

```
In [135]:
import numpy as npy
import pandas as pd
from scipy import stats
import seaborn as sns
import matplotlib.pyplot as plt

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

In [136]:
df.head()

Out[136]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation  S
0         0  1001  35.0      1.0      0.0              2
1         1  1002  18.0      0.0      0.0              1
2         2  1003  NaN      0.0      2.0              3
3         3  1004  17.0      1.0      NaN              3
4         4  1005  17.0      NaN      0.0              2

In [137]:
df.head(10)

Out[137]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation  S
0         0  1001  35.0      1.0      0.0              2
1         1  1002  18.0      0.0      0.0              1
2         2  1003  NaN      0.0      2.0              3
3         3  1004  17.0      1.0      NaN              3
4         4  1005  17.0      NaN      0.0              2
5         5  1006  18.0      0.0      0.0              1
6         6  1007  15.0      0.0      1.0              1
7         7  1008  15.0      1.0      NaN              4
8         8  1009  17.0      0.0      0.0              0
9         9  1010  16.0      1.0      0.0              1

In [138]:
df.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                 2391 non-null  float64
2   Gender              2391 non-null  float64
3   Ethnicity           2390 non-null  float64
4   ParentalEducation   2392 non-null  int64
5   StudyTimeWeekly     2391 non-null  float64
6   Absences            2391 non-null  float64
7   Tutoring            2392 non-null  int64
8   ParentalSupport     2391 non-null  float64
9   Extracurricular     2392 non-null  int64
10  Sports              2391 non-null  float64
11  Music               2391 non-null  float64
12  Volunteering        2392 non-null  float64
13  GPA                 2390 non-null  float64
14  GradeClass          2391 non-null  float64
dtypes: float64(10), int64(5)
memory usage: 280.4 KB

In [139]:
df.describe()

Out[139]:
   StudentID      Age      Gender      Ethnicity  ParentalEducation  S
count  2392.000000    2391.000000    2391.000000    2390.000000
mean    2196.500000    16.484316      0.510665    0.877878
std      690.655244      1.244765      0.499991    1.028747
min    1001.000000    15.000000      0.000000    0.000000
25%    1598.750000    15.000000      0.000000    0.000000
50%    2196.500000    16.000000      1.000000    0.000000
75%    2794.250000    17.000000      1.000000    2.000000
max    3392.000000    35.000000      1.000000    3.000000

In [140]:
df.isnull()

Out[140]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
0         0  False  False  False      False
1         1  False  False  False      False
2         2  False   True  False      False
3         3  False  False  False       True
4         4  False  False   True      False
...
2387        6  False  False  False      False
2388        7  False  False  False      False
2389        8  False  False  False      False
2390        9  False  False  False      False
2391       10  False  False  False      False
2392 rows x 15 columns

In [141]:
df.isnull().sum()

Out[141]:
StudentID      0
Age            1
Gender         1
Ethnicity      2
ParentalEducation  0
StudyTimeWeekly  1
Absences       1
Tutoring       0
ParentalSupport  1
Extracurricular  0
Sports         1
Music          1
Volunteering   0
GPA            2
GradeClass     1
dtype: int64

In [142]:
df.notnull()

Out[142]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
0         0   True   True   True      True
1         1   True   True   True      True
2         2   True  False   True      True
3         3   True   True   True      False
4         4   True   True  False      True
...
2387       6   True   True   True      True
2388       7   True   True   True      True
2389       8   True   True   True      True
2390       9   True   True   True      True
2391      10   True   True   True      True
2392 rows x 15 columns

In [143]:
df.notnull().sum()

Out[143]:
StudentID      2392
Age            2391
Gender         2391
Ethnicity      2390
ParentalEducation  2390
StudyTimeWeekly  2391
Absences       2391
Tutoring       2392
ParentalSupport  2391
Extracurricular  2392
Sports         2391
Music          2391
Volunteering   2392
GPA            2390
GradeClass     2391
dtype: int64

In [144]:
df.isna()

Out[144]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
0         0  False  False  False      False
1         1  False  False  False      False
2         2  False   True  False      False
3         3  False  False  False       True
4         4  False  False   True      False
...
2387       6  False  False  False      False
2388       7  False  False  False      False
2389       8  False  False  False      False
2390       9  False  False  False      False
2391      10  False  False  False      False
2392 rows x 15 columns

In [145]:
df.fillna(1)

Out[145]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
0         0  1001  35.0      1.0      0.0              2
1         1  1002  18.0      0.0      0.0              1
2         2  1003   1.0      0.0      2.0              3
3         3  1004  17.0      1.0      1.0              3
4         4  1005  17.0      1.0      0.0              2
...
2387       6  3388  18.0      1.0      0.0              3
2388       7  3389  17.0      0.0      0.0              1
2389       8  3390  16.0      1.0      0.0              2
2390       9  3391  16.0      1.0      1.0              0
2391      10  3392  16.0      1.0      0.0              2
2392 rows x 15 columns

In [146]:
df.ffill()

Out[146]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
0         0  1001  35.0      1.0      0.0              2
1         1  1002  18.0      0.0      0.0              1
2         2  1003  18.0      0.0      2.0              3
3         3  1004  17.0      1.0      2.0              3
4         4  1005  17.0      1.0      0.0              2
...
2387       6  3388  18.0      1.0      0.0              3
2388       7  3389  17.0      0.0      0.0              1
2389       8  3390  16.0      1.0      0.0              2
2390       9  3391  16.0      1.0      1.0              0
2391      10  3392  16.0      1.0      0.0              2
2392 rows x 15 columns

In [147]:
df.bfill()

Out[147]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
0         0  1001  35.0      1.0      0.0              2
1         1  1002  18.0      0.0      0.0              1
2         2  1003  17.0      0.0      2.0              3
3         3  1004  17.0      1.0      0.0              3
4         4  1005  17.0      0.0      0.0              2
...
2387       6  3388  18.0      1.0      0.0              3
2388       7  3389  17.0      0.0      0.0              1
2389       8  3390  16.0      1.0      0.0              2
2390       9  3391  16.0      1.0      1.0              0
2391      10  3392  16.0      1.0      0.0              2
2392 rows x 15 columns

In [148]:
df.dropna()

Out[148]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
0         0  1001  35.0      1.0      0.0              2
1         1  1002  18.0      0.0      0.0              1
9        10  1010  16.0      1.0      0.0              1
10       11  1011  35.0      0.0      0.0              1
11       12  1012  17.0      0.0      0.0              1
...
2387       6  3388  18.0      1.0      0.0              3
2388       7  3389  17.0      0.0      0.0              1
2389       8  3390  16.0      1.0      0.0              2
2390       9  3391  16.0      1.0      1.0              0
2391      10  3392  16.0      1.0      0.0              2
2385 rows x 15 columns

In [149]:
df.drop('Age',axis = 1)

Out[149]:
   StudentID  Gender  Ethnicity  ParentalEducation  Stud
0         0  1001      1.0      0.0              2
1         1  1002      0.0      0.0              1
2         2  1003      0.0      2.0              3
3         3  1004      1.0      NaN              3
4         4  1005      NaN      0.0              2
...
2387       6  3388      1.0      0.0              3
2388       7  3389      0.0      0.0              1
2389       8  3390      1.0      0.0              2
2390       9  3391      1.0      1.0              0
2391      10  3392      1.0      0.0              2
2392 rows x 14 columns

In [150]:
df.replace(to_replace = npy.nan,value = '1' )

Out[150]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
0         0  1001  35.0      1.0      0.0              2
1         1  1002  18.0      0.0      0.0              1
2         2  1003   1      0.0      2.0              3
3         3  1004  17.0      1.0      1              3
4         4  1005  17.0      1      0.0              2
...
2387       6  3388  18.0      1.0      0.0              3
2388       7  3389  17.0      0.0      0.0              1
2389       8  3390  16.0      1.0      0.0              2
2390       9  3391  16.0      1.0      1.0              0
2391      10  3392  16.0      1.0      0.0              2
2392 rows x 15 columns

In [151]:
df.interpolate()

Out[151]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
0         0  1001  35.0      1.0      0.0              2
1         1  1002  18.0      0.0      0.0              1
2         2  1003  17.5      0.0      2.0              3
3         3  1004  17.0      1.0      1.0              3
4         4  1005  17.0      0.5      0.0              2
...
2387       6  3388  18.0      1.0      0.0              3
2388       7  3389  17.0      0.0      0.0              1
2389       8  3390  16.0      1.0      0.0              2
2390       9  3391  16.0      1.0      1.0              0
2391      10  3392  16.0      1.0      0.0              2
2392 rows x 15 columns

In [152]:
z_scores = stats.zscore(df['StudentID'])

In [153]:
print(z_scores)

[-1.73132686 -1.72987865 -1.72843045 ...  1.72843045
 1.72987865
 1.73132686]

In [154]:
abs_z_scores = npy.abs(z_scores)

In [155]:
abs_z_scores

Out[155]:
array([1.73132686, 1.72987865, 1.72843045, ..., 1.72843045,
       1.72987865, 1.73132686], shape=(2392,))

In [ ]:

In [156]:
outliers = (z_scores > 1.73)

In [157]:
print(outliers)

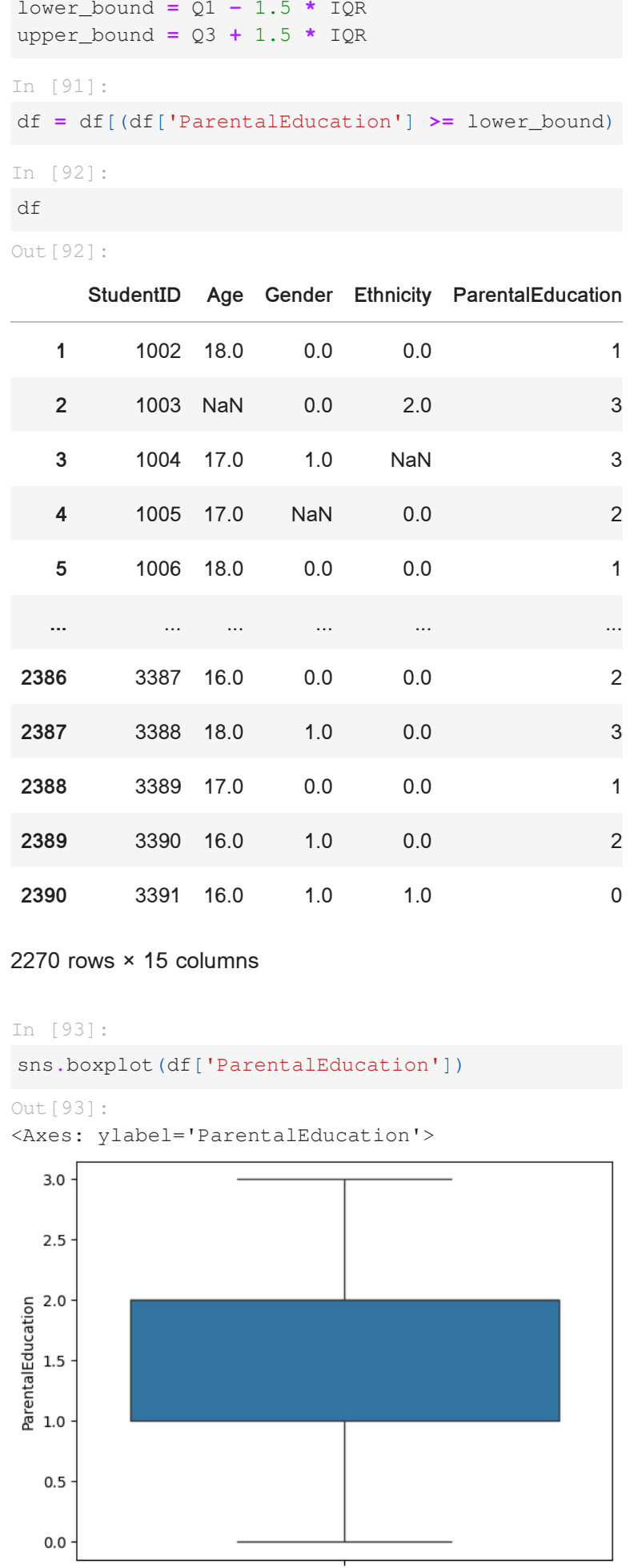
[False False False ... False False  True]

In [162]:
outliers = outliers[:len(df)]
df = df[~outliers]

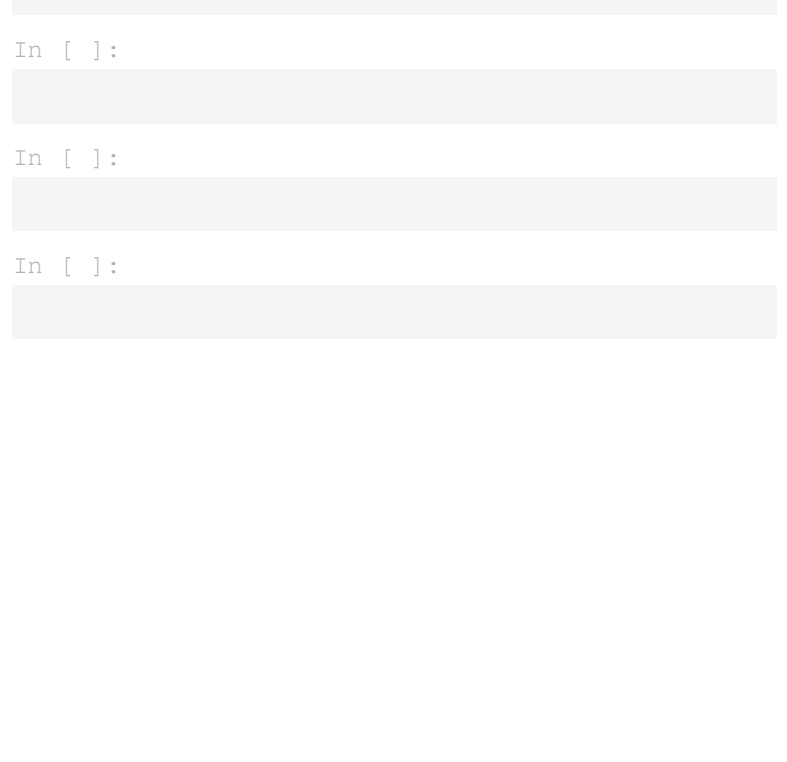
In [163]:
print(outliers)

[False False False ... False False False]

In [161]:
sns.boxplot(df['StudentID'])

Out[161]:
<Axes: ylabel='StudentID'>


In [87]:
sns.boxplot(df['ParentalEducation'])

Out[87]:
<Axes: ylabel='ParentalEducation'>


In [88]:
Q1 = df['ParentalEducation'].quantile(0.25)

In [89]:
Q3 = df['ParentalEducation'].quantile(0.75)

In [90]:
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

In [91]:
df = df[(df['ParentalEducation'] >= lower_bound)

In [92]:
df

Out[92]:
   StudentID  Age  Gender  Ethnicity  ParentalEducation
1         1  1002  18.0      0.0      0.0              1
2         2  1003  NaN      0.0      2.0              3
3         3  1004  17.0      1.0      NaN              3
4         4  1005  17.0      NaN      0.0              2
5         5  1006  18.0      0.0      0.0              1
...
2386       6  3387  16.0      0.0      0.0              2
2387       7  3388  18.0      1.0      0.0              3
2388       8  3389  17.0      0.0      0.0              1
2389       9  3390  16.0      1.0      0.0              2
2390      10  3391  16.0      1.0      1.0              0
2270 rows x 15 columns

In [93]:
sns.boxplot(df['ParentalEducation'])

Out[93]:
<Axes: ylabel='ParentalEducation'>


In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

```