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

df = pd.read_csv("Expanded_data_with_more_features.csv")
df.head()
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 30641,\n  \"fields\":
[\n    {\n      \"column\": \"Unnamed: 0\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 288,\n        \"min\": 0,\n
\"max\": 999,\n        \"num_unique_values\": 1000,\n
\"samples\": [\n          549,\n          773,\n          776\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Gender\",\n      \"properties\":
{\n        \"dtype\": \"category\",\n        \"num_unique_values\":
2,\n        \"samples\": [\n          \"male\",\n          \"female\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"EthnicGroup\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 5,\n        \"samples\":
[\n          \"group B\",\n          \"group E\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"ParentEduc\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 6,\n        \"samples\": [\n
\"bachelor's degree\",\n          \"some college\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"LunchType\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n
\"free/reduced\",\n          \"standard\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"TestPrep\",\n      \"properties\":
{\n        \"dtype\": \"category\",\n        \"num_unique_values\":
2,\n        \"samples\": [\n          \"completed\",\n
\"none\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"ParentMaritalStatus\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 4,\n        \"samples\":
[\n          \"single\",\n          \"divorced\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"PracticeSport\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n
\"regularly\",\n          \"sometimes\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"IsFirstChild\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n          \"no\",\n
```

\"yes\"\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"NrSiblings\",\n   \"properties\": {\n   \"dtype\":
\"number\",\n   \"std\": 1.4582424759684511,\n   \"min\":
0.0,\n   \"max\": 7.0,\n   \"num_unique_values\": 8,\n
\"samples\": [\n   0.0,\n   5.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"TransportMeans\",\n
\"properties\": {\n   \"dtype\": \"category\",\n
\"num_unique_values\": 2,\n   \"samples\": [\n
\"private\",\n   \"school_bus\"\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"WklyStudyHours\",\n
\"properties\": {\n   \"dtype\": \"category\",\n
\"num_unique_values\": 3,\n   \"samples\": [\n   \"< 5\",\
n   \"5 - 10\"\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"MathScore\",\n   \"properties\": {\n   \"dtype\":
\"number\",\n   \"std\": 15,\n   \"min\": 0,\n
\"max\": 100,\n   \"num_unique_values\": 95,\n
\"samples\": [\n   36,\n   70\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"ReadingScore\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
14,\n   \"min\": 10,\n   \"max\": 100,\n
\"num_unique_values\": 90,\n   \"samples\": [\n   48,\n
65\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"WritingScore\",\n   \"properties\": {\n   \"dtype\":
\"number\",\n   \"std\": 15,\n   \"min\": 4,\n
\"max\": 100,\n   \"num_unique_values\": 93,\n
\"samples\": [\n   10,\n   76\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   }\n]\n}","type":"dataframe","variable_name":"df"}

```
df.describe()
```

{"summary":"{\n \"name\": \"df\",\n \"rows\": 8,\n \"fields\": [\n
{\n   \"column\": \"Unnamed: 0\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 10671.681928672426,\n
\"min\": 0.0,\n   \"max\": 30641.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
499.5566071603407,\n   500.0,\n   30641.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"NrSiblings\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
10276.60508653049,\n   \"min\": 0.0,\n   \"max\": 29069.0,\n
\"num_unique_values\": 8,\n   \"samples\": [\n
2.1458942516082424,\n   2.0,\n   29069.0\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\

```
n    },\n    {\n      \"column\": \"MathScore\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
10813.938124618964,\n        \"min\": 0.0,\n        \"max\": 30641.0,\
n        \"num_unique_values\": 8,\n        \"samples\": [\n
66.5584021409223,\n          67.0,\n          30641.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"ReadingScore\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
10812.912200605591,\n        \"min\": 10.0,\n        \"max\":
30641.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n
69.37753337032082,\n          70.0,\n          30641.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"WritingScore\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
10813.383566214232,\n        \"min\": 4.0,\n        \"max\": 30641.0,\
n        \"num_unique_values\": 8,\n        \"samples\": [\n
68.41862210763357,\n          69.0,\n          30641.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe"}
```


```
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
```

Checking for total number of null values in each column

```
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 Unnamed Column

```
df = df.drop("Unnamed: 0", axis = 1)
df.head()
```


```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 30641,\n  \"fields\":
[\n    {\n      \"column\": \"Gender\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 2,\n
\"samples\": [\n          \"male\",\n          \"female\"\n        ],\
n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"EthnicGroup\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n          \"group
B\",\n          \"group E\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"ParentEduc\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 6,\n
\"samples\": [\n          \"bachelor's degree\",\n          \"some
college\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"LunchType\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 2,\n        \"samples\":
[\n          \"free/reduced\",\n          \"standard\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"TestPrep\",\n      \"properties\":
{\n        \"dtype\": \"category\",\n        \"num_unique_values\":
2,\n        \"samples\": [\n          \"completed\",\n
\"none\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"ParentMaritalStatus\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 4,\n        \"samples\":
[\n          \"single\",\n          \"divorced\"\n        ],\n
```

```
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"PracticeSport\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n
\"regularly\",\n          \"sometimes\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"IsFirstChild\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n          \"no\",\n
\"yes\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"NrSiblings\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 1.4582424759684511,\n        \"min\":
0.0,\n        \"max\": 7.0,\n        \"num_unique_values\": 8,\n
\"samples\": [\n          0.0,\n          5.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"TransportMeans\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n
\"private\",\n          \"school_bus\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"WklyStudyHours\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n          \"< 5\",\
n          \"5 - 10\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"MathScore\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 15,\n        \"min\": 0,\n
\"max\": 100,\n        \"num_unique_values\": 95,\n
\"samples\": [\n          36,\n          70\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"ReadingScore\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
14,\n        \"min\": 10,\n        \"max\": 100,\n
\"num_unique_values\": 90,\n        \"samples\": [\n          48,\n
65\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"WritingScore\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 15,\n        \"min\": 4,\n
\"max\": 100,\n        \"num_unique_values\": 93,\n
\"samples\": [\n          10,\n          76\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```


Gender distribution

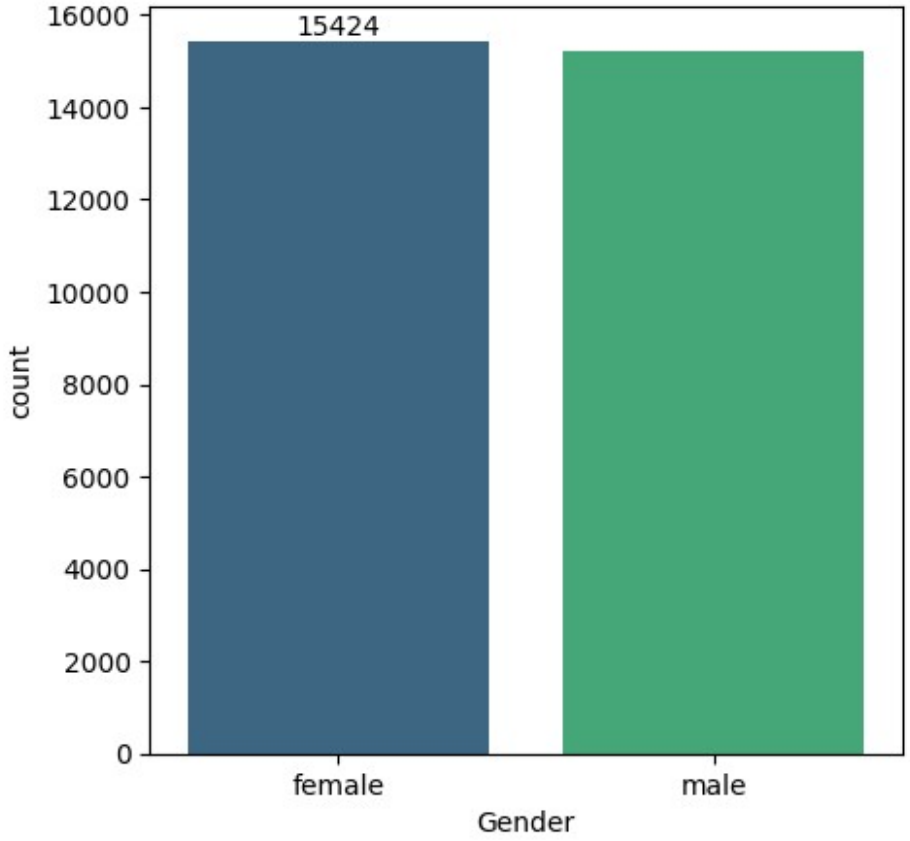
```
plt.figure(figsize = (5,5))
ax = sns.countplot(data = df, x = "Gender", palette="viridis")
```

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

```
<ipython-input-31-dd9f18cea7bf>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  ax = sns.countplot(data = df, x = "Gender", palette="viridis")
```



Conclusion: From above chart we analyzed that: Number of females are more then number of males.

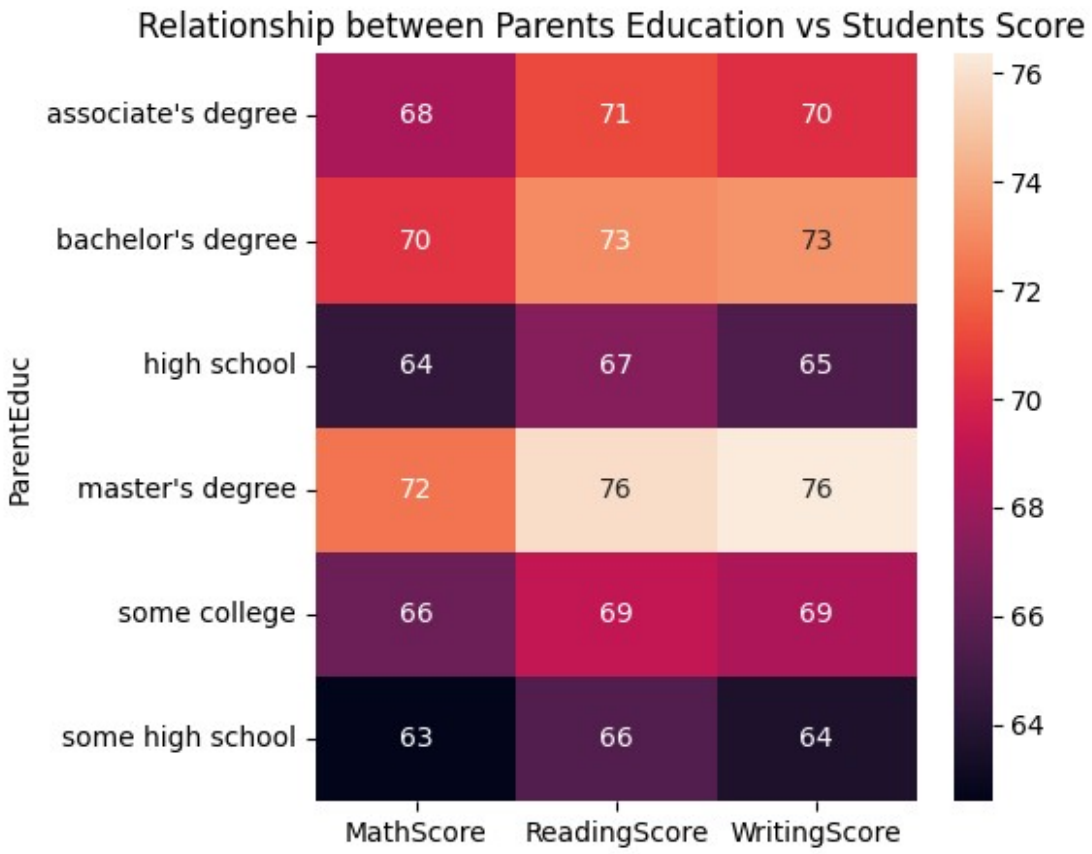
Parent Education role in Students Marks

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

```
{"summary":"{\n  \"name\": \"gb\",\n  \"rows\": 6,\n  \"fields\": [\n
{\n      \"column\": \"ParentEduc\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 6,\n
\"samples\": [\n          \"associate's degree\",\n
```

\"bachelor's degree\",\n   \"some high school\"\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"MathScore\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
3.6795770950348223,\n   \"min\": 62.58401305057096,\n
\"max\": 72.33613445378151,\n   \"num_unique_values\": 6,\n
\"samples\": [\n   68.36558558558559,\n
70.46662728883639,\n   62.58401305057096\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"ReadingScore\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
3.8114035417911296,\n   \"min\": 65.51078484683705,\n
\"max\": 75.83292140385566,\n   \"num_unique_values\": 6,\n
\"samples\": [\n   71.12432432432432,\n
73.06202008269344,\n   65.51078484683705\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"WritingScore\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
4.782187685280533,\n   \"min\": 63.63240891789016,\n
\"max\": 76.35689569945626,\n   \"num_unique_values\": 6,\n
\"samples\": [\n   70.2990990990991,\n
73.33106910809214,\n   63.63240891789016\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   }\n]\n}","type":"dataframe","variable_name":"gb"}

```
plt.figure(figsize = (5,5))
sns.heatmap(data = gb, annot = True)
plt.title("Relationship between Parents Education vs Students Score ")
plt.show()
```



Conclusion: In above char we concluded that: Education of parents have good impact on students score.

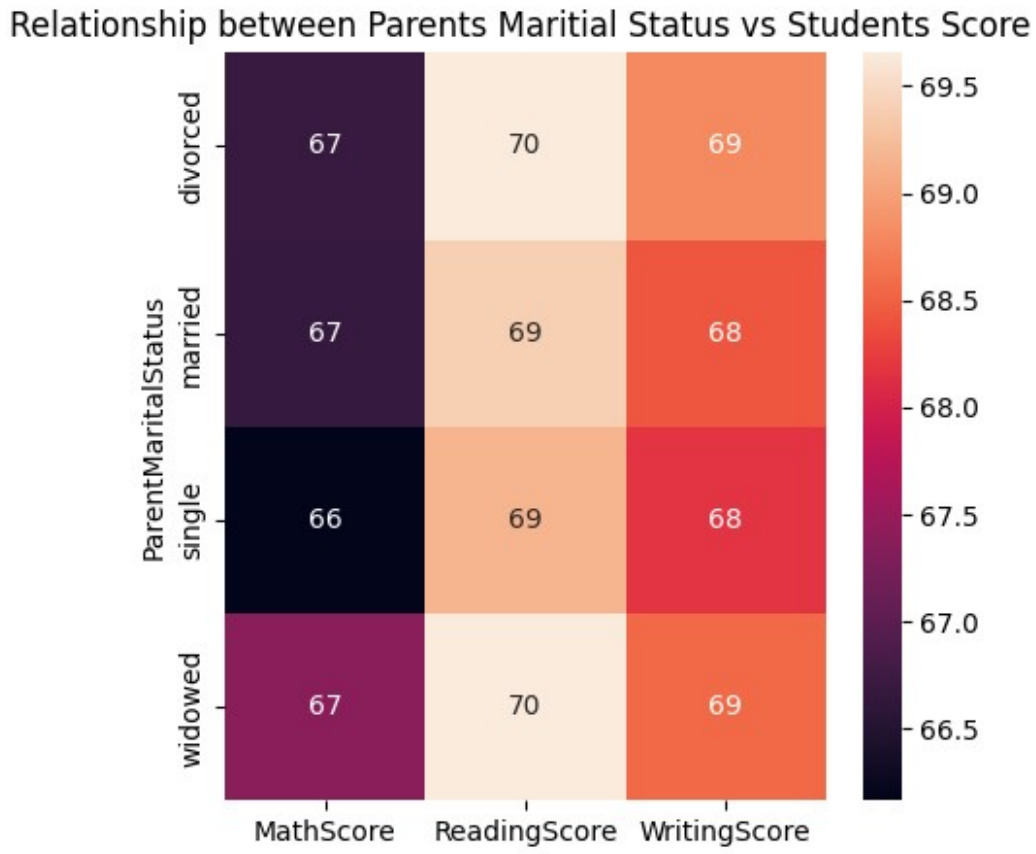
Role Parents Maritial Status in Student education

```
gb1 = df.groupby("ParentMaritalStatus").agg({"MathScore":'mean',
"ReadingScore":'mean', "WritingScore":'mean'})
gb1
```

```
{"summary":"{\n  \"name\": \"gb1\",\n  \"rows\": 4,\n  \"fields\": [\n
{\n      \"column\": \"ParentMaritalStatus\",\n      \"properties\":
{\n        \"dtype\": \"string\",\n        \"num_unique_values\": 4,\n
\"samples\": [\n          \"married\",\n          \"widowed\",\n
\"divorced\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"MathScore\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0.4943099533587517,\n        \"min\":
66.16570381851487,\n        \"max\": 67.3688663282572,\n
\"num_unique_values\": 4,\n        \"samples\": [\n
66.65732605081928,\n          67.3688663282572,\n
66.69119739784509\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"ReadingScore\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0.2389221929621977,\n        \"min\":
```

```
69.15724954206003,\n        \"max\": 69.65501118113438,\n
\"num_unique_values\": 4,\n        \"samples\": [\n
69.38957492282118,\n          69.65143824027072,\n
69.65501118113438\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"WritingScore\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0.2616023471332318,\n        \"min\":
68.17443990418487,\n        \"max\": 68.79914616792031,\n
\"num_unique_values\": 4,\n        \"samples\": [\n
68.42098076466398,\n          68.56345177664974,\n
68.79914616792031\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"gb1"}
```

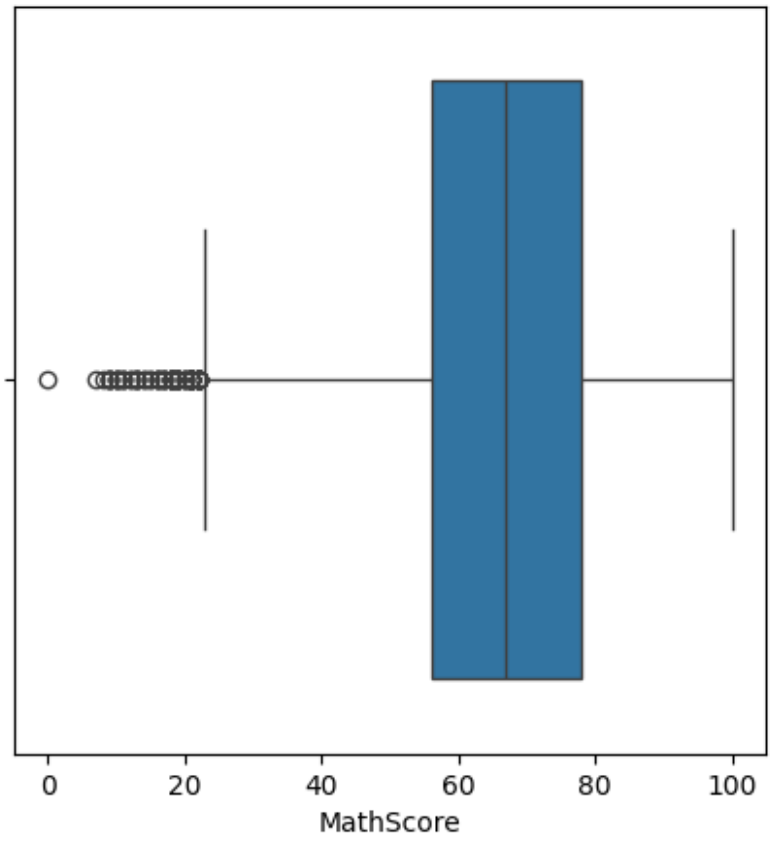

```
plt.figure(figsize = (5,5))
sns.heatmap(data = gb1, annot = True)
plt.title("Relationship between Parents Maritial Status vs Students
Score ")
plt.show()
```



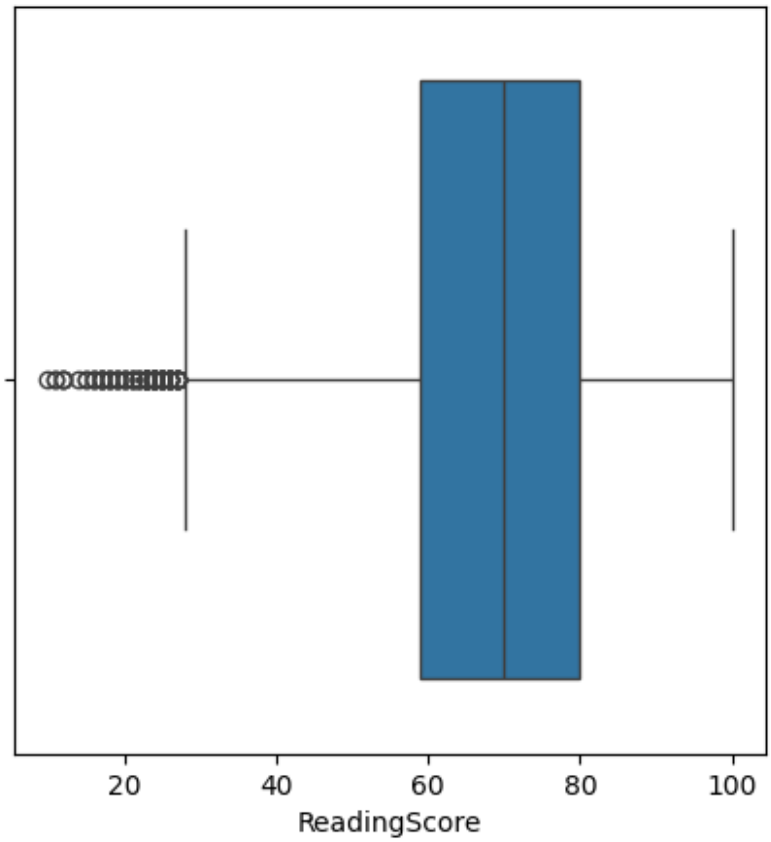
Conclusion: From above heatmap we can conclude that: there is no impact of student score based on their parents maritial status.

To Findout Outliers

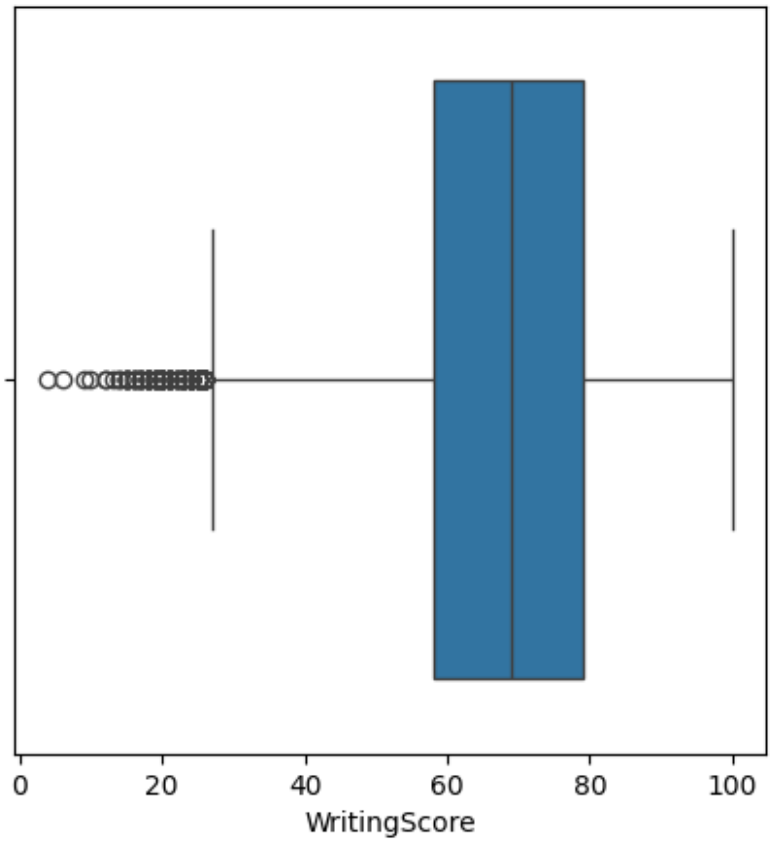
```
plt.figure(figsize = (5,5))
sns.boxplot(data = df, x = "MathScore")
plt.show()
```




```
plt.figure(figsize = (5,5))
sns.boxplot(data = df, x = "ReadingScore")
plt.show()
```




```
plt.figure(figsize = (5,5))
sns.boxplot(data = df, x = "WritingScore")
plt.show()
```




```
print(df["EthnicGroup"].unique())
```

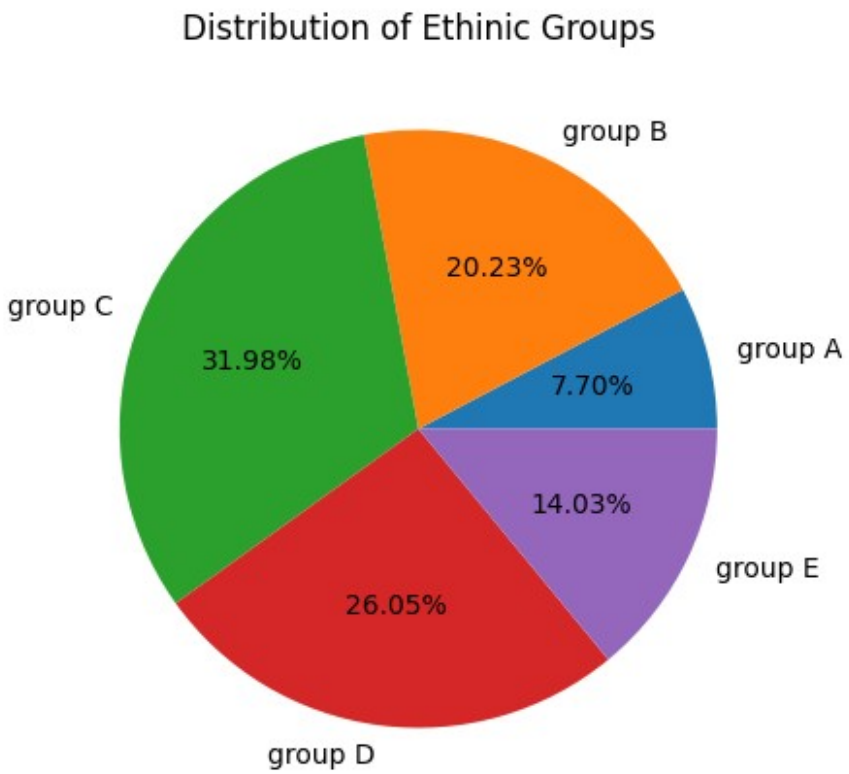
```
[nan 'group C' 'group B' 'group A' 'group D' 'group E']
```

Distribution of Ethinic Groups

```
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()

labe = ["group A", "group B","group C","group D","group E"]
mlist = [groupA["EthnicGroup"], groupB["EthnicGroup"],
groupC["EthnicGroup"], groupD["EthnicGroup"], groupE["EthnicGroup"]]

plt.figure(figsize = (5,5))
plt.pie(mlist, labels = labe, autopct = "%1.2f%%")
plt.title("Distribution of Ethinic Groups")
plt.show()
```



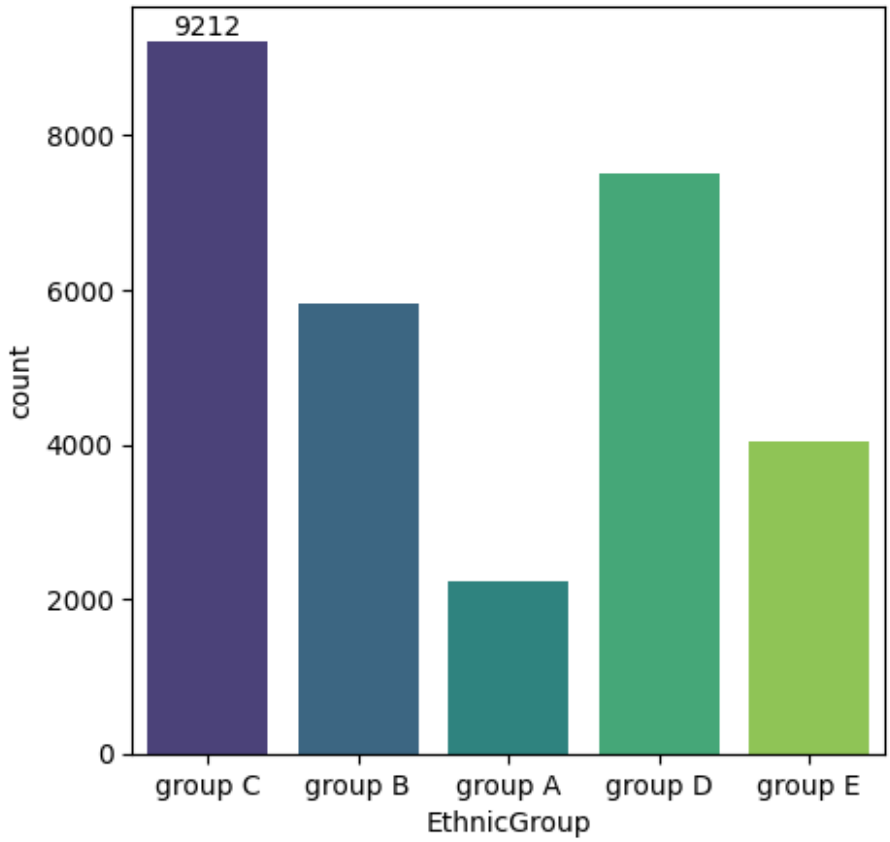
Count of Ethnic Group

```
plt.figure(figsize=(5,5))
ax = sns.countplot(data = df, x = "EthnicGroup", palette="viridis")
ax.bar_label(ax.containers[0])
plt.show()
```

```
<ipython-input-21-d658afd65d22>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  ax = sns.countplot(data = df, x = "EthnicGroup", palette="viridis")
```



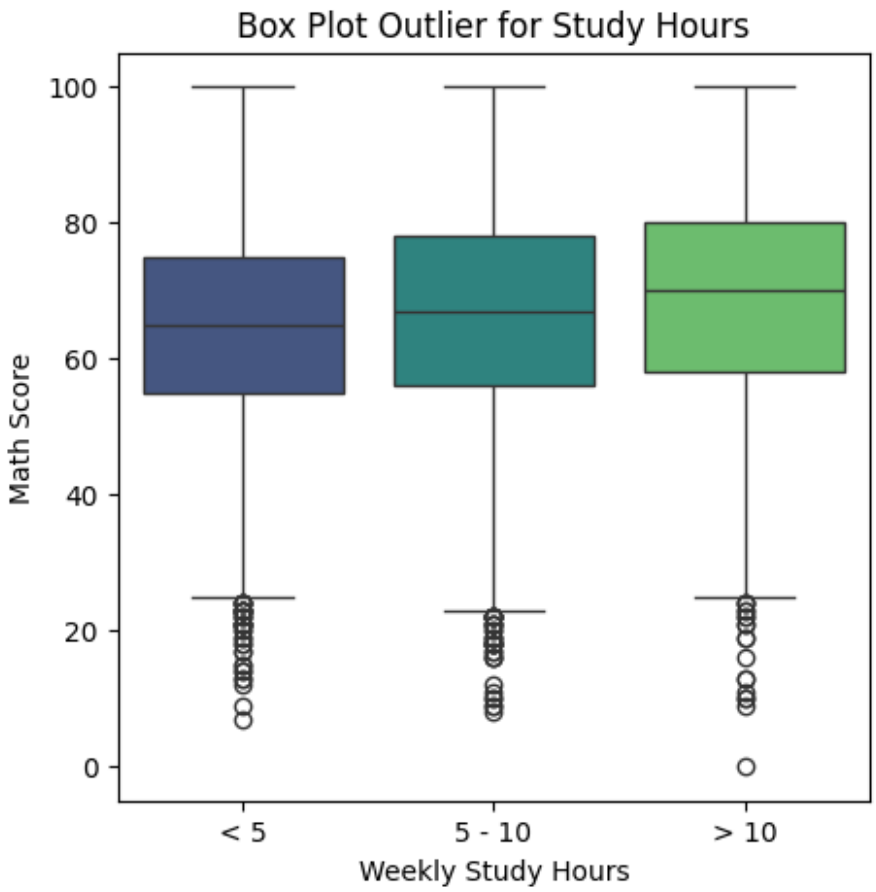
Study hours vs Maths Score

```
plt.figure(figsize=(5,5))
sns.boxplot(data=df,x='WklyStudyHours',y='MathScore',
palette="viridis")
plt.xlabel('Weekly Study Hours')
plt.ylabel('Math Score')
plt.title('Box Plot Outlier for Study Hours')
plt.show()
```

```
<ipython-input-32-fc412e76c70b>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.boxplot(data=df,x='WklyStudyHours',y='MathScore',
palette="viridis")
```



Conclusion: From above box plot we came to know that:Student who studied for more then 10hr per week got 0 marks in maths