In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [18]:
```
df = pd.read_csv("C:/Users/PRATIM PAL/Downloads/Expanded_data_with_more_features.csv.zip")
print(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   freereduced     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            NaN         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
```

In [19]:
```
df.isnull().sum()
```

Out[19]:
```
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 unnamed column

In [21]:
```
df.describe
```

Out[21]:
```
<bound method NDFrame.describe of        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   freereduced
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   freereduced
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         1.0
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]>
```

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

change weekly study hours column

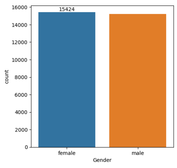
In [22]:
```
df["WklyStudyHours"] = df["WklyStudyHours"].str.replace("05-Oct","5-10")
df.head()
```

Out[22]:

	Unnamed: 0	Gender	EthnicGroup	ParentEduc	LunchType	TestPrep	ParentMaritalStatus	PracticeSport	IsFirstChild	NrSiblings	TransportMeans	WklyStudyHours	MathScore	ReadingScore	WritingScore
0	0	female	NaN	bachelor's degree	standard	none	married	regularly	yes	3.0	school_bus	< 5	71	71	74
1	1	female	group C	some college	standard	NaN	married	sometimes	yes	0.0	NaN	5 - 10	69	90	88
2	2	female	group B	master's degree	standard	none	single	sometimes	yes	4.0	school_bus	< 5	87	93	91
3	3	male	group A	associate's degree	freereduced	none	married	never	no	1.0	NaN	5 - 10	45	56	42
4	4	male	group C	some college	standard	none	married	sometimes	yes	0.0	school_bus	5 - 10	76	78	75

Gender distribution

In [38]:
```
plt.figure(figsize =(5,5))
ax = sns.countplot(data = df , x = "Gender", hue ="Gender")
ax.bar_label(ax.containers[0])
plt.show()
```

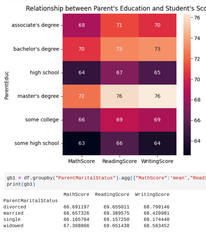


In [38]:
```
# From this above chart we have analysed that:-
# the number of female in the data is more than the number of males.
```

In [39]:
```
gb = df.groupby("ParentEduc").agg({"MathScore":'mean',"ReadingScore":"mean","WritingScore":"mean"})
print(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
```

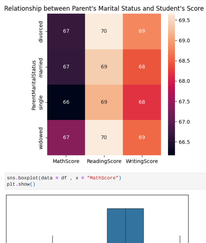
In [50]:
```
plt.figure(figsize =(5,5))
sns.heatmap(gb, annot = True)
plt.title("Relationship between Parent's Education and Student's Score")
plt.show()
```



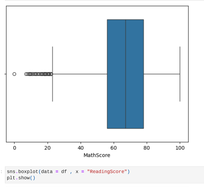
In [51]:
```
gbs = df.groupby("ParentMaritalStatus").agg({"MathScore":'mean',"ReadingScore":"mean","WritingScore":"mean"})
print(gbs)
```

```
                   MathScore  ReadingScore  WritingScore
ParentMaritalStatus
divorced           66.691197     69.655011     68.799146
married            66.657326     69.389575     68.420981
single             66.165704     69.157258     68.174440
widowed            67.368866     69.451438     68.503452
```

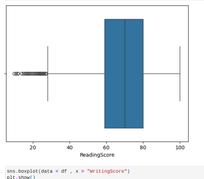
In [54]:
```
plt.figure(figsize =(5,5))
sns.heatmap(gbs, annot = True)
plt.title("Relationship between Parent's Marital Status and Student's Score")
plt.show()
```



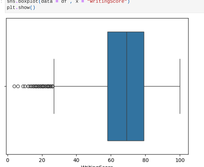
In [56]:
```
sns.boxplot(data = df , x = "MathScore")
plt.show()
```



In [58]:
```
sns.boxplot(data = df , x = "ReadingScore")
plt.show()
```



In [57]:
```
sns.boxplot(data = df , x = "WritingScore")
plt.show()
```



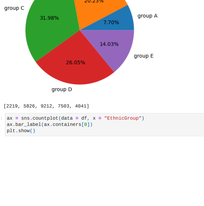
In [59]:
```
print(df["EthnicGroup"].unique())
```

```
[nan 'group C' 'group B' 'group A' 'group D' 'group E']
```

Distribution of Ethnic Groups

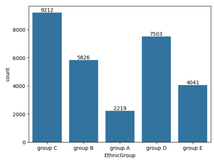
In [60]:
```
groupA = df.loc[(df['EthnicGroup'] == "group A")].count()
groupB = df.loc[(df['EthnicGroup'] == "group B")].count()
groupC = df.loc[(df['EthnicGroup'] == "group C")].count()
groupD = df.loc[(df['EthnicGroup'] == "group D")].count()
groupE = df.loc[(df['EthnicGroup'] == "group E")].count()

l = ["group A","group B","group C","group D","group E"]
mlist = [groupA["EthnicGroup"],groupB["EthnicGroup"],groupC["EthnicGroup"],groupD["EthnicGroup"],groupE["EthnicGroup"]]
plt.pie(mlist, labels = l, autopct= "%1.2f%%")
plt.title("Distribution of Ethnic Groups")
plt.show()
print(mlist)
#print(groupA)
```




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

In [60]:
```
ax = sns.countplot(data = df, x = "EthnicGroup")
ax.bar_label(ax.containers[0])
plt.show()
```


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