```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read csv(r"C:\Users\Salman Ramzan\Downloads\archive (3)\
Expanded data with more features.csv")
df.head()
   Unnamed: 0 Gender EthnicGroup
                                             ParentEduc
                                                            LunchType
TestPrep
            0
               female
                               NaN
                                     bachelor's degree
                                                             standard
none
            1 female
                                          some college
                                                             standard
1
                           group C
NaN
            2 female
                                       master's degree
                                                             standard
2
                           group B
none
            3
                 male
                           group A associate's degree free/reduced
none
                 male
                           group C
                                          some college
4
                                                             standard
none
  ParentMaritalStatus PracticeSport IsFirstChild NrSiblings
TransportMeans
              married
                           regularly
                                                           3.0
                                               yes
school_bus
                                                           0.0
1
              married
                           sometimes
                                               yes
NaN
                                                           4.0
               single
                           sometimes
                                               yes
school bus
3
              married
                                                           1.0
                               never
                                                no
NaN
                                                           0.0
              married
                           sometimes
                                               yes
school bus
  WklyStudyHours
                              ReadingScore
                                            WritingScore
                  MathScore
0
                                                       74
             < 5
                          71
                                        71
          5 - 10
                          69
                                        90
                                                       88
1
2
                                        93
                                                       91
             < 5
                          87
3
          5 - 10
                          45
                                        56
                                                       42
4
          5 - 10
                          76
                                        78
                                                       75
df.shape
(30641, 15)
df.isnull().sum()
Unnamed: 0
                           0
Gender
                           0
```

EthnicGroup ParentEduc LunchType TestPrep ParentMaritalSt PracticeSport IsFirstChild NrSiblings TransportMeans WklyStudyHours MathScore ReadingScore WritingScore	tatus	1840 1845 0 1830 1190 631 904 1572 3134 955 0		
dtype: int64				
<pre>df.duplicated()</pre>				
0 False 1 False 2 False 3 False 4 False				
30636 False 30637 False 30638 False 30639 False 30640 False Length: 30641,	dtype:	bool		
<pre>df.describe()</pre>				
Unname WritingScore	ed: 0	NrSiblings	MathScore	ReadingScore
count 30641.00 30641.000000	00000 2	29069.000000	30641.000000	30641.000000
mean 499.55 68.418622	56607	2.145894	66.558402	69.377533
std 288.74	17894	1.458242	15.361616	14.758952
15.443525 min 0.00	00000	0.000000	0.000000	10.000000
4.000000				
25% 249.00 58.000000	00000	1.000000	56.000000	59.000000
50% 500.00	00000	2.000000	67.000000	70.000000
69.000000 75% 750.00	00000	3.000000	78.000000	80.000000
79.000000 max 999.00	00000	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 Dtvpe \_ 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 ParentMaritalStatus 29451 non-null object 6 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 30641 non-null 14 WritingScore int64 dtypes: float64(1), int64(4), object(10) memory usage: 3.5+ MB df.head() Unnamed: 0 Gender EthnicGroup ParentEduc LunchType TestPrep female NaN bachelor's degree standard none 1 1 female group C some college standard NaN 2 female group B master's degree standard 2 none 3 male group A associate's degree free/reduced none male group C some college standard none ParentMaritalStatus PracticeSport IsFirstChild NrSiblings TransportMeans \ married regularly 3.0 ves school bus 1 married sometimes 0.0 yes NaN 4.0 2 single sometimes yes school bus 1.0 3 married never no NaN

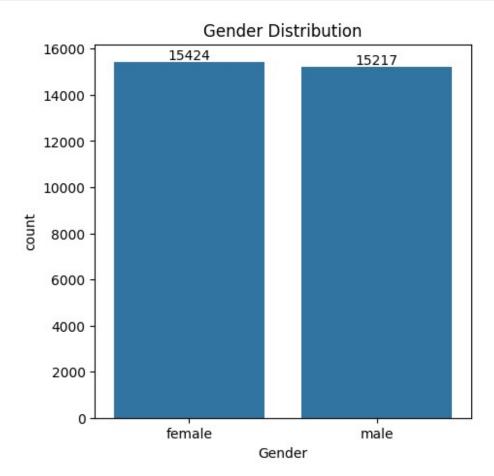
4 s c	mar chool_bus	ried so	metimes	yes	0.0
0 1 2 3 4	WklyStudyHours	MathScore 71 69 87 45 76	ReadingScore 71 90 93 56 78	WritingScore 74 88 91 42 75	

### **Data Cleaning**

```
#df=df.drop("Unnamed: 0",axis=1)
df.head()
   Gender EthnicGroup
                                ParentEduc
                                                LunchType TestPrep \
                                                 standard
  female
                   NaN
                         bachelor's degree
                                                               none
1
  female
              group C
                              some college
                                                 standard
                                                                NaN
  female
              group B
                           master's degree
                                                 standard
                                                               none
3
              group A associate's degree
                                             free/reduced
     male
                                                               none
4
     male
                              some college
                                                 standard
              group C
                                                               none
  ParentMaritalStatus PracticeSport IsFirstChild NrSiblings
TransportMeans
                           regularly
              married
                                                            3.0
                                               yes
school bus
                                                            0.0
              married
                           sometimes
                                               yes
NaN
                single
                           sometimes
                                                            4.0
                                               yes
school bus
              married
                               never
                                                no
                                                            1.0
NaN
                                                            0.0
              married
                           sometimes
                                               yes
school_bus
                                             WritingScore
  WklyStudyHours
                  MathScore
                              ReadingScore
0
                                                        74
             < 5
                          71
                                         71
          5 - 10
                          69
                                         90
                                                        88
1
2
             < 5
                          87
                                         93
                                                        91
3
          5 - 10
                          45
                                         56
                                                        42
4
          5 - 10
                          76
                                                        75
# For replacing values
# df["WklyStudyHours"]=df["WklyStudyHours"].str.replace("5 - 10,
89"),, waise hi
```

#### **EDA**

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



From the above analysis we analysed that the number of female in the data is more than male

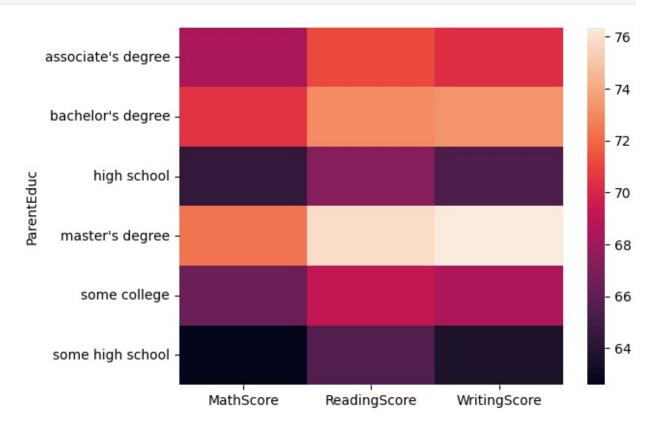
#### Score on the base of Parent's Education

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

sns.heatmap(gb)

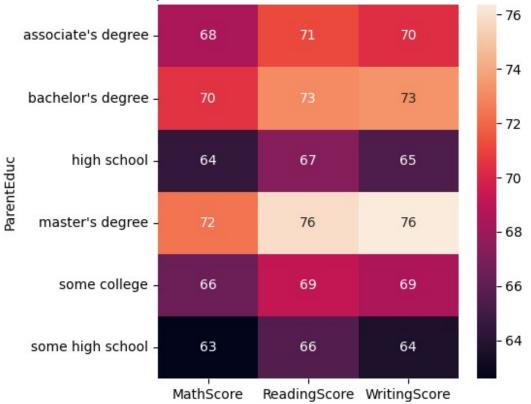
<Axes: ylabel='ParentEduc'>



```
#for showing values in heatmap
plt.figure(figsize=(5,5))
sns.heatmap(gb, annot = True)
plt.title("Relationship between Parent Education and Student's Score")

Text(0.5, 1.0, "Relationship between Parent Education and Student's Score")
```





From the above chart we analysed the education of parent has a good impact on the student's score

#### Score on the base of Parent Marital Status

```
qb1=
df.groupby("ParentMaritalStatus").agg({"MathScore": "mean", "ReadingScor
e":"mean","WritingScore":"mean"})
print(qb1)
                     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=(5,5))
sns.heatmap(gb1, annot=True)
plt.title("Relationship between Parent Marital Status and Student's
Score")
plt.show()
```

## Relationship between Parent Marital Status and Student's Score



From above chart we can analysis that there is no/negligible impact of parent marital status on Student's score

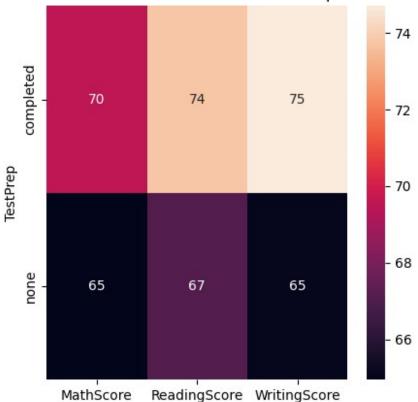
1								
df	.head()							
0	female		bachelor's d	_	sta	ndard		\
1 2 3 4	female female male	group B group A	associate's d	degree degree	sta free/re		none none	
			some co	_				
	ansport!		Fracticesport	121112	CHILL	MI STOC	riigs	
0		married	regularly		yes		3.0	
SC	hool_bus	s married	sometimes		1/05		0 0	
Na	N	marrieu	Sometimes		yes		0.0	
2		single	sometimes		yes		4.0	
sc 3	hool_bus	s married	novor		no		1.0	
Na	N	marrieu	never		no		1.0	
4		married	sometimes		yes		0.0	

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

### Score on the base of Test Preparation

```
gb2 =
df.groupby("TestPrep").agg({"MathScore":"mean","ReadingScore":"mean","
WritingScore":"mean"})
print(gb2)
           MathScore ReadingScore WritingScore
TestPrep
completed
            69.54666
                         73.732998
                                       74.703265
none
            64.94877
                         67.051071
                                       65.092756
plt.figure(figsize=(5,5))
sns.heatmap(gb2, annot=True)
plt.title("Student's Score on the base of Test Prepartion")
plt.show()
```

#### Student's Score on the base of Test Prepartion

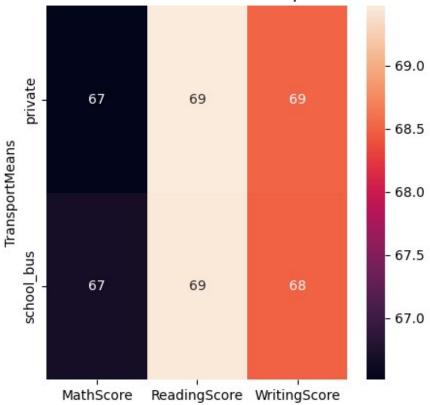


From above analysis we can see there is large impact of "Test Preparation" on the student's score

# Score on the base of Transport Means

```
ab3=
df.groupby("TransportMeans").agg({"MathScore":"mean","ReadingScore":"m
ean","WritingScore":"mean"})
print(qb3)
                MathScore ReadingScore WritingScore
TransportMeans
private
                66.511354
                              69.472364
                                            68.509593
school bus
                66.674636
                              69.446206
                                            68.492351
plt.figure(figsize=(5,5))
sns.heatmap(gb3, annot=True)
plt.title("Student Score on the base of Transport Means")
plt.show()
```

# Student Score on the base of Transport Means



From above chart we can analysis that there is no/negligible impact of "Transport Mean" status on Student's score

df	.head()							
0 1 2 3 4	Gender E female female male male	EthnicGroup NaN group C group B group A group C	bachelor's d some co	llege egree egree	sta sta sta free/re	ndard Indard Indard	TestPrep none NaN none none	\
	ParentMaı ansportMe		PracticeSport	IsFirst	tChild	NrSibl	ings	
0 sc	hool_bus	married	regularly		yes		3.0	
1 Na	_	married	sometimes		yes		0.0	
2	hool bus	single	sometimes		yes		4.0	
3 Na	_	married	never		no		1.0	
1N a	IV	married	sometimes		yes		0.0	

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

#### Score on the base on No. of Siblings

```
ab4 =
df.groupby("NrSiblings").agg({"MathScore":"mean","ReadingScore":"mean"
, "WritingScore": "mean"})
print(gb4)
            MathScore
                       ReadingScore WritingScore
NrSiblings
            66.819449
                           69.547812
                                         68.746515
0.0
1.0
            66.473896
                           69.259097
                                         68.245345
2.0
            66.554934
                           69.472018
                                         68.522533
3.0
            66.719092
                           69.488159
                                         68,650498
            66.245495
4.0
                           69.144169
                                         68.073444
5.0
                                         68.282576
            66.630303
                           69.453788
6.0
            65.917219
                           68.801325
                                         67.860927
7.0
            67.615120
                          69.828179
                                         68.986254
plt.figure(figsize=(5,5))
sns.heatmap(qb4, annot=True)
plt.title("Student Score on the base of No.of Siblings")
plt.show()
```

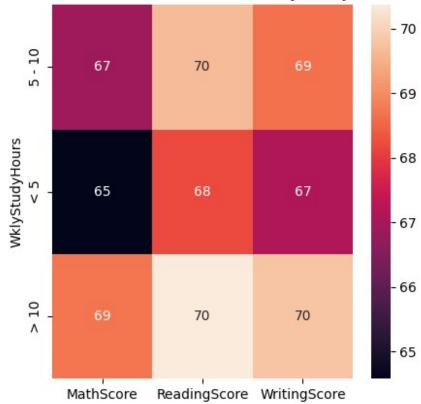
#### Student Score on the base of No.of Siblings



From above chart we can analysis that there is no/negligible impact of "No.of Siblings" on Student's score

```
qb5 =
df.groupby("WklyStudyHours").agg({"MathScore":"mean","ReadingScore":"m
ean","WritingScore":"mean"})
print(gb5)
                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
plt.figure(figsize=(5,5))
sns.heatmap(gb5, annot=True)
plt.title("Student Score on the base of Weekly Study Hours")
plt.show()
```

## Student Score on the base of Weekly Study Hours

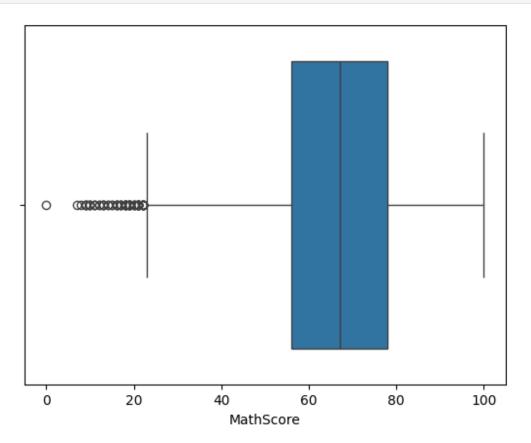


From above chart we can analysis that there is no/negligible impact of "Weekly Study Hours" on Student's score

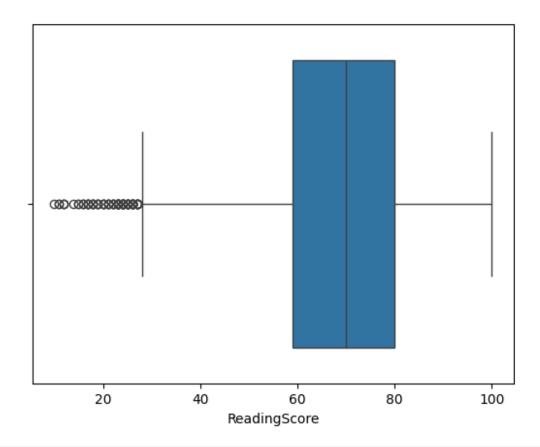
# **Checking Outliers**

Circulary							
<pre>df.head()</pre>							
Gender female female female male male	<b>5</b> 1	Paren bachelor's d some co master's d associate's d some co	egree llege egree egree	st st st free/r	andard andard andard	TestPrep none NaN none none	\
ParentMa TransportM		PracticeSport	IsFirs <sup>.</sup>	tChild	NrSibl	lings	
0	married	regularly		yes		3.0	
school_bus						0 0	
1 NaN	married	sometimes		yes		0.0	
2	single	sometimes		yes		4.0	
school_bus				-		1.0	
3	married	never		no		1.0	

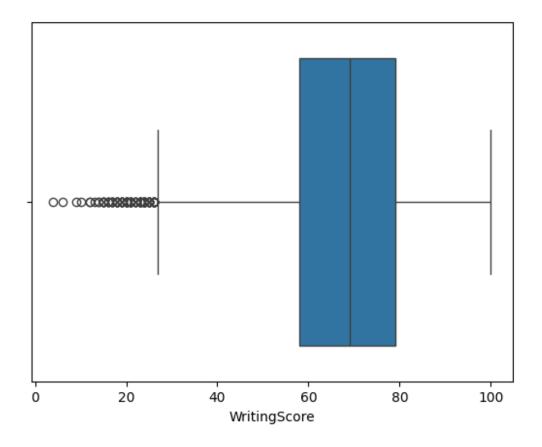
```
NaN
                           sometimes
                                                            0.0
              married
                                               yes
4
school_bus
  WklyStudyHours
                  MathScore
                              ReadingScore
                                             WritingScore
             < 5
                          71
                                         71
                                                       74
                                         90
1
2
                          69
                                                       88
          5 - 10
                                         93
                          87
                                                       91
             < 5
3
          5 - 10
                          45
                                         56
                                                       42
          5 - 10
4
                                                       75
                          76
                                         78
sns.boxplot(x="MathScore", data=df)
plt.show()
```



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

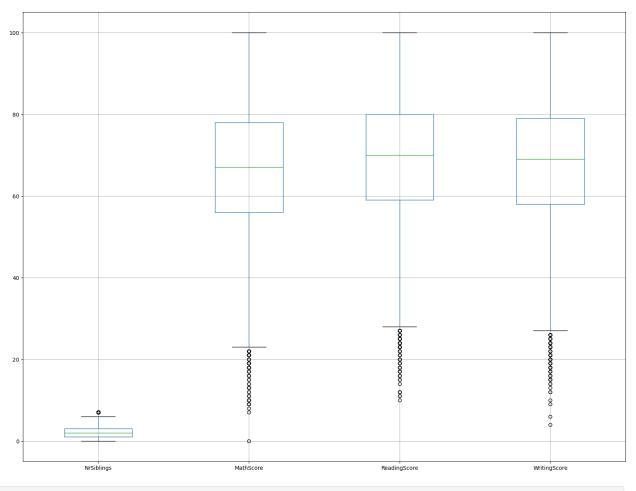


sns.boxplot(x="WritingScore",data=df)
plt.show()



From above graphs we analysed that students are comparitively weak in "Math" instead of reading and writing score

```
df.boxplot(figsize=(20,15))
plt.show()
```



df	.head()							
0 1 2 3 4	Gender female female female male	group B	bachelor's c	ollege degree degree	st st st free/r	andard andard andard	none	\
	ParentMa ansportM		PracticeSport	IsFirs	tChild	NrSibl	ings	
0	hool bus	married	regularly		yes		3.0	
1 Na	_	married	sometimes		yes		0.0	
2	hool bus	single	sometimes		yes		4.0	
3 Na	_	married	never		no		1.0	
4	hool_bus	married	sometimes		yes		0.0	

```
WklyStudyHours MathScore
                                ReadingScore
                                               WritingScore
0
              < 5
                           71
                                                           74
                                           71
1
           5 - 10
                           69
                                           90
                                                          88
2
              < 5
                           87
                                           93
                                                          91
3
           5 - 10
                           45
                                           56
                                                           42
           5 - 10
                                                           75
4
                           76
                                           78
df["EthnicGroup"].unique()
array([nan, 'group C', 'group B', 'group A', 'group D', 'group E'],
      dtype=object)
GroupA = df.loc[(df["EthnicGroup"]== "group A")].count()
GroupB = df.loc[(df["EthnicGroup"]== "group B")].count()
GroupC = df.loc[(df["EthnicGroup"]== "group C")].count()
GroupD = df.loc[(df["EthnicGroup"]== "group D")].count()
GroupE = df.loc[(df["EthnicGroup"]== "group E")].count()
l = ["GroupA", "GroupB", "GroupC", "GroupD", "GroupE"]
mlist
=(GroupA["EthnicGroup"],GroupB["EthnicGroup"],GroupC["EthnicGroup"],Gr
oupD["EthnicGroup"],GroupE["EthnicGroup"])
plt.pie(mlist, labels = l,autopct="%1.2f%%")
plt.title("distribution of Ethnic Group")
plt.show()
```

#### distribution of Ethnic Group

