PROBLEM STATEMENT: Create a bar chart or histogram to visualize the distribution of a categorical or continuous variable, such as the distribution of ages or gender in a population Dataset : Student performance About the dataset : The student dataset contains demographic details (age and gender), study habits (study hours per week), parental involvement (low, medium, high), participation in extracurricular activities (yes, no), and academic performance (grades or scores). It includes both continuous variables (age, study hours per week, academic performance) and categorical variables (gender, parental involvement, extracurricular activities). The dataset is useful for analyzing factors influencing academic performance and the impact of study habits, parental involvement, and extracurricular participation on students outcomes.

In [ ]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
# Load the Students performance dataset
file_path = 'c:\\Users\\Admin\\Downloads\\SP.csv'
data = pd.read_csv(file_path)
print(data)
```

```
      StudentID  Age  Gender  Ethnicity  ParentalEducation  StudyTimeWeekly  \
0          1001   17       1          0                  2        19.833723
1          1002   18       0          0                  1        15.408756
2          1003   15       0          2                  3         4.210570
3          1004   17       1          0                  3        10.028829
4          1005   17       1          0                  2         4.672495
...         ...  ...     ...        ...                ...              ...
2387       3388   18       1          0                  3        10.680555
2388       3389   17       0          0                  1         7.583217
2389       3390   16       1          0                  2         6.805500
2390       3391   16       1          1                  0        12.416653
2391       3392   16       1          0                  2        17.819907

      Absences  Tutoring  ParentalSupport  Extracurricular  Sports  Music  \
0            7         1                2                0       0      1
1            0         0                1                0       0      0
2           26         0                2                0       0      0
3           14         0                3                1       0      0
4           17         1                3                0       0      0
...        ...       ...              ...              ...     ...    ...
2387         2         0                4                1       0      0
2388         4         1                4                0       1      0
2389        20         0                2                0       0      0
2390        17         0                2                0       1      1
2391        13         0                2                0       0      0

      Volunteering       GPA  GradeClass
0                0  2.929196         2.0
1                0  3.042915         1.0
2                0  0.112602         4.0
3                0  2.054218         3.0
4                0  1.288061         4.0
...            ...       ...         ...
2387             0  3.455509         0.0
2388             0  3.279150         4.0
2389             1  1.142333         2.0
2390             0  1.803297         1.0
2391             1  2.140014         1.0

[2392 rows x 15 columns]
```

EDA

In [ ]:
```
# Check data types and missing values
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2392 entries, 0 to 2391
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   StudentID          2392 non-null   int64  
 1   Age                2392 non-null   int64  
 2   Gender             2392 non-null   int64  
 3   Ethnicity          2392 non-null   int64  
 4   ParentalEducation  2392 non-null   int64  
 5   StudyTimeWeekly    2392 non-null   float64
 6   Absences           2392 non-null   int64  
 7   Tutoring           2392 non-null   int64  
 8   ParentalSupport    2392 non-null   int64  
 9   Extracurricular    2392 non-null   int64  
 10  Sports             2392 non-null   int64  
 11  Music              2392 non-null   int64  
 12  Volunteering       2392 non-null   int64  
 13  GPA                2392 non-null   float64
 14  GradeClass         2392 non-null   float64
dtypes: float64(3), int64(12)
memory usage: 280.4 KB
None
```

In [ ]:
```
#  Descriptive statistics
print(data.describe())
```

	StudentID	Age	Gender	Ethnicity	ParentalEducation	\
count	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000	
mean	2196.500000	16.468645	0.510870	0.877508	1.746237	
std	690.655244	1.123798	0.499986	1.028476	1.000411	
min	1001.000000	15.000000	0.000000	0.000000	0.000000	
25%	1598.750000	15.000000	0.000000	0.000000	1.000000	
50%	2196.500000	16.000000	1.000000	0.000000	2.000000	
75%	2794.250000	17.000000	1.000000	2.000000	2.000000	
max	3392.000000	18.000000	1.000000	3.000000	4.000000	

	StudyTimeWeekly	Absences	Tutoring	ParentalSupport	\
count	2392.000000	2392.000000	2392.000000	2392.000000	
mean	9.771992	14.541388	0.301421	2.122074	
std	5.652774	8.467417	0.458971	1.122813	
min	0.001057	0.000000	0.000000	0.000000	
25%	5.043079	7.000000	0.000000	1.000000	
50%	9.705363	15.000000	0.000000	2.000000	
75%	14.408410	22.000000	1.000000	3.000000	
max	19.978094	29.000000	1.000000	4.000000	

	Extracurricular	Sports	Music	Volunteering	GPA	\
count	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000	
mean	0.383361	0.303512	0.196906	0.157191	1.906186	
std	0.486307	0.459870	0.397744	0.364057	0.915156	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	1.174803	
50%	0.000000	0.000000	0.000000	0.000000	1.893393	
75%	1.000000	1.000000	0.000000	0.000000	2.622216	
max	1.000000	1.000000	1.000000	1.000000	4.000000	

	GradeClass
count	2392.000000
mean	2.983696
std	1.233908
min	0.000000
25%	2.000000
50%	4.000000
75%	4.000000
max	4.000000

In [ ]:
```
#Selecting the numerical columns:
print("The Integer columns are: ")
data_numerical=data.select_dtypes(np.number)
data_numerical
```

The Integer columns are:

Out[ ]:

	StudentID	Age	Gender	Ethnicity	ParentalEducation	StudyTimeWeekly	Absences
0	1001	17	1	0	2	19.833723	7
1	1002	18	0	0	1	15.408756	0
2	1003	15	0	2	3	4.210570	26
3	1004	17	1	0	3	10.028829	14
4	1005	17	1	0	2	4.672495	17
...	...	...	...	...	...	...	...
2387	3388	18	1	0	3	10.680555	2
2388	3389	17	0	0	1	7.583217	4
2389	3390	16	1	0	2	6.805500	20
2390	3391	16	1	1	0	12.416653	17
2391	3392	16	1	0	2	17.819907	13

2392 rows × 15 columns

In [ ]:
```
#  Correlation matrix (for continuous variables)
correlation_matrix = data_numerical.corr()
correlation_matrix
```

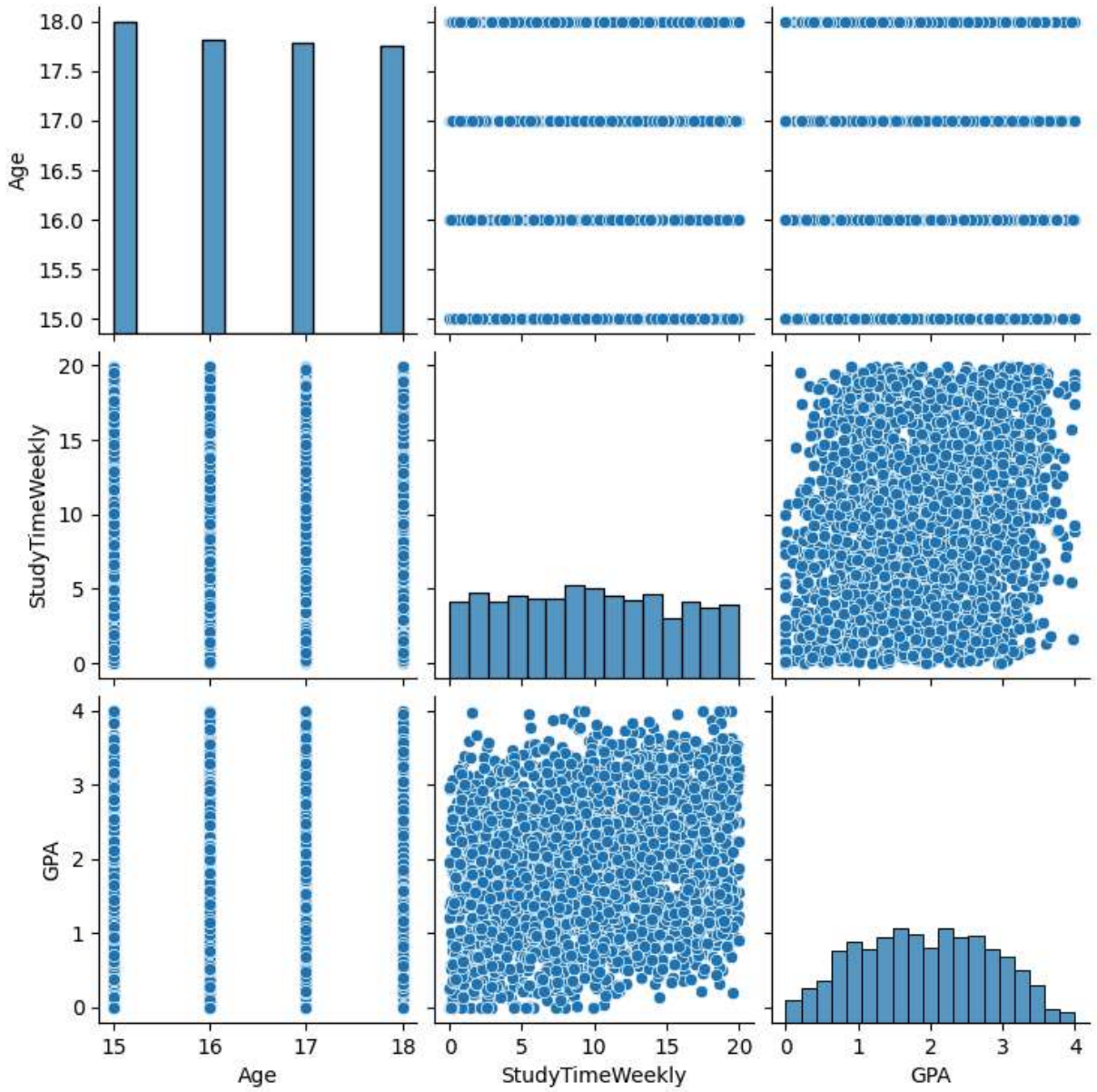
Out[ ]:

	StudentID	Age	Gender	Ethnicity	ParentalEducation	StudyTim
StudentID	1.000000	-0.042255	-0.014625	-0.012990	-0.002307	
Age	-0.042255	1.000000	0.044895	-0.028473	0.025099	-
Gender	-0.014625	0.044895	1.000000	0.016010	0.006771	
Ethnicity	-0.012990	-0.028473	0.016010	1.000000	0.033595	
ParentalEducation	-0.002307	0.025099	0.006771	0.033595	1.000000	-
StudyTimeWeekly	0.026976	-0.006800	0.011469	0.007184	-0.011051	
Absences	0.014841	-0.011511	0.021479	-0.025712	0.036518	
Tutoring	-0.007834	-0.012076	-0.031597	-0.017440	-0.017340	
ParentalSupport	0.003016	0.033197	0.008065	0.020922	-0.017463	
Extracurricular	-0.003611	-0.025061	-0.005964	-0.008927	0.007479	-
Sports	-0.020703	-0.046320	-0.008897	-0.004484	0.002029	
Music	-0.005468	-0.003492	0.007109	-0.014627	0.039439	
Volunteering	0.008011	0.013074	-0.000200	0.013468	0.011960	-
GPA	-0.002697	0.000275	-0.013360	0.027760	-0.035854	
GradeClass	-0.098500	-0.006250	0.022998	-0.023326	0.041031	-

In [ ]:
```
#pairplot
sns.pairplot(data=data, vars=["Age","StudyTimeWeekly", "GPA"])
```

```
c:\Users\Admin\Desktop\TRIAL\.venv\lib\site-packages\seaborn\axisgrid.py:123: UserWa
rning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[ ]: <seaborn.axisgrid.PairGrid at 0x29fdeb64520>



Checking the categorical features

In [ ]:
```
data.columns
```

Out[ ]:
```
Index(['StudentID', 'Age', 'Gender', 'Ethnicity', 'ParentalEducation',
       'StudyTimeWeekly', 'Absences', 'Tutoring', 'ParentalSupport',
       'Extracurricular', 'Sports', 'Music', 'Volunteering', 'GPA',
       'GradeClass'],
      dtype='object')
```

In [ ]:
```
data['Gender'].nunique()
```

Out[ ]:
```
2
```

In [ ]:
```
data['ParentalSupport'].nunique()
```

Out[ ]:
```
5
```

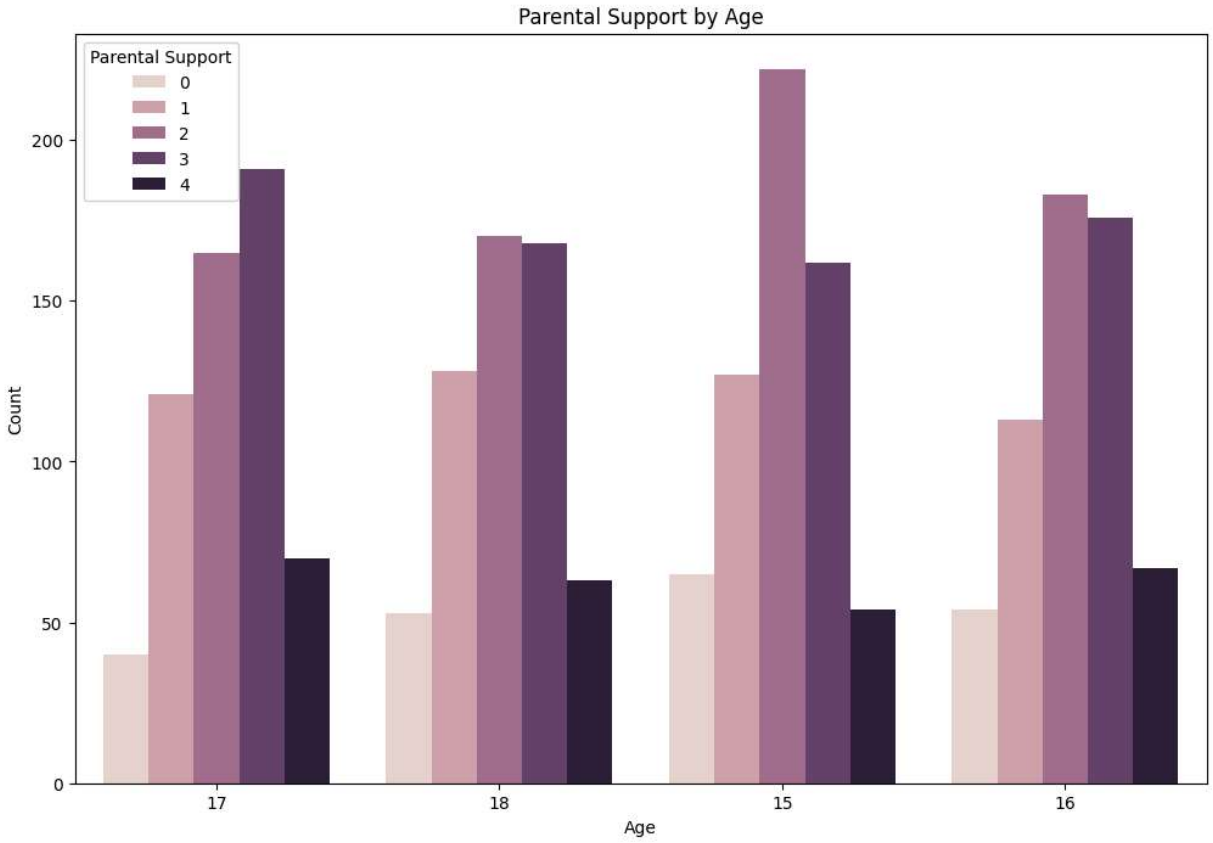
In [ ]:
```
data['Extracurricular'].nunique()
```

Out[ ]:
```
2
```

In [ ]:
```
# Frequency of each region
data['ParentalSupport'].value_counts()
```

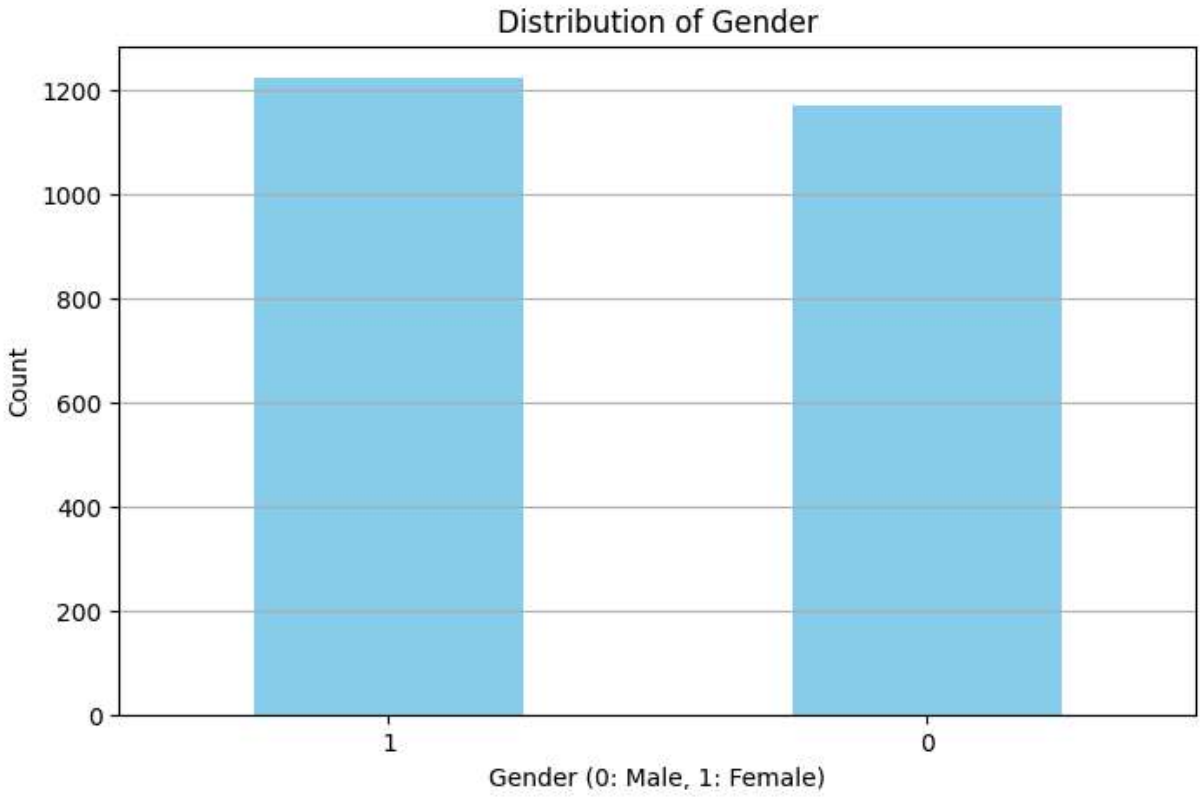
Out[ ]:
```
ParentalSupport
2    740
3    697
1    489
4    254
0    212
Name: count, dtype: int64
```

In [ ]:
```
data['Age'] = data['Age'].astype(str)
plt.figure(figsize=(12, 8))
sns.countplot(data=data, x='Age', hue='ParentalSupport')
plt.title('Parental Support by Age')
plt.xlabel('Age')
plt.ylabel('Count')
plt.legend(title='Parental Support')
plt.show()
```



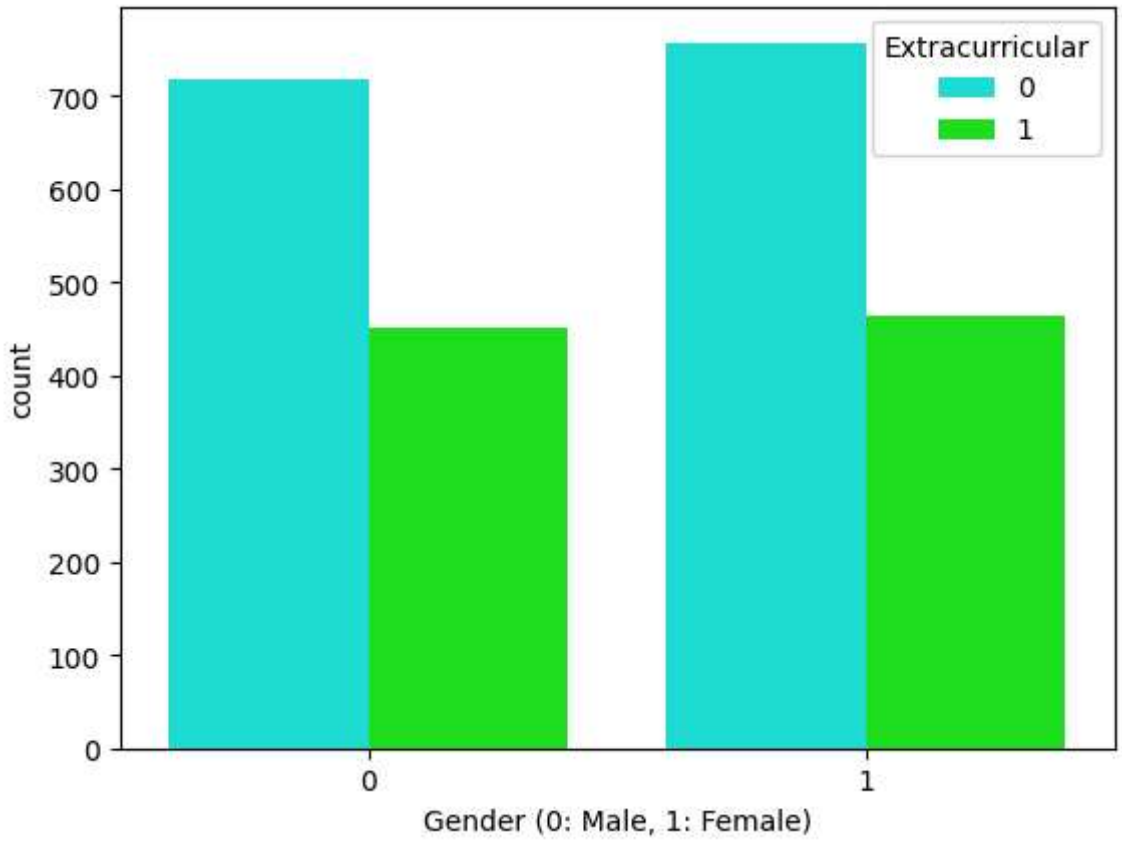
In [ ]:
```
#Distribution of gender
gender_column_name = 'Gender'
gender_counts = data[gender_column_name].value_counts()
plt.figure(figsize=(8, 5))
gender_counts.plot(kind='bar', color='skyblue')
```

```
plt.xlabel('Gender (0: Male, 1: Female)')
plt.ylabel('Count')
plt.title('Distribution of Gender')
plt.xticks(rotation=0)
plt.grid(axis='y')
plt.show()
```



In [ ]:
```
sns.countplot(data=data,x='Gender',hue='Extracurricular',palette=['#00FFF0', '#00FF
plt.xlabel('Gender (0: Male, 1: Female)')
```

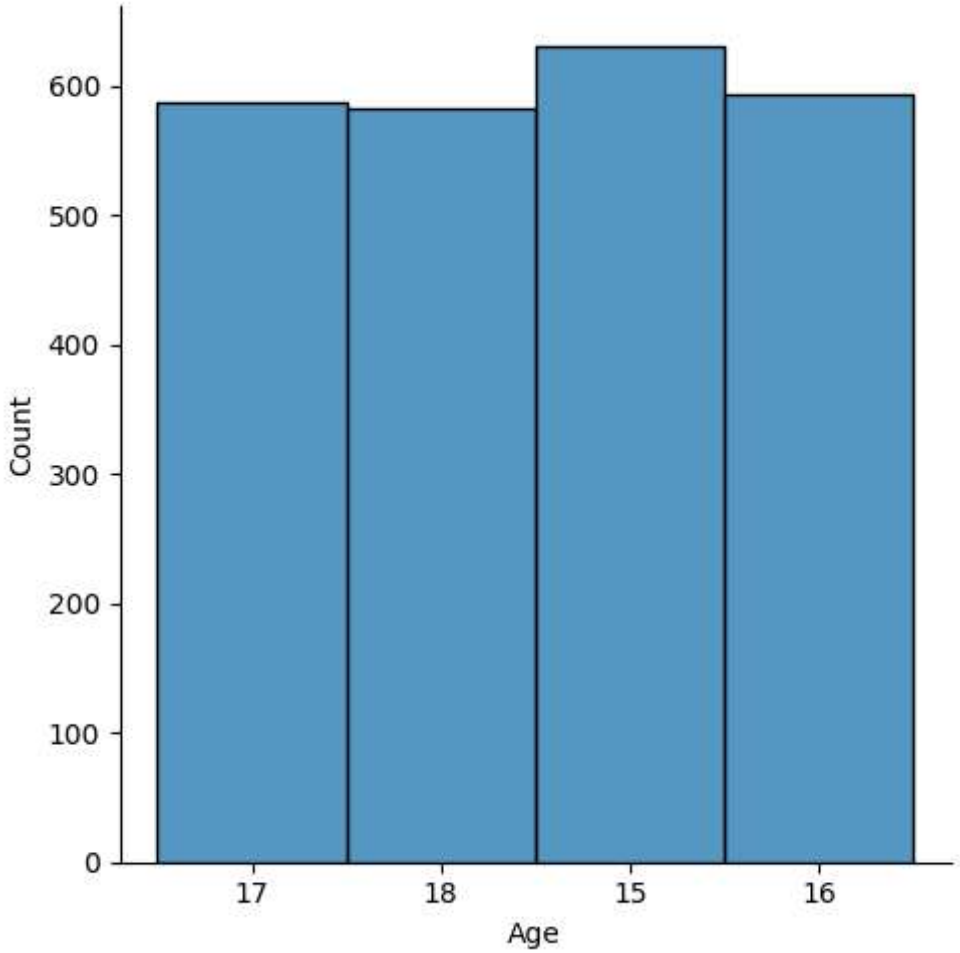
Out[ ]:
```
Text(0.5, 0, 'Gender (0: Male, 1: Female)')
```



In [ ]:
```
#Histogram for continuous variable
sns.displot(data['Age'])
```

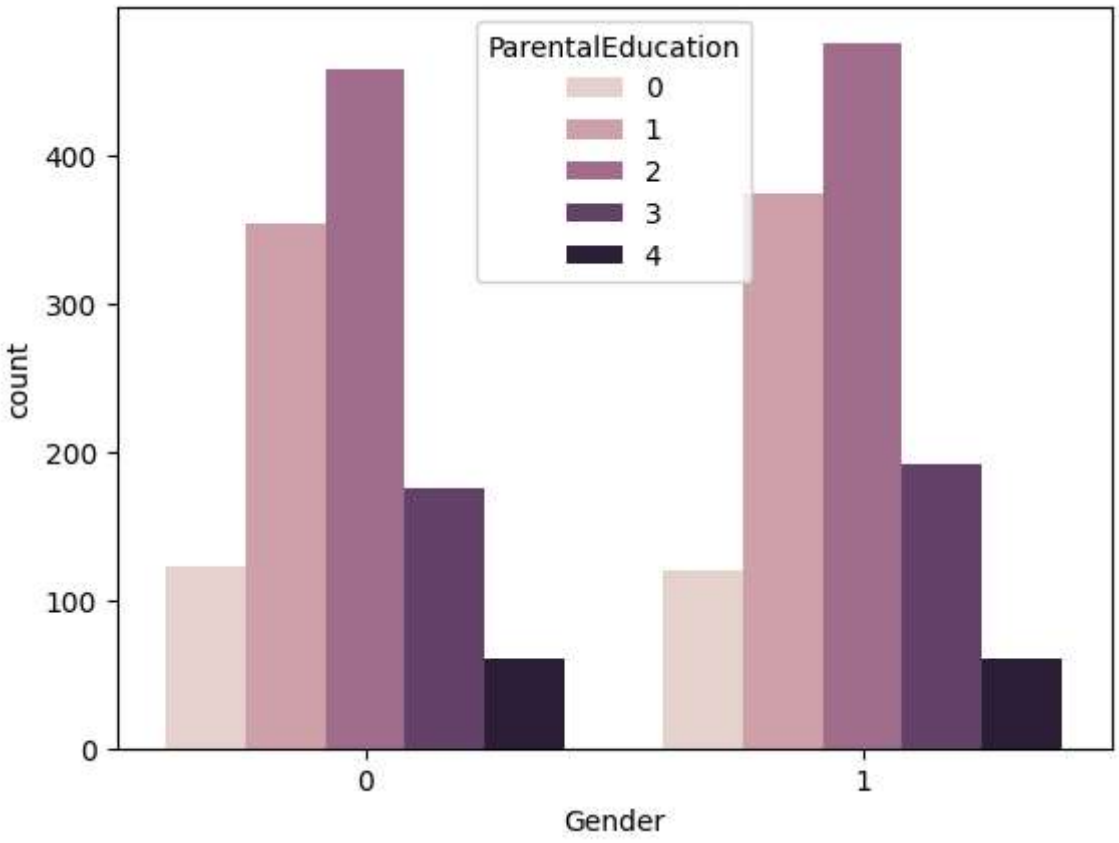
```
c:\Users\Admin\Desktop\TRIAL\.venv\lib\site-packages\seaborn\axisgrid.py:123: UserWa
rning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[ ]: <seaborn.axisgrid.FacetGrid at 0x29fe5153e20>

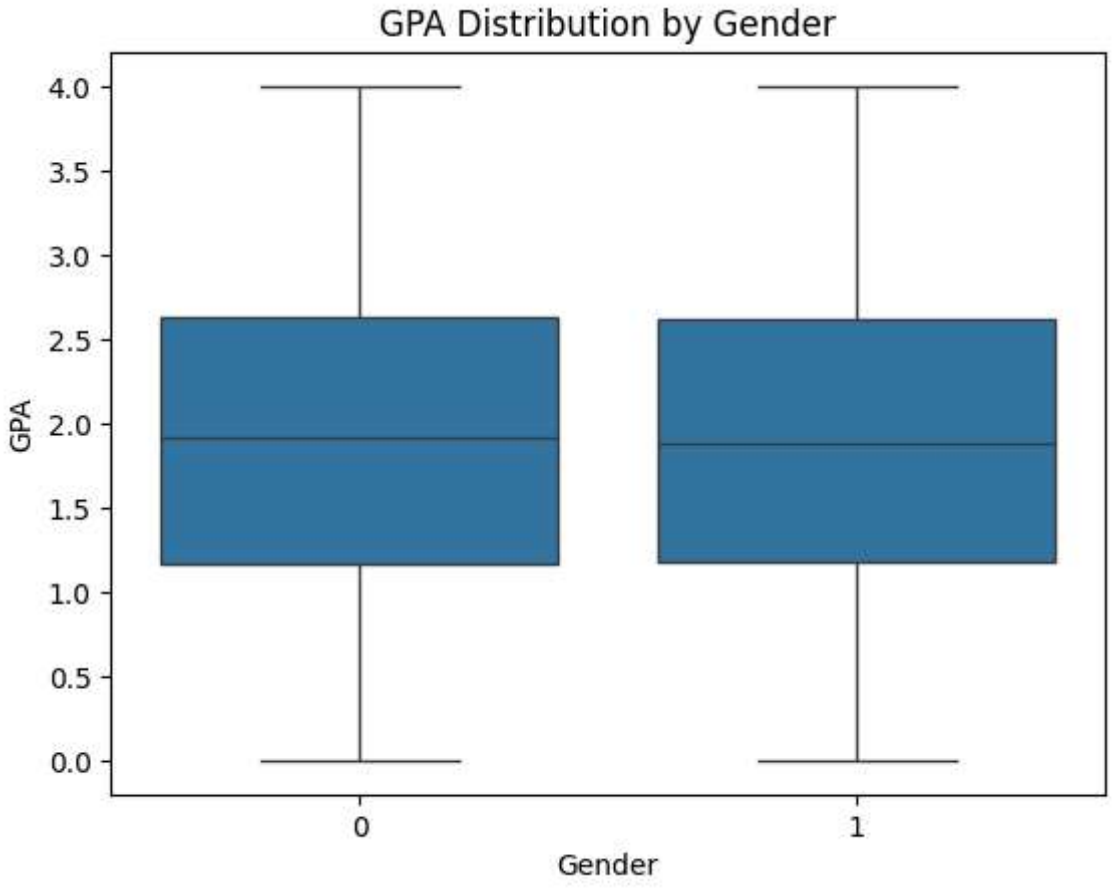


In [ ]:
```
#  Distribution of Gender by Parental Education Level
sns.countplot(data=data,x='Gender',hue='ParentalEducation')
```

Out[ ]:
```
<Axes: xlabel='Gender', ylabel='count'>
```



In [ ]:
```
#GPA distribution by gender
sns.boxplot(x='Gender', y='GPA', data=data)
plt.title('GPA Distribution by Gender')
plt.xlabel('Gender')
plt.ylabel('GPA')
plt.show()
```


GPA Distribution by Gender
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
GPA
0
1
Gender

In [ ]: