EXPLORATORY DATA ANALYSIS

```
#importing various libraries
import pandas as pd
import numpy as np
import seaborn as sns
import statistics as sc

#importing the csv file
data = pd.read_csv("students.csv")
```

(A) Understanding the Data

```
data.head()
```

```
   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
```

```
data.tail()
```

```
     gender race/ethnicity parental level of education
lunch  \
995  female        group E             master's degree      standard

996    male        group C                high school  free/reduced

997  female        group C                high school  free/reduced

998  female        group D               some college      standard

999  female        group D               some college  free/reduced


    test preparation course  math score  reading score  writing score

995               completed          88             99             95
```

```
996                   none          62             55             55

997              completed          59             71             65

998              completed          68             78             77

999                   none          77             86             86
```

```
data.shape
```

```
(1000, 8)
```

```
data.nunique()
```

```
gender                         2
race/ethnicity                 5
parental level of education    6
lunch                          2
test preparation course        2
math score                    81
reading score                 72
writing score                 77
dtype: int64
```

```
couple_columns= data[['gender', 'math score', 'reading score',
'writing score']]
couple_columns
```

```
     gender  math score  reading score  writing score
0    female          72             72             74
1    female          69             90             88
2    female          90             95             93
3      male          47             57             44
4      male          76             78             75
..      ...         ...            ...            ...
995  female          88             99             95
996    male          62             55             55
997  female          59             71             65
998  female          68             78             77
999  female          77             86             86

[1000 rows x 4 columns]
```

(B) Types of EDA with Examples

1. Univariate Non- Graphical

```
#MEAN, STANDARD DEVIATION, QUARTILES
```

```
data.describe()
```

```
       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
```

```
#MEDIAN

math = data['math score'].median()
reading = data['reading score'].median()
writing = data['writing score'].median()

print('math score median = ', math)
print('reading score median = ', reading)
print('writing score median = ', writing )
```

```
math score median =  66.0
reading score median =  70.0
writing score median =  69.0
```

```
#MODE

math1 = data['math score'].mode()
reading1 = data['reading score'].mode()
writing1 = data['writing score'].mode()

print('math score mode = ', math1)
print('reading score mode = ', reading1)
print('writing score mode = ', writing1)
```

```
math score mode =  0    65
Name: math score, dtype: int64
reading score mode =  0    72
Name: reading score, dtype: int64
writing score mode =  0    74
Name: writing score, dtype: int64
```

```
#MEAN
math2 = data['math score'].mean()
reading2 = data['reading score'].mean()
writing2 = data['writing score'].mean()

print('math score mode = ', math2)
print('reading score mode = ', reading2)
print('writing score mode = ', writing2)
```

```
math score mode =  66.089
reading score mode =  69.169
writing score mode =  68.054
```

2. Multivariate Non - Graphical

```
M = pd.read_csv("students.csv")
M
```

```
     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

..      ...            ...                         ...           ...

995  female        group E             master's degree      standard

996    male        group C                 high school  free/reduced

997  female        group C                 high school  free/reduced

998  female        group D                some college      standard

999  female        group D                some college  free/reduced


    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

..                      ...         ...            ...            ...
```

```
995                completed          88              99              95

996                     none          62              55              55

997                completed          59              71              65

998                completed          68              78              77

999                     none          77              86              86


[1000 rows x 8 columns]
```

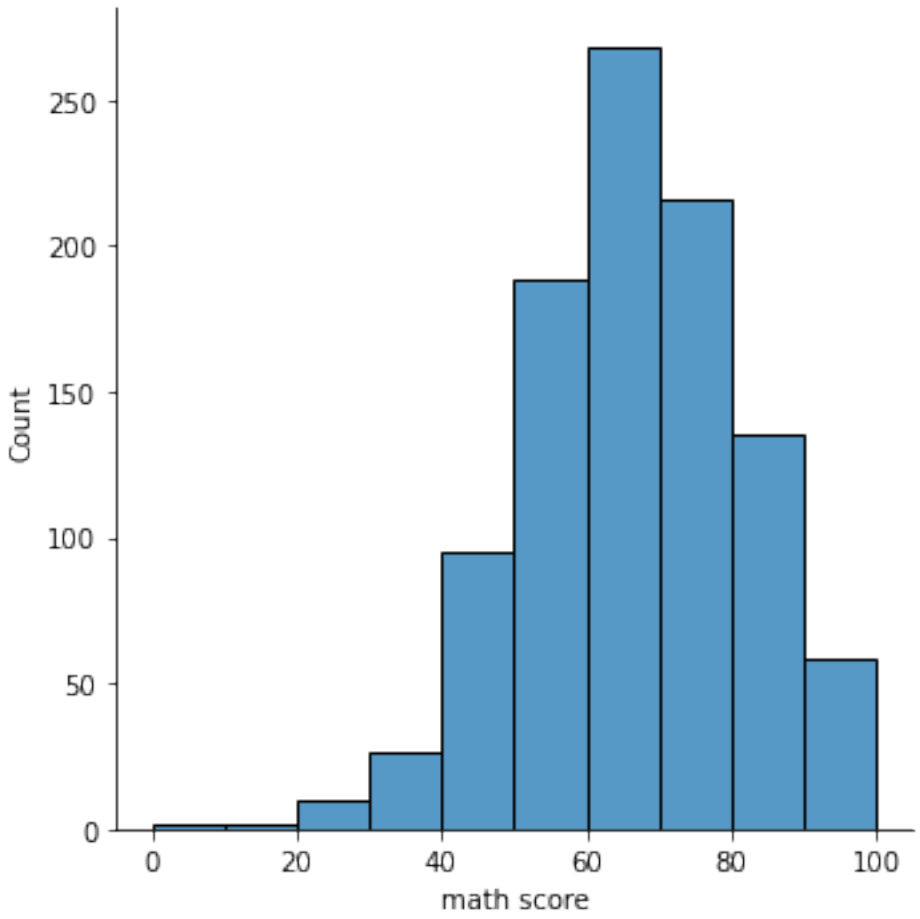
```
cross_tab = pd.crosstab(M.gender,M.lunch, normalize = 'index')
cross_tab
```

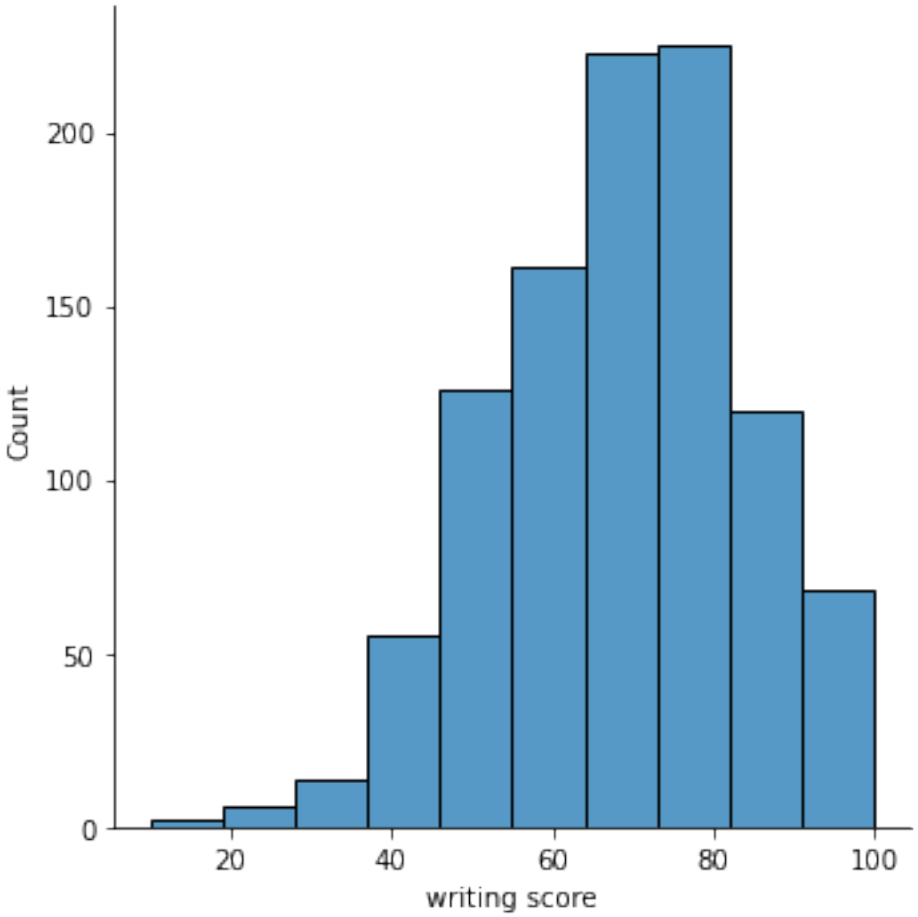
```
lunch   free/reduced  standard
gender
female      0.364865  0.635135
male        0.344398  0.655602
```

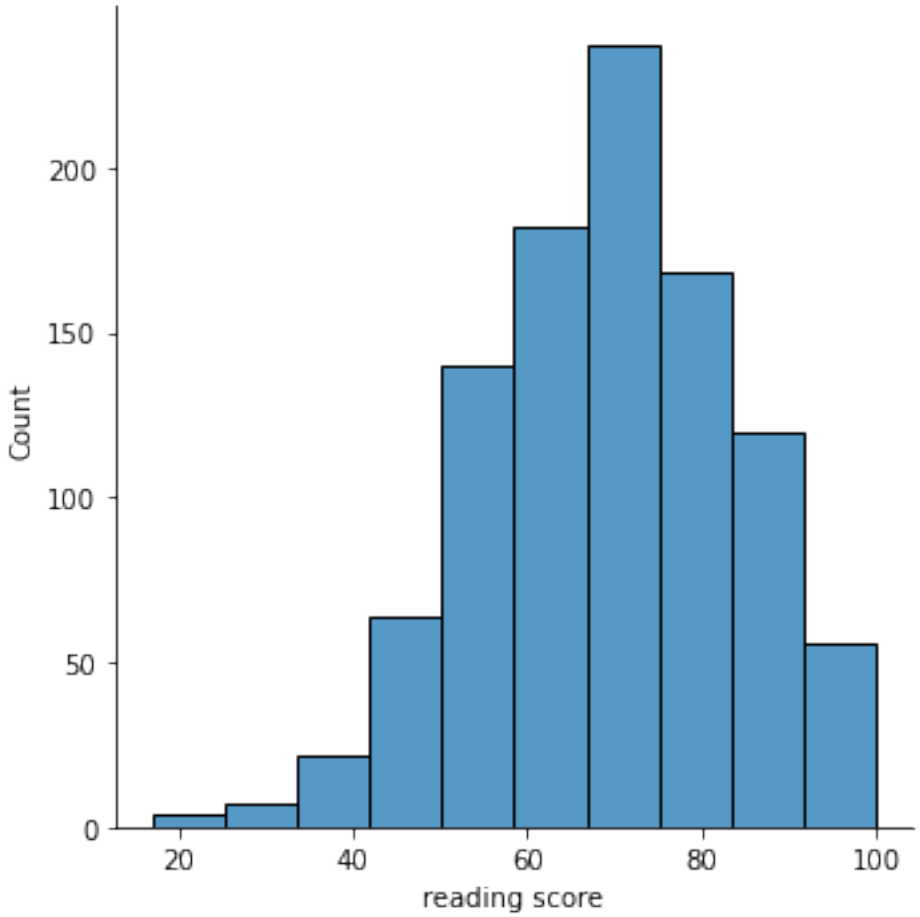
3. Univariate Graphical

```
#HISTOGRAM

sns.displot(data['math score'], bins = 10)
sns.displot(data['writing score'], bins = 10)
sns.displot(data['reading score'], bins = 10)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fa4f0c8a970>
```


250
200
150
100
50
0
Count
0
20
40
60
80
100
math score

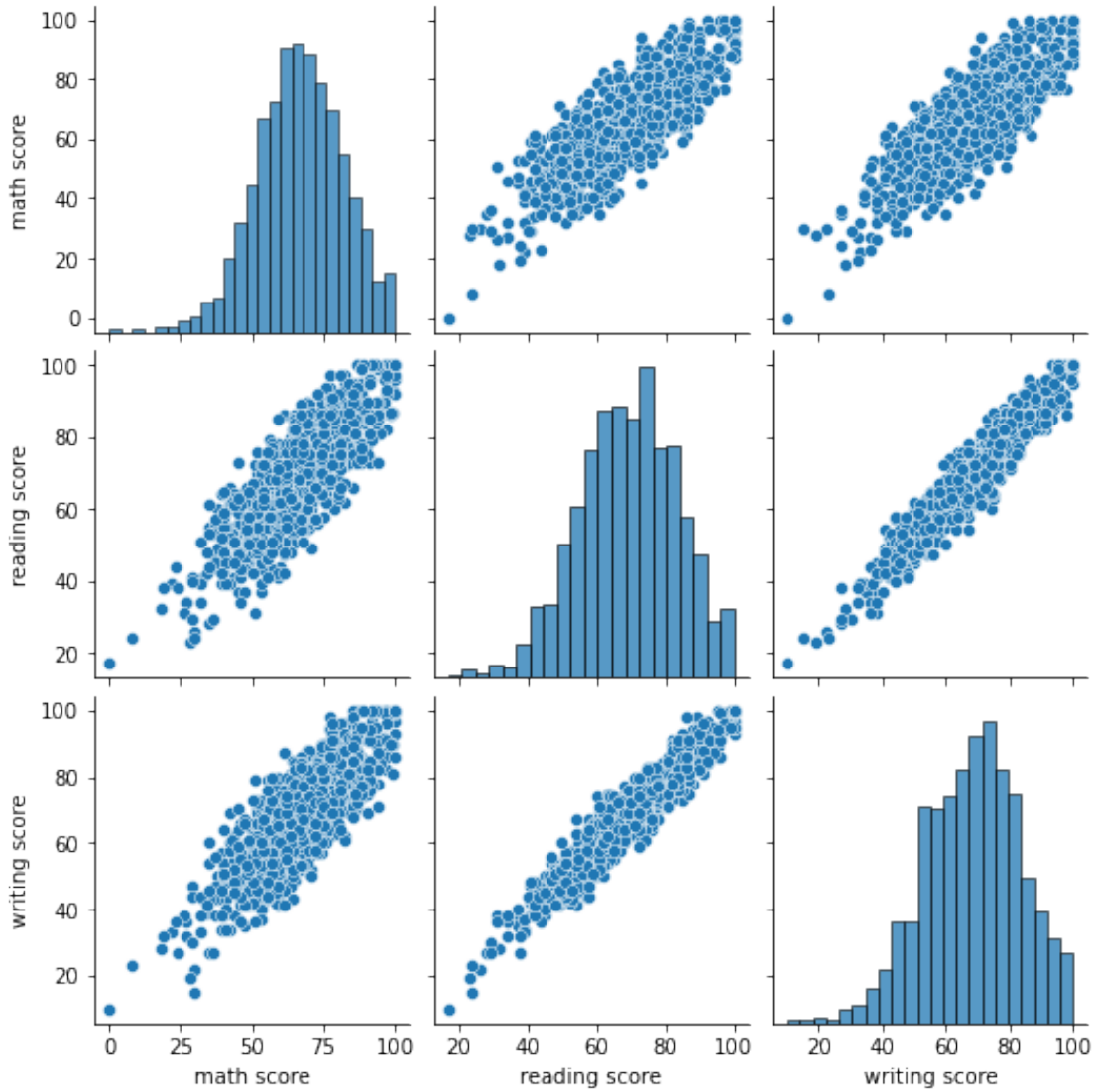

Count
200
150
100
50
0
20
40
60
80
100
writing score




```
#PAIR PLOT
```

```
sns.pairplot(data)
```

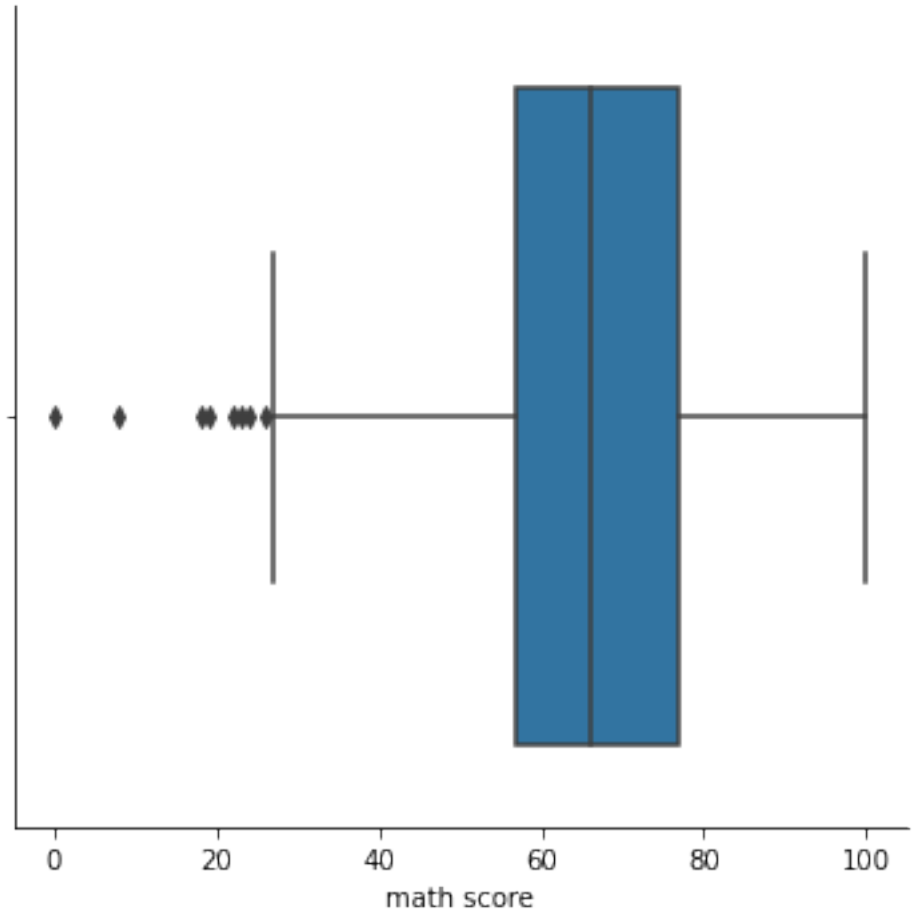
```
<seaborn.axisgrid.PairGrid at 0x7fa5231a24c0>
```




```
#BOX PLOT
```

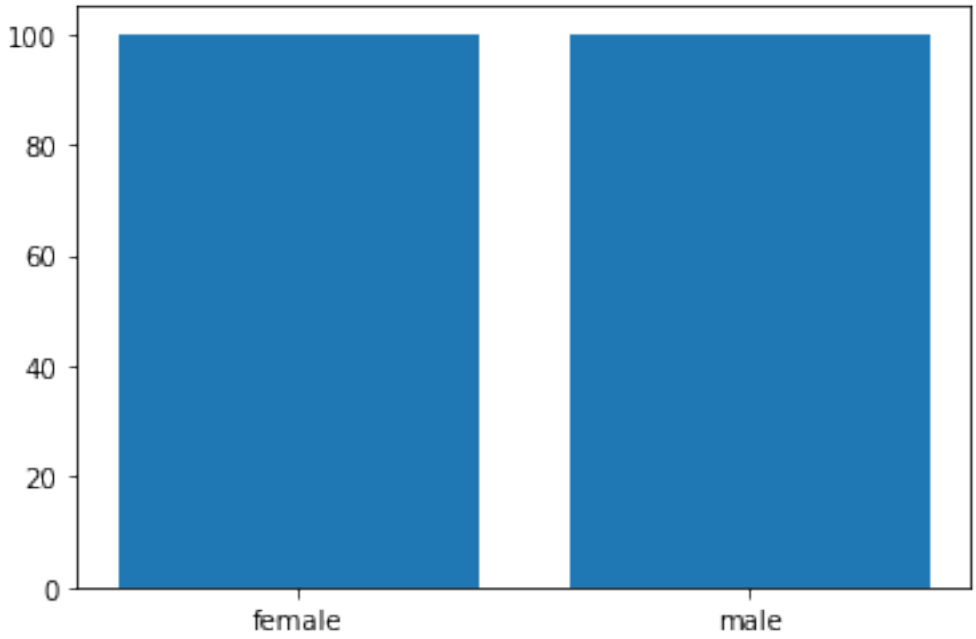
```
sns.catplot(x = 'math score', kind = 'box', data= data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fa512941790>
```



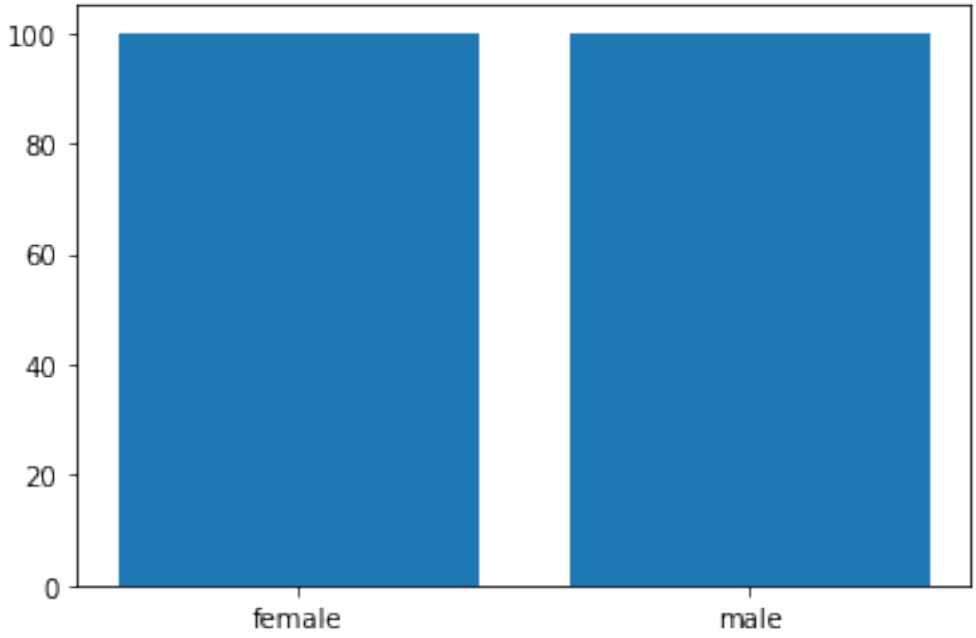

```
import matplotlib.pyplot as plt

#BAR GRAPHS
x = data['gender']
y = data['math score']
plt.bar(x,y)
```

```
<BarContainer object of 1000 artists>
```



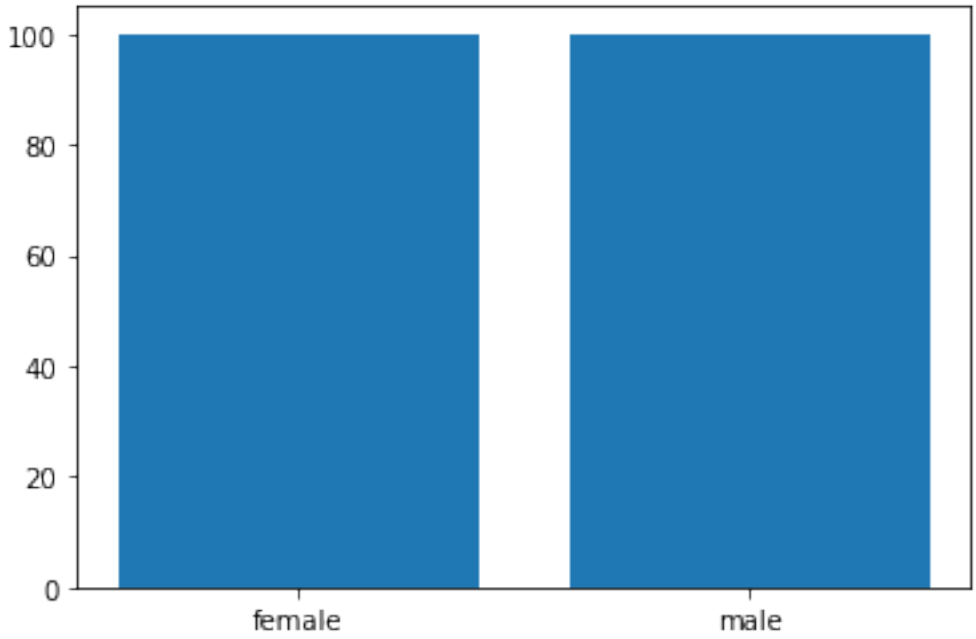

```
x = data['gender']
y = data['reading score']
plt.bar(x,y)
```

```
<BarContainer object of 1000 artists>
```




```
x = data['gender']
y = data['writing score']
plt.bar(x,y)
```

```
<BarContainer object of 1000 artists>
```

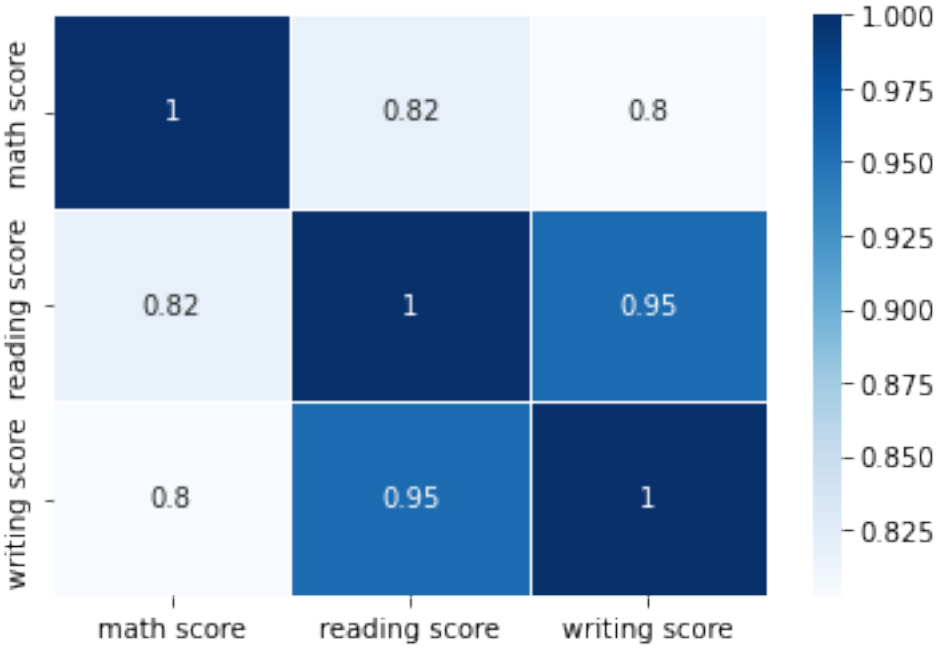


4. Multivariate Graphical

#HEAT MAP

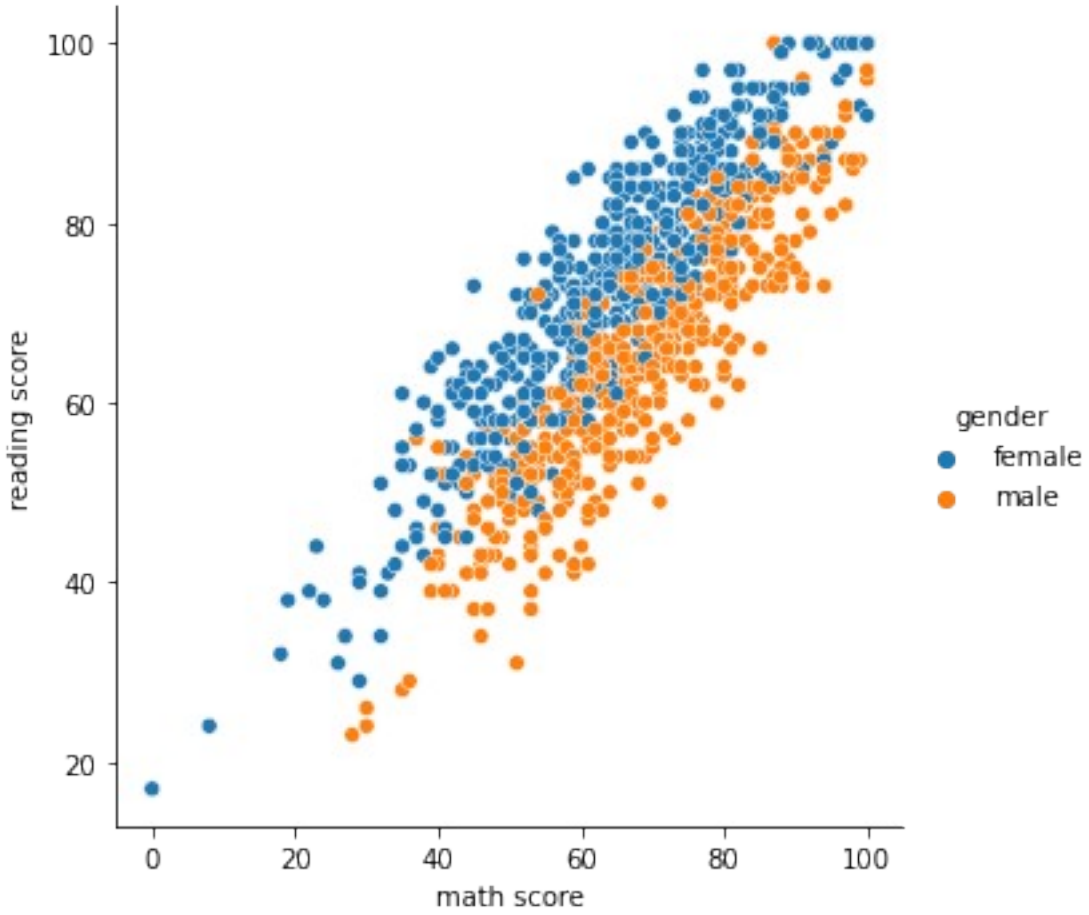
```
sns.heatmap(data.corr(), annot = True, linewidth = 0.5, cmap =
'Blues')
```

```
<AxesSubplot:>
```

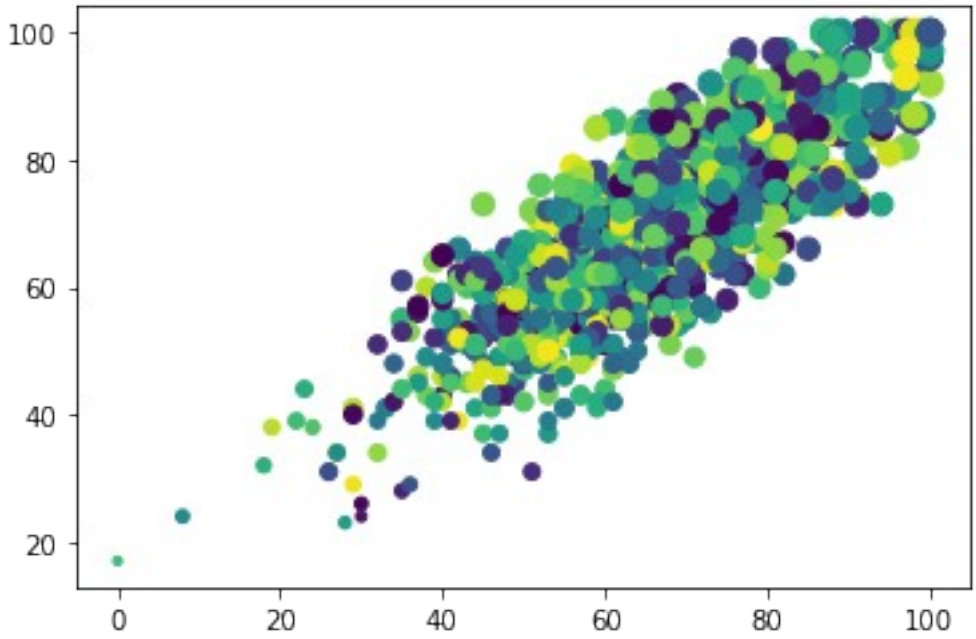


```
#SCATTER PLOT

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

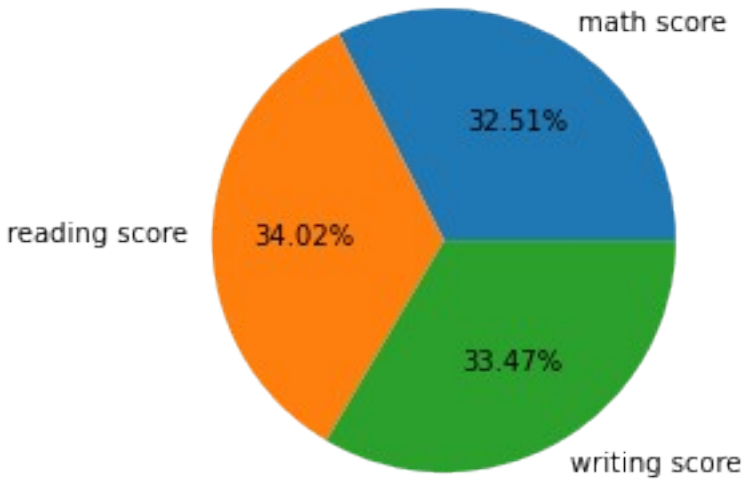
```
<seaborn.axisgrid.FacetGrid at 0x7fa5015ce070>
```




```
#BUBBLE CHART
m = data['math score']
r = data['reading score']
w = data['writing score']
colours = np.random.rand(1000)
plt.scatter(m,r,w, c = colours)
plt.show()
```




```
#Pie Chart
#(for the average score in all subjects)

val_ue = (math2, reading2, writing2)
lab_el = ('math score', 'reading score', 'writing score')

plt.pie(val_ue, labels = lab_el, autopct = '%2.2f%%')
plt.show()
```



(C) CONCLUSION

1. About the Data: A sample data set 'students.csv' was downloaded from Kaggle. The data set gives basic information related to 999 students along with marks in three different domains. A total of 8 columns make up the data set.

2. Exploratory data analysis was performed on the data and the following conclusions were drawn:

####2.1)The entire class of students performed the best in Reading followed by Writing and then Math. The lowest mark in math is 0, for the other two subjects, the lowest mark > 0. Hence, the mean for maths is the least.However, in each subject, the maximum marks obtained is 100 [Mean and Pie Chart]

####2.2)Most people have scored between 60 to 70 marks in Maths, 60-80 in case of Writing and 70 - 80 in case of Reading [Mode and Histogram]

####2.3)No gender disparity was seen in terms of scoring highest marks in any subject [Bar Graph]

####2.4)However, female collectivey have performed better in case of Reading as compared to Maths. For male, the relationship is vice-versa [Scatter Plot]

####2.5)Majority of the class population has scored above 40 in all subjects. Implying that, if we assume 40 to be the cut off marks for qualifying the exams, the majority of class has passed the exams [Bubble Chart]