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

```
df = pd.read_csv("/content/academic.csv")
```

```
df.shape
```

```
(480, 17)
```

```
df.size
```

```
8160
```

```
df.columns
```

```
Index(['gender', 'NationalITy', 'PlaceofBirth', 'StageID', 'GradeID',
       'SectionID', 'Topic', 'Semester', 'Relation', 'raisedhands',
       'VisITedResources', 'AnnouncementsView', 'Discussion',
       'ParentAnsweringSurvey', 'ParentschoolSatisfaction',
       'StudentAbsenceDays', 'Class'],
      dtype='object')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   gender                    478 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    object
 9   raisedhands               478 non-null    float64
 10  VisITedResources          480 non-null    int64
 11  AnnouncementsView         480 non-null    int64
 12  Discussion                480 non-null    int64
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: float64(1), int64(3), object(13)
memory usage: 63.9+ KB
```

```
df.describe()
```

	raisedhands	VisITedResources	AnnouncementsView	Discussion
count	478.000000	480.000000	480.000000	480.000000
mean	46.939331	54.797917	37.918750	43.283333
std	31.375699	33.080007	26.611244	27.637735
min	0.000000	0.000000	0.000000	1.000000
25%	15.000000	20.000000	14.000000	20.000000
50%	50.000000	65.000000	33.000000	39.000000
75%	75.000000	84.000000	58.000000	70.000000
max	170.000000	99.000000	98.000000	99.000000

```
df.isna().sum()
```

```
gender                      2
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     0
SectionID                   0
Topic                       0
Semester                    0
Relation                    0
raisedhands                 2
VisITedResources            0
AnnouncementsView           0
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
Class                       0
dtype: int64
```

```
df['gender'].fillna(df['gender'].mode()[0], inplace = True)
df['raisedhands'].fillna(df['raisedhands'].mean(), inplace = True)
```

```
df.isna().sum()
```

```
gender                      0
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     0
SectionID                   0
Topic                       0
Semester                    0
Relation                    0
raisedhands                 0
VisITedResources            0
AnnouncementsView           0
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
Class                       0
dtype: int64
```

```
def DetectOutlier(df, var):
  Q1 = df[var].quantile(0.25)
  Q3 = df[var].quantile(0.75)
  IQR = Q3 - Q1

  high = Q3 + 1.5 * IQR
  low = Q1 - 1.5 * IQR
  print("Highest allowed variable: ", var, high)
  print("Lowest allowed variable: ", var, low)

  count = df[(df[var] > high) | (df[var] < low)][var].count()
  print("Total outliers in: ", var, ':', count)

  df1 = df[((df[var] < low) | (df[var] > high))]
  print("Outliers: \n", len(df1))
  print(df1.T)

  df = df[((df[var] >= low) & (df[var] <= high))]
  return (df)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
```

```
 0   gender                    480 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    object
 9   raisedhands               480 non-null    float64
 10  VisITedResources          480 non-null    int64
 11  AnnouncementsView         480 non-null    int64
 12  Discussion                480 non-null    int64
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: float64(1), int64(3), object(13)
memory usage: 63.9+ KB
```

```
df['Relation'] = df['Relation'].astype('category')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   gender                    480 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    category
 9   raisedhands               480 non-null    float64
 10  VisITedResources          480 non-null    int64
 11  AnnouncementsView         480 non-null    int64
 12  Discussion                480 non-null    int64
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: category(1), float64(1), int64(3), object(12)
memory usage: 60.7+ KB
```

```
df.head()
```

	GradeID	SectionID	Topic	Semester	Relation	raisedhands	VisITedResources	Announ
	G-04	A	IT	F	Father	15.000000	16	
	G-04	A	IT	F	Father	46.939331	20	
	G-04	A	IT	F	Father	10.000000	7	
	G-04	A	IT	F	Father	30.000000	25	
	G-04	A	IT	F	Father	0.000000	50	

Next steps: Generate code with df View recommended plots

```
df['Relation'] = df['Relation'].cat.codes
```

```
df.head()
```

Relation	raisedhands	VisITedResources	AnnouncementsView	Discussion	ParentAnswer
0	15.000000	16	2	20	
0	46.939331	20	3	25	
0	10.000000	7	0	30	
0	30.000000	25	5	35	
0	0.000000	50	12	50	

Next steps: Generate code with df View recommended plots

```
df['ParentAnsweringSurvey'].replace(['Yes', 'No'], [1, 0], inplace =  True)
```

```
df.head()
```

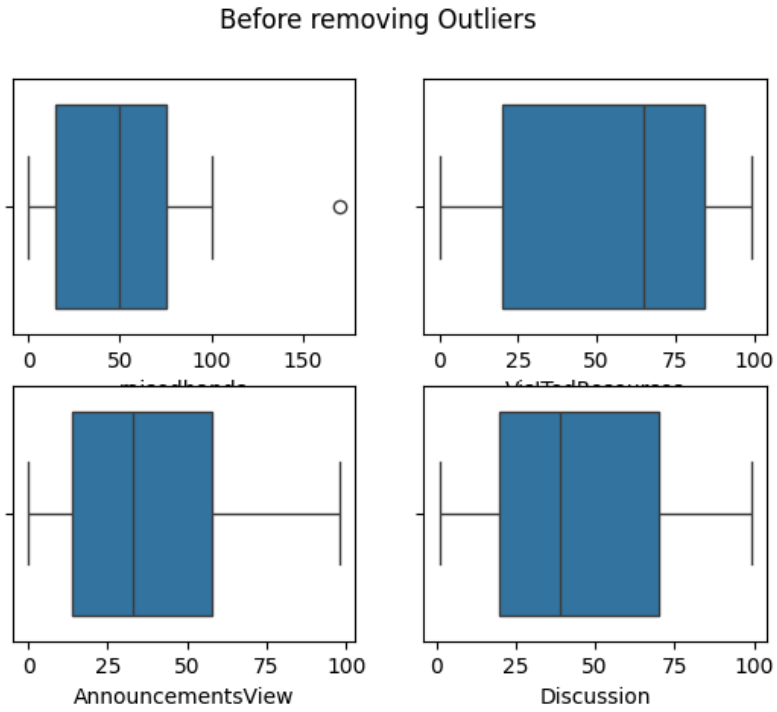
s	VisITedResources	AnnouncementsView	Discussion	ParentAnsweringSurvey	Parentschoo
0	16	2	20	1	
1	20	3	25	1	
0	7	0	30	0	
0	25	5	35	0	
0	50	12	50	0	

Next steps: Generate code with df View recommended plots

Outliers can be visualized using boxplot

using seaborn library we can plot the boxplot

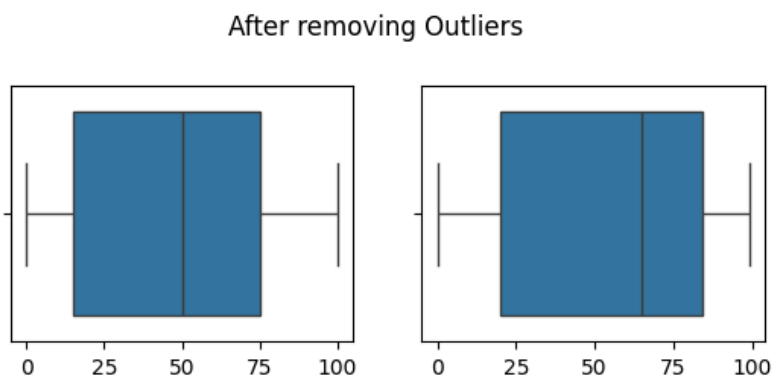
```
fig, axes = plt.subplots(2, 2)
fig.suptitle('Before removing Outliers')
sns.boxplot(data = df, x = 'raisedhands', ax=axes[0,0])
sns.boxplot(data = df, x = 'VisITedResources', ax=axes[0,1])
sns.boxplot(data = df, x = 'AnnouncementsView', ax=axes[1, 0])
sns.boxplot(data = df, x = 'Discussion', ax=axes[1,1])
plt.show()
```




```
df = DetectOutlier(df, 'raisedhands')
```

```
Highest allowed variable:  raisedhands 165.0
Lowest allowed variable:  raisedhands -75.0
Total outliers in:  raisedhands : 1
Outliers:
 1
                                  28
gender                             M
NationalITy                       KW
PlaceofBirth                  KuwaIT
StageID                   MiddleSchool
GradeID                        G-08
SectionID                         A
Topic                       Science
Semester                          F
Relation                          0
raisedhands                   170.0
VisITedResources                 85
AnnouncementsView                52
Discussion                       43
ParentAnsweringSurvey             1
ParentschoolSatisfaction       Good
StudentAbsenceDays          Under-7
Class                             M
```

```
fig, axes = plt.subplots(2,2)
fig.suptitle('After removing Outliers')
sns.boxplot(data = df, x ='raisedhands', ax=axes[0,0])
sns.boxplot(data = df, x ='VisITedResources', ax=axes[0,1])
sns.boxplot(data = df, x ='AnnouncementsView', ax=axes[1,0])
sns.boxplot(data = df, x ='Discussion', ax=axes[1,1])
plt.show()
```




```
df = DetectOutlier(df, 'Discussion')
```

```
Highest allowed variable:  Discussion 145.0
Lowest allowed variable:  Discussion -55.0
Total outliers in:  Discussion : 0
Outliers:
 0
Empty DataFrame
Columns: []
Index: [gender, NationalITy, PlaceofBirth, StageID, GradeID, SectionID, Topic, Semester, Relation, raisedhands, Vi
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 479 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   gender                    479 non-null    object
 1   NationalITy               479 non-null    object
 2   PlaceofBirth              479 non-null    object
 3   StageID                   479 non-null    object
 4   GradeID                   479 non-null    object
 5   SectionID                 479 non-null    object
 6   Topic                     479 non-null    object
 7   Semester                  479 non-null    object
 8   Relation                  479 non-null    int8
 9   raisedhands               479 non-null    float64
 10  VisITedResources          479 non-null    int64
 11  AnnouncementsView         479 non-null    int64
 12  Discussion                479 non-null    int64
 13  ParentAnsweringSurvey     479 non-null    int64
 14  ParentschoolSatisfaction  479 non-null    object
 15  StudentAbsenceDays        479 non-null    object
 16  Class                     479 non-null    object
dtypes: float64(1), int64(4), int8(1), object(11)
```