ALY6010 22162 Probability Theory & Introductory Statistics SEC 17 Winter 2022 CPS [SJO-B-HY]

Module 2 Practice Assignment

Submitted by: Sourabh D Khot

NUID: 002754952

Submitted to: Professor Behzad Ahmadi

Date: March 11, 2022

Introduction

In this exercise, I have used the 'Students Performance in Exams' dataset from Kaggle, which has some numerical attributes of marks secured and some categorical attributes of the students. We will clean the data, analyze it, draw interpretations and address why we use a specific form of tables and plots.

Data Preparation

In [1]:

```
import pandas as pd                 #importing libraries
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
df = pd.read_csv('StudentsPerformance.csv')
```

In [3]:

```
df.head(3)
```

Out[3]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93

In [4]:

```
for col in df[['gender','race/ethnicity','parental level of education','lunch',\
               'test preparation course']]:
    print(col,df[col].unique())         #finding unique in categorical attributes
```

```
gender ['female' 'male']
race/ethnicity ['group B' 'group C' 'group A' 'group D' 'group E']
parental level of education ["bachelor's degree" 'some college' "master's degree" "associa
te's degree"
 'high school' 'some high school']
lunch ['standard' 'free/reduced']
test preparation course ['none' 'completed']
```

In [5]:

```
df.rename(columns = {'race/ethnicity':'race_ethnicity','lunch':'lunch_plan',\
                     'parental level of education':'parental_education',\
                     'test preparation course':'prep_course_completed',\
                     'math score':'math_score','reading score':'reading_score',\
                     'writing score':'writing_score'}, inplace = True)
```

In [6]:

```
#setting variable types
df.gender = df.gender.astype('category')
df.race_ethnicity = df.race_ethnicity.astype('category')
df.lunch_plan = df.lunch_plan.astype('category')

#setting order of parental education from low to high
from pandas.api.types import CategoricalDtype
df.parental_education = df.parental_education.astype(CategoricalDtype(\
                ['some high school','high school','some college',"associate's degree",\
                 "bachelor's degree","master's degree"], ordered=True))
```

In [7]:

```
df.prep_course_completed = df.prep_course_completed.map(\
                                 {'completed': True,'none':False})
```

In [8]:

```
#adding computed variable of total score
df['total_score'] = df.math_score + df.reading_score + df.writing_score
```

In [9]:

```
df.info()     #there are no null values in the dataset
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype   
---  ------                 --------------  -----   
 0   gender                 1000 non-null   category
 1   race_ethnicity         1000 non-null   category
 2   parental_education     1000 non-null   category
 3   lunch_plan             1000 non-null   category
 4   prep_course_completed  1000 non-null   bool    
 5   math_score             1000 non-null   int64   
 6   reading_score          1000 non-null   int64   
 7   writing_score          1000 non-null   int64   
 8   total_score            1000 non-null   int64   
dtypes: bool(1), category(4), int64(4)
memory usage: 36.9 KB
```

In [10]:

```
df.tail(3)
```

Out[10]:

	gender	race_ethnicity	parental_education	lunch_plan	prep_course_completed	math_score	reading_score	wr
997	female	group C	high school	free/reduced	True	59	71	
998	female	group D	some college	standard	True	68	78	
999	female	group D	some college	free/reduced	False	77	86	

Data is ready for analysis.

1. Descriptive Statistics

I will begin with descriptive statistics of overall subject scores and the computed total score. Then I will split the data into categorical values, inspect the descriptive statistics and provide my interpretations.

(a) Analysis

In [11]:

```
print('\033[1m' + 'Table 1: Descriptive Statistics (DS) of All Scores' +'\033[0m')
stats = df.iloc[:,[5,6,7,8]].describe()
stats.loc['var'] = df.iloc[:,[5,6,7,8]].var().tolist()
stats.loc['skew'] = df.iloc[:,[5,6,7,8]].skew().tolist()
stats.loc['kurt'] = df.iloc[:,[5,6,7,8]].kurtosis().tolist()
stats.transpose()
```

```
Table 1: Descriptive Statistics (DS) of All Scores
```

Out[11]:

	count	mean	std	min	25%	50%	75%	max	var	skew	kurt
math_score	1000.0	66.089	15.163080	0.0	57.00	66.0	77.0	100.0	229.918998	-0.278935	0.274964
reading_score	1000.0	69.169	14.600192	17.0	59.00	70.0	79.0	100.0	213.165605	-0.259105	-0.068265
writing_score	1000.0	68.054	15.195657	10.0	57.75	69.0	79.0	100.0	230.907992	-0.289444	-0.033365
total_score	1000.0	203.312	42.771978	27.0	175.00	205.0	233.0	300.0	1829.442098	-0.299057	0.125843

As per Table 1, reading_score has the highest mean and median, followed by writing_score and lastly math_score. The minimum of math_score is zero, while other scores have non-zero min. All scores have negligible skewness (less than 0.5) and low kurtosis/peak (much less than 3).

In [12]:

```
print('\033[1m' + 'Table 2: DS - Total Score by Gender' +'\033[0m')
df.iloc[:,[8,0]].groupby('gender').describe()
```

```
Table 2: DS - Total Score by Gender
```

Out[12]:

	total_score							
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
gender								
female	518.0	208.708494	43.625427	27.0	182.0	211.0	236.00	300.0
male	482.0	197.512448	41.096520	69.0	168.0	199.0	228.75	300.0

The sample has more females; also females have higher mean and median of total_score. However, the standard deviation of females is higher, which also explains why the minimum of females is lower than that to males.

In [13]:

```
print('\033[1m' + 'Table 3: DS - Total Score by Race/Ethnic Groups' +'\033[0m')
df.iloc[:,[8,1]].groupby('race_ethnicity').describe()
```

```
Table 3: DS - Total Score by Race/Ethnic Groups
```

Out[13]:

	total_score							
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
race_ethnicity								
group A	89.0	188.977528	43.333794	70.0	156.0	184.0	219.00	289.0
group B	190.0	196.405263	44.196399	55.0	170.0	195.0	230.50	290.0
group C	319.0	201.394984	41.616633	27.0	173.0	205.0	231.00	296.0
group D	262.0	207.538168	39.758327	93.0	181.0	210.0	235.75	297.0
group E	140.0	218.257143	43.695047	78.0	194.0	220.5	247.25	300.0

The sample has the highest represenation from group C and the least from A. The mean, median, 75% and maximum increase as we go through groups A, B, C, D and E. However the standard deviation and the mininum show no such trend.

In [14]:

```
print('\033[1m' + 'Table 4: DS - Total Score by Parental Education' +'\033[0m')
df.iloc[:,[8,2]].groupby('parental_education').describe()
```

```
Table 4: DS - Total Score by Parental Education
```

Out[14]:

	total_score							
	count	mean	std	min	25%	50%	75%	max
parental_education								
some high school	179.0	195.324022	44.952235	27.0	167.00	200.0	229.5	297.0
high school	196.0	189.290816	40.531749	55.0	161.75	195.0	218.0	287.0
some college	226.0	205.429204	41.132921	70.0	180.00	206.0	234.0	297.0
associate's degree	222.0	208.707207	41.012743	95.0	176.00	209.0	237.0	300.0
bachelor's degree	118.0	215.771186	41.839827	117.0	192.25	213.5	242.0	300.0
master's degree	59.0	220.796610	40.803051	134.0	189.50	220.0	256.5	293.0

More number of students having a moderate level of parental education than lower and higher levels. Excluding 'some high school', students with other parental education show a steady increase in mean and median scores at higher level of parental education.

In [15]:

```
print('\033[1m' + 'Table 5: DS - Total Score by Lunch Plan' +'\033[0m')
df.iloc[:,[8,3]].groupby(['lunch_plan']).describe()
```

```
Table 5: DS - Total Score by Lunch Plan
```

Out[15]:

	total_score							
	count	mean	std	min	25%	50%	75%	max
lunch_plan								
free/reduced	355.0	186.597183	43.374971	27.0	158.5	188.0	217.5	293.0
standard	645.0	212.511628	39.559515	78.0	187.0	214.0	239.0	300.0

Students with standard lunch plan have all descriptive statistics (except standard deviation) higher than those with free/reduced plan.

In [16]:

```
print('\033[1m' + 'Table 6: DS - Total Score by Completion of Prep Course' +'\033[0m')
df.iloc[:,[8,4]].groupby('prep_course_completed').describe()
```

```
Table 6: DS - Total Score by Completion of Prep Course
```

Out[16]:

	total_score							
	count	mean	std	min	25%	50%	75%	max
prep_course_completed								
False	642.0	195.116822	42.560121	27.0	166.25	196.0	225.0	300.0
True	358.0	218.008380	39.110881	103.0	195.00	220.5	246.5	300.0

Almost twice as many students did not complete the course as those who did. Students who completed the preparation course have higher mean, mininum, 25 precentile, median, 75 precentile and maximum of total_score. They also have a lower standard deviation.

(b) Description of Three-line Table Format

Three-line Table Format is the standard layout for formatting tables adopted in white papers. It has the following characteristics:
• Table should have only three horizontal lines with column headers between the first and second lines and records in between the second and third line. There should be no vertical lines.
• Table title should be above the first line, be clear and brief and mention table number ordered as cited.
• The table is centered, with a blank after the table and before the text.
• All units should be indicated and be in the SI units.
This format allows for easy reading and interpretation across the scientific community orldwide.

(c) Interpretations

As per the sample data and descriptive statistics analysis, following is my interpretation:
• From table 1, all subject scores are distributed almost symmetrically with much broader peaks and thickened tails. Math scores are lower than the writing and reading scores in general.
• From table 2, females have higher total scores in general but deviate more.
• From table 3, race/ethnic groups E has the highest scores in general, followed by D, C, B and A in order. Group C and D dominates the count.
• From table 4, students with higher parental education have slightly higher scores, which may or may not be statistically significant.
• From table 5, students with standard lunch plan perform better in general, while others have lower scores. Correlation is not equal to causation, and lower scores may be caused by other parameters more prevalent in students with free/reduced lunch plan.
• From table 6, only one-third of students completed the test preparation course, and those who completed have greater scores in general.

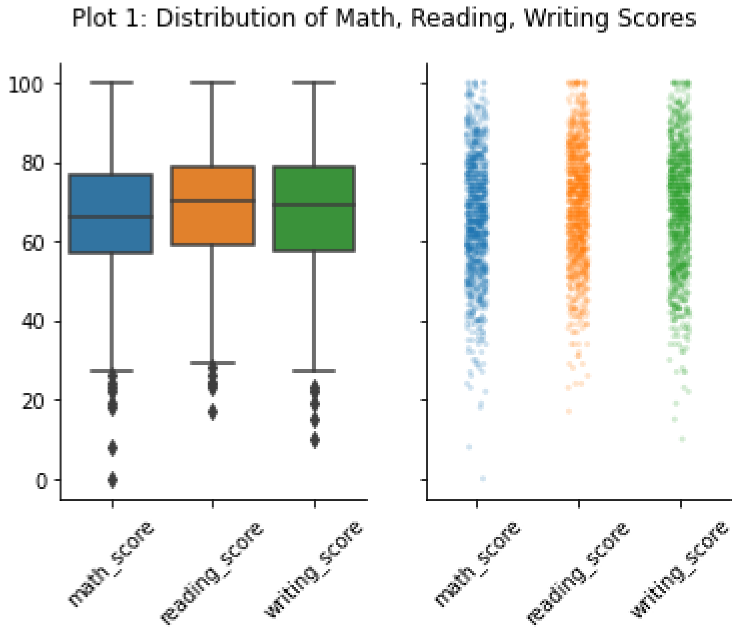
2. Visualizations

In this section, I will examine in more detail the scores and impact of categorical attributes through visualizations.

(a) Analysis

In [17]:

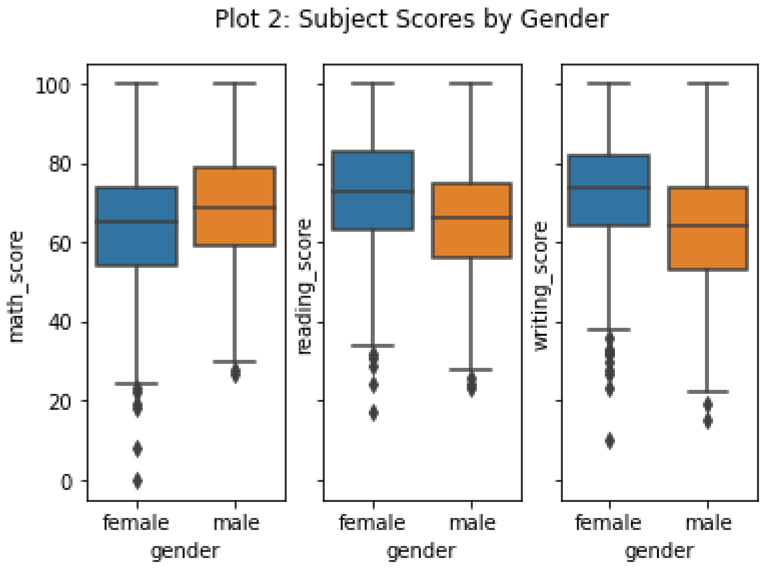
```
fig, axs = plt.subplots(1,2,sharey=True)
fig.suptitle('Plot 1: Distribution of Math, Reading, Writing Scores')

sns.boxplot(ax=axs[0], data=df.iloc[:,[5,6,7]]);
sns.stripplot(ax=axs[1], data=df.iloc[:,[5,6,7]], size=3, alpha=.2, jitter=True, edgecolor
sns.despine()

for tick in axs[0].get_xticklabels():
    tick.set_rotation(45)
for tick in axs[1].get_xticklabels():
    tick.set_rotation(45)
```



Math scores have many outliers toward the lower side, followed by writing scores. Reading scores are slightly less dense.

In [18]:

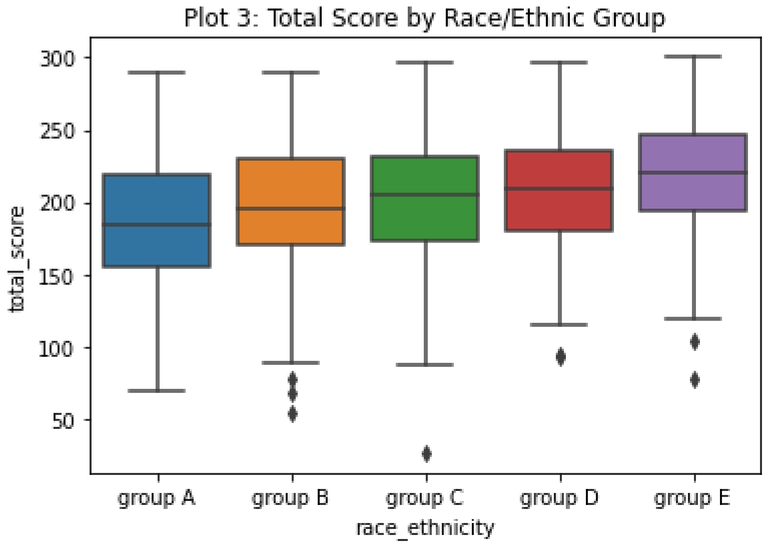
```
fig, axs = plt.subplots(1,3,sharey=True)
fig.suptitle('Plot 2: Subject Scores by Gender')

sns.boxplot(ax=axs[0], x=df.gender,y=df.math_score)
sns.boxplot(ax=axs[1], x=df.gender,y=df.reading_score)
sns.boxplot(ax=axs[2], x=df.gender,y=df.writing_score);
```



Males have higher math scores, while females have higher reading and writing scores. Most lower outliers belong to the female gender, while there are no higher outliers.

In [19]:

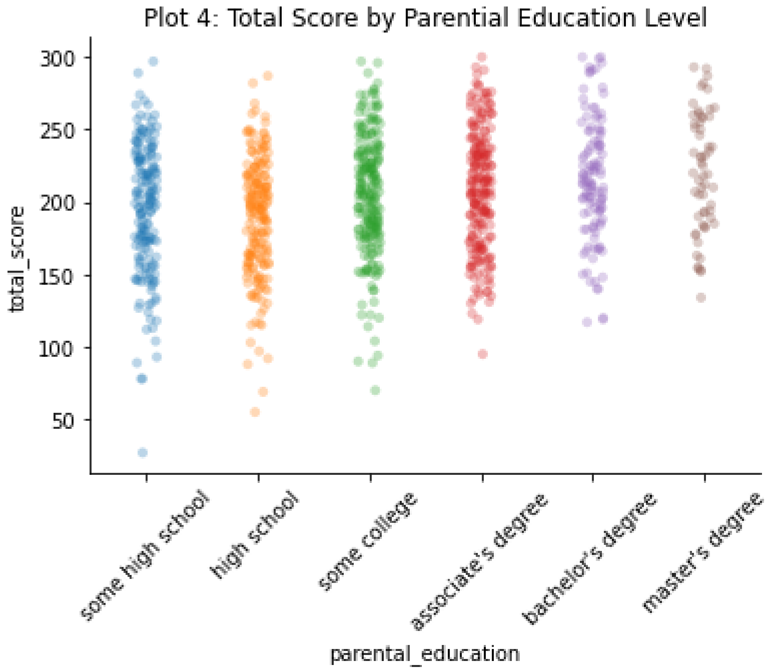
```
sns.boxplot(x=df.race_ethnicity,y=df.total_score)
plt.title('Plot 3: Total Score by Race/Ethnic Group');
```



As we move from group A to group E, groups on the right have all quartiles higher. There are few outliers in group B and E.

In [20]:

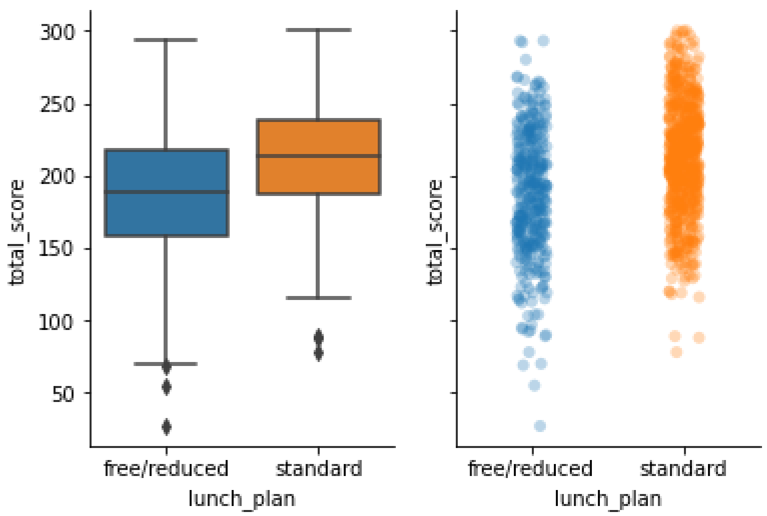
```
sns.stripplot(x='parental_education',y='total_score',data=df, size=5, alpha=.3, jitter=Tru
plt.xticks(rotation = 45)
sns.despine()
plt.title('Plot 4: Total Score by Parential Education Level');
```



As parental education increases, the total score increases and becomes less spread.

In [21]:

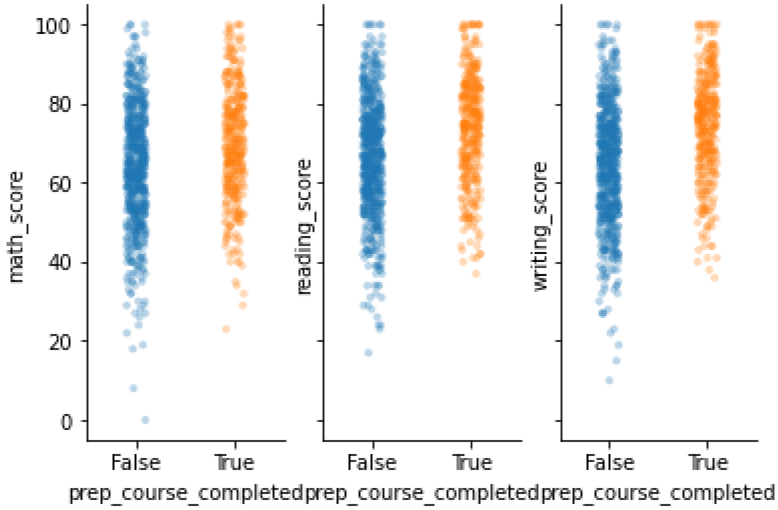
```
fig, axs = plt.subplots(1,2,sharey=True)
fig.suptitle('Plot 5: Total Score by Lunch Plan')
sns.boxplot(ax=axs[0], x=df.lunch_plan,y=df.total_score)
sns.stripplot(ax=axs[1],x='lunch_plan',y='total_score',data=df, size=6, alpha=.3, jitter=T
sns.despine()
```


Plot 5: Total Score by Lunch Plan

Students with standard meals have higher total scores, less outliers and less scattered distribution.

In [22]:

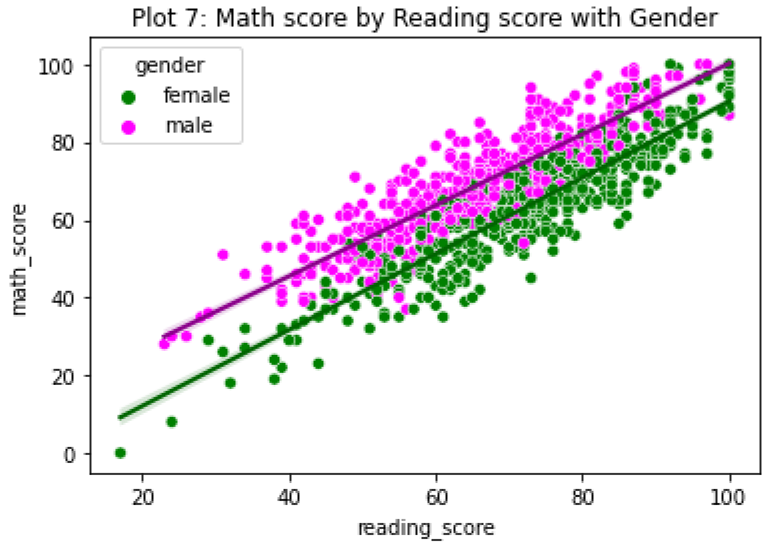
```
fig, axs = plt.subplots(1,3,sharey=True)
fig.suptitle('Plot 6: Total Score by Completion of Prep Course')
sns.stripplot(ax=axs[0],x='prep_course_completed',y='math_score',data=df, size=4, alpha=.3
sns.stripplot(ax=axs[1],x='prep_course_completed',y='reading_score',data=df, size=4, alpha
sns.stripplot(ax=axs[2],x='prep_course_completed',y='writing_score',data=df, size=4, alpha
sns.despine()
```


Plot 6: Total Score by Completion of Prep Course

Those who completed the preparation course have higher and denser scores in all subjects.

In [23]:

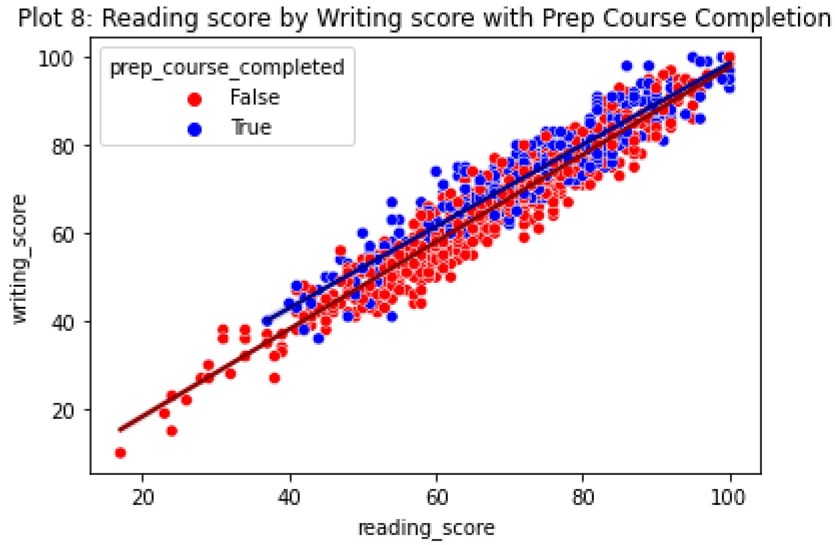
```
sns.regplot(x='reading_score', y='math_score', data=df[df['gender']=='female'],\
            color='darkgreen', scatter_kws={"s": 0})
sns.regplot(x='reading_score', y='math_score', data=df[df['gender']=='male'],\
            color='darkmagenta', scatter_kws={"s": 0})
sns.scatterplot(x='reading_score', y='math_score', data=df, hue = 'gender',\
                palette=['green','magenta'])
plt.title('Plot 7: Math score by Reading score with Gender');
```


Plot 7: Math score by Reading score with Gender

Reading score may not be highly correlated with math score, and the broader trend of a higher math score with a higher reading score may be linked with student's capability in general. However, from the color and regression line, males with the same reading score as females, have higher math scores, and vice versa, females with the same math score as males have higher reading score.

In [24]:

```
sns.regplot(x='reading_score', y='writing_score', data=df[df['prep_course_completed']==Fal
            color='darkred', scatter_kws={"s": 0})
sns.regplot(x='reading_score', y='writing_score', data=df[df['prep_course_completed']==Tru
            color='darkblue', scatter_kws={"s": 0})
sns.scatterplot(x='reading_score', y='writing_score', data=df, hue = 'prep_course_complete
                palette=['red','blue'])
plt.title('Plot 8: Reading score by Writing score with Prep Course Completion');
```


Plot 8: Reading score by Writing score with Prep Course Completion

Reading and writing scores are highly correlated. Those who completed the preparation course are in blue and improved writing more than reading.

(b) Jitter chart & Box plots

When to use jitter charts?

- Jitter plots are helpful to uderstand the dispersion of one-dimensional data, overall or within groups.
- Strip plots also show dispersion, but jitter plots randomly spread the distribution across a small width of the x-axis. This helps to better examine the density in case of overplotting.

How to detect outliers in box plots?
- Box plots contain median, upper quartile and lower quartile indicated by lines on a solid rectange.
- The regular minimum and maximum are calculated as 1.5 times of interquartile range outside the two quartiles and any data outside these are defined as outliers.
- Whiskers visually indicate the minimum and maximum, and the data points outside these whiskers can visually be detected as outliers.

(b) Interpretations

From the visualizations and analysis, following is my interpretation:
- From plot 1, reading scores are higher on average and math scores are lower with many low outliers. Perhaps the math test was difficult, or students focused less on math.
- From plots 2 and 7, males are good at maths in general, while females are good at reading and writing, although female data has more low outliers in all subjects.
- From plot 3, scores increases as we move from group A through E.
- From plot 4, higher parental education is associated with a higher and certain (less dense) total scores.
- From plot 5, students in standard lunch plan score more than those in free/reduced lunch plan. It may be due to the standard plan being more nutritious, or those enrolled in the free plan skip meals at school and are less active or have to spend more time at home cooking.
- From plots 6 and 8, those completing the preparation course scored more for sure in all subjects, with more improvement in writing.
- From plots 7 and 8, a student with better reading will also have better writing and vice versa as they are highly correlated.

Summary

We have completed the analysis and interpretation of 'Students Performance in Exams' dataset. The details give next stand out from the interpretation.

As observed, math subject has lower scores in general including zero and needs to be focused more. Reading and writing can be improved together.

Among attributes that cannot be changed, females have a higher total score in general and men are better in math. Groups E scored the highest, while higher the parental education, higher is the total score of their children in general.

Among choices, those enrolled in standard meal plan have higher scores in general. However, those who completed the test preparation course have certainty of better scores in all subjects, especially writing.

Biliography

Dataset Source:
kaggle.com/spscientist/students-performance-in-exams

Library Documentation:
pandas.pydata.org/docs/getting_started/intro_tutorials/03_subset_data.html
pandas.pydata.org/pandas-docs/stable/user_guide/categorical.html#categoricaldtype
matplotlib.org/stable/gallery/subplots_axes_and_figures/subplots_demo.html
matplotlib.org/3.1.0/api/markers_api.html

StackOverFlow.com:

stackoverflow.com/questions/30601830/when-to-use-category-rather-than-object
stackoverflow.com/questions/23959207/advanced-describe-pandas
stackoverflow.com/questions/34023918/make-new-column-in-panda-dataframe-by-adding-values-from-other-columns
stackoverflow.com/questions/49554139/boxplot-of-multiple-columns-of-a-pandas-dataframe-on-the-same-figure-seaborn
stackoverflow.com/questions/22408237/named-colors-in-matplotlib

Other Foruns:

geeksforgeeks.org/how-to-rename-columns-in-pandas-dataframe/
researchgate.net/figure/The-classification-of-parental-education-derived-from-the-highest-level-of-education_tbl4_6687684
dataviztalk.blogspot.com/2016/02/how-to-add-jitter-to-plot-using-pythons.html
dev.to/thalesbruno/subplotting-with-matplotlib-and-seaborn-5ei8

Resources on tables and plots:

coursehero.com/file/p788n7io/Table-1-The-Three-Line-Table-Format-Notice-that-the-number-of-the-table-should/ datavizproject.com/data-type/jitter-plot/ thedataschool.co.uk/michael-mcfadden/tableau-tutorials-build-jitter-plot