EDA of Students Performeance in deifferent school or college

In [*]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Import required libraries for EDA.

In [3]:

```
data = pd.read_csv('../input/students-performance-in-exams/StudentsPerformance.csv')
```

In [4]:

```
data.head(10)
```

Out[4]:

	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
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
5	female	group B	associate's degree	standard	none	71	83	78
6	female	group B	some college	standard	completed	88	95	92
7	male	group B	some college	free/reduced	none	40	43	39
8	male	group D	high school	free/reduced	completed	64	64	67
9	female	group B	high school	free/reduced	none	38	60	50

In [5]:

```
data.shape
```

Out[5]:

```
(1000, 8)
```

In [6]:

```
data.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
```

Dataset contains data about different students at a school/college, and their scores in 3 subjects.

In [8]:

```
data.describe(include='all')
```

Out[8]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	
count	1000	1000	1000	1000	1000	1000.00000	1000.000000	1000.
unique	2	5	6	2	2	NaN	NaN	
top	female	group C	some college	standard	none	NaN	NaN	
freq	518	319	226	645	642	NaN	NaN	
mean	NaN	NaN	NaN	NaN	NaN	66.08900	69.169000	68.
std	NaN	NaN	NaN	NaN	NaN	15.16308	14.600192	15.
min	NaN	NaN	NaN	NaN	NaN	0.00000	17.000000	10.
25%	NaN	NaN	NaN	NaN	NaN	57.00000	59.000000	57.
50%	NaN	NaN	NaN	NaN	NaN	66.00000	70.000000	69.
75%	NaN	NaN	NaN	NaN	NaN	77.00000	79.000000	79.
max	NaN	NaN	NaN	NaN	NaN	100.00000	100.000000	100.

Check for missing values.

In [11]:

```
data.isnull().sum()
```

Out[11]:

```
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
```

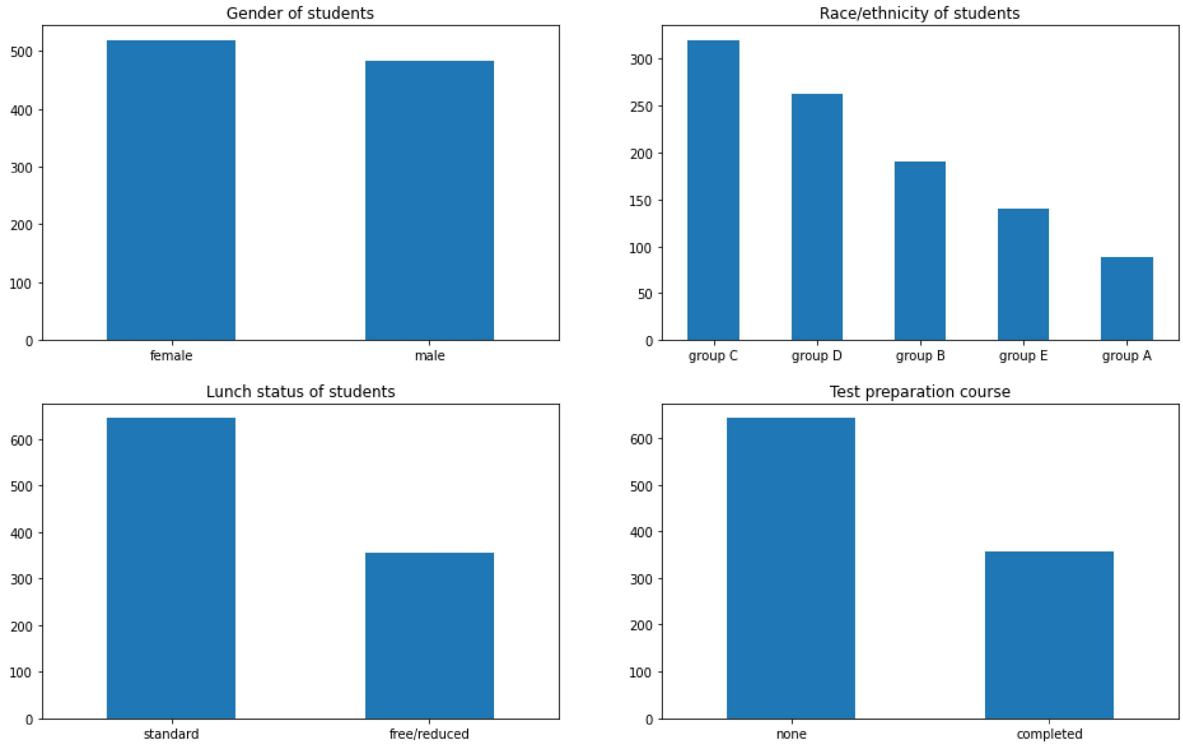
There are no missing values in the given dataset.

Lets see the graphical representation of dataset.

In [16]:

```
plt.subplot(221)
data['gender'].value_counts().plot(kind='bar', title='Gender of students', figsize=(16,10))
plt.xticks(rotation=0)

plt.subplot(222)
data['race/ethnicity'].value_counts().plot(kind='bar', title='Race/ethnicity of students')
plt.xticks(rotation=0)

plt.subplot(223)
data['lunch'].value_counts().plot(kind='bar', title='Lunch status of students')
plt.xticks(rotation=0)

plt.subplot(224)
data['test preparation course'].value_counts().plot(kind='bar', title='Test preparation cou
plt.xticks(rotation=0)

plt.show()
```

```
/opt/conda/lib/python3.7/site-packages/pandas/plotting/_matplotlib/tools.py:
400: MatplotlibDeprecationWarning:
The is_first_col function was deprecated in Matplotlib 3.4 and will be remov
ed two minor releases later. Use ax.get_subplotspec().is_first_col() instea
d.
  if ax.is_first_col():
/opt/conda/lib/python3.7/site-packages/pandas/plotting/_matplotlib/tools.py:
400: MatplotlibDeprecationWarning:
The is_first_col function was deprecated in Matplotlib 3.4 and will be remov
ed two minor releases later. Use ax.get_subplotspec().is_first_col() instea
d.
  if ax.is_first_col():
/opt/conda/lib/python3.7/site-packages/pandas/plotting/_matplotlib/tools.py:
400: MatplotlibDeprecationWarning:
The is_first_col function was deprecated in Matplotlib 3.4 and will be remov
ed two minor releases later. Use ax.get_subplotspec().is_first_col() instea
d.
  if ax.is_first_col():
```



1. From **gender graph** we can see that school contains more number of female students.
2. From **ethnicity graph** majority of the students belong to groups C and D.
3. From **lunch graph** more than 60% of the students have a standard lunch at school.
4. From **test preparation** graphmore than 60% of students have not taken any test preparation course.

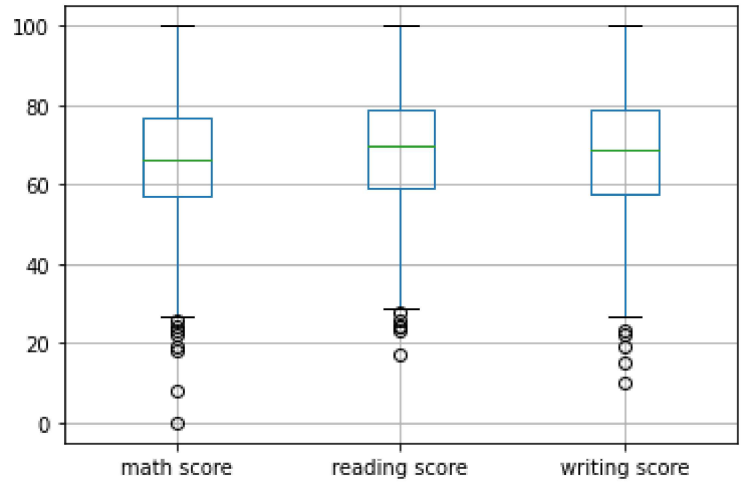
We use boxplot to identify the outliers.

In [17]:

```
data.boxplot()
```

Out[17]:

<AxesSubplot:>



Hollow circles near the tails represent outliers in the dataset.
It is very much possible for a student to score extremely low marks in a test, we will not remove these outliers.

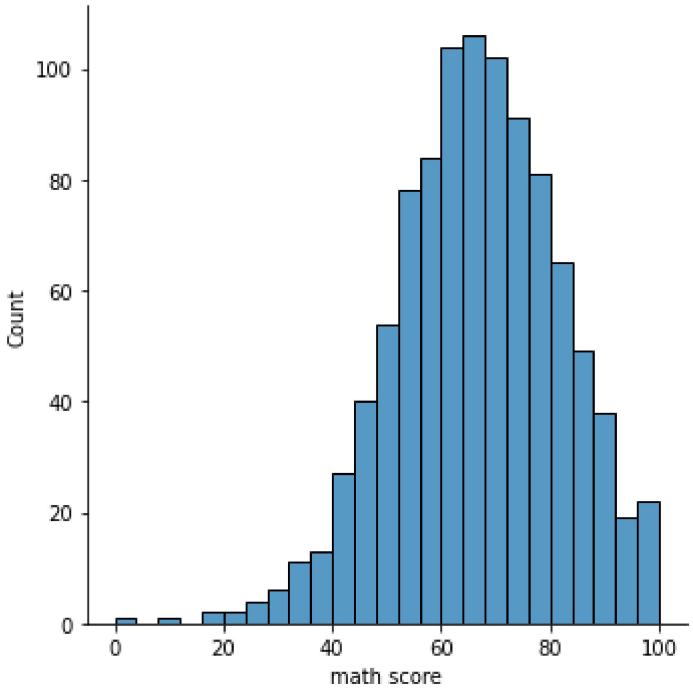
We make distribution plot of the students math score.

In [22]:

```
sns.displot(data['math score'])
```

Out[22]:

```
<seaborn.axisgrid.FacetGrid at 0x7f976ddc2d50>
```

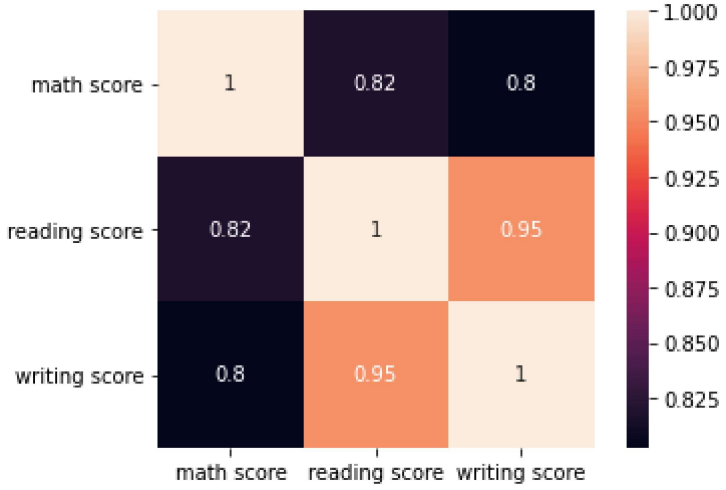


The peak is at around 65 marks, the mean of the math score of the students in the dataset.

Representation of correlation between the 3 scores with the help of a heatmap.

In [24]:

```
cor = data.corr()
sns.heatmap(cor, annot = True, square = True)
plt.yticks(rotation=0)
plt.show()
```



The heatmap shows that the 3 scores are highly correlated. Reading score has a correlation coefficient of 0.95 with the writing score. Math score has a correlation coefficient of 0.82 with the reading score, and 0.80 with the writing score.

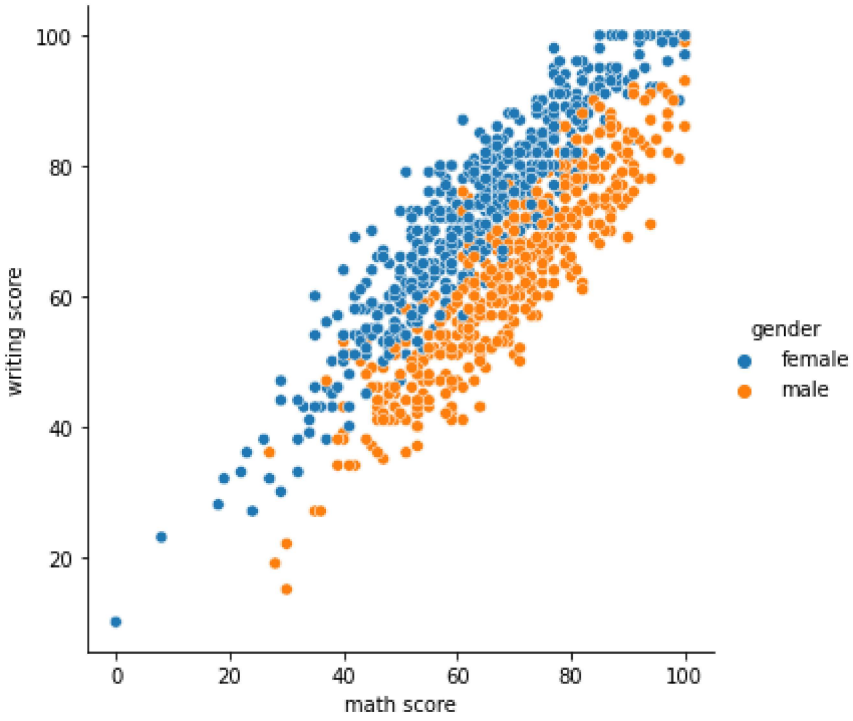
Relational plot to understand the relationship between 2 variables.

In [25]:

```
sns.relplot(x='math score', y='writing score', hue='gender', data=data)
```

Out[25]:

```
<seaborn.axisgrid.FacetGrid at 0x7f976dafbd90>
```



graph shows a difference in scores between the male and female students.
For the same math score, female students are more likely to have a higher writing score than male students.
For the same writing score, male students are expected to have a higher math score than female students.

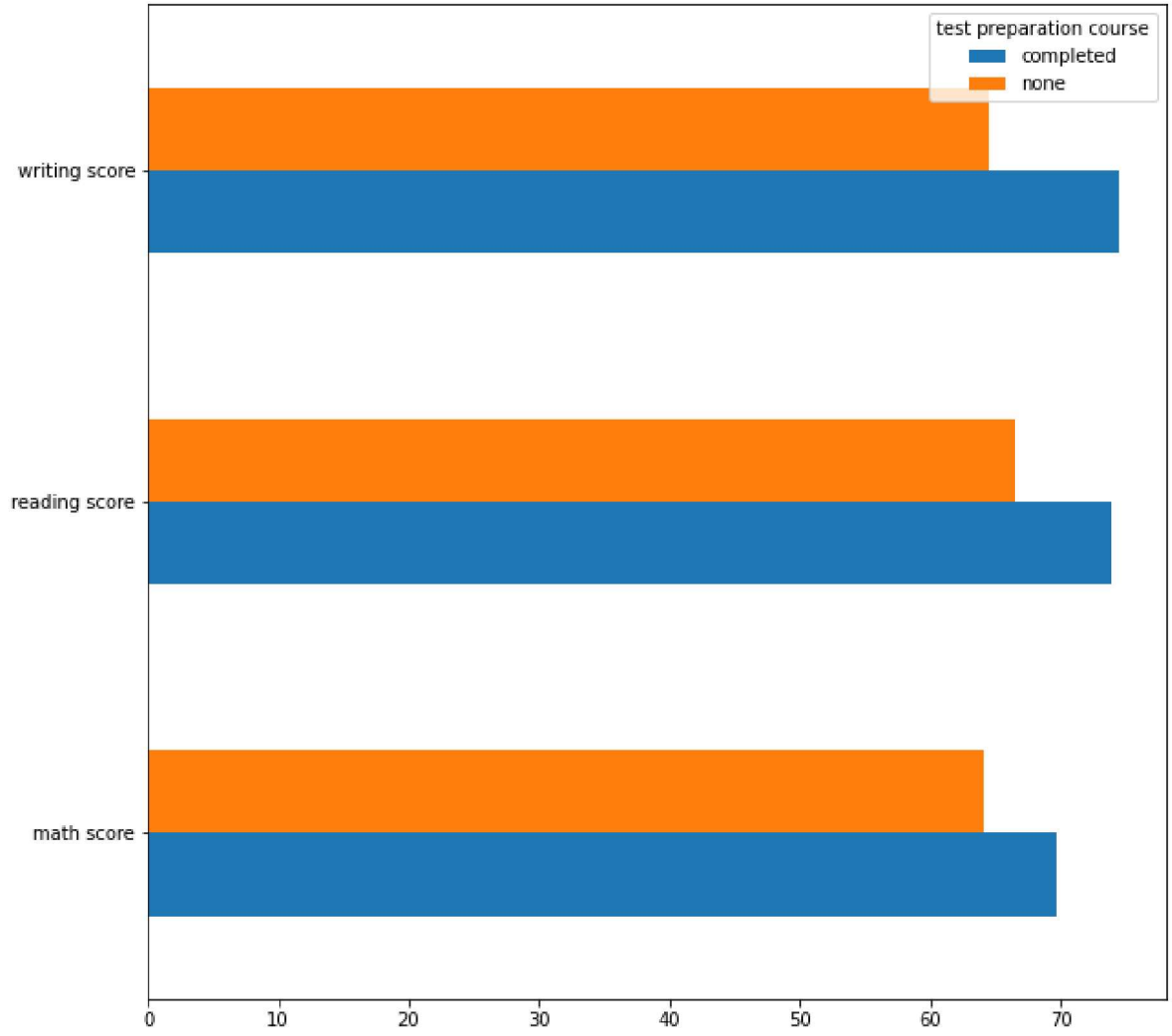
Impact of the test preparation course on students' performance using a horizontal bar graph.

In [26]:

```
data.groupby('test preparation course')[['math score', 'reading score', 'writing score']].m
```

Out[26]:

<AxesSubplot:>



students who have completed the test preparation course have performed better, on average, as compared to students who have not opted for the course.