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

df = pd.read_csv("C:\\Users\\Sarvadnya\\OneDrive\\Desktop\\ROHIT
STUIDY MATERIALS\\Student_dataset\\
Expanded_data_with_more_features.csv")

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30641 entries, 0 to 30640
Data columns (total 15 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Unnamed: 0           30641 non-null  int64  
 1   Gender               30641 non-null  object 
 2   EthnicGroup          28801 non-null  object 
 3   ParentEduc           28796 non-null  object 
 4   LunchType            30641 non-null  object 
 5   TestPrep             28811 non-null  object 
 6   ParentMaritalStatus  29451 non-null  object 
 7   PracticeSport        30010 non-null  object 
 8   IsFirstChild         29737 non-null  object 
 9   NrSiblings           29069 non-null  float64
 10  TransportMeans       27507 non-null  object 
 11  WklyStudyHours       29686 non-null  object 
 12  MathScore            30641 non-null  int64  
 13  ReadingScore         30641 non-null  int64  
 14  WritingScore         30641 non-null  int64  
dtypes: float64(1), int64(4), object(10)
memory usage: 3.5+ MB
```

```python
df.head()
```

```
   Unnamed: 0  Gender EthnicGroup          ParentEduc     LunchType
TestPrep  \
0           0  female         NaN   bachelor's degree      standard
none   
1           1  female     group C        some college      standard
NaN   
2           2  female     group B     master's degree      standard
none   
3           3    male     group A  associate's degree  free/reduced
none   
4           4    male     group C        some college      standard
none   

  ParentMaritalStatus PracticeSport IsFirstChild  NrSiblings
```

```
TransportMeans  \
0              married     regularly          yes         3.0
school_bus
1              married     sometimes          yes         0.0
NaN
2               single     sometimes          yes         4.0
school_bus
3              married         never           no         1.0
NaN
4              married     sometimes          yes         0.0
school_bus

  WklyStudyHours  MathScore  ReadingScore  WritingScore
0            < 5         71            71            74
1         5 - 10         69            90            88
2            < 5         87            93            91
3         5 - 10         45            56            42
4         5 - 10         76            78            75
```

```
df.describe()
```

```
         Unnamed: 0    NrSiblings     MathScore  ReadingScore
WritingScore
count  30641.000000  29069.000000  30641.000000  30641.000000
30641.000000
mean     499.556607      2.145894     66.558402     69.377533
68.418622
std      288.747894      1.458242     15.361616     14.758952
15.443525
min        0.000000      0.000000      0.000000     10.000000
4.000000
25%      249.000000      1.000000     56.000000     59.000000
58.000000
50%      500.000000      2.000000     67.000000     70.000000
69.000000
75%      750.000000      3.000000     78.000000     80.000000
79.000000
max      999.000000      7.000000    100.000000    100.000000
100.000000
```

```
df.isnull().sum()
```

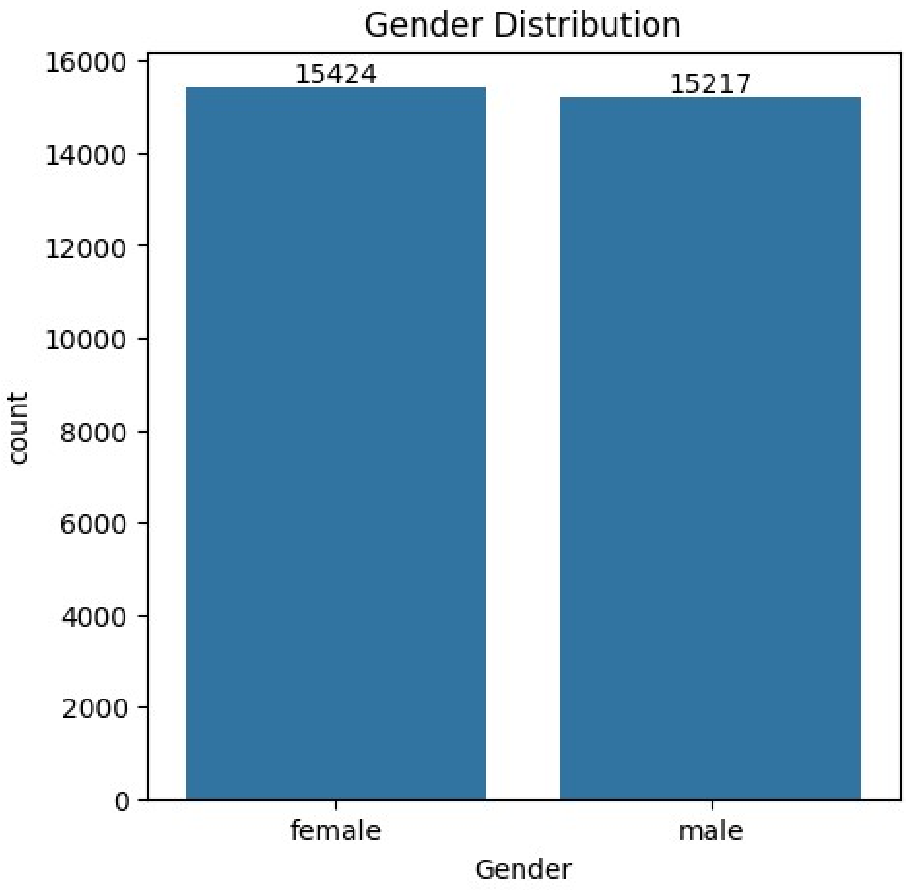
```
Unnamed: 0                0
Gender                    0
EthnicGroup            1840
ParentEduc             1845
LunchType                 0
TestPrep               1830
ParentMaritalStatus    1190
PracticeSport           631
```

```
IsFirstChild            904
NrSiblings             1572
TransportMeans         3134
WklyStudyHours          955
MathScore                 0
ReadingScore              0
WritingScore              0
dtype: int64
```

```
df.head()
```

```
   Gender EthnicGroup          ParentEduc     LunchType TestPrep  \
0  female         NaN   bachelor's degree      standard     none
1  female     group C        some college      standard      NaN
2  female     group B     master's degree      standard     none
3    male     group A  associate's degree  free/reduced     none
4    male     group C        some college      standard     none

  ParentMaritalStatus PracticeSport IsFirstChild  NrSiblings
TransportMeans  \
0             married     regularly          yes         3.0
school_bus
1             married     sometimes          yes         0.0
NaN
2              single     sometimes          yes         4.0
school_bus
3             married         never           no         1.0
NaN
4             married     sometimes          yes         0.0
school_bus

  WklyStudyHours  MathScore  ReadingScore  WritingScore
0            < 5         71            71            74
1         5 - 10         69            90            88
2            < 5         87            93            91
3         5 - 10         45            56            42
4         5 - 10         76            78            75
```

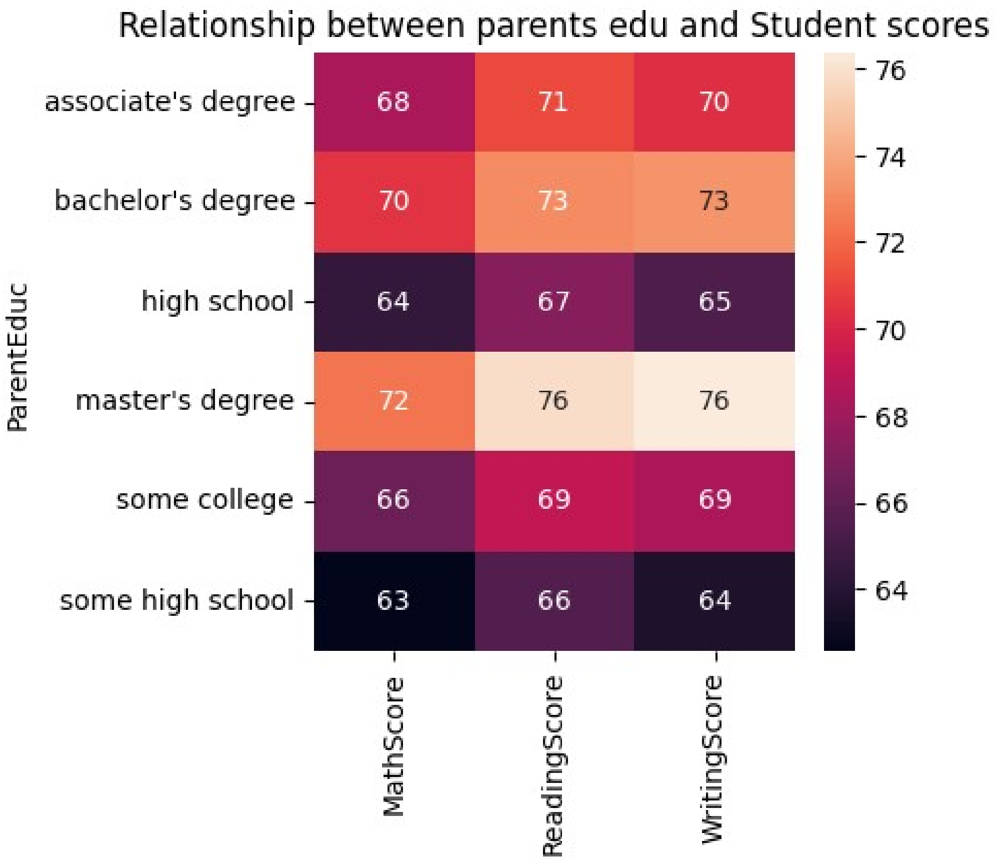
```
#gender Distrubution
plt.figure(figsize = (5,5))
ax = sns.countplot( data = df, x = "Gender")
ax.bar_label(ax.containers[0])
plt.title("Gender Distribution")
plt.show()
```



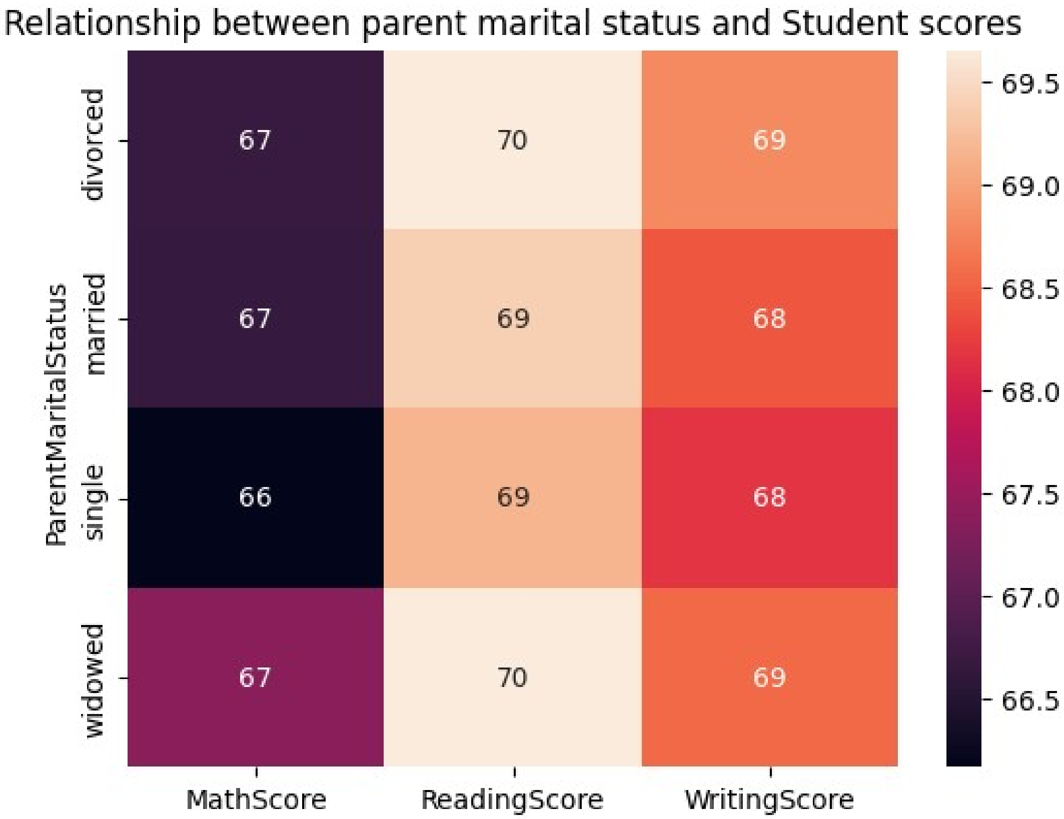
```
# from the above chart female is dominant as compare to male

gb = df.groupby("ParentEduc").agg({"MathScore" : 'mean' ,
"ReadingScore" : 'mean' , "WritingScore" : 'mean'})
gb
```

```
                    MathScore  ReadingScore  WritingScore
ParentEduc
associate's degree  68.365586     71.124324     70.299099
bachelor's degree   70.466627     73.062020     73.331069
high school         64.435731     67.213997     65.421136
master's degree     72.336134     75.832921     76.356896
some college        66.390472     69.179708     68.501432
some high school    62.584013     65.510785     63.632409
```

```
plt.figure(figsize = (4,4))
sns.heatmap(gb , annot = True)
plt.title("Relationship between parents edu and Student scores")
plt.show()
```



```
# from the above we conclude the parent eduction is good impact their
scores

gp = df.groupby("ParentMaritalStatus").agg({"MathScore" : 'mean' ,
"ReadingScore" : 'mean' , "WritingScore" : 'mean'})
gp

                     MathScore  ReadingScore  WritingScore
ParentMaritalStatus
divorced             66.691197     69.655011     68.799146
married              66.657326     69.389575     68.420981
single               66.165704     69.157250     68.174440
widowed              67.368866     69.651438     68.563452

sns.heatmap(gp , annot = True)
plt.title("Relationship between parent marital status and Student
scores")
plt.plot()

[]
```



```
# The above chart shown the parent marital status is not heavily
impact on the scores

df.head()

   Gender EthnicGroup          ParentEduc     LunchType TestPrep  \
0  female         NaN   bachelor's degree      standard     none
1  female     group C        some college      standard      NaN
2  female     group B     master's degree      standard     none
3    male     group A  associate's degree  free/reduced     none
4    male     group C        some college      standard     none

  ParentMaritalStatus PracticeSport IsFirstChild  NrSiblings
TransportMeans  \
0             married     regularly          yes         3.0
school_bus
1             married     sometimes          yes         0.0
NaN
2              single     sometimes          yes         4.0
school_bus
3             married         never           no         1.0
NaN
4             married     sometimes          yes         0.0
school_bus
```

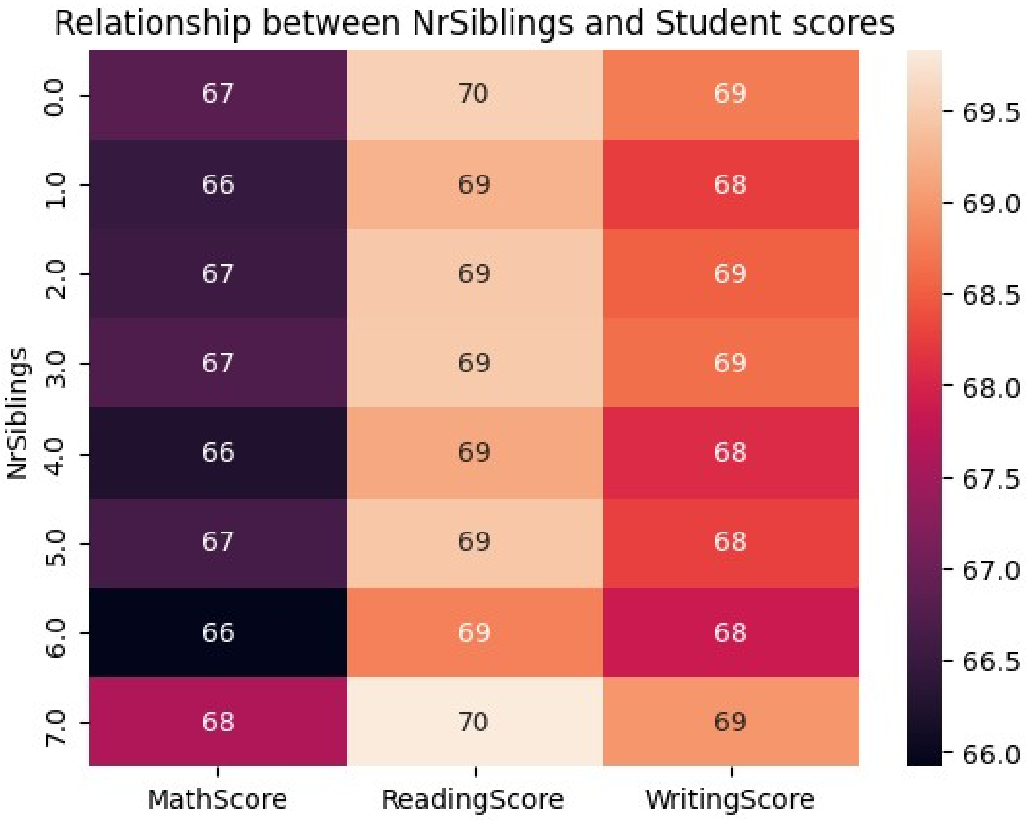
```
  WklyStudyHours  MathScore  ReadingScore  WritingScore
0            < 5         71            71            74
1         5 - 10         69            90            88
2            < 5         87            93            91
3         5 - 10         45            56            42
4         5 - 10         76            78            75
```

```
ga = df.groupby("NrSiblings").agg({"MathScore" : 'mean' ,
"ReadingScore" : 'mean' , "WritingScore" : 'mean'})
ga
```

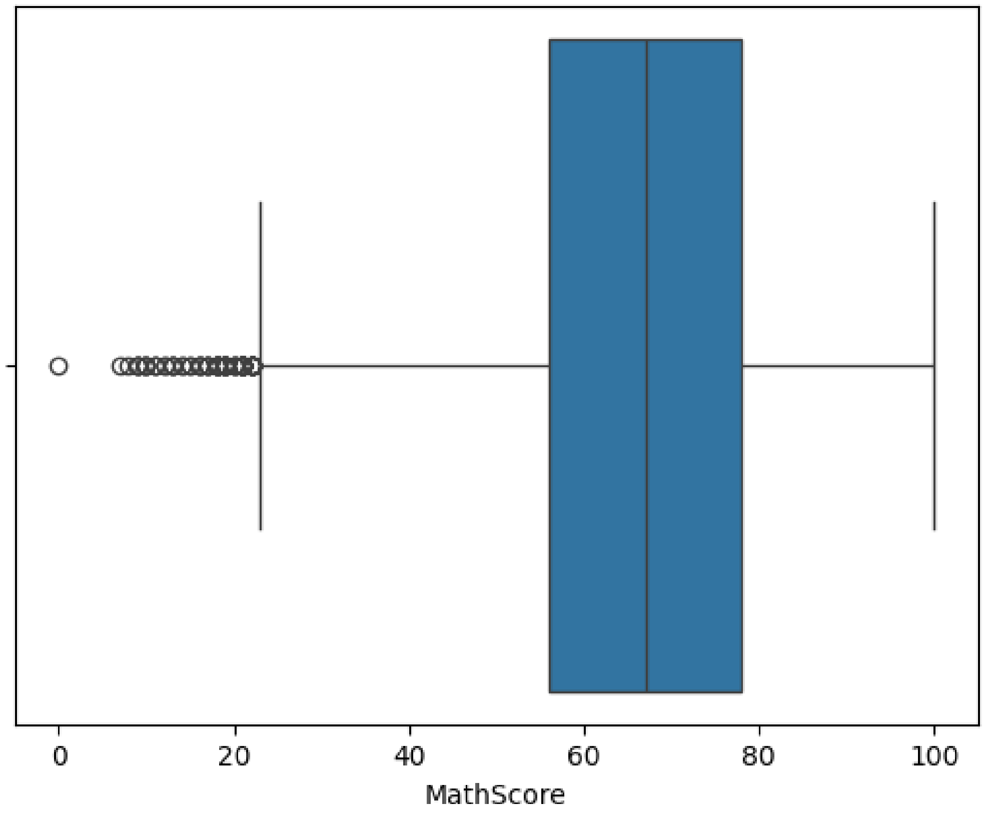
```
            MathScore  ReadingScore  WritingScore
NrSiblings
0.0         66.819449     69.547812     68.746515
1.0         66.473896     69.259097     68.245345
2.0         66.554934     69.472018     68.522533
3.0         66.719092     69.488159     68.650498
4.0         66.245495     69.144169     68.073444
5.0         66.630303     69.453788     68.282576
6.0         65.917219     68.801325     67.860927
7.0         67.615120     69.828179     68.986254
```

```
sns.heatmap(ga , annot = True)
plt.title("Relationship between NrSiblings and Student scores")
plt.plot()
```

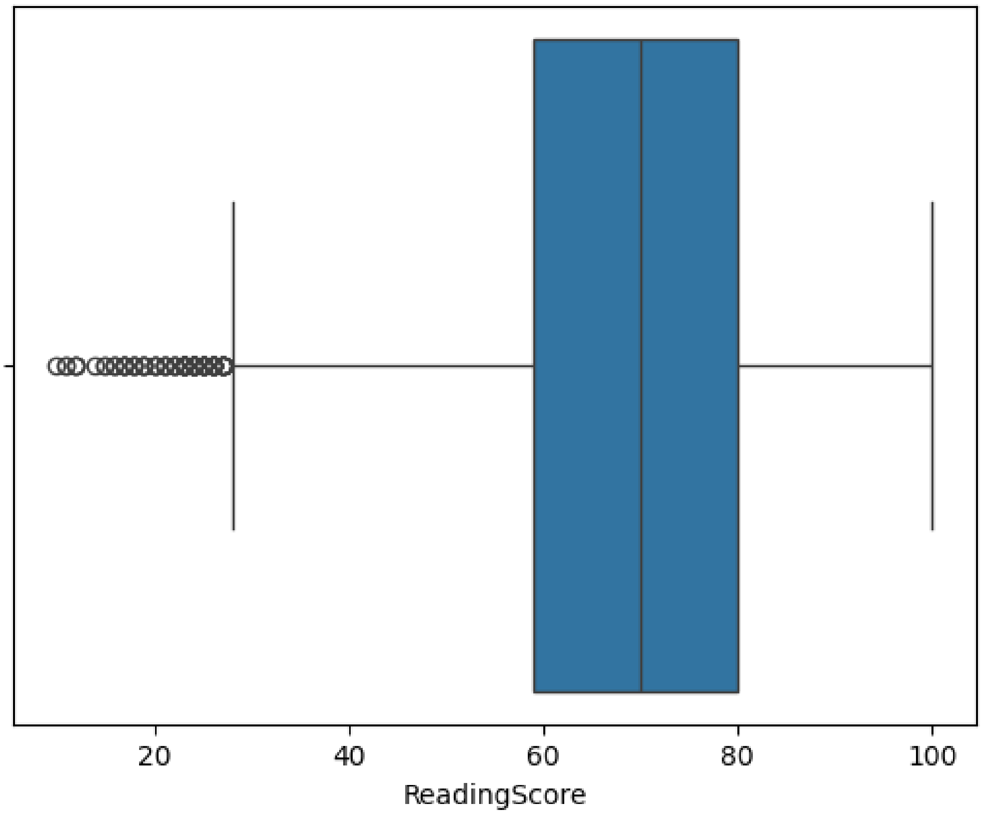
```
[]
```



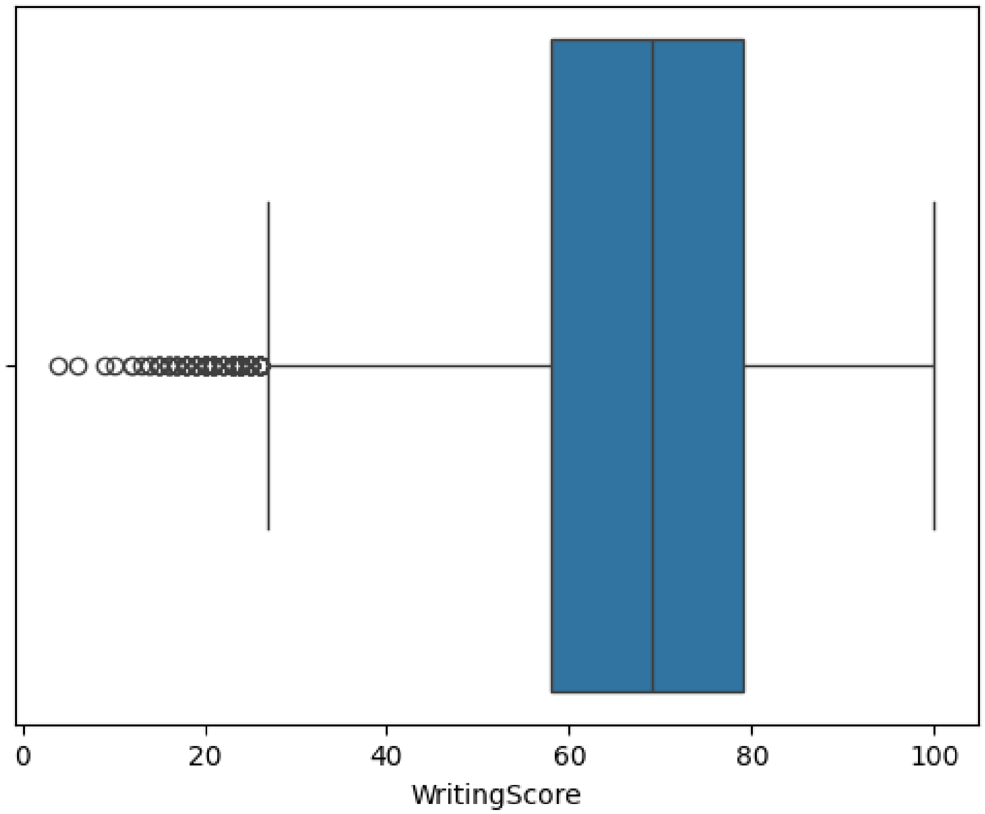
```
sns.boxplot(data = df , x = 'MathScore')
plt.plot()

[]
```



```
sns.boxplot(data = df , x = 'ReadingScore')
plt.plot()

[]
```



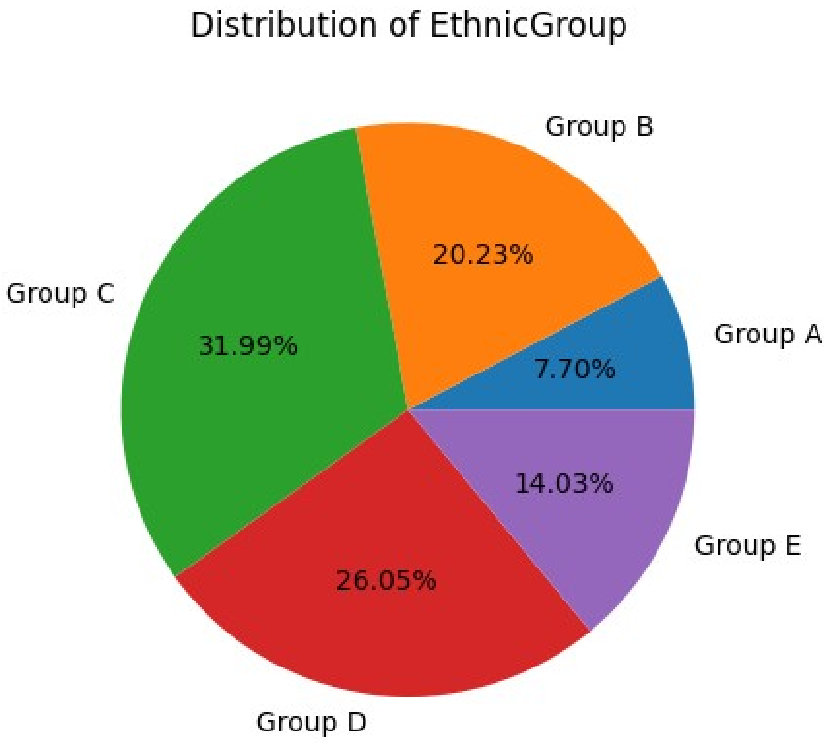
```
sns.boxplot(data = df , x = 'WritingScore')
plt.plot()

[]
```



```
df['EthnicGroup'].unique()
```

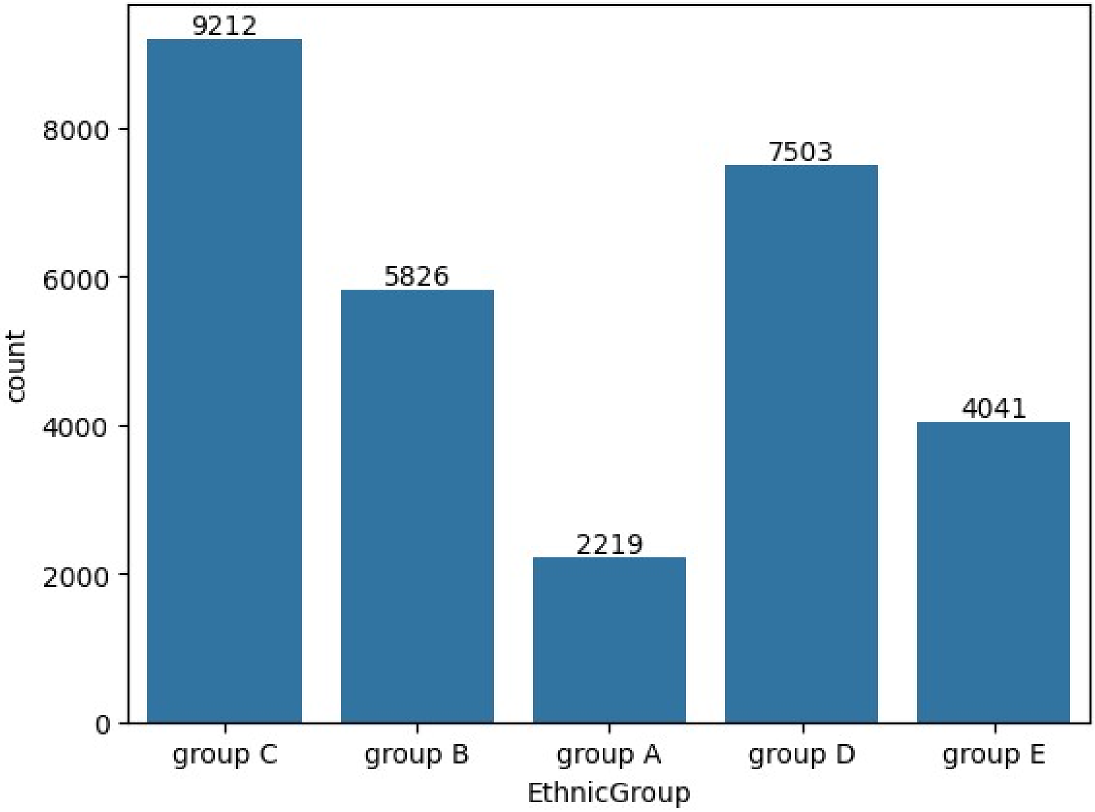
```
array([nan, 'group C', 'group B', 'group A', 'group D', 'group E'],
      dtype=object)
```

```
#distribution of ethnicGroup
GroupA = df[df['EthnicGroup'] == 'group A'].count()
GroupB = df[df['EthnicGroup'] == 'group B'].count()
GroupC = df[df['EthnicGroup'] == 'group C'].count()
GroupD = df[df['EthnicGroup'] == 'group D'].count()
GroupE = df[df['EthnicGroup'] == 'group E'].count()

mylist = [GroupA["EthnicGroup"] , GroupB["EthnicGroup"] ,
GroupC["EthnicGroup"] , GroupD["EthnicGroup"] , GroupE["EthnicGroup"]]
l = ["Group A" , "Group B" , "Group C" , "Group D" , "Group E"]

plt.pie(mylist , labels = l , autopct = "%1.2f%%")
plt.title("Distribution of EthnicGroup")
plt.show()
```



```
ax = sns.countplot(data = df , x = "EthnicGroup")
ax.bar_label(ax.containers[0])
```

```
[Text(0, 0, '9212'),
 Text(0, 0, '5826'),
 Text(0, 0, '2219'),
 Text(0, 0, '7503'),
 Text(0, 0, '4041')]
```

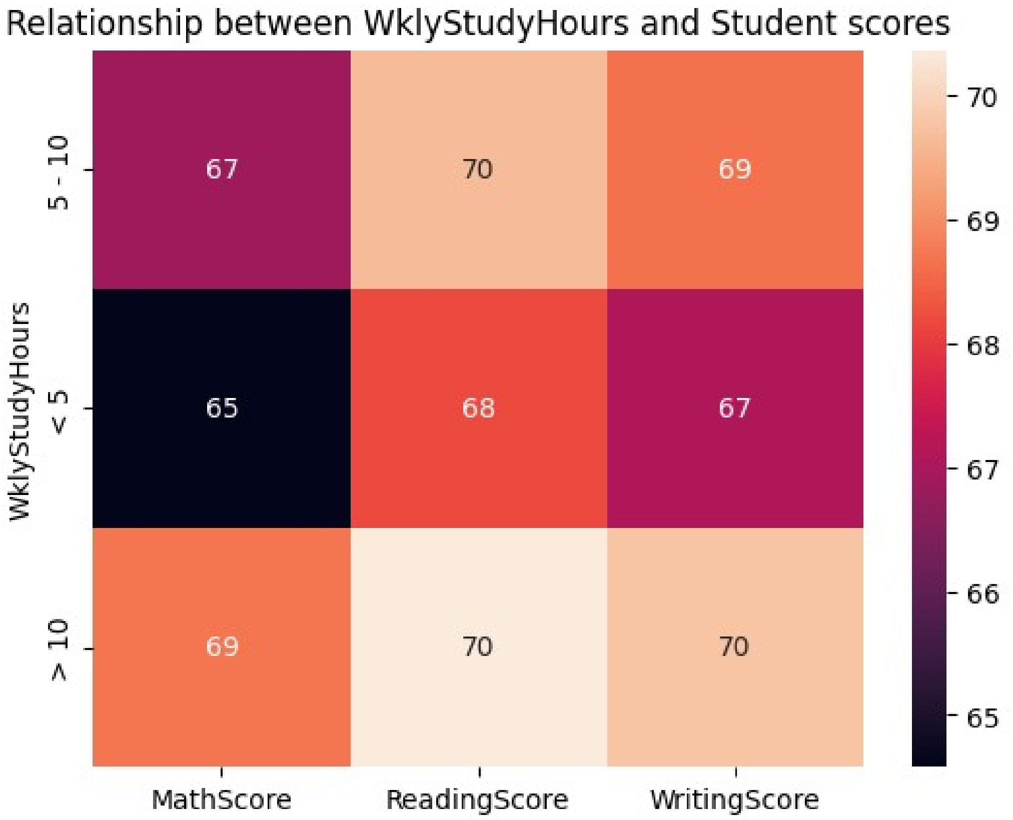


```
gs = df.groupby("WklyStudyHours").agg({"MathScore" : 'mean' ,
"ReadingScore" : 'mean' , "WritingScore" : 'mean'})
gs
```

```
                MathScore  ReadingScore  WritingScore
WklyStudyHours
5 - 10          66.870491     69.660532     68.636280
< 5             64.580359     68.176135     67.090192
> 10            68.696655     70.365436     69.777778
```

```
sns.heatmap(gs , annot = True)
plt.title("Relationship between WklyStudyHours and Student scores")
plt.plot()
```

```
[]
```

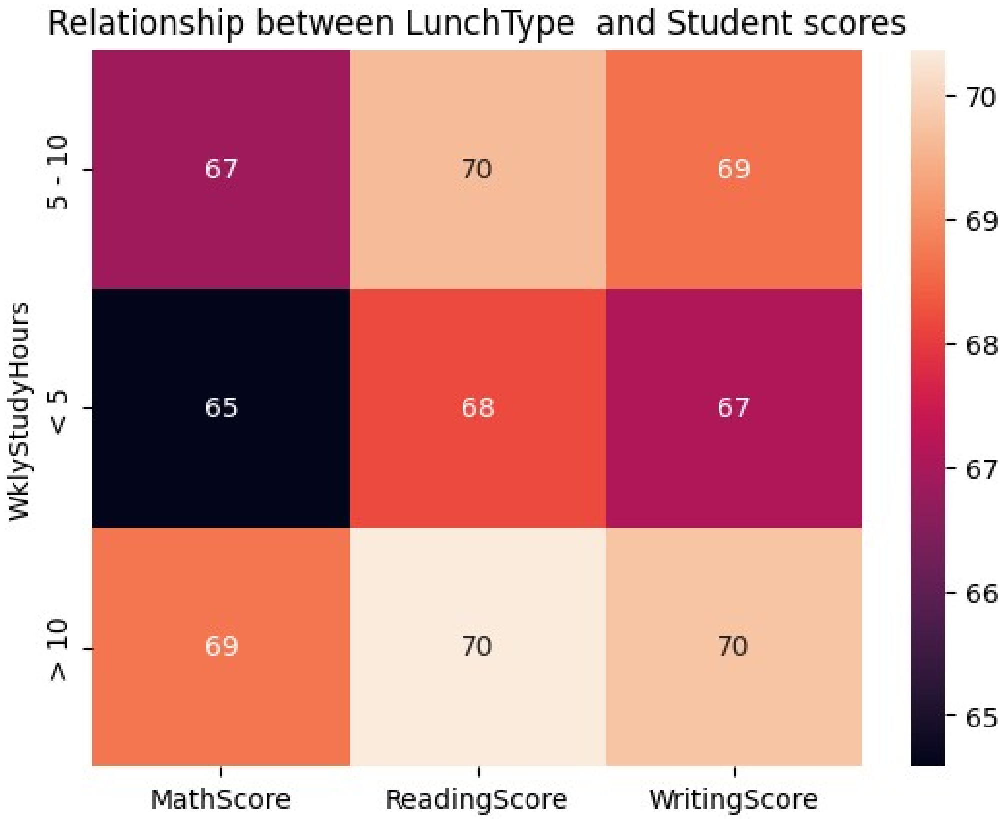


```
gl = df.groupby("LunchType").agg({"MathScore" : 'mean' ,
"ReadingScore" : 'mean' , "WritingScore" : 'mean'})
gl
```

```
              MathScore  ReadingScore  WritingScore
LunchType
free/reduced  58.862332     64.189735     62.650522
standard      70.709370     72.175634     71.529716
```

```
sns.heatmap(gs , annot = True)
plt.title("Relationship between LunchType  and Student scores")
plt.plot()
```

```
[]
```

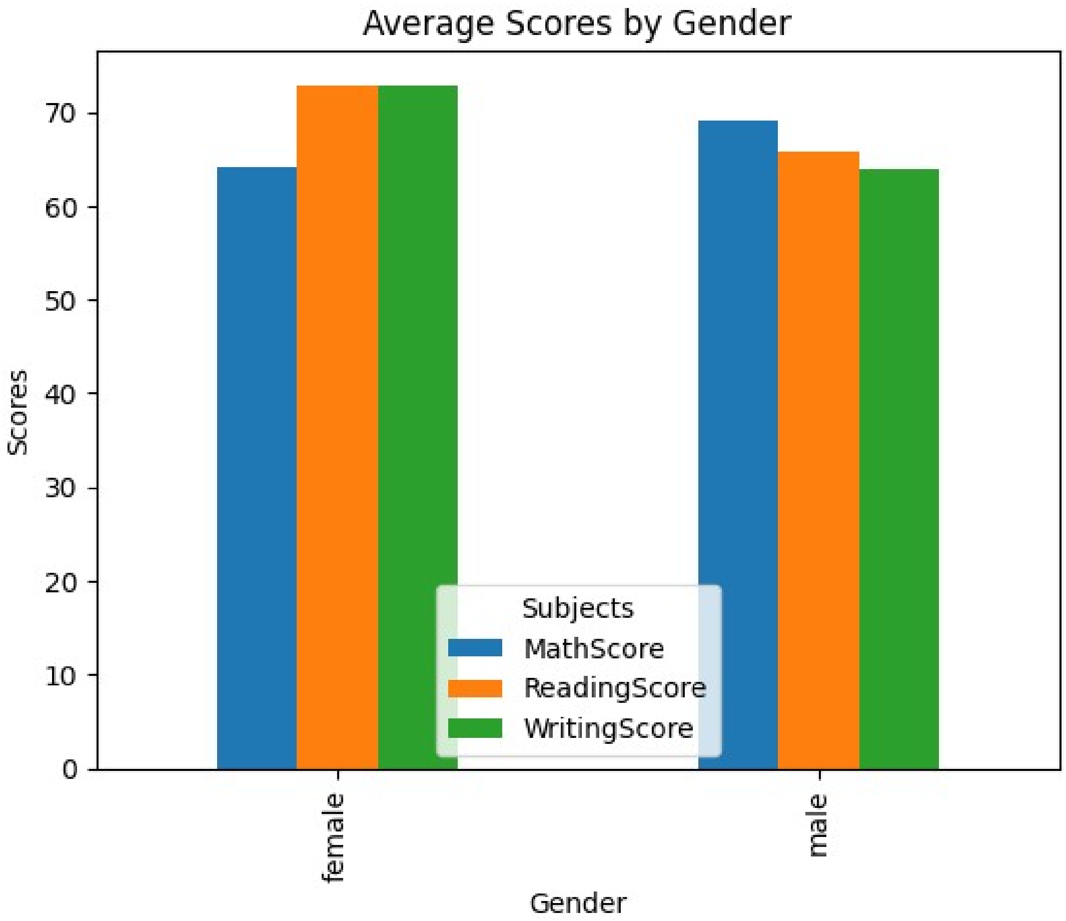


```
#Students with standard lunch scored higher than those with free or
reduced lunch.

gn = df.groupby("Gender").agg({"MathScore" : 'mean' , "ReadingScore" :
'mean' , "WritingScore" : 'mean'})
gn
```

```
        MathScore  ReadingScore  WritingScore
Gender
female  64.080654     72.853216     72.856457
male    69.069856     65.854571     63.920418
```

```
gn.plot(kind='bar')
plt.title("Average Scores by Gender")
plt.ylabel("Scores")
plt.xlabel("Gender")
plt.legend(title="Subjects")
```

```
<matplotlib.legend.Legend at 0x21cdd679bd0>
```



```
# The above chart say Female students gain more scores as compare to
male students
```