In [1]:
```
#statement:-Data Wrangling II
'''Create an “Academic performance” dataset of students and perform the following
using Python.
1. Scan all variables for missing values and inconsistencies. If there are missin
and/or inconsistencies, use any of the suitable techniques to deal with them.
2. Scan all numeric variables for outliers. If there are outliers, use any of the
techniques to deal with them.
3. Apply data transformations on at least one of the variables. The purpose of th
transformation should be one of the following reasons: to change the scale for be
understanding of the variable, to convert a non-linear relation into a linear one
decrease the skewness and convert the distribution into a normal distribution.
Reason and document your approach properly.'''
```

Out[1]:
```
'Create an “Academic performance” dataset of students and perform the following
operations\nusing Python.\n1. Scan all variables for missing values and inconsi
stencies. If there are missing values\nand/or inconsistencies, use any of the s
uitable techniques to deal with them.\n2. Scan all numeric variables for outlie
rs. If there are outliers, use any of the suitable\ntechniques to deal with the
m.\n3. Apply data transformations on at least one of the variables. The purpose
of this\ntransformation should be one of the following reasons: to change the s
cale for better\nunderstanding of the variable, to convert a non-linear relatio
n into a linear one, or to\ndecrease the skewness and convert the distribution
into a normal distribution.\nReason and document your approach properly.'
```

In [2]:
```
import pandas as pd
```

In [3]:
```
import numpy as np
```

In [4]:
```
df=pd.read_csv(r'C:\Users\user\Downloads\archive (6)\xAPI-Edu-Data.csv')
```

In [5]:
```
df
```

Out[5]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Rela
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fa
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fa
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fa
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fa
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fa
...	...	...	...	...	...	...	...	...	
475	F	Jordan	Jordan	MiddleSchool	G-08	A	Chemistry	S	Fa
476	F	Jordan	Jordan	MiddleSchool	G-08	A	Geology	F	Fa
477	F	Jordan	Jordan	MiddleSchool	G-08	A	Geology	S	Fa
478	F	Jordan	Jordan	MiddleSchool	G-08	A	History	F	Fa
479	F	Jordan	Jordan	MiddleSchool	G-08	A	History	S	Fa

480 rows × 17 columns

In [6]:
```
df.shape
```

Out[6]:
```
(480, 17)
```

In [7]:
```
df.columns
```

Out[7]:
```
Index(['gender', 'NationalITy', 'PlaceofBirth', 'StageID', 'GradeID',
       'SectionID', 'Topic', 'Semester', 'Relation', 'raisedhands',
       'VisITedResources', 'AnnouncementsView', 'Discussion',
       'ParentAnsweringSurvey', 'ParentschoolSatisfaction',
       'StudentAbsenceDays', 'Class'],
      dtype='object')
```

In [8]:
```
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    int64 
 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: int64(4), object(13)
memory usage: 63.9+ KB
```

In [9]:
```
df.describe()
```

Out[9]:

	raisedhands	VisITedResources	AnnouncementsView	Discussion
count	480.000000	480.000000	480.000000	480.000000
mean	46.775000	54.797917	37.918750	43.283333
std	30.779223	33.080007	26.611244	27.637735
min	0.000000	0.000000	0.000000	1.000000
25%	15.750000	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	100.000000	99.000000	98.000000	99.000000

In [10]:
```
df.corr()
```

Out[10]:

	raisedhands	VisITedResources	AnnouncementsView	Discussion
raisedhands	1.000000	0.691572	0.643918	0.339386
VisITedResources	0.691572	1.000000	0.594500	0.243292
AnnouncementsView	0.643918	0.594500	1.000000	0.417290
Discussion	0.339386	0.243292	0.417290	1.000000

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
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 [ ]:
```

```

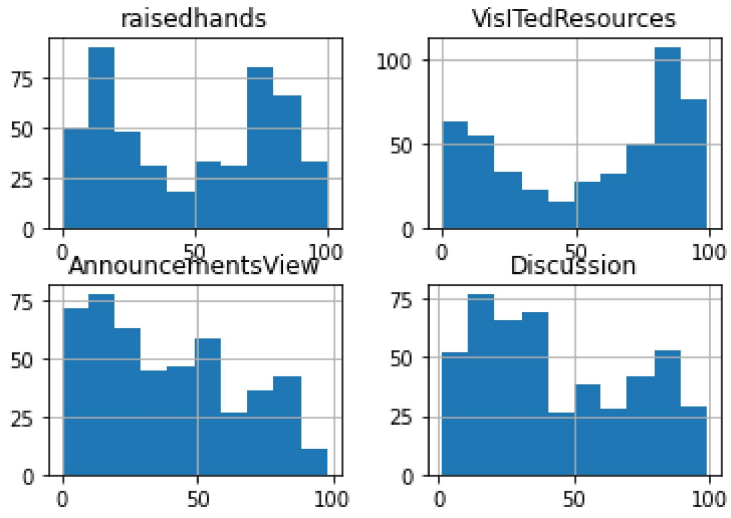
In [12]:
```
df.dropna(inplace=True)
df.isnull().sum()
```

Out[12]:
```
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 [13]:
```
from matplotlib import pyplot as plt
```

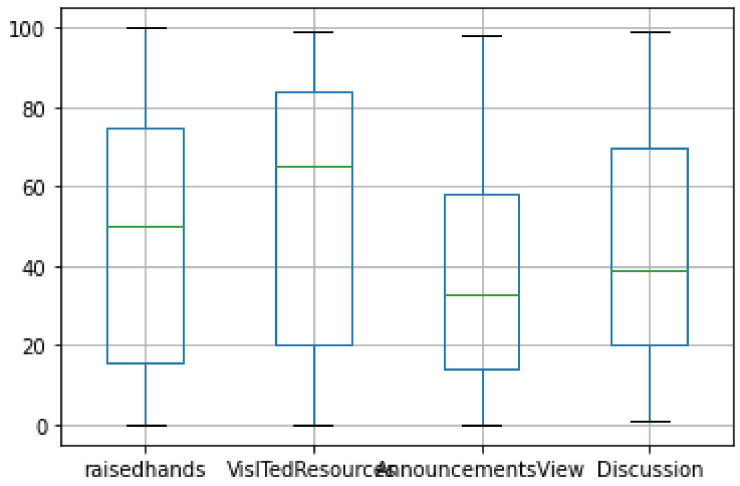
In [14]:
```
df.hist()
```

Out[14]:
```
array([[<AxesSubplot:title={'center':'raisedhands'}>,
        <AxesSubplot:title={'center':'VisITedResources'}>],
       [<AxesSubplot:title={'center':'AnnouncementsView'}>,
        <AxesSubplot:title={'center':'Discussion'}>]], dtype=object)
```



In [15]:
```
df.boxplot()
```

Out[15]: <AxesSubplot:>



In [16]:
```
#if it has outlier then
'''
df.dropna(inplace=True)
print(df.isnull().sum())'''
```

Out[16]: '\ndf.dropna(inplace=True)\nprint(df.isnull().sum())'

In [17]:
```
#df['clm']=df['clm'].replace(np.NaN,df['clm'].mean())
```

In [ ]:
```

```