project-1

May 7, 2024

[27]:

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

[2]:

```
df = pd.read_csv("D:\\Personal␣
 ↪Project\\project1\\Expanded_data_with_more_features.csv")
df
```

[2]:

```
       Unnamed: 0  Gender EthnicGroup          ParentEduc     LunchType  \
0               0  female         NaN   bachelor's degree      standard
1               1  female     group C        some college      standard
2               2  female     group B     master's degree      standard
3               3    male     group A  associate's degree  free/reduced
4               4    male     group C        some college      standard
…           …     …         …                 …            …
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  \
0           none              married     regularly          yes         3.0
1            NaN              married     sometimes          yes         0.0
2           none               single     sometimes          yes         4.0
3           none              married         never           no         1.0
4           none              married     sometimes          yes         0.0
…          …                 …           …          …         …
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
0         school_bus            < 5         71            71            74
```

```
1                NaN         5 - 10        69            90            88
2         school_bus            < 5        87            93            91
3                NaN         5 - 10        45            56            42
4         school_bus         5 - 10        76            78            75
…               …            …        …           …            …
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

[30641 rows x 15 columns]
```

[3]:
```
df.describe()
```

[3]:
```
         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
```

[4]:
```
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
```

```
[5]: df.isnull().sum()
```

```
[5]: 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
```

```
[10]: df = df.drop("Unnamed: 0",axis=1)
      df
```

```
[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
      …       …         …                 …           …        …
      30636  female     group D         high school      standard       none
      30637    male     group E         high school      standard       none
      30638  female         NaN         high school  free/reduced  completed
      30639  female     group D  associate's degree      standard  completed
      30640    male     group B        some college      standard       none

            ParentMaritalStatus PracticeSport IsFirstChild  NrSiblings  \
      0                 married     regularly          yes         3.0
      1                 married     sometimes          yes         0.0
      2                  single     sometimes          yes         4.0
      3                 married         never           no         1.0
      4                 married     sometimes          yes         0.0
      …                   …           …          …         …
      30636              single     sometimes           no         2.0
      30637              single     regularly           no         1.0
```

```
30638             married     sometimes           no         1.0
30639             married     regularly           no         3.0
30640             married         never           no         1.0

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

[30641 rows x 14 columns]
```

[11]:

```
df.head(5)
```

[11]:

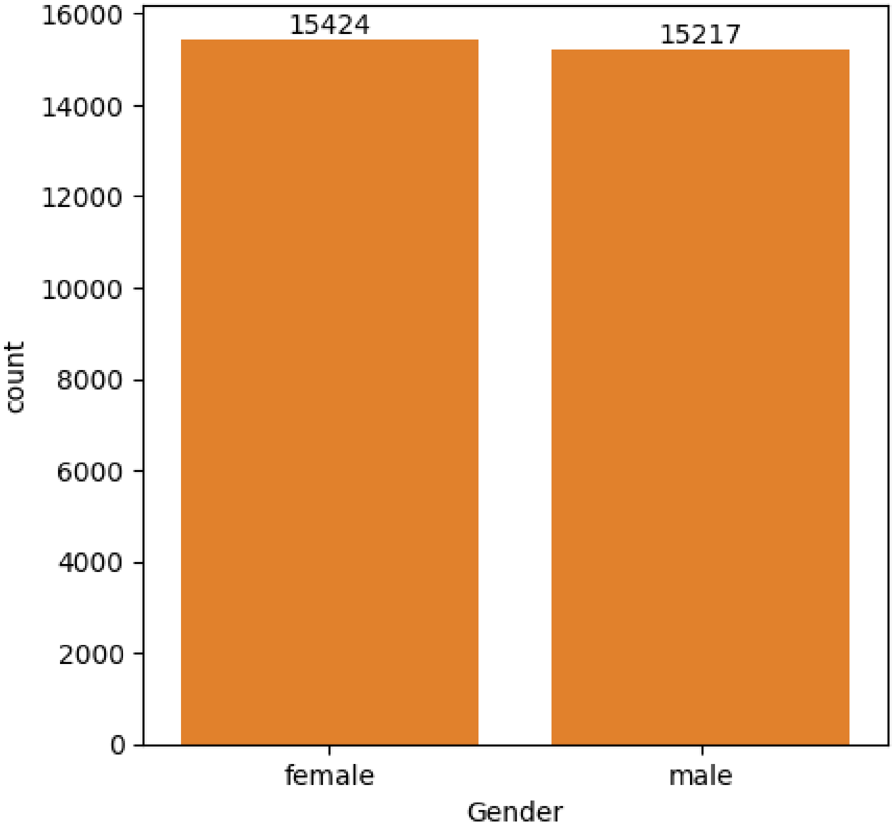
```
   Gender EthnicGroup          ParentEduc     LunchType TestPrep  \
0  female         NaN   bachelor's degree      standard     none
1  female     group C        some college      standard      NaN
2  female     group B     master's degree      standard     none
3    male     group A  associate's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

THE NUMBER OF FEMALES ARE MORE

[35]:

```
plt.figure(figsize=(5,5))
sns.countplot(data=df,x="Gender")
ax = sns.countplot(data=df,x="Gender")
```

```python
ax.bar_label(ax.containers[0])
plt.show()
```



[36]:

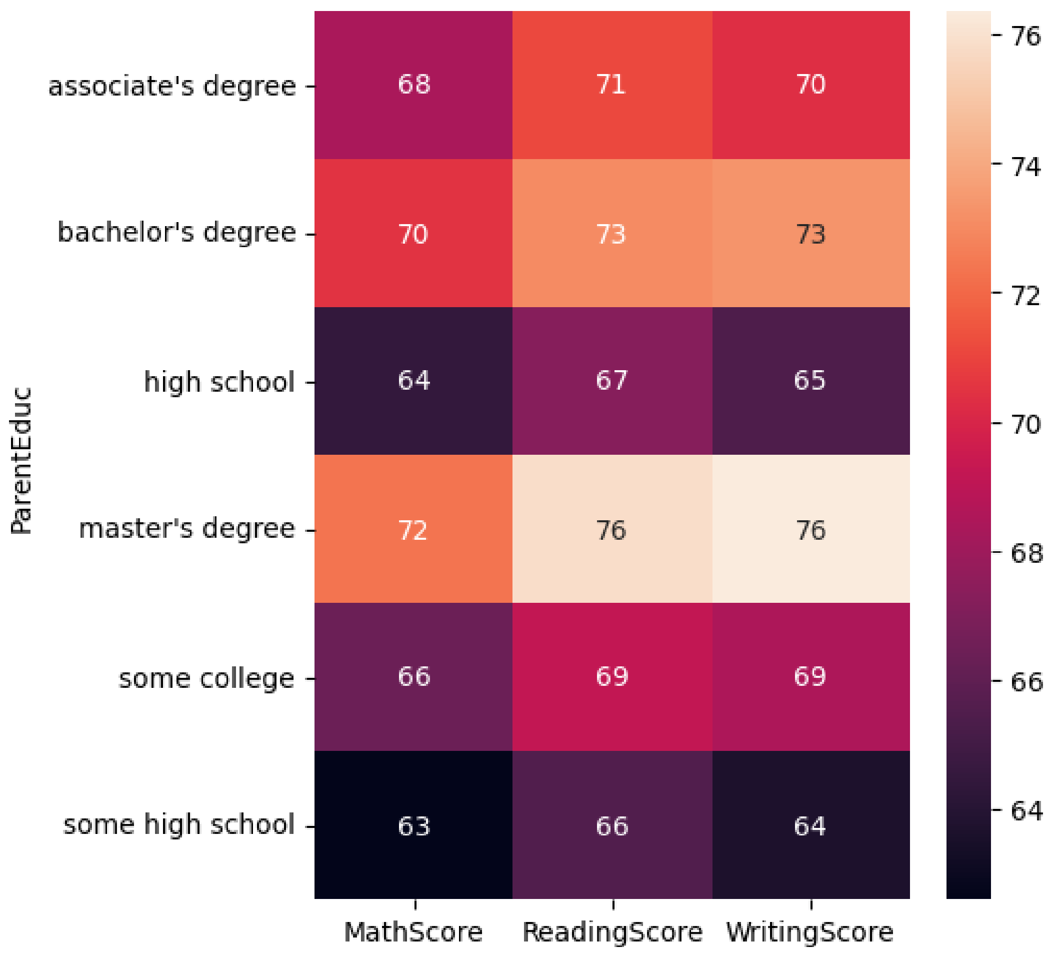
```python
gp = df.groupby("ParentEduc").agg({"MathScore":"mean","ReadingScore":
↪"mean","WritingScore":"mean"})
gp
```

[36]:

```
                    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
```

[44]:

```python
plt.figure(figsize=(5,6))
sns.heatmap(gp,annot=True)
plt.show()
```



FROM ABOVE CHART WE CAN SEE THAT THE STUDENTS WHOSE PARENTS HAVE MASTER'S DEGREE HAVE HIGHEST SCORES IN ALL THE SUBJECTS

[45]:

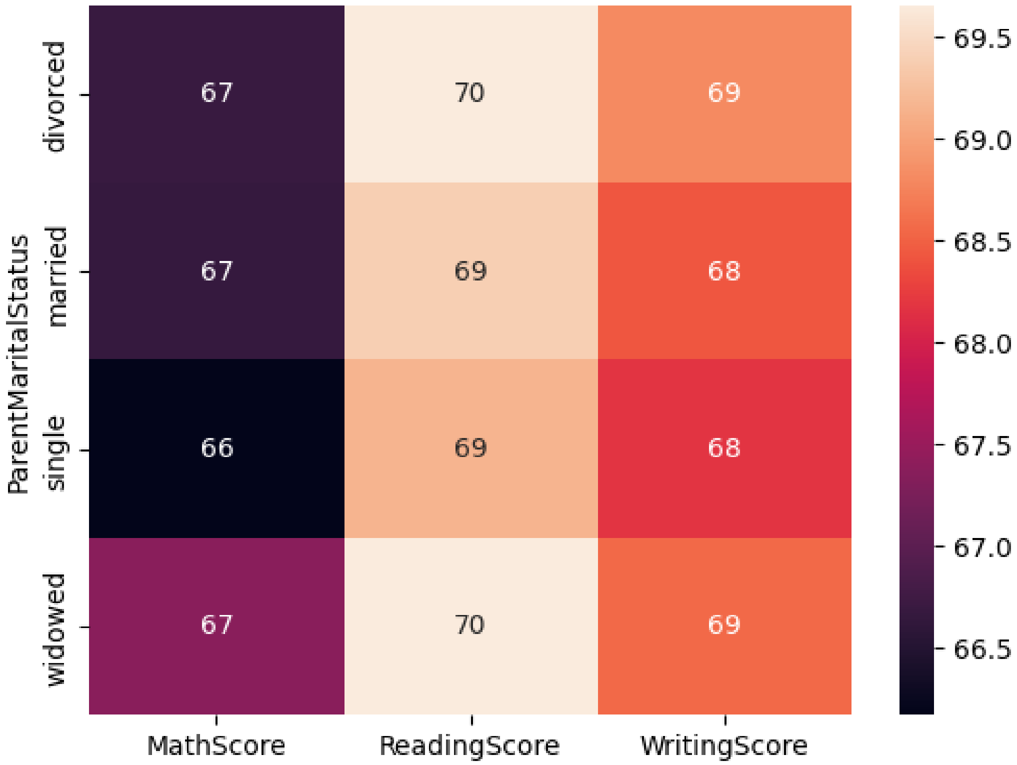
```
gp1 = df.groupby("ParentMaritalStatus").agg({"MathScore":"mean","ReadingScore":
  ↪"mean","WritingScore":"mean"})
gp1
```

[45]:

```
                     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
```

[47]:

```
sns.heatmap(gp1,annot=True)
plt.show()
```

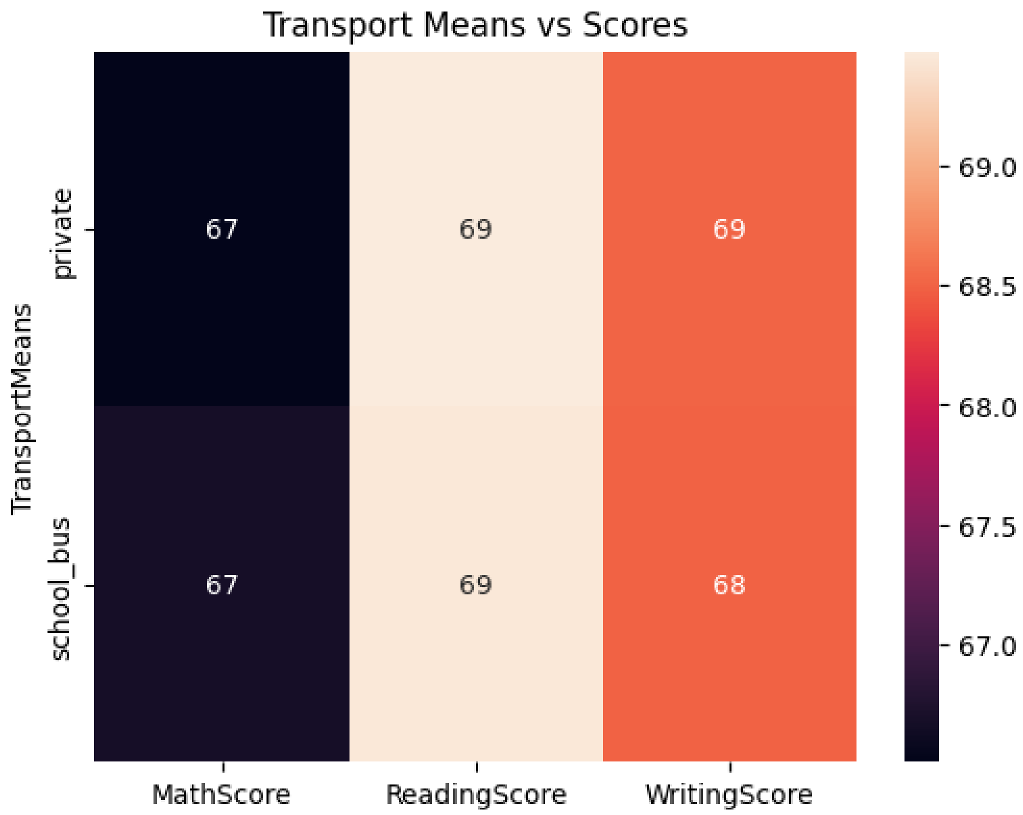


IS IT NOT MORE EFFECT ON THE SCORES OF THE STUDENTS

```
[48]: gp3 = df.groupby("TransportMeans").agg({"MathScore":"mean","ReadingScore":
       ↪"mean","WritingScore":"mean"})
      gp3
```

```
[48]:                  MathScore  ReadingScore  WritingScore
      TransportMeans
      private          66.511354     69.472364     68.509593
      school_bus       66.674636     69.446206     68.492351
```

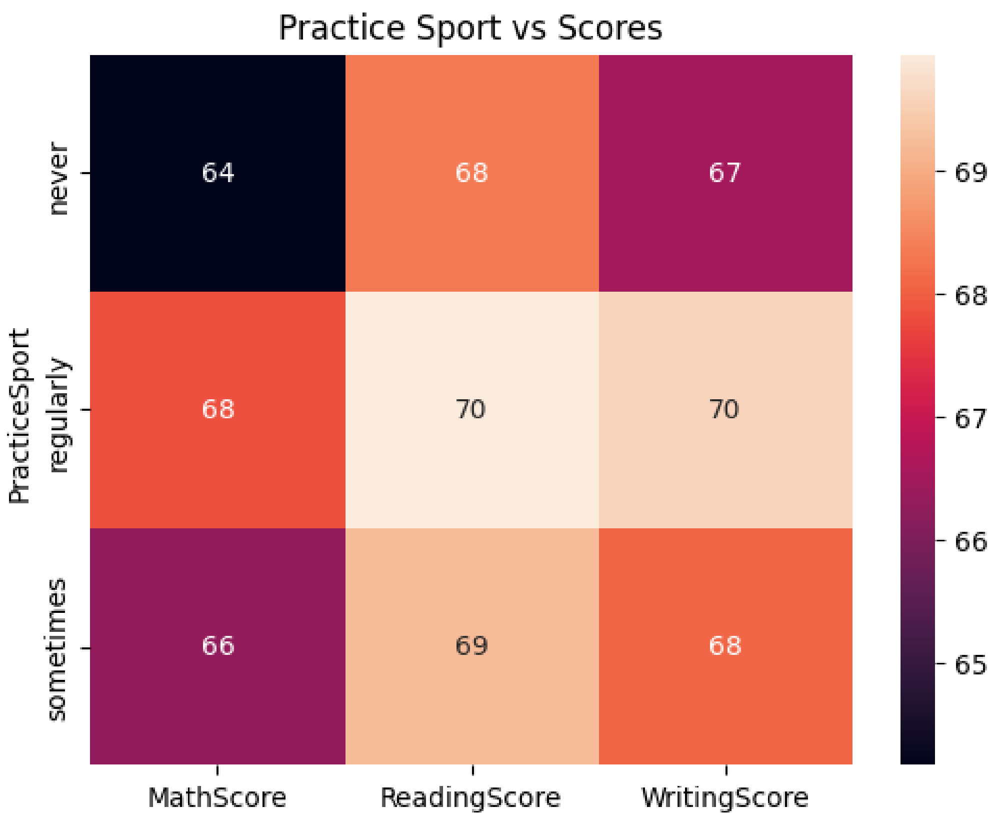
```
[50]: sns.heatmap(gp3,annot=True)
      plt.title("Transport Means vs Scores")
      plt.show()
```



[51]:
```
gp4 = df.groupby("PracticeSport").agg({"MathScore":"mean","ReadingScore":
  ↪"mean","WritingScore":"mean"})
gp4
```

[51]:
```
                 MathScore  ReadingScore  WritingScore
PracticeSport
never            64.171079     68.337662     66.522727
regularly        67.839155     69.943019     69.604003
sometimes        66.274831     69.241307     68.072438
```

[52]:
```
sns.heatmap(gp4,annot=True)
plt.title("Practice Sport vs Scores")
plt.show()
```

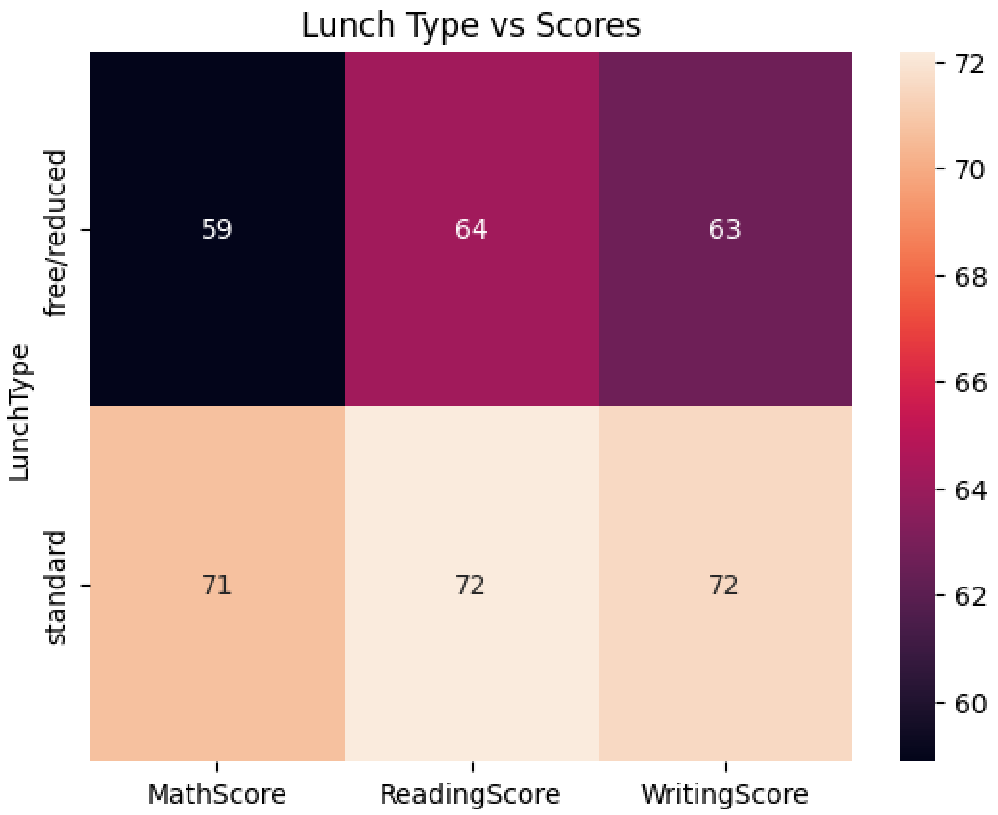


FROM ABOVE CHART WE CAN SEE THAT THE STUDENTS WHO PRACTICE SPORTS HAVE HIGHER SCORES

[53]:
```
gp5 = df.groupby("LunchType").agg({"MathScore":"mean","ReadingScore":
  ↪"mean","WritingScore":"mean"})
gp5
```

[53]:
```
               MathScore  ReadingScore  WritingScore
LunchType
free/reduced   58.862332     64.189735     62.650522
standard       70.709370     72.175634     71.529716
```

[54]:
```
sns.heatmap(gp5,annot=True)
plt.title("Lunch Type vs Scores")
plt.show()
```

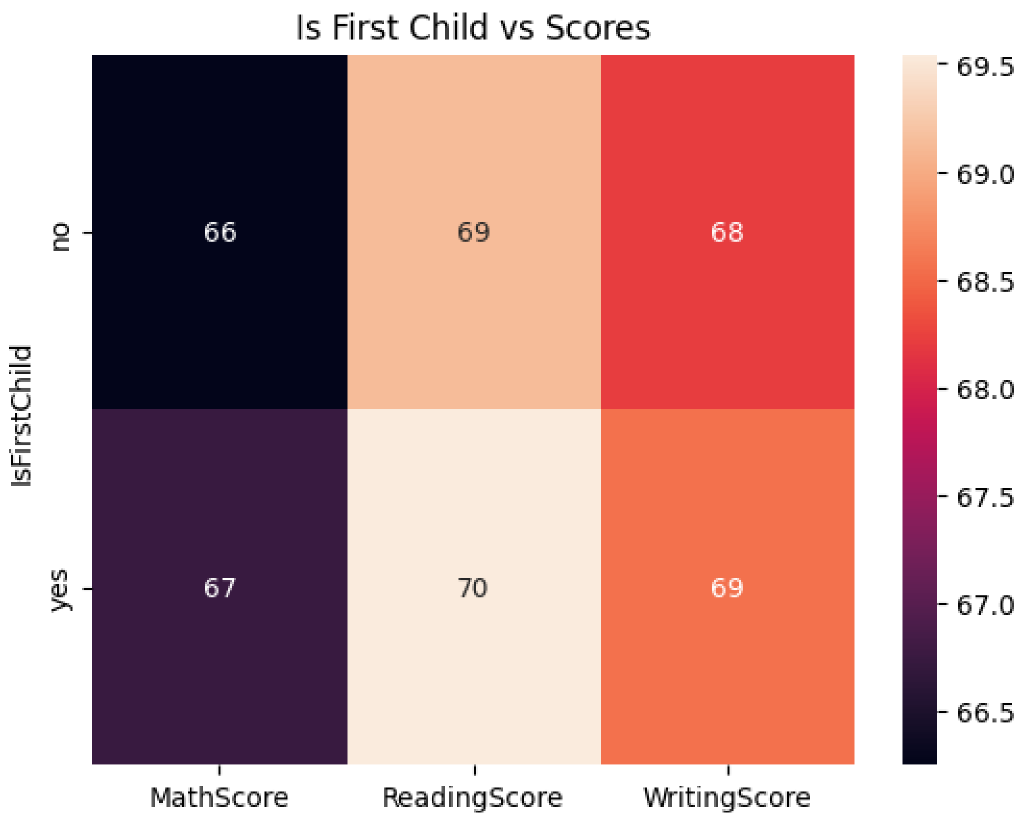


FROM ABOVE CHART WE CAN SEE THAT THE STUDENTS WHO HAVE STANDARD LUNCH HAVE HIGHER SCORES

[55]:
```
gp6 =df.groupby("IsFirstChild").agg({"MathScore":"mean","ReadingScore":
 ↪"mean","WritingScore":"mean"})
gp6
```

[55]:
```
              MathScore  ReadingScore  WritingScore
IsFirstChild
no            66.246832     69.132614     68.210887
yes           66.740646     69.542553     68.558484
```

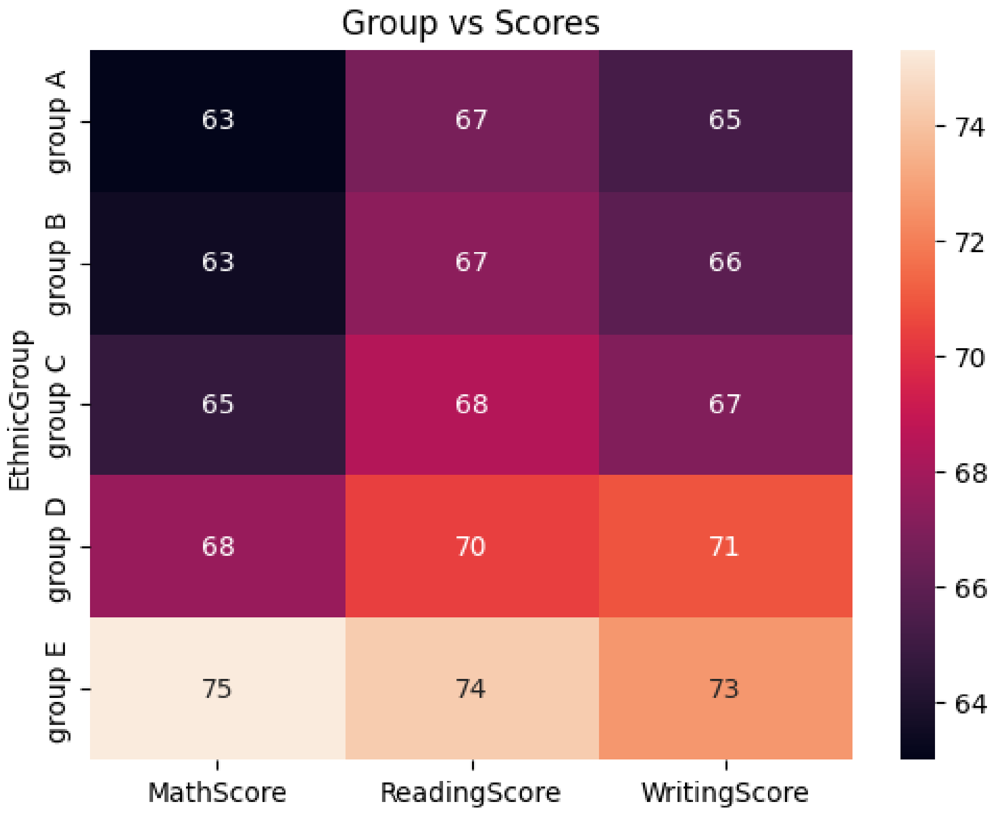
[56]:
```
sns.heatmap(gp6,annot=True)
plt.title("Is First Child vs Scores")
plt.show()
```



```
[58]: gp7 = df.groupby("EthnicGroup").agg({"MathScore":"mean","ReadingScore":
        ↪"mean","WritingScore":"mean"})
      gp7
```

```
[58]:              MathScore  ReadingScore  WritingScore
      EthnicGroup
      group A      62.991888     66.787742     65.251915
      group B      63.490216     67.320460     65.895125
      group C      64.695723     68.438233     66.999240
      group D      67.666400     70.382247     70.890844
      group E      75.298936     74.251423     72.677060
```

```
[60]: sns.heatmap(gp7,annot=True)
      plt.title("Group vs Scores")
      plt.show()
```

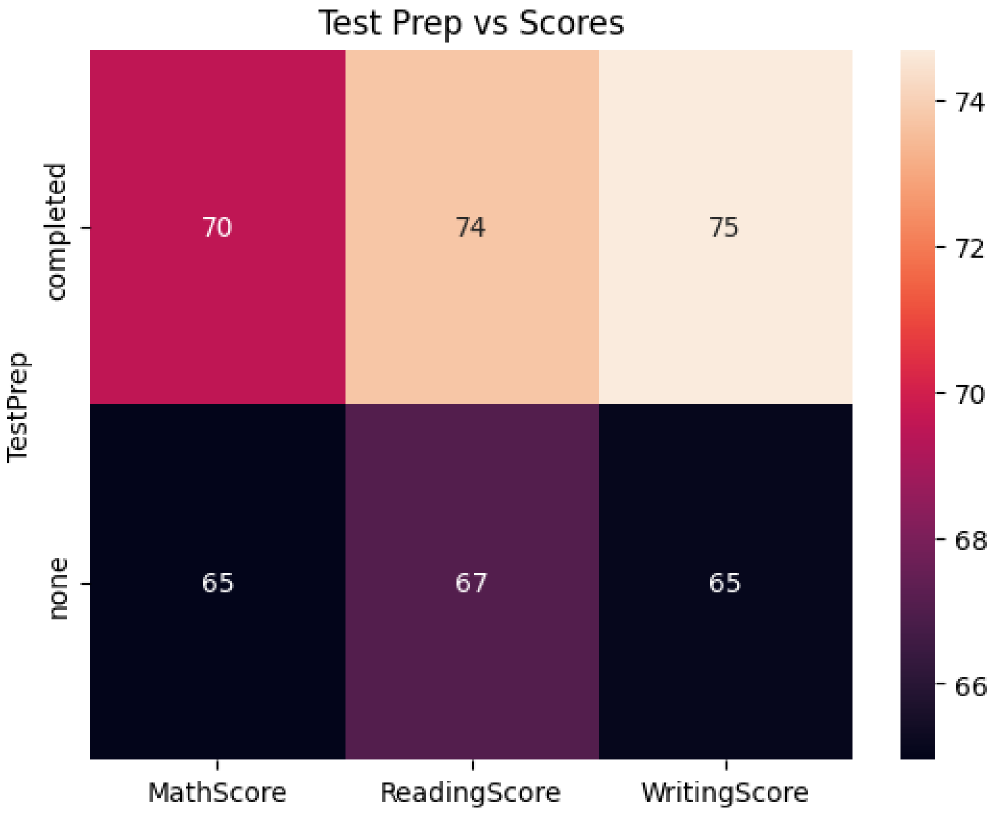


FROM ABOVE CHART WE CAN SEE THAT THE STUDENTS WHO BELONG TO GROUP E HAVE HIGHER SCORES

[64]:
```
gp8 = df.groupby("TestPrep").agg({"MathScore":"mean","ReadingScore":
  ↪"mean","WritingScore":"mean"})
gp8
```

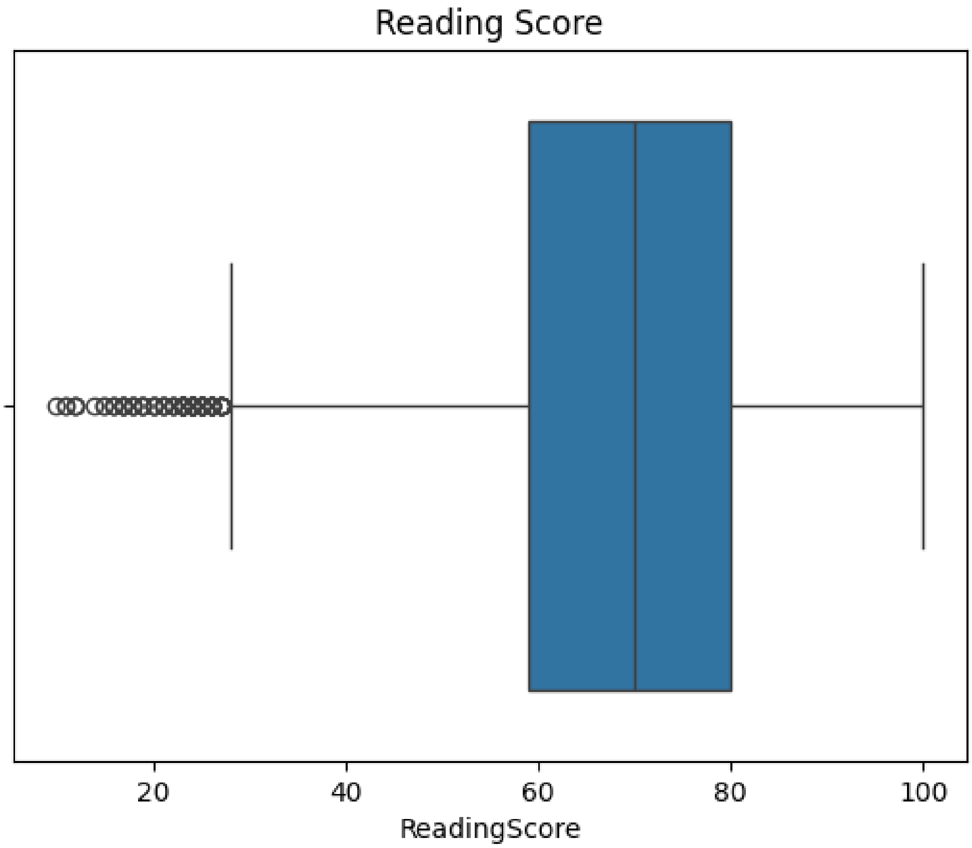
[64]:
```
            MathScore  ReadingScore  WritingScore
TestPrep
completed    69.54666     73.732998     74.703265
none         64.94877     67.051071     65.092756
```

[65]:
```
sns.heatmap(gp8,annot=True)
plt.title("Test Prep vs Scores")
plt.show()
```



FROM ABOVE CHART WE CAN SEE THAT THE STUDENTS WHO HAVE COMPLETED THE TEST PREP COURSE HAVE HIGHER SCORES

[68]:
```
sns.boxplot(data=df,x="ReadingScore")
plt.title("Reading Score")
plt.show()
```

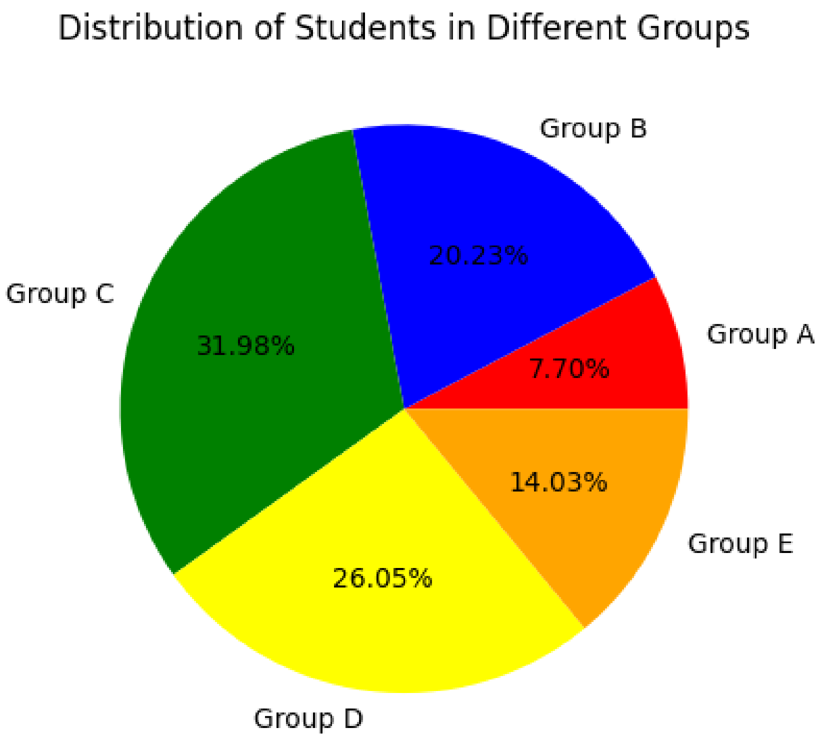


FROM ABOVE CHART WE CAN SEE THAT THE MEDIAN OF THE READING SCORE IS AROUND 70

[81]:
```
goupa = df.loc[(df['EthnicGroup']=='group A')].count()
goupb = df.loc[(df['EthnicGroup']=='group B')].count()
goupc = df.loc[(df['EthnicGroup']=='group C')].count()
goupd = df.loc[(df['EthnicGroup']=='group D')].count()
goupe = df.loc[(df['EthnicGroup']=='group E')].count()

mlist =␣
  ↪[goupa["EthnicGroup"],goupb["EthnicGroup"],goupc["EthnicGroup"],goupd["EthnicGroup"],goupe[
print(mlist)
plt.pie(mlist,labels=["Group A","Group B","Group C","Group D","Group␣
  ↪E"],autopct='%1.2f%%',colors=["red","blue","green","yellow","orange"])
plt.title("Distribution of Students in Different Groups")
```

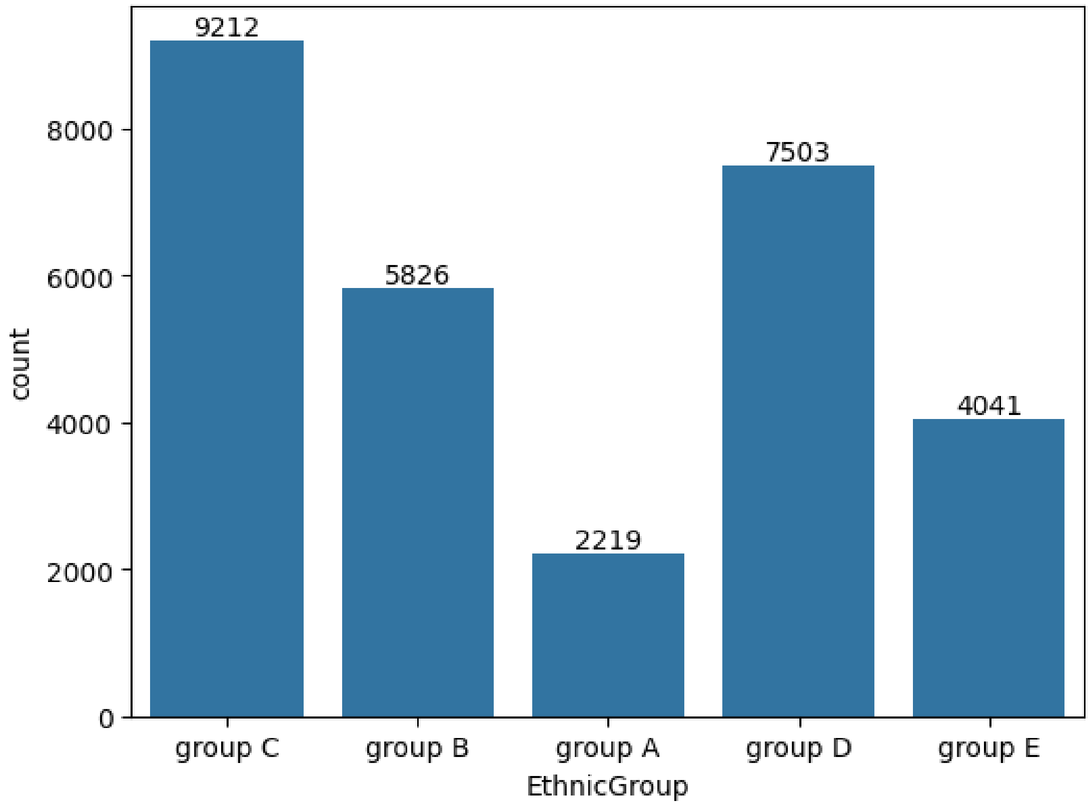
```
[2219, 5826, 9212, 7503, 4041]
```

[81]:
```
Text(0.5, 1.0, 'Distribution of Students in Different Groups')
```



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

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


9212
5826
2219
7503
4041
8000
6000
4000
2000
0
count
group C
group B
group A
group D
group E
EthnicGroup

[ ]: