In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
Data1 = pd.read_csv('AcademicPerformance.csv')
print(Data1)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04       NaN
2        M          KW       KuwaIT    lowerlevel    G-04         A
3      NaN          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

          Topic Semester Relation  raisedhands  VisITedResources  \
0            IT        F   Father         15.0              16.0
1            IT        F   Father         20.0              20.0
2            IT      NaN   Father         10.0               7.0
3            IT        F   Father         30.0              25.0
4            IT        F   Father         40.0              50.0
..          ...      ...      ...          ...               ...
475  Chemistry        S   Father          5.0               4.0
476    Geology        F   Father         50.0              77.0
477    Geology        S   Father         55.0              74.0
478    History        F   Father         30.0              17.0
479    History        S   Father         35.0              14.0

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                  2.0          20                   Yes
1                  3.0          25                   Yes
2                  0.0          30                    No
3                  5.0          35                    No
4                 12.0          50                    No
..                 ...         ...                   ...
475                5.0           8                    No
476               14.0          28                    No
477               25.0          29                    No
478               14.0          57                    No
479               23.0          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good           Under-7     M
1                       Good           Under-7     M
2                        Bad           Above-7     L
3                        Bad           Above-7     L
4                        Bad           Above-7     M
..                       ...               ...   ...
475                      Bad           Above-7     L
476                      Bad           Under-7     M
477                      Bad           Under-7     M
478                      Bad           Above-7     L
479                      Bad           Above-7     L

[480 rows x 17 columns]
```

In [3]:

```
Data1.shape
```

Out[3]:
```
(480, 17)
```

In [4]:
```
print(Data1.isnull().sum())
```

```
gender                     10
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     0
SectionID                   6
Topic                       0
Semester                    9
Relation                    0
raisedhands                10
VisITedResources            5
AnnouncementsView           4
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
Class                       0
dtype: int64
```

In [5]:
```
Data1.dropna(inplace=True)
print(Data1.isnull().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
```

In [6]:
```
import pandas as pd
import numpy as np
Data1 = pd.read_csv('AcademicPerformance.csv')
print(Data1)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04       NaN
2        M          KW       KuwaIT    lowerlevel    G-04         A
3      NaN          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
```

```
477      F      Jordan       Jordan  MiddleSchool     G-08         A
478      F      Jordan       Jordan  MiddleSchool     G-08         A
479      F      Jordan       Jordan  MiddleSchool     G-08         A

          Topic Semester Relation  raisedhands  VisITedResources  \
0            IT        F   Father         15.0              16.0
1            IT        F   Father         20.0              20.0
2            IT      NaN   Father         10.0               7.0
3            IT        F   Father         30.0              25.0
4            IT        F   Father         40.0              50.0
..          ...      ...      ...          ...               ...
475  Chemistry        S   Father          5.0               4.0
476    Geology        F   Father         50.0              77.0
477    Geology        S   Father         55.0              74.0
478    History        F   Father         30.0              17.0
479    History        S   Father         35.0              14.0

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                  2.0          20                   Yes
1                  3.0          25                   Yes
2                  0.0          30                    No
3                  5.0          35                    No
4                 12.0          50                    No
..                 ...         ...                   ...
475                5.0           8                    No
476               14.0          28                    No
477               25.0          29                    No
478               14.0          57                    No
479               23.0          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good           Under-7     M
1                       Good           Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       ...                ...   ...
475                      Bad            Above-7     L
476                      Bad           Under-7     M
477                      Bad           Under-7     M
478                      Bad            Above-7     L
479                      Bad            Above-7     L

[480 rows x 17 columns]
```

In [7]:
```
Data1['raisedhands'] = Data1['raisedhands'].replace(np.NaN,Data1['raisedhands'].mean()
```

In [8]:
```
print(Data1['raisedhands'][:10])
```

```
0    15.0
1    20.0
2    10.0
3    30.0
4    40.0
5    42.0
6    35.0
7    50.0
8    12.0
9    70.0
Name: raisedhands, dtype: float64
```

In [9]:

```
import numpy as np
Data1 = pd.read_csv('AcademicPerformance.csv')
print(Data1)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04       NaN
2        M          KW       KuwaIT    lowerlevel    G-04         A
3      NaN          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

          Topic Semester Relation  raisedhands  VisITedResources  \
0            IT        F   Father         15.0              16.0
1            IT        F   Father         20.0              20.0
2            IT      NaN   Father         10.0               7.0
3            IT        F   Father         30.0              25.0
4            IT        F   Father         40.0              50.0
..          ...      ...      ...          ...               ...
475  Chemistry        S   Father          5.0               4.0
476    Geology        F   Father         50.0              77.0
477    Geology        S   Father         55.0              74.0
478    History        F   Father         30.0              17.0
479    History        S   Father         35.0              14.0

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                  2.0          20                   Yes
1                  3.0          25                   Yes
2                  0.0          30                    No
3                  5.0          35                    No
4                 12.0          50                    No
..                 ...         ...                   ...
475                5.0           8                    No
476               14.0          28                    No
477               25.0          29                    No
478               14.0          57                    No
479               23.0          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good           Under-7     M
1                       Good           Under-7     M
2                        Bad           Above-7     L
3                        Bad           Above-7     L
4                        Bad           Above-7     M
..                       ...               ...   ...
475                      Bad           Above-7     L
476                      Bad           Under-7     M
477                      Bad           Under-7     M
478                      Bad           Above-7     L
479                      Bad           Above-7     L

[480 rows x 17 columns]
```

In [10]:
```
Data1['raisedhands'] = Data1['raisedhands'].replace(np.NaN,Data1['raisedhands'].mediar
```

In [11]:
```
print(Data1['raisedhands'][:10])
```

```
0    15.0
1    20.0
2    10.0
3    30.0
4    40.0
5    42.0
6    35.0
7    50.0
8    12.0
9    70.0
Name: raisedhands, dtype: float64
```

In [12]:

```
import statistics
Data1 = pd.read_csv('AcademicPerformance.csv')
print(Data1)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04       NaN
2        M          KW       KuwaIT    lowerlevel    G-04         A
3      NaN          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

          Topic Semester Relation  raisedhands  VisITedResources  \
0            IT        F   Father         15.0              16.0
1            IT        F   Father         20.0              20.0
2            IT      NaN   Father         10.0               7.0
3            IT        F   Father         30.0              25.0
4            IT        F   Father         40.0              50.0
..          ...      ...      ...          ...               ...
475  Chemistry        S   Father          5.0               4.0
476    Geology        F   Father         50.0              77.0
477    Geology        S   Father         55.0              74.0
478    History        F   Father         30.0              17.0
479    History        S   Father         35.0              14.0

      AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                   2.0          20                   Yes
1                   3.0          25                   Yes
2                   0.0          30                    No
3                   5.0          35                    No
4                  12.0          50                    No
..                  ...         ...                   ...
475                 5.0           8                    No
476                14.0          28                    No
477                25.0          29                    No
478                14.0          57                    No
479                23.0          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good            Under-7     M
1                       Good            Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       ...                ...   ...
475                      Bad            Above-7     L
```

```
476                  Bad          Under-7      M
477                  Bad          Under-7      M
478                  Bad          Above-7      L
479                  Bad          Above-7      L

[480 rows x 17 columns]
```

In [13]:

```
Data1['raisedhands'] = Data1['raisedhands'].replace(np.NaN,statistics.mode(Data1['rais
print(Data1['raisedhands'][:10])
```

```
0    15.0
1    20.0
2    10.0
3    30.0
4    40.0
5    42.0
6    35.0
7    50.0
8    12.0
9    70.0
Name: raisedhands, dtype: float64
```

In [14]:

```
Data1 = pd.read_csv('AcademicPerformance.csv')
print(Data1)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04       NaN
2        M          KW       KuwaIT    lowerlevel    G-04         A
3      NaN          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

          Topic Semester Relation  raisedhands  VisITedResources  \
0            IT        F   Father         15.0              16.0
1            IT        F   Father         20.0              20.0
2            IT      NaN   Father         10.0               7.0
3            IT        F   Father         30.0              25.0
4            IT        F   Father         40.0              50.0
..          ...      ...      ...          ...               ...
475  Chemistry        S   Father          5.0               4.0
476    Geology        F   Father         50.0              77.0
477    Geology        S   Father         55.0              74.0
478    History        F   Father         30.0              17.0
479    History        S   Father         35.0              14.0

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                  2.0          20                   Yes
1                  3.0          25                   Yes
2                  0.0          30                    No
3                  5.0          35                    No
4                 12.0          50                    No
..                 ...         ...                   ...
475                5.0           8                    No
476               14.0          28                    No
477               25.0          29                    No
478               14.0          57                    No
479               23.0          62                    No
```

```
    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good            Under-7     M
1                       Good            Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       ...                ...   ...
475                      Bad            Above-7     L
476                      Bad            Under-7     M
477                      Bad            Under-7     M
478                      Bad            Above-7     L
479                      Bad            Above-7     L

[480 rows x 17 columns]
```

In [15]:
```
Data1.isnull().sum()
```

Out[15]:
```
gender                    10
NationalITy                0
PlaceofBirth               0
StageID                    0
GradeID                    0
SectionID                  6
Topic                      0
Semester                   9
Relation                   0
raisedhands               10
VisITedResources           5
AnnouncementsView          4
Discussion                 0
ParentAnsweringSurvey      0
ParentschoolSatisfaction   0
StudentAbsenceDays         0
Class                      0
dtype: int64
```

In [16]:
```
#Perform the imputation on value
Data1['gender'] = Data1['gender'].fillna('T')
```

In [17]:
```
Data1.isnull().sum()
```

Out[17]:
```
gender                    0
NationalITy               0
PlaceofBirth              0
StageID                   0
GradeID                   0
SectionID                 6
Topic                     0
Semester                  9
Relation                  0
raisedhands              10
VisITedResources          5
AnnouncementsView         4
Discussion                0
ParentAnsweringSurvey     0
ParentschoolSatisfaction  0
StudentAbsenceDays        0
Class                     0
dtype: int64
```

In [18]:
```
print(Data1)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04       NaN
2        M          KW       KuwaIT    lowerlevel    G-04         A
3        T          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

          Topic Semester Relation  raisedhands  VisITedResources  \
0            IT        F   Father         15.0              16.0
1            IT        F   Father         20.0              20.0
2            IT      NaN   Father         10.0               7.0
3            IT        F   Father         30.0              25.0
4            IT        F   Father         40.0              50.0
..          ...      ...      ...          ...               ...
475  Chemistry        S   Father          5.0               4.0
476    Geology        F   Father         50.0              77.0
477    Geology        S   Father         55.0              74.0
478    History        F   Father         30.0              17.0
479    History        S   Father         35.0              14.0

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                  2.0          20                   Yes
1                  3.0          25                   Yes
2                  0.0          30                    No
3                  5.0          35                    No
4                 12.0          50                    No
..                 ...         ...                   ...
475                5.0           8                    No
476               14.0          28                    No
477               25.0          29                    No
478               14.0          57                    No
479               23.0          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good           Under-7     M
1                       Good           Under-7     M
2                        Bad           Above-7     L
3                        Bad           Above-7     L
4                        Bad           Above-7     M
..                       ...               ...   ...
475                      Bad           Above-7     L
476                      Bad           Under-7     M
477                      Bad           Under-7     M
478                      Bad           Above-7     L
479                      Bad           Above-7     L

[480 rows x 17 columns]
```

In [19]:
```
Data2 = pd.read_csv('AcademicPerformance.csv')
Data2['raisedhands'][:20]
```

Out[19]:
```
0     15.0
1     20.0
2     10.0
3     30.0
```

```
4     40.0
5     42.0
6     35.0
7     50.0
8     12.0
9     70.0
10    50.0
11    19.0
12     5.0
13    20.0
14     NaN
15    30.0
16    36.0
17    55.0
18    69.0
19    70.0
Name: raisedhands, dtype: float64
```

In [20]:
```
Data2.isnull().sum()
```

Out[20]:
```
gender                    10
NationalITy                0
PlaceofBirth               0
StageID                    0
GradeID                    0
SectionID                  6
Topic                      0
Semester                   9
Relation                   0
raisedhands               10
VisITedResources           5
AnnouncementsView          4
Discussion                 0
ParentAnsweringSurvey      0
ParentschoolSatisfaction   0
StudentAbsenceDays         0
Class                      0
dtype: int64
```

In [21]:
```
#Last Obeservation Carried forward
Data2['raisedhands'] = Data2['raisedhands'].fillna(method ='ffill')
Data2.isnull().sum()
```

Out[21]:
```
gender                    10
NationalITy                0
PlaceofBirth               0
StageID                    0
GradeID                    0
SectionID                  6
Topic                      0
Semester                   9
Relation                   0
raisedhands                0
VisITedResources           5
AnnouncementsView          4
Discussion                 0
ParentAnsweringSurvey      0
ParentschoolSatisfaction   0
StudentAbsenceDays         0
Class                      0
dtype: int64
```

In [22]:

```
# example of summarizing the number of missing values for each variable
from pandas import read_csv
# load the dataset
dataset = read_csv('AcademicPerformance.csv', header=None)
# count the number of missing values for each column
num_missing = (dataset[[1,2,3,4,5]] == 0).sum()
# report the results
print(num_missing)
```

```
1    0
2    0
3    0
4    0
5    0
dtype: int64
```

In [23]:

```
# example of review rows from the dataset with missing values marked
from numpy import nan
from pandas import read_csv
# load the dataset
dataset = read_csv('AcademicPerformance.csv', header=None)
# replace '0' values with 'nan'
dataset[[1,2,3,4,5]] = dataset[[1,2,3,4,5]].replace(0, nan)
# print the first 20 rows of data
print(dataset.head(20))
```

```
        0           1            2            3        4          5   \
0   gender  NationalITy  PlaceofBirth      StageID  GradeID  SectionID
1        M          KW        KuwaIT   lowerlevel     G-04          A
2        M          KW        KuwaIT   lowerlevel     G-04        NaN
3        M          KW        KuwaIT   lowerlevel     G-04          A
4      NaN          KW        KuwaIT   lowerlevel     G-04          A
5        M          KW        KuwaIT   lowerlevel     G-04          A
6      NaN          KW        KuwaIT   lowerlevel     G-04          A
7        M          KW        KuwaIT  MiddleSchool     G-07        NaN
8        M          KW        KuwaIT  MiddleSchool     G-07          A
9        F          KW        KuwaIT  MiddleSchool     G-07          A
10     NaN          KW        KuwaIT  MiddleSchool     G-07          B
11       M          KW        KuwaIT  MiddleSchool     G-07          A
12       M          KW        KuwaIT  MiddleSchool     G-07          B
13     NaN          KW        KuwaIT   lowerlevel     G-04        NaN
14       M      lebanon       lebanon  MiddleSchool     G-08          A
15       F          KW        KuwaIT  MiddleSchool     G-08          A
16       F          KW        KuwaIT  MiddleSchool     G-06          A
17     NaN          KW        KuwaIT  MiddleSchool     G-07          B
18       M          KW        KuwaIT  MiddleSchool     G-07          A
19       F          KW        KuwaIT  MiddleSchool     G-07        NaN

        6        7         8           9                10  \
0   Topic  Semester  Relation  raisedhands  VisITedResources
1      IT         F    Father           15                16
2      IT         F    Father           20                20
3      IT       NaN    Father           10                 7
4      IT         F    Father           30                25
5      IT         F    Father           40                50
6      IT         F    Father           42                30
7    Math       NaN    Father           35                12
8    Math         F    Father           50                10
9    Math         F    Father           12                21
10     IT         F    Father           70                80
11   Math         F    Father           50                88
12   Math         F    Father           19                 6
13      0         F    Father            5                 1
```

```
14   Math       NaN    Father          20               14
15   Math         F       Mum         NaN               70
16     IT         F    Father          30               40
17     IT         F    Father          36               30
18   Math         F    Father          55               13
19     IT         F       Mum          69               15

                  11          12                   13  \
0   AnnouncementsView  Discussion  ParentAnsweringSurvey
1                   2          20                  Yes
2                   3          25                  Yes
3                   0          30                   No
4                   5          35                   No
5                  12          50                   No
6                  13          70                  Yes
7                   0          17                   No
8                  15          22                  Yes
9                  16          50                  Yes
10                 25          70                  Yes
11                 30          80                  Yes
12                 19          12                  Yes
13                NaN          11                   No
14                 12          19                   No
15                 44          60                   No
16                 22          66                  Yes
17                 20          80                   No
18                 35          90                   No
19                 36          96                  Yes

                         14                15     16
0   ParentschoolSatisfaction  StudentAbsenceDays  Class
1                       Good            Under-7      M
2                       Good            Under-7      M
3                        Bad            Above-7      L
4                        Bad            Above-7      L
5                        Bad            Above-7      M
6                        Bad            Above-7      M
7                        Bad            Above-7      L
8                       Good            Under-7      M
9                       Good            Under-7      M
10                      Good            Under-7      M
11                      Good            Under-7      H
12                      Good            Under-7      M
13                       Bad            Above-7      L
14                       Bad            Above-7      L
15                       Bad            Above-7      H
16                      Good            Under-7      M
17                       Bad            Above-7      M
18                       Bad            Above-7      M
19                      Good            Under-7      M
```

In [24]:

```
#Inconsistancies in dataset
Data1 = pd.read_csv('AcademicPerformance.csv')
print(Data1)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04       NaN
2        M          KW       KuwaIT    lowerlevel    G-04         A
3      NaN          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
```

```
476      F      Jordan      Jordan  MiddleSchool    G-08          A
477      F      Jordan      Jordan  MiddleSchool    G-08          A
478      F      Jordan      Jordan  MiddleSchool    G-08          A
479      F      Jordan      Jordan  MiddleSchool    G-08          A

          Topic Semester Relation  raisedhands  VisITedResources  \
0            IT        F   Father         15.0              16.0
1            IT        F   Father         20.0              20.0
2            IT      NaN   Father         10.0               7.0
3            IT        F   Father         30.0              25.0
4            IT        F   Father         40.0              50.0
..          ...      ...      ...          ...               ...
475  Chemistry        S   Father          5.0               4.0
476    Geology        F   Father         50.0              77.0
477    Geology        S   Father         55.0              74.0
478    History        F   Father         30.0              17.0
479    History        S   Father         35.0              14.0

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                  2.0          20                   Yes
1                  3.0          25                   Yes
2                  0.0          30                    No
3                  5.0          35                    No
4                 12.0          50                    No
..                 ...         ...                   ...
475                5.0           8                    No
476               14.0          28                    No
477               25.0          29                    No
478               14.0          57                    No
479               23.0          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good           Under-7     M
1                       Good           Under-7     M
2                        Bad           Above-7     L
3                        Bad           Above-7     L
4                        Bad           Above-7     M
..                       ...               ...   ...
475                      Bad           Above-7     L
476                      Bad           Under-7     M
477                      Bad           Under-7     M
478                      Bad           Above-7     L
479                      Bad           Above-7     L

[480 rows x 17 columns]
```

In [25]:
```
Data1['ParentAnsweringSurvey'].isnull()
```

Out[25]:
```
0      False
1      False
2      False
3      False
4      False
       ...
475    False
476    False
477    False
478    False
479    False
Name: ParentAnsweringSurvey, Length: 480, dtype: bool
```

In [26]:
```
#Detecting number
import numpy as np
```

```
cnt = 0
for row in Data1['ParentAnsweringSurvey']:
    try:
        int(row)
        Data1.loc[cnt, 'ParentAnsweringSurvey']=np.nan
    except ValueError:
        pass
    cnt+=1
```

In [27]:

```
Data1['ParentAnsweringSurvey'].isnull()
print(Data1)
```

```
    gender NationalITy PlaceofBirth       StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04       NaN
2        M          KW       KuwaIT    lowerlevel    G-04         A
3      NaN          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

          Topic Semester Relation  raisedhands  VisITedResources  \
0            IT        F   Father         15.0              16.0
1            IT        F   Father         20.0              20.0
2            IT      NaN   Father         10.0               7.0
3            IT        F   Father         30.0              25.0
4            IT        F   Father         40.0              50.0
..          ...      ...      ...          ...               ...
475  Chemistry        S   Father          5.0               4.0
476    Geology        F   Father         50.0              77.0
477    Geology        S   Father         55.0              74.0
478    History        F   Father         30.0              17.0
479    History        S   Father         35.0              14.0

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                  2.0          20                   Yes
1                  3.0          25                   Yes
2                  0.0          30                    No
3                  5.0          35                    No
4                 12.0          50                    No
..                 ...         ...                   ...
475                5.0           8                    No
476               14.0          28                    No
477               25.0          29                    No
478               14.0          57                    No
479               23.0          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good           Under-7     M
1                       Good           Under-7     M
2                        Bad           Above-7     L
3                        Bad           Above-7     L
4                        Bad           Above-7     M
..                       ...               ...   ...
475                      Bad           Above-7     L
476                      Bad           Under-7     M
477                      Bad           Under-7     M
478                      Bad           Above-7     L
```

```
479 . . . . . Bad . . . . . Above-7 . . . L

[480 rows x 17 columns]
```

Feature Engineering – How to Detect and Remove

Z-score treatment :

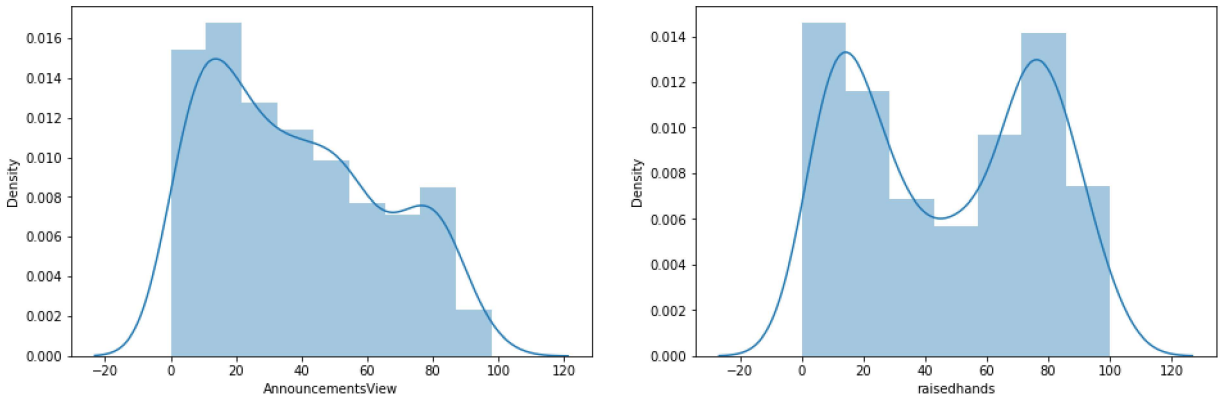
In [28]:
```
#Step-1: Importing Necessary Dependencies
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [29]:
```
#Step-2: Read and Load the Dataset
df = pd.read_csv('AcademicPerformance.csv')
df.sample(10)
```

Out[29]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relati
229	M	KW	KuwaIT	MiddleSchool	G-08	A	Spanish	S	Fath
73	F	KW	KuwaIT	MiddleSchool	G-07	A	English	F	Fath
206	M	KW	KuwaIT	MiddleSchool	G-08	B	Arabic	S	Fath
107	M	KW	KuwaIT	lowerlevel	G-02	B	IT	F	Fath
180	F	SaudiArabia	SaudiArabia	lowerlevel	G-02	B	French	S	Fath
151	M	SaudiArabia	USA	HighSchool	G-11	A	Science	S	Fath
182	M	KW	KuwaIT	MiddleSchool	G-08	A	Arabic	S	Mu
401	M	Jordan	Jordan	MiddleSchool	G-07	A	Biology	S	Fath
21	F	KW	KuwaIT	MiddleSchool	G-07	B	IT	F	Fath
46	M	KW	KuwaIT	lowerlevel	G-05	A	English	F	Fath

In [37]:
```
#Step-3: Plot the Distribution plots for the features
import warnings
warnings.filterwarnings('ignore')
plt.figure(figsize=(16,5))
plt.subplot(1,2,1)
sns.distplot(df['AnnouncementsView'])
plt.subplot(1,2,2)
sns.distplot(df['raisedhands'])
plt.show()
```



In [38]:
```
# Finding the Boundary Values
print("Highest allowed",df['raisedhands'].mean() + 3*df['raisedhands'].std())
print("Lowest allowed",df['raisedhands'].mean() - 3*df['raisedhands'].std())
```

```
Highest allowed 138.95605953669565
Lowest allowed -43.94329357924884
```

In [39]:
```
#Step-5: Finding the Outliers
df[(df['raisedhands'] > 138.95) | (df['raisedhands'] < -43.94)]
```

Out[39]:

gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedh

IQR based filtering :

In [40]:
```
#Used when our data distribution is skewed.
#Step-1: Import necessary dependencies
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

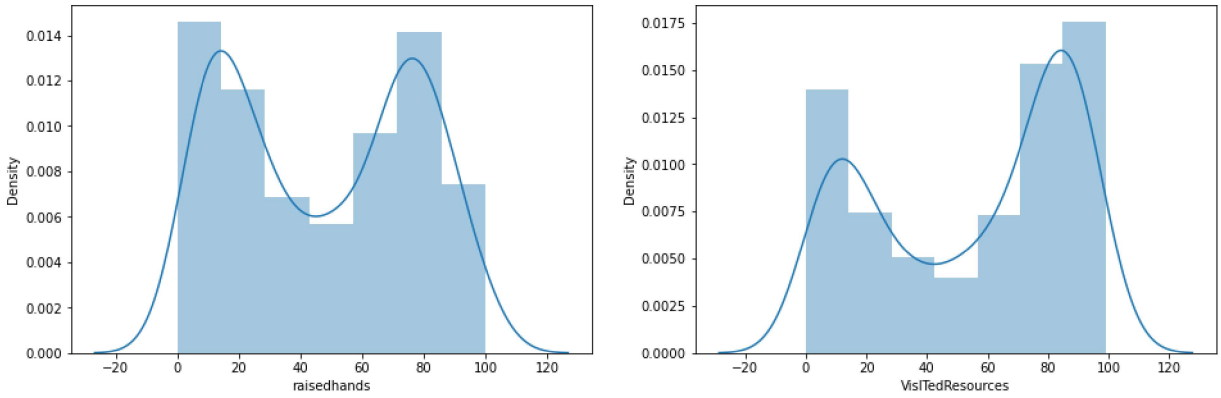
In [41]:
```
#Step-2: Read and Load the Dataset
df = pd.read_csv('AcademicPerformance.csv')
df.head()
```

Out[41]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	raise
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	
1	M	KW	KuwaIT	lowerlevel	G-04	NaN	IT	F	Father	
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	NaN	Father	
3	NaN	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	

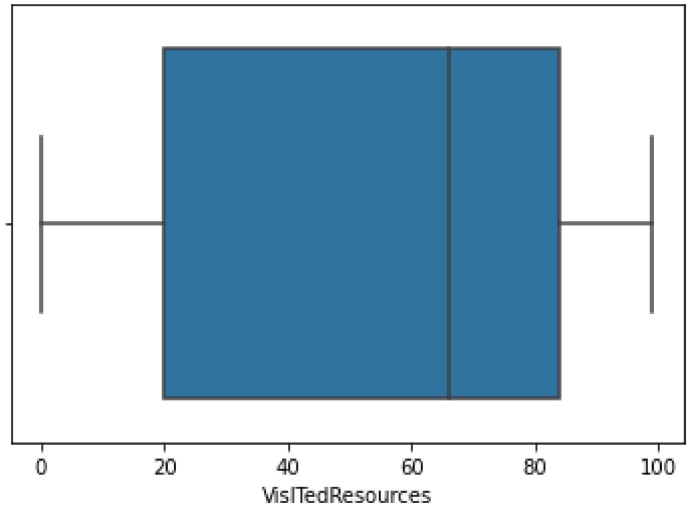
In [42]:
```
#Step-3: Plot the distribution plot for the features
plt.figure(figsize=(16,5))
plt.subplot(1,2,1)
sns.distplot(df['raisedhands'])
```

```
plt.subplot(1,2,2)
sns.distplot(df['VisITedResources'])
plt.show()
```



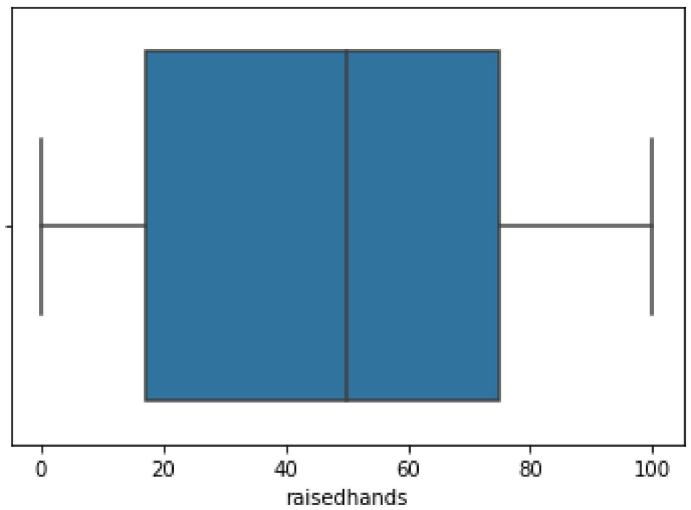
In [43]:
```
#Step-4: Form a Box-plot for the skewed feature
sns.boxplot(df['VisITedResources'])
```

Out[43]: <AxesSubplot:xlabel='VisITedResources'>



In [44]:
```
#Step-4: Form a Box-plot for the skewed feature
sns.boxplot(df['raisedhands'])
```

Out[44]: <AxesSubplot:xlabel='raisedhands'>



In [45]:
```
#Step-5: Finding the IQR
percentile25 = df['VisITedResources'].quantile(0.25)
percentile75 = df['VisITedResources'].quantile(0.75)
```

In [46]:
```
#Step-6: Finding upper and lower limit
iqr=0.0
upper_limit = percentile75 + 1.5 * iqr
lower_limit = percentile25 - 1.5 * iqr
```

In [ ]:
```

```