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

df= pd.read_csv("expanded_data_with_more_Features.csv")

df.head(10)
```

```
   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
5  female     group B  associate's degree      standard       none
6  female     group B        some college      standard  completed
7    male     group B        some college  free/reduced       none
8    male     group D         high school  free/reduced  completed
9  female     group B         high school  free/reduced       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
5             married     regularly          yes         1.0
school_bus
6             widowed         never           no         1.0
private
7             married     sometimes          yes         1.0
private
8              single     sometimes           no         3.0
private
9             married     regularly          yes         NaN
private

  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
```

```
5         5 - 10         73            84            79
6         5 - 10         85            93            89
7           > 10         41            43            39
8           > 10         65            64            68
9            < 5         37            59            50

df.tail()

       Unnamed: 0  Gender EthnicGroup          ParentEduc
LunchType  \
30636         816  female     group D         high school
standard
30637         890    male     group E         high school
standard
30638         911  female         NaN         high school
free/reduced
30639         934  female     group D  associate's degree
standard
30640         960    male     group B        some college
standard

        TestPrep ParentMaritalStatus PracticeSport IsFirstChild
NrSiblings  \
30636       none              single     sometimes           no
2.0
30637       none              single     regularly           no
1.0
30638  completed             married     sometimes           no
1.0
30639  completed             married     regularly           no
3.0
30640       none             married         never           no
1.0

      TransportMeans WklyStudyHours  MathScore  ReadingScore
WritingScore
30636     school_bus         5 - 10         59            61
65
30637        private         5 - 10         58            53
51
30638        private         5 - 10         61            70
67
30639     school_bus         5 - 10         82            90
93
30640     school_bus         5 - 10         64            60
58

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.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
```

```
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
```

```
# Drop unamed column
df = df.drop("Unnamed: 0", axis=1)


print(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
```

```
# change wrong value (5 oct) in correct one(hours) in  WklyStudyHours
```

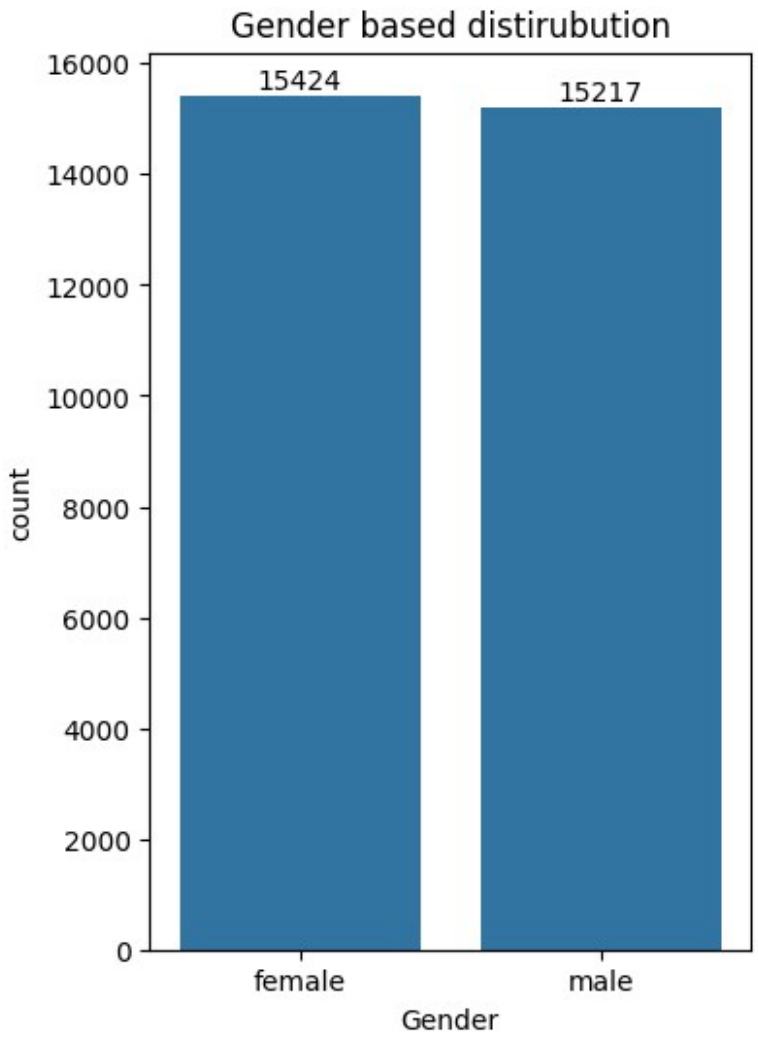
```
df["WklyStudyHours"] = df["WklyStudyHours"].str.replace("05-Oct","5-
10")
df.head()
```

```
   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 Distirbution
plt.figure(figsize = (4,6))
ax=sns.countplot(data = df,x="Gender")
plt.title("Gender based distirubution")
ax.bar_label(ax.containers[0])
plt.show()
```



From the above chart we have found that females are more than male

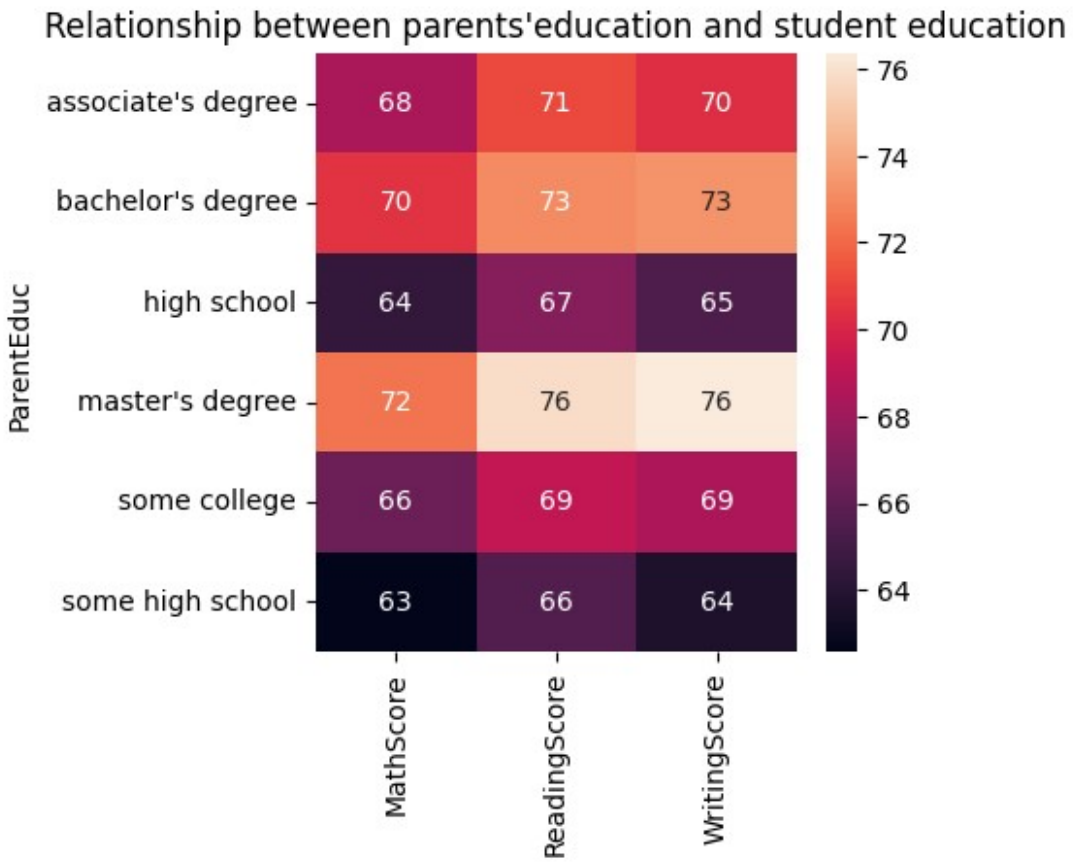
```
af=
df.groupby("ParentEduc").agg({"MathScore":'mean',"ReadingScore":'mean'
,"WritingScore":'mean'})
print(af)
```

```
                    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(af,annot=True)
```

```
plt.title("Relationship between parents'education and student
education")
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



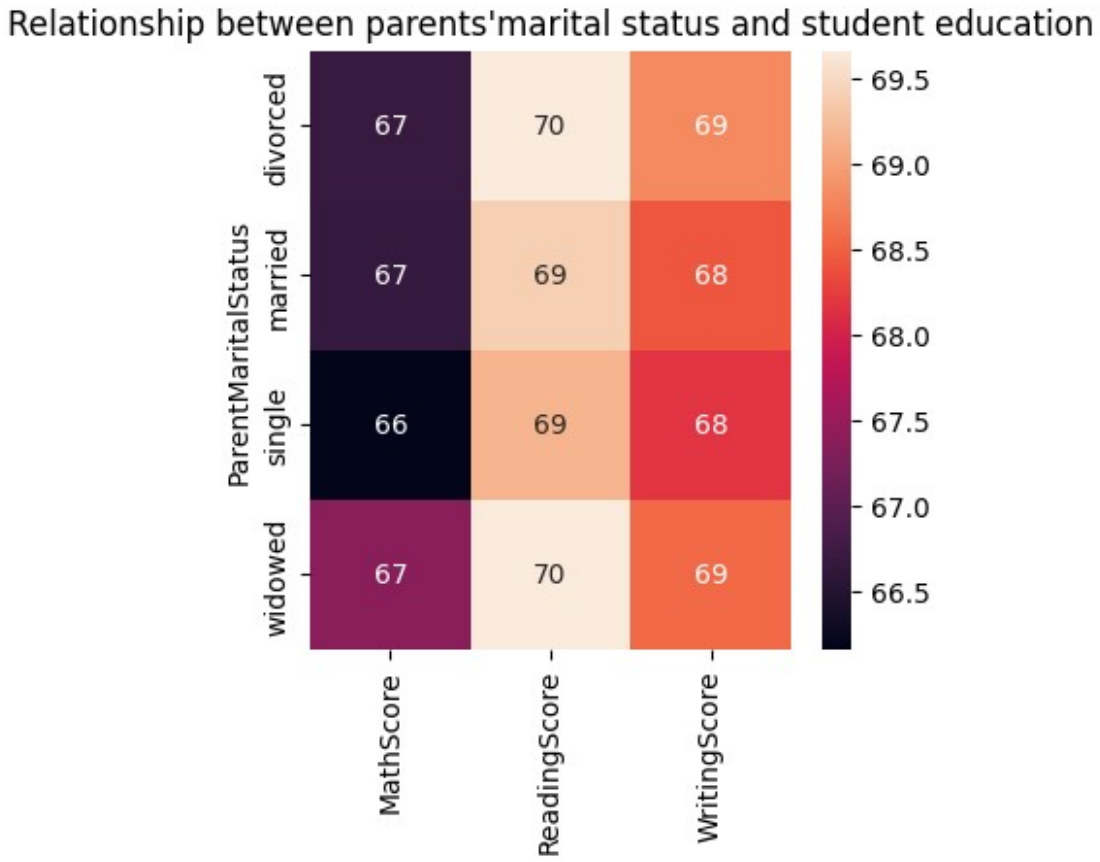
From the above chart we can see that parents'education have major imact on children'score

```
af1 =
df.groupby("ParentMaritalStatus").agg({"MathScore":'mean',"ReadingScor
e":'mean',"WritingScore":'mean'})
print(af1)
```

```
                     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
```

```
plt.figure(figsize = (4,4))
sns.heatmap(af1,annot = True)
plt.title("Relationship between parents'marital status and student
```

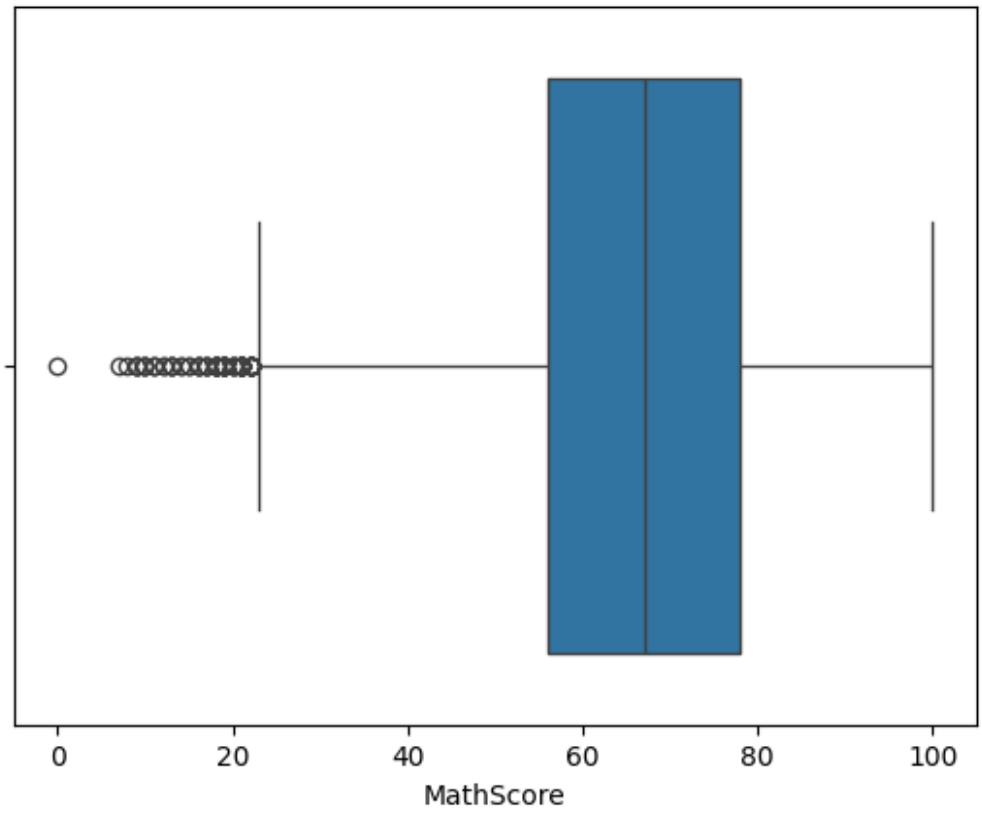
```
education")
plt.show()
```



From above chart we can see that parent's martial has nothing to do with childern'score

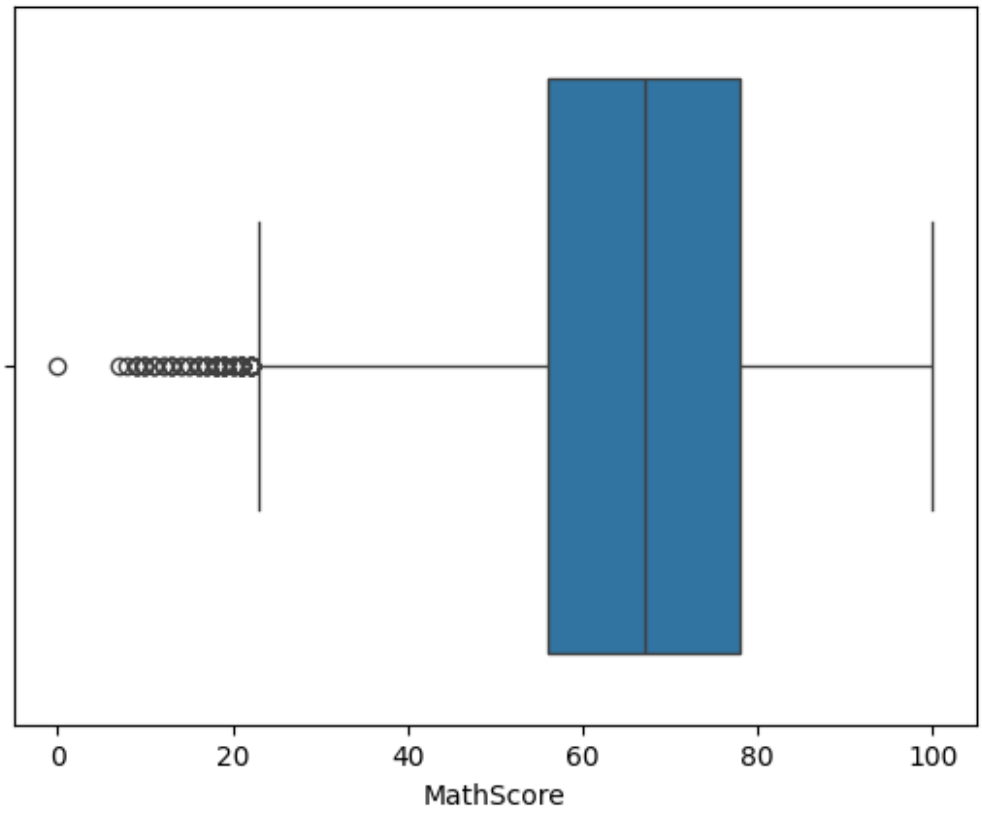
```
sns.boxplot(data =df,x= "MathScore")
```

```
<Axes: xlabel='MathScore'>
```



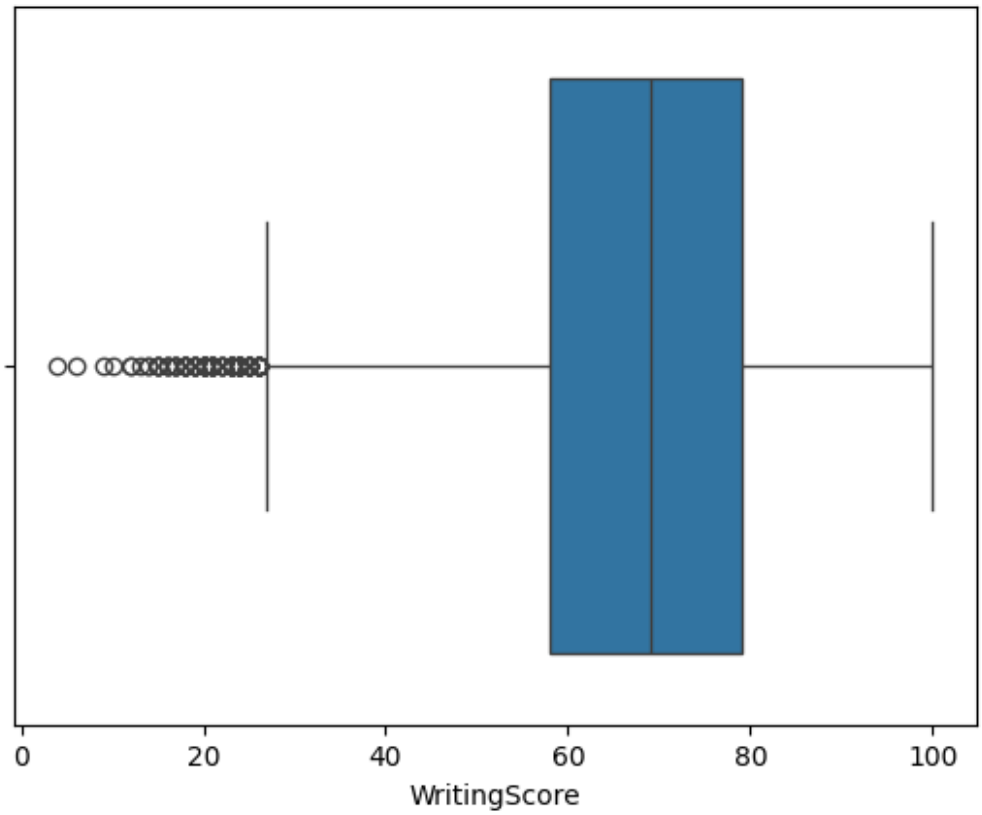

```
sns.boxplot(data =df,x= "ReadingScore")
```

```
<Axes: xlabel='MathScore'>
```




```
sns.boxplot(data =df,x= "WritingScore")
```

```
<Axes: xlabel='WritingScore'>
```



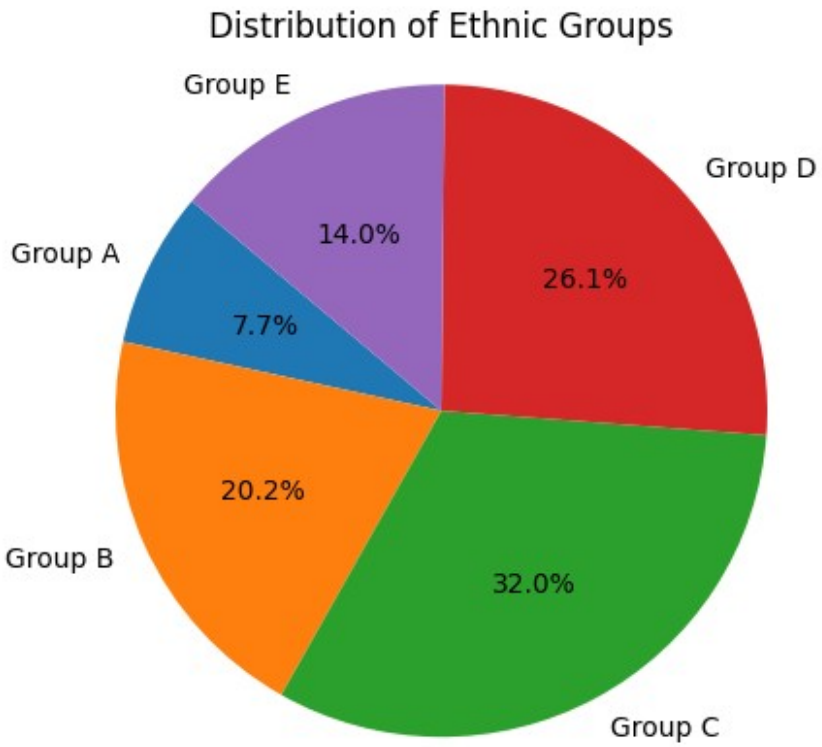
disitribution of EthnicGroup

```
labels = ['Group A', 'Group B', 'Group C', 'Group D', 'Group E']

groupA = df[df['EthnicGroup'] == "group A"]['EthnicGroup'].count()
groupB = df[df['EthnicGroup'] == "group B"]['EthnicGroup'].count()
groupC = df[df['EthnicGroup'] == "group C"]['EthnicGroup'].count()
groupD = df[df['EthnicGroup'] == "group D"]['EthnicGroup'].count()
groupE = df[df['EthnicGroup'] == "group E"]['EthnicGroup'].count()

mylist = [groupA, groupB, groupC, groupD, groupE]


plt.pie(mylist, labels=labels, autopct='%1.1f%%', startangle=140)
plt.title("Distribution of Ethnic Groups")
plt.axis('equal')  # Equal aspect ratio ensures pie is a circle.
plt.show()
```



The pie chart shows that Group C comprises 30% of the students, making it the largest ethnic group in the dataset, while Group E accounts for only 10%