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

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

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  free/reduced
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            < 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
```

```
df.tail(10)
```

```
       Unnamed: 0  Gender EthnicGroup          ParentEduc
LunchType  \
30631         765    male     group E    some high school
standard
30632         778  female     group D        some college
standard
30633         783  female     group C     master's degree
standard
```

```
30634         785    male     group A  associate's degree
free/reduced
30635         794    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  \
30631       none              married     sometimes           no
3.0
30632       none              married     regularly           no
3.0
30633  completed              married         never           no
2.0
30634  completed                  NaN     sometimes           no
2.0
30635       none              married     regularly           no
2.0
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
30631     school_bus            < 5         80            65
66
30632        private         5 - 10         82            88
97
30633     school_bus         5 - 10         84            99
99
30634     school_bus         5 - 10         65            60
60
30635     school_bus         5 - 10         58            53
```

```
49
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.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.shape
```

```
(30641, 15)
```

```
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.columns
```

```
Index(['Unnamed: 0', 'Gender', 'EthnicGroup', 'ParentEduc',
'LunchType',
       'TestPrep', 'ParentMaritalStatus', 'PracticeSport',
'IsFirstChild',
       'NrSiblings', 'TransportMeans', 'WklyStudyHours', 'MathScore',
       'ReadingScore', 'WritingScore'],
      dtype='object')
```

```
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
```

```
df.duplicated().sum()
```

```
np.int64(0)
```

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

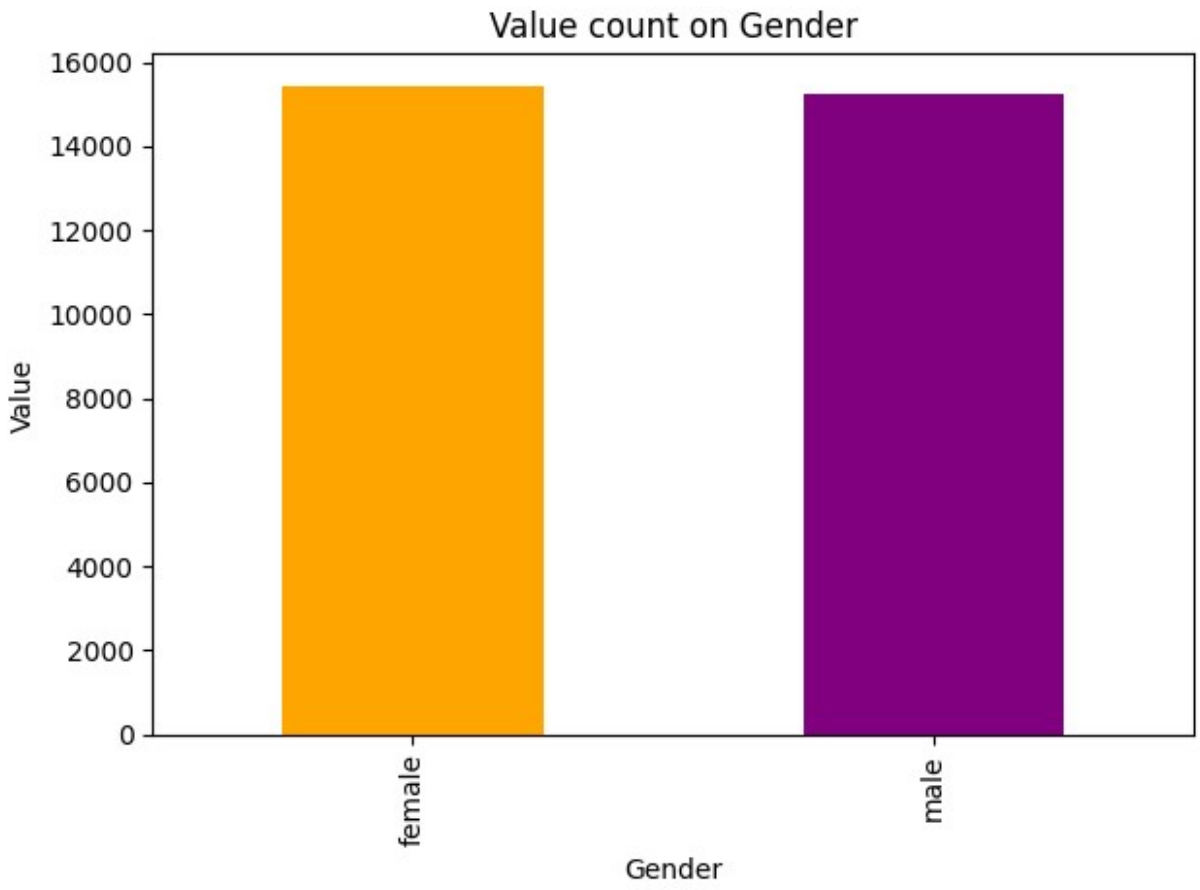
```
df.head(10)
```

```
   Gender EthnicGroup          ParentEduc     LunchType   TestPrep  \
0  female         NaN   bachelor's degree      standard       none
1  female     group C        some college      standard        NaN
```

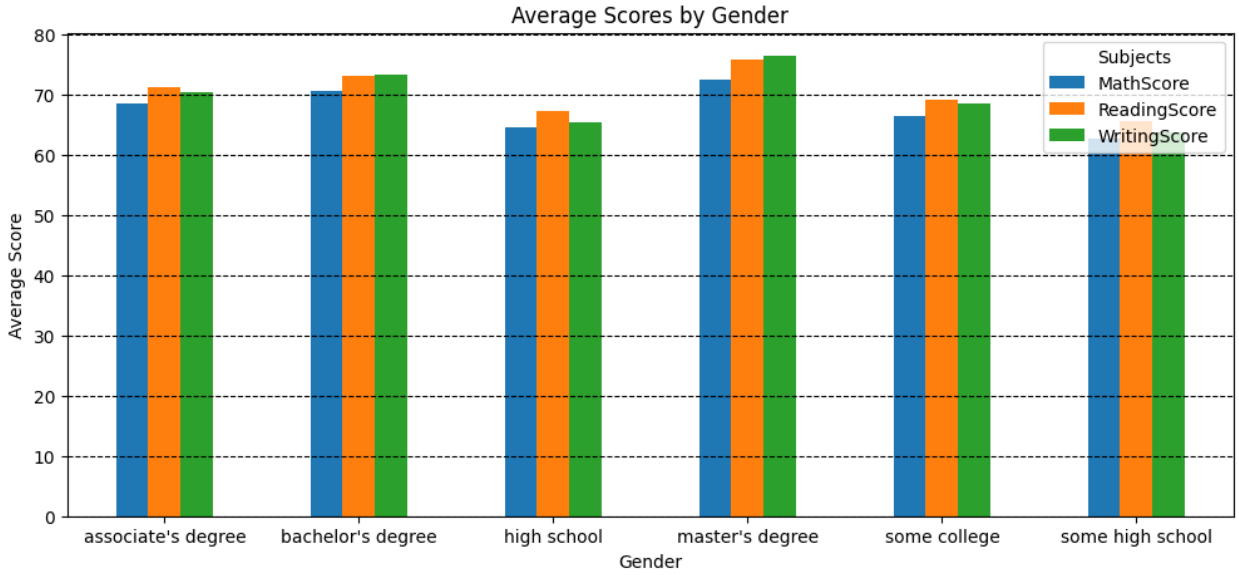
```
2  female     group B     master's degree      standard       none
3    male     group A  associate's degree  free/reduced       none
4    male     group C        some college      standard       none
5  female     group B  associate's degree      standard       none
6  female     group B        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["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
```

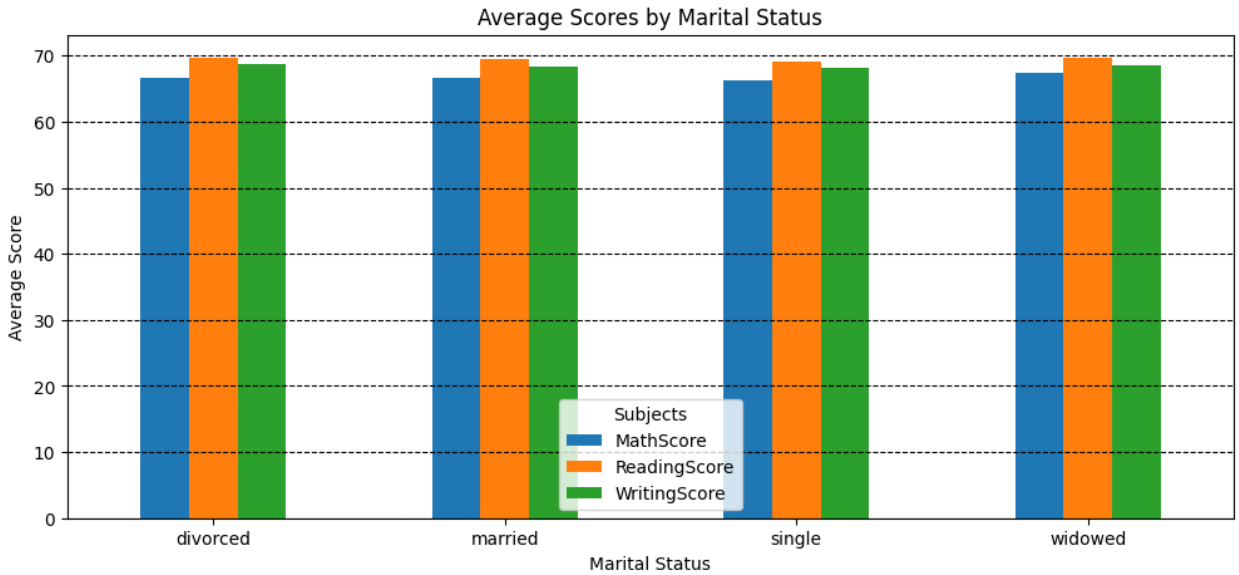
```
df["Gender"].value_counts().plot(kind = "bar" , color = ["orange",
"purple"])
plt.title("Value count on Gender")
plt.xlabel("Gender")
plt.ylabel("Value")
plt.tight_layout()
plt.show()
```




```
gender_scores = df.groupby("ParentEduc")[["MathScore", "ReadingScore",
"WritingScore"]].mean()
gender_scores.plot(kind="bar", figsize=(12,5))
plt.title("Average Scores by Gender")
plt.ylabel("Average Score")
plt.xlabel("Gender")
plt.xticks(rotation=0)  # Keep x labels horizontal
plt.legend(title="Subjects")
plt.grid(axis = "y", linestyle = "--", color = "black")
plt.show()
```



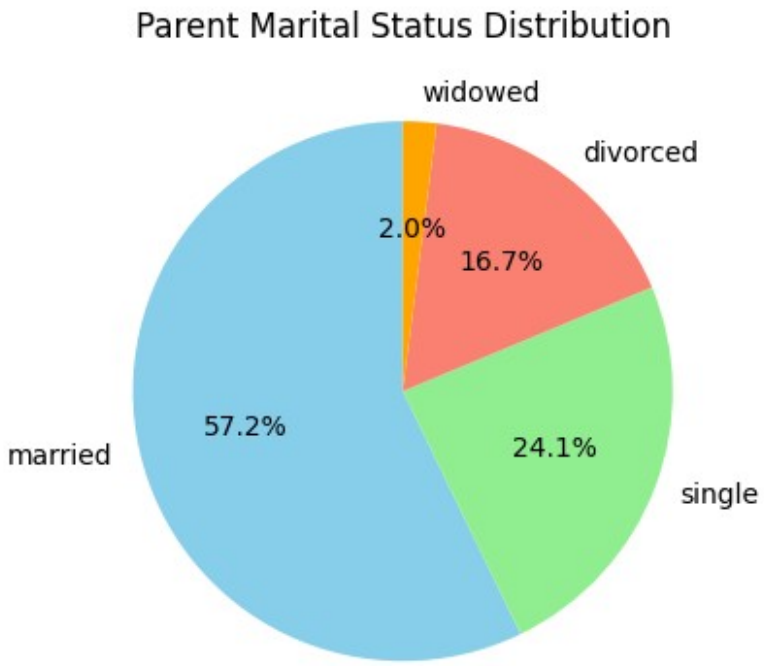

```
# from the abobe graph we have calculated that the education of
parents have a good impact

gender_scores = df.groupby("ParentMaritalStatus")[["MathScore",
"ReadingScore", "WritingScore"]].mean()
gender_scores.plot(kind="bar", figsize=(12,5))
plt.title("Average Scores by Marital Status")
plt.ylabel("Average Score")
plt.xlabel("Marital Status")
plt.xticks(rotation=0)  # Keep x labels horizontal
plt.legend(title="Subjects")
plt.grid(axis = "y", linestyle = "--", color = "black")
plt.show()
```



```
# From the above chart we have concluded that there is no impact on
the student`s score duto there parent married Status

marital = df["ParentMaritalStatus"].value_counts()
print(marital)
```

```
ParentMaritalStatus
married     16844
single       7097
divorced     4919
widowed       591
Name: count, dtype: int64
```

```
plt.figure(figsize=(12,4))
plt.pie(marital,
        labels=marital.index,
        autopct='%1.1f%%',   # show percentages
        startangle=90,       # rotate for better view
        colors=["skyblue", "lightgreen", "salmon", "orange"])  #
optional colors

plt.title("Parent Marital Status Distribution")
plt.tight_layout()
plt.show()
```




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