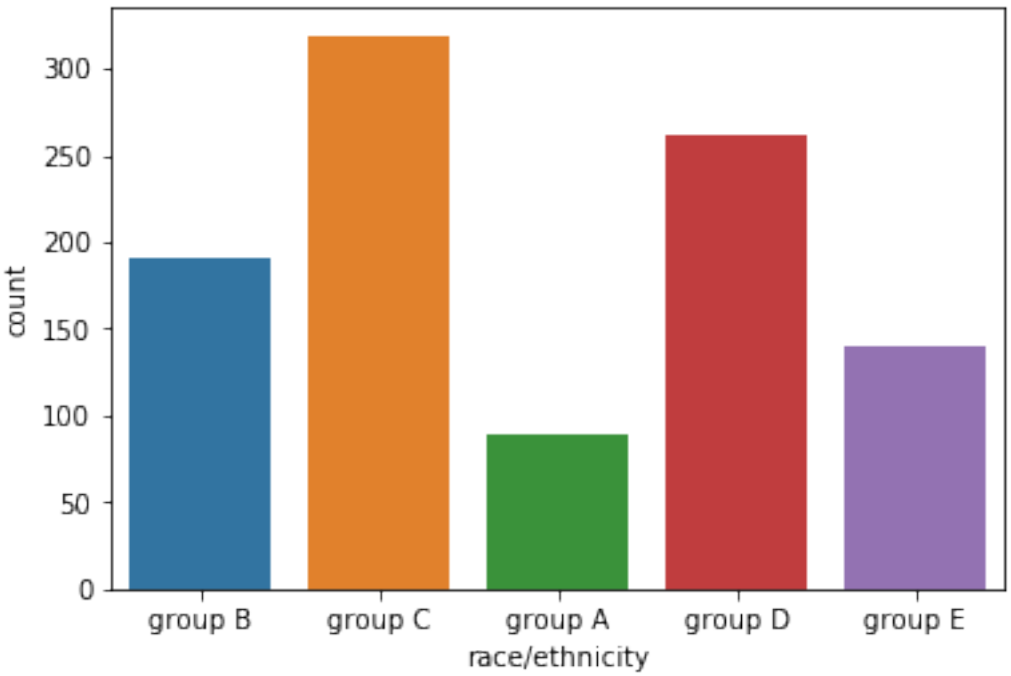


From the below plot we see that the the group C race are more dominant followed by the group D which is followed by group B then group E and lastly A

```
In [6]: print(data['race/ethnicity'].value_counts())

        sns.countplot(data['race/ethnicity'])
```

```
group C    319
group D    262
group B    190
group E    140
group A     89
Name: race/ethnicity, dtype: int64
```

```
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x7f9375f5bdd8>
```

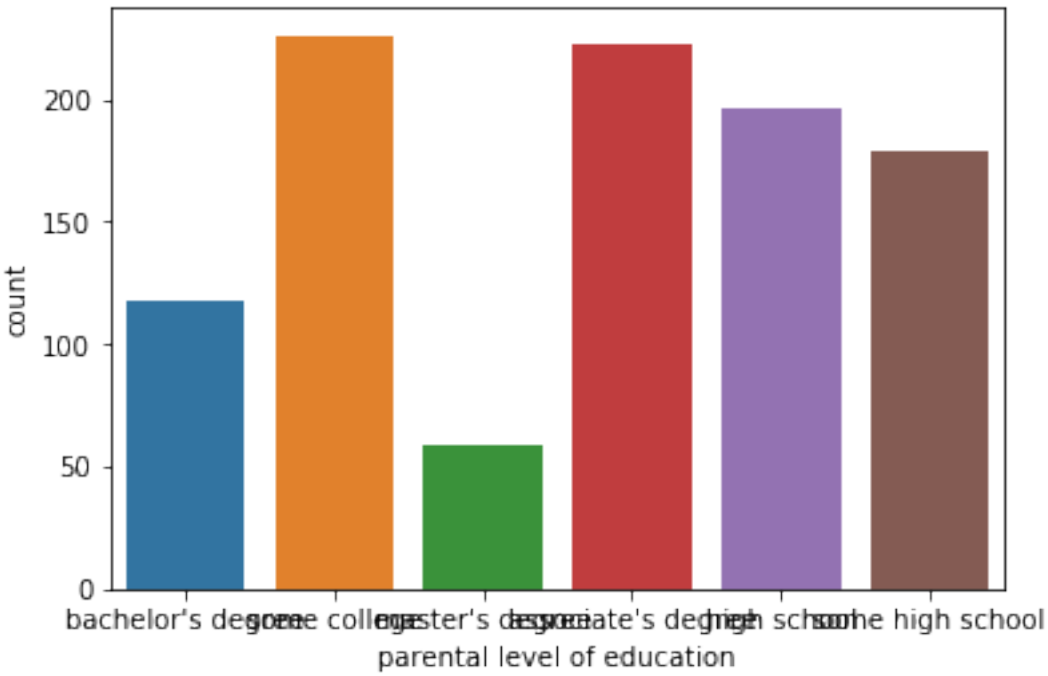


From the below plot we see that parent's with some college degree are more dominant that the rest, with parents with master's degree having being the least dominant.

```
In [7]: #get the number representing each categorical value
        print(data['parental level of education'].value_counts())

        sns.countplot(data['parental level of education'])
```

```
some college          226
associate's degree    222
high school           196
some high school      179
bachelor's degree     118
master's degree        59
Name: parental level of education, dtype: int64
```

```
Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x7f936e345588>
```

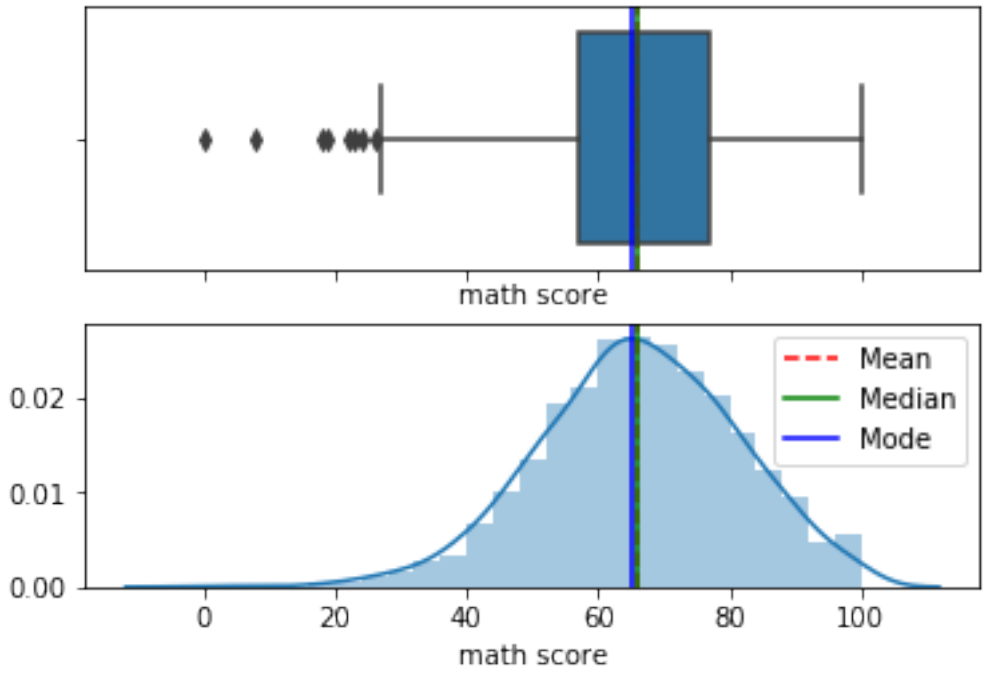


The box plot and histogram below both shows the distribution of math scores. The histogram is a normal distribution because it is symmetrical and that also tells us that our mean equals mode equals median. The boxplot shows that we have some outliers.

```
In [8]: f, (ax_box, ax_hist) = plt.subplots(2, sharex=True)
        mean=data['math score'].mean()
        median=data['math score'].median()
        mode=data['math score'].mode().get_values()[0]

        sns.boxplot(data['math score'], ax=ax_box)
        ax_box.axvline(mean, color='r', linestyle='--')
        ax_box.axvline(median, color='g', linestyle='-')
        ax_box.axvline(mode, color='b', linestyle='-')


        sns.distplot(data['math score'], ax=ax_hist)
        ax_hist.axvline(mean, color='r', linestyle='--')
        ax_hist.axvline(median, color='g', linestyle='-')
        ax_hist.axvline(mode, color='b', linestyle='-')

        plt.legend({'Mean':mean,'Median':median,'Mode':mode})

        print('Skewness of distribution is ' + str(skew(data['math score'])))
        print('Kurtosis of distribution is ' + str(kurtosis(data['math score'])))
```

```
/home/godwin/anaconda3/lib/python3.6/site-packages/matplotlib/axes/_axes.py:6462: UserWarning:
```

```
  warnings.warn("The 'normed' kwarg is deprecated, and has been "


Skewness of distribution is -0.278516571914075
Kurtosis of distribution is 0.26759715461497846
```

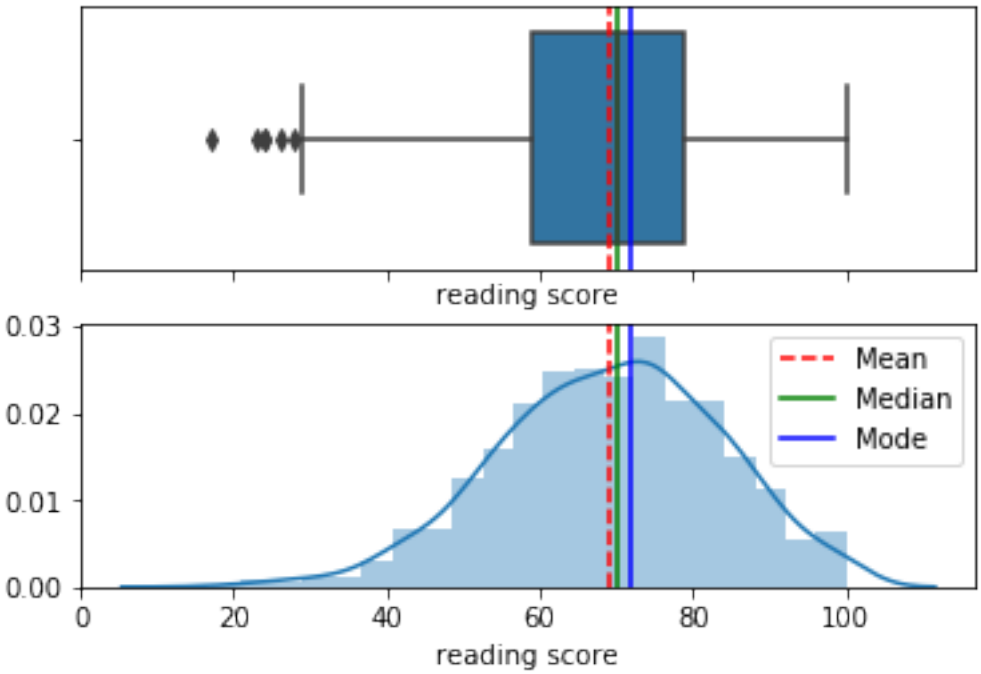


The histogram below shows the reading scores are negatively skewed, the mode score greater than the median, median score greater than the mean score. We can also se we have some outliers which are low scores.

```
In [9]: f, (ax_box, ax_hist) = plt.subplots(2, sharex=True)
        mean=data['reading score'].mean()
        median=data['reading score'].median()
        mode=data['reading score'].mode().get_values()[0]

        sns.boxplot(data['reading score'], ax=ax_box)
        ax_box.axvline(mean, color='r', linestyle='--')
        ax_box.axvline(median, color='g', linestyle='-')
        ax_box.axvline(mode, color='b', linestyle='-')

        sns.distplot(data['reading score'], ax=ax_hist)
        ax_hist.axvline(mean, color='r', linestyle='--')
        ax_hist.axvline(median, color='g', linestyle='-')
        ax_hist.axvline(mode, color='b', linestyle='-')
```

```
        plt.legend({'Mean':mean,'Median':median,'Mode':mode})

        print('Skewness of distribution is ' + str(skew(data['reading score'])))
        print('Kurtosis of distribution is ' + str(kurtosis(data['reading score'])))
```

```
/home/godwin/anaconda3/lib/python3.6/site-packages/matplotlib/axes/_axes.py:6462: UserWarning:
  warnings.warn("The 'normed' kwarg is deprecated, and has been "


Skewness of distribution is -0.25871569927829347
Kurtosis of distribution is -0.07391861478331307
```



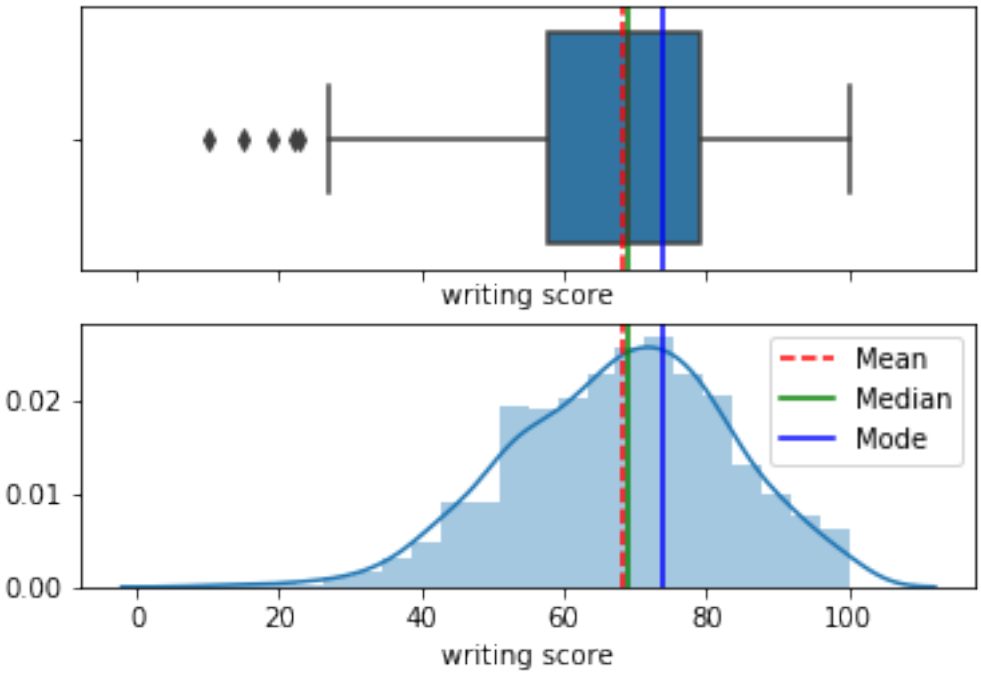
The histogram above shows wrtitng scores are negatively skewed, there are some outliers below the minimum score of the boxplot.

```
In [10]: f, (ax_box, ax_hist) = plt.subplots(2, sharex=True)
         mean=data['writing score'].mean()
         median=data['writing score'].median()
         mode=data['writing score'].mode().get_values()[0]

         sns.boxplot(data['writing score'], ax=ax_box)
         ax_box.axvline(mean, color='r', linestyle='--')
         ax_box.axvline(median, color='g', linestyle='-')
         ax_box.axvline(mode, color='b', linestyle='-')
```

```
        sns.distplot(data['writing score'], ax=ax_hist)
        ax_hist.axvline(mean, color='r', linestyle='--')
        ax_hist.axvline(median, color='g', linestyle='-')
        ax_hist.axvline(mode, color='b', linestyle='-')

        plt.legend({'Mean':mean,'Median':median,'Mode':mode})

        print('Skewness of distribution is ' + str(skew(data['writing score'])))
        print('Kurtosis of distribution is ' + str(kurtosis(data['writing score'])))
```

```
/home/godwin/anaconda3/lib/python3.6/site-packages/matplotlib/axes/_axes.py:6462: UserWarning:
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

```
Skewness of distribution is -0.28900962452114176
Kurtosis of distribution is -0.03919203131162252
```



5 MultiVariate

The data description shows us the statistical analysis of our numerical columns, we can see the mean and standard deviation for our scores there and also the percentiles of the each columns.

We can see the percentiles for our various columns which are the 25%, 50% and 75% The 25 percentile for math scores is 66 which tells us that 25% of our math scores fall below the 57 score.

The 50 percentile for our math score tells us that's the median of our scores is 50. The 75 percentile for our math score tells us that 75% of our dataset falls below 77 score.

```
In [11]: #data description
         data.describe()
```

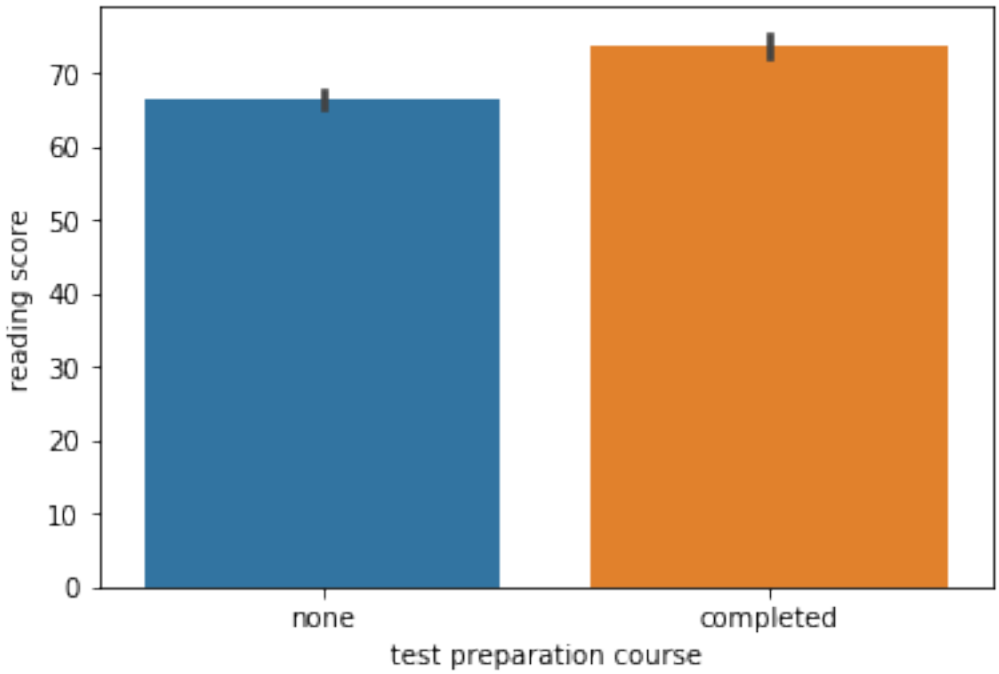
Out[11]:

	math score	reading score	writing score
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000

The below barplots compares the the various scores against the test preparation variables and we can see that overall, those who completed the test preparation courses had better scores.

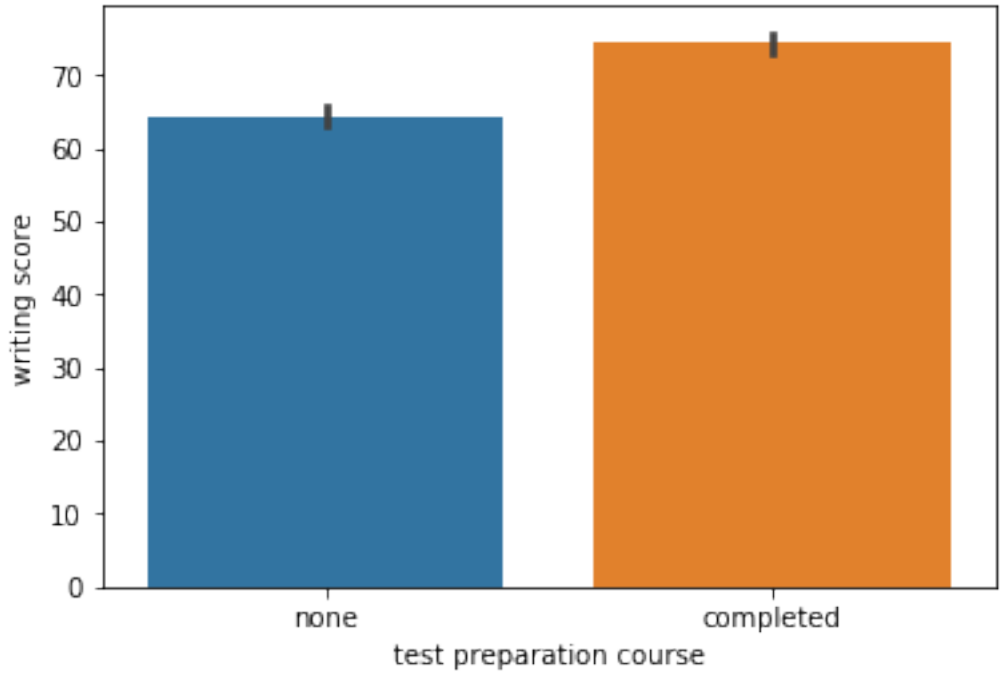
```
In [12]: sns.barplot(data['test preparation course'], data['reading score'])
```

```
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x7f936e122d30>
```



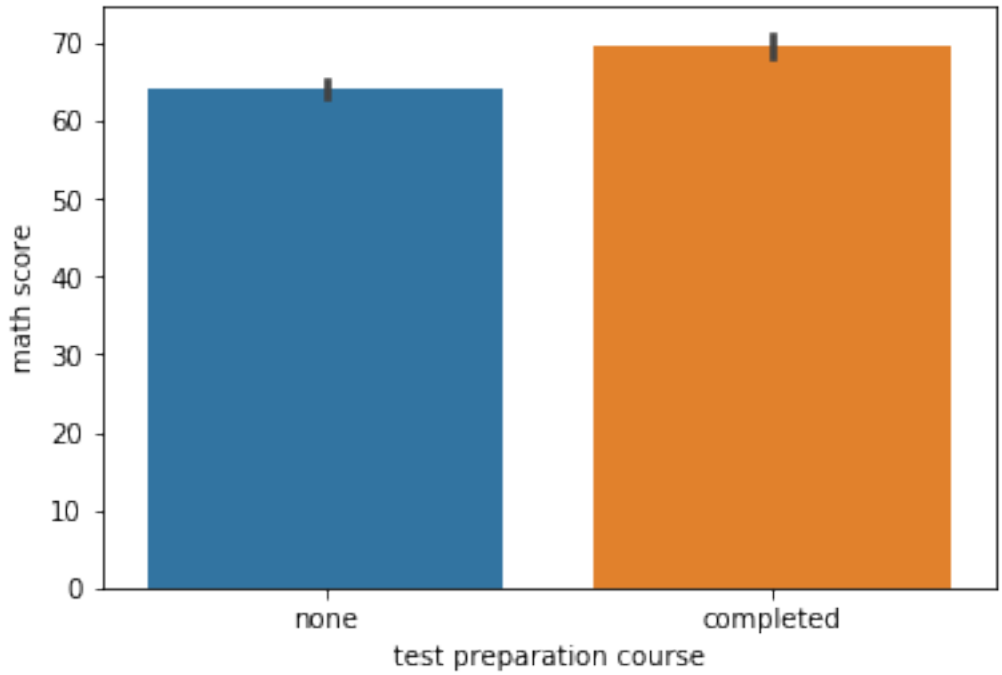

```
In [13]: sns.barplot(data['test preparation course'], data['writing score'])
```

```
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x7f936e0240b8>
```




```
In [14]: sns.barplot(data['test preparation course'], data['math score'])
```

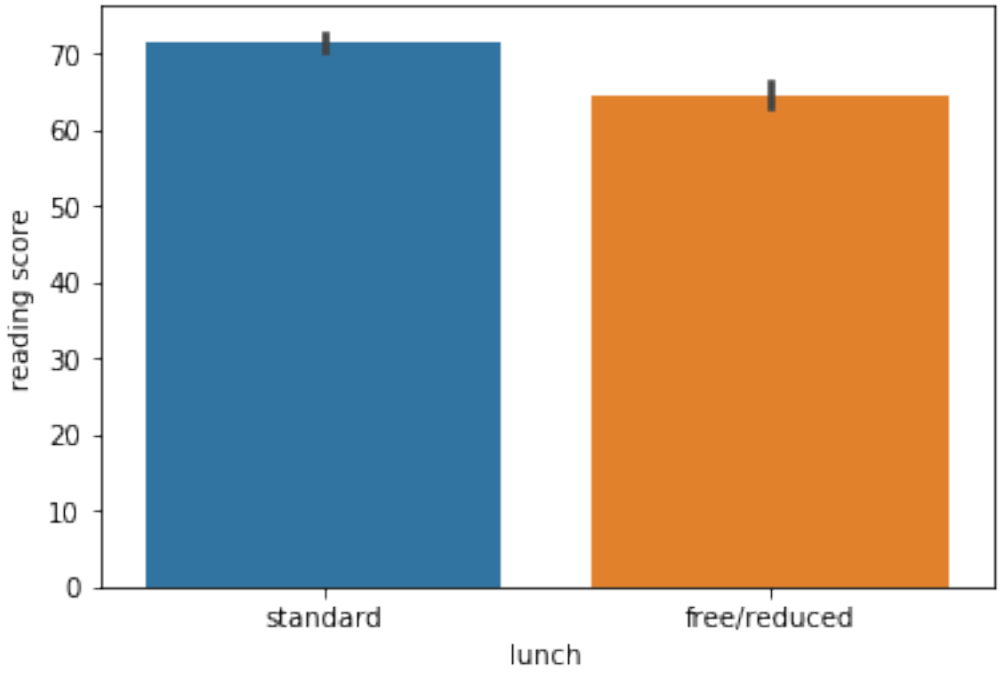
```
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x7f936e1336a0>
```



The below barplots compares the type of launch eaten by each student and their various scores. The plot shows that students who had the standard launch did better.

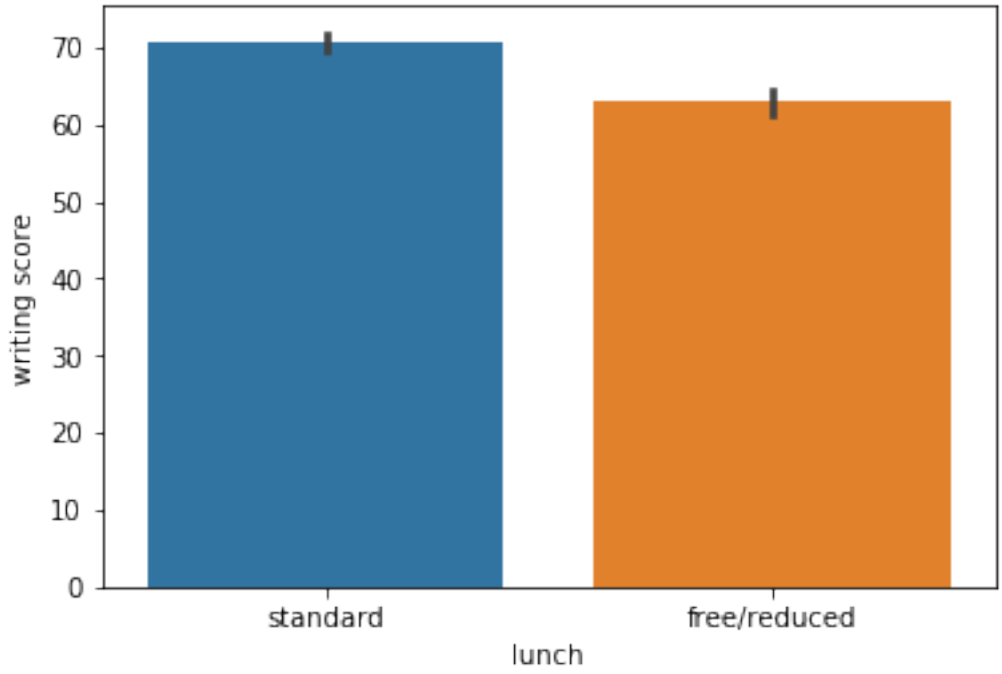
```
In [15]: sns.barplot(data['lunch'], data['reading score'])
```

```
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x7f936e04c5f8>
```



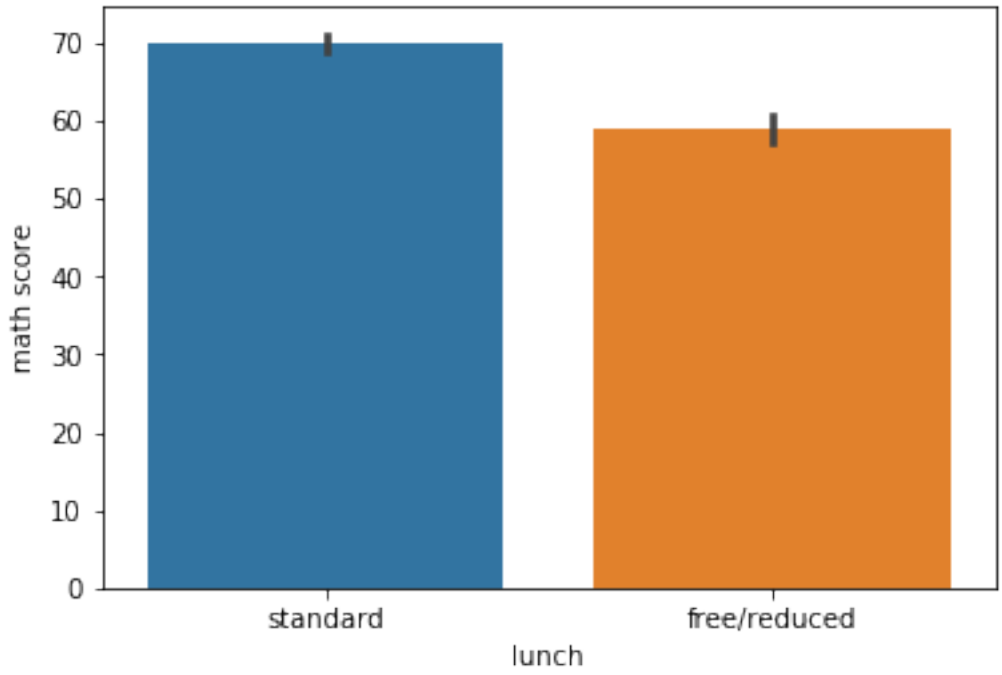

```
In [16]: sns.barplot(data['lunch'], data['writing score'])
```

```
Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x7f936df00978>
```




```
In [17]: sns.barplot(data['lunch'], data['math score'])
```

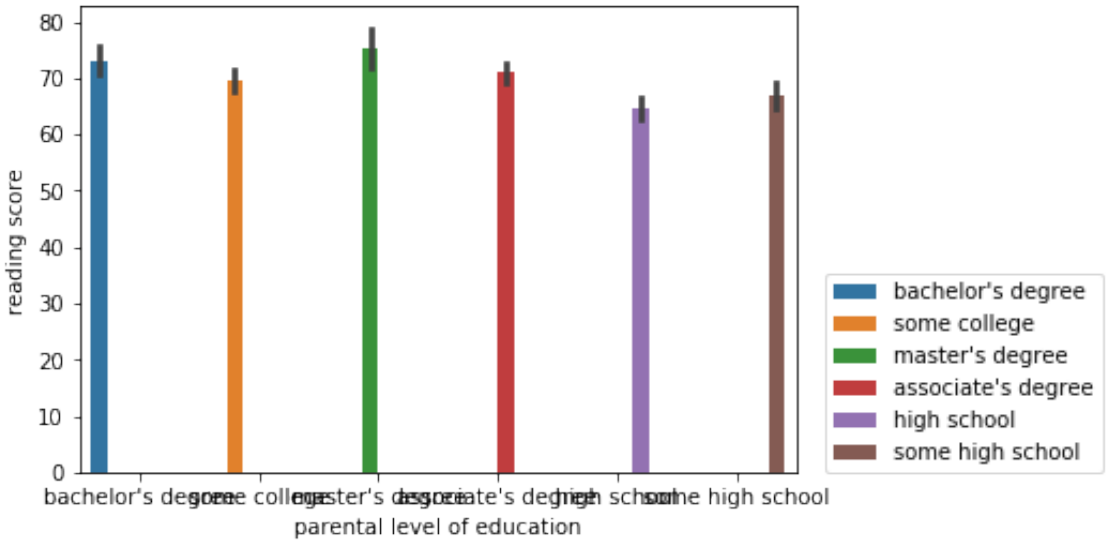
```
Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x7f936e176080>
```



The below plots shows that students with parents who had better academic background performed better than the rest. Although the difference is not much.

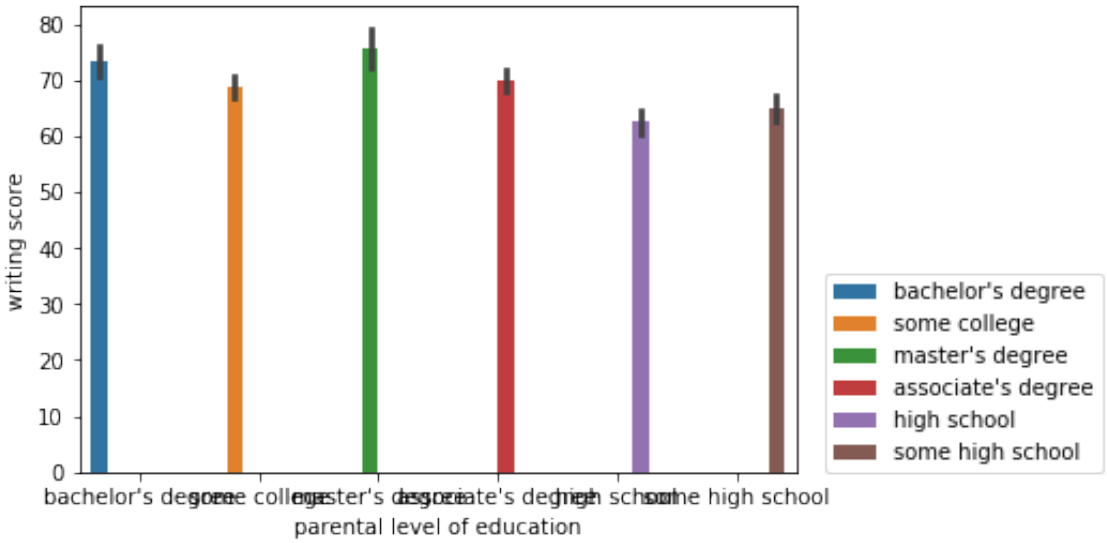
```
In [18]: ax =sns.barplot(data['parental level of education'], data['reading score'], hue=data[
         ax.legend(loc=(1.04, 0))
```

```
Out[18]: <matplotlib.legend.Legend at 0x7f936e0c2860>
```

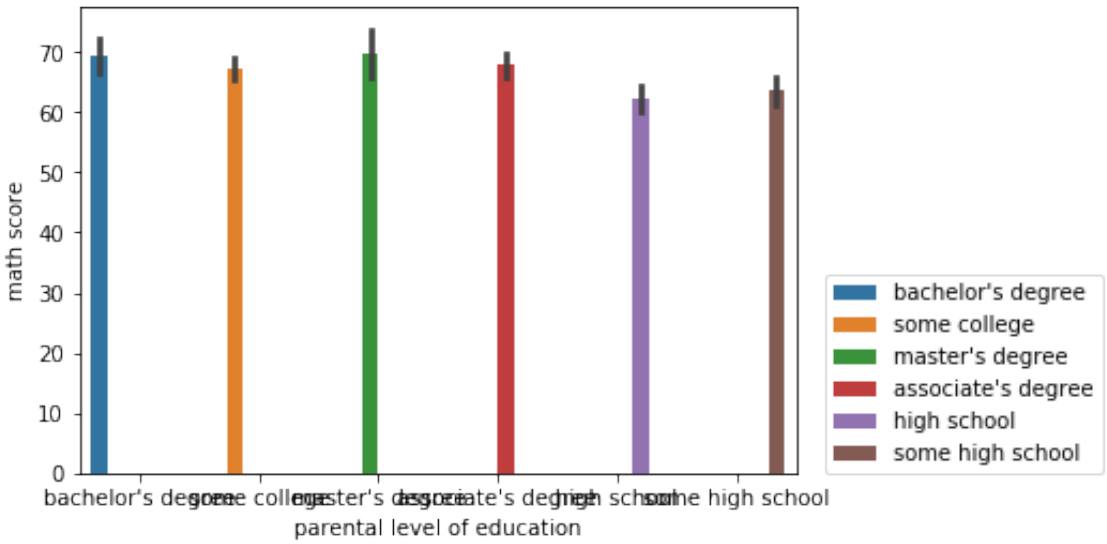



```
In [19]: ax =sns.barplot(data['parental level of education'], data['writing score'], hue=data[
         ax.legend(loc=(1.04, 0))
```

```
Out[19]: <matplotlib.legend.Legend at 0x7f936dea4710>
```

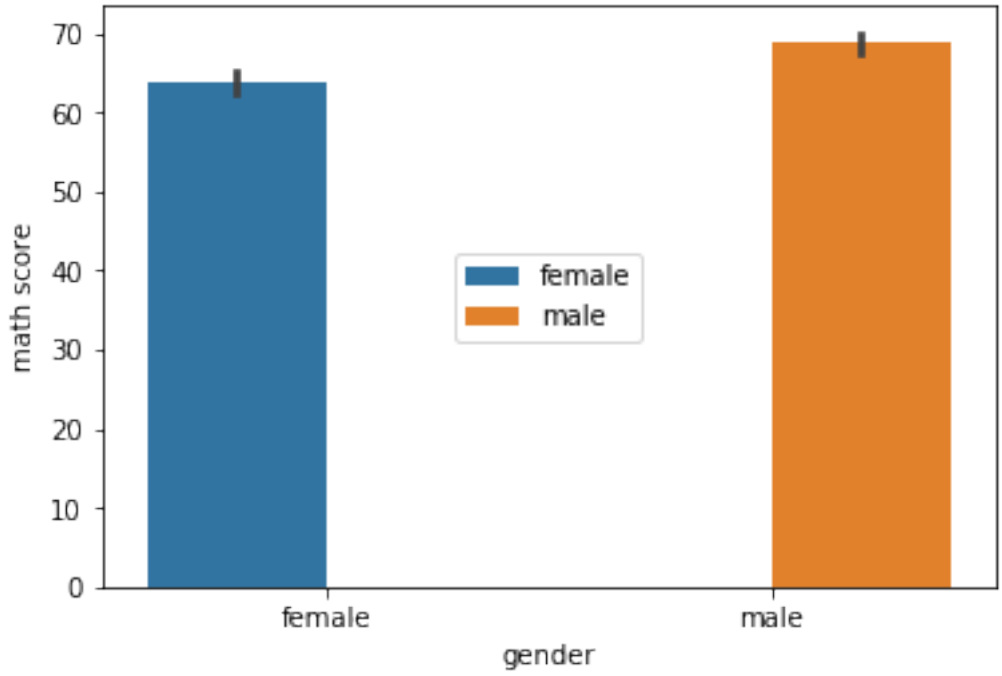


```
In [20]: ax = sns.barplot(data['parental level of education'], data['math score'], hue=data['pa
         ax.legend(loc=(1.04, 0))

Out[20]: <matplotlib.legend.Legend at 0x7f936ddba5c0>
```



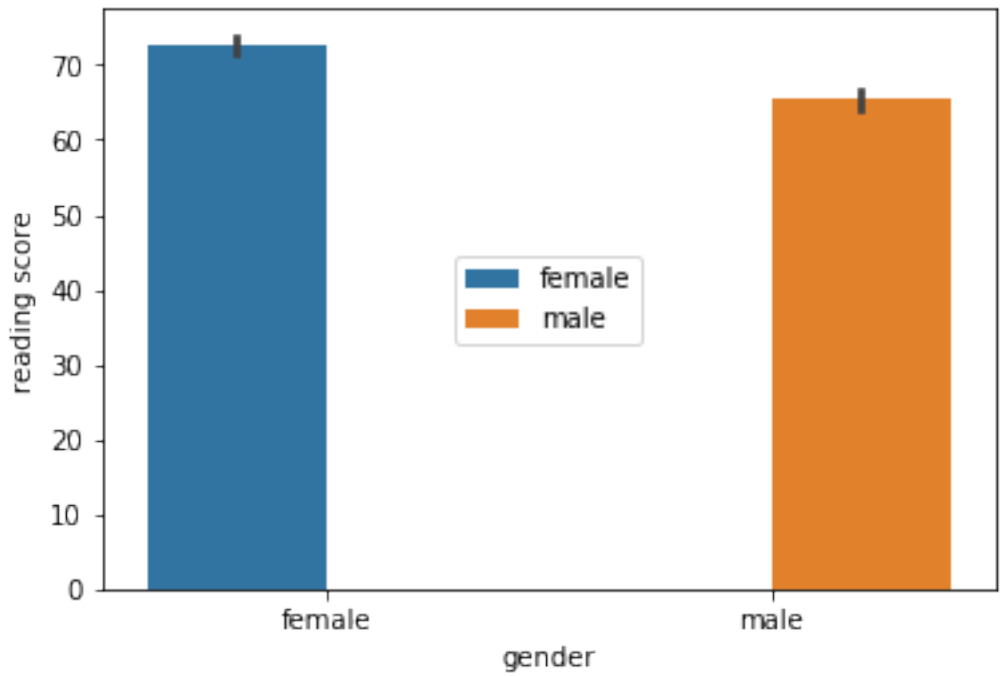
The below bar plots tries to check which gender performs best. We can see that the male gender had better performance from the female gender in maths.While the female performed better than the male in both readind and writing scores.

```
In [21]: ax = sns.barplot(data['gender'], data['math score'], hue=data['gender'])
         ax.legend(loc='center')

Out[21]: <matplotlib.legend.Legend at 0x7f936dc66128>
```



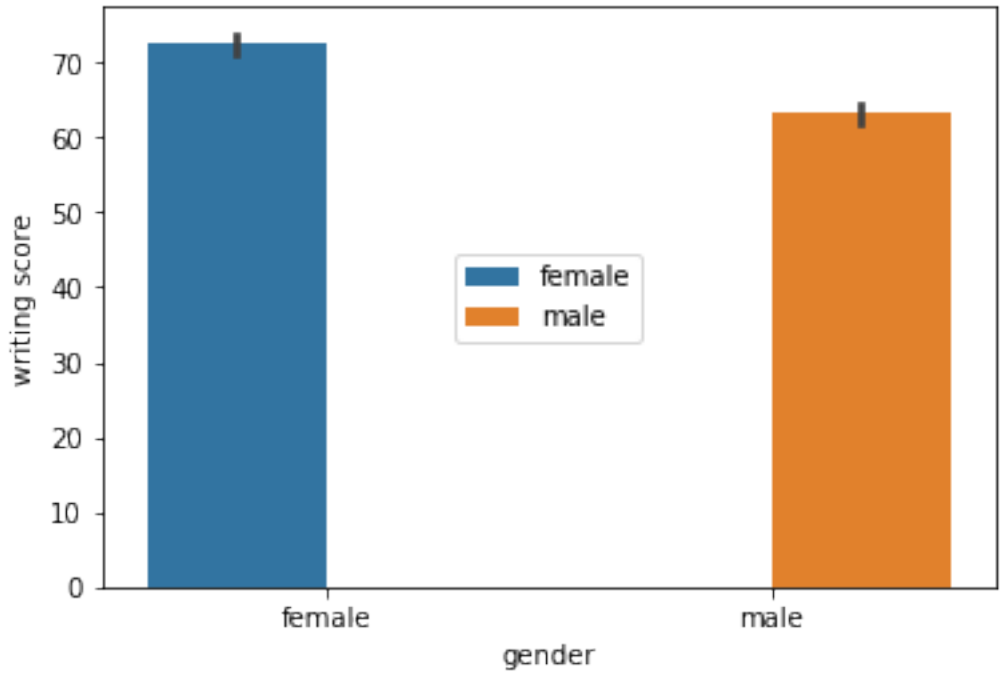

```
In [22]: ax = sns.barplot(data['gender'], data['reading score'], hue=data['gender'])
         ax.legend(loc='center')
```

```
Out[22]: <matplotlib.legend.Legend at 0x7f936db87d68>
```



```
In [23]: ax = sns.barplot(data['gender'], data['writing score'], hue=data['gender'])
         ax.legend(loc='center')
```

```
Out[23]: <matplotlib.legend.Legend at 0x7f936daf2240>
```



From the below correlation coefficient. There is a high positive relationship between all scores. Which shows that as one's score in a subject improves, the scores of the other subjects improves too.

```
In [24]: corr, _ = pearsonr(data['reading score'], data['writing score'])
         print(corr)
```

```
0.9545980771462478
```

```
In [25]: corr, _ = pearsonr(data['reading score'], data['math score'])
         print(corr)
```

```
0.817579663672054
```

```
In [26]: corr, _ = pearsonr(data['math score'], data['writing score'])
         print(corr)
```

```
0.8026420459498078
```

6 Conclusion and Recommendation

In conclusion we see that there are a lot of relationships between our variables and how students perform in the various subjects.

I will recommend that to improve the performance of the student lacking. They should be given the same launch, standard launch and also be encouraged and monitored to complete the test preparaton courses.