In [1]:

```
pip install pandas
```

```
Requirement already satisfied: pandas in c:\users\mahima-pc\appdata\local\programs\py
thon\python38-32\lib\site-packages (1.4.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\mahima-pc\appdata\local\progr
ams\python\python38-32\lib\site-packages (from pandas) (2020.5)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\users\mahima-pc\appdata\l
ocal\programs\python\python38-32\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: numpy>=1.18.5 in c:\users\mahima-pc\appdata\local\prog
rams\python\python38-32\lib\site-packages (from pandas) (1.20.3)
Requirement already satisfied: six>=1.5 in c:\users\mahima-pc\appdata\roaming\python
\python38\site-packages (from python-dateutil>=2.8.1->pandas) (1.14.0)
Note: you may need to restart the kernel to use updated packages.
```

```
WARNING: You are using pip version 21.1.2; however, version 22.0.4 is available.
You should consider upgrading via the 'c:\users\mahima-pc\appdata\local\programs\pyth
on\python38-32\python.exe -m pip install --upgrade pip' command.
```

In [14]:

```
import pandas as pd
url="edudata.csv"
```

In [16]:

```
import os
```

In [18]:

```
os.getcwd()
```

Out[18]:

```
'C:\\Users\\Mahima-PC\\2'
```

In [20]:

```
df = pd.read_csv(url)
print(df)
```

```
   gender NationalITy PlaceofBirth       StageID GradeID SectionID   Topic  \
0     NaN          KW       KuwaIT    lowerlevel    G-04         A      IT
1       M          KW          NaN    lowerlevel    G-04         A     NaN
2       M          KW       KuwaIT           NaN    G-04         A      IT
3       M          KW       KuwaIT    lowerlevel    G-04         A      IT
4     NaN          KW       KuwaIT    lowerlevel    G-04         A      IT
5       F          KW       KuwaIT    lowerlevel    G-04         A      IT
6       M          KW       KuwaIT  MiddleSchool    G-07         A     NaN
7       M          KW          NaN  MiddleSchool    G-07         A    Math
8       F          KW       KuwaIT  MiddleSchool    G-07         A    Math
9       F          KW       KuwaIT  MiddleSchool    G-07         B      IT
10      M          KW       KuwaIT  MiddleSchool    G-07         A    Math
11      M          KW       KuwaIT  MiddleSchool    G-07         B    Math
12      M          KW       KuwaIT    lowerlevel     NaN         A      IT
13      M     lebanon      lebanon           NaN    G-08         A    Math
14      F          KW       KuwaIT  MiddleSchool    G-08         A    Math
15      F          KW       KuwaIT  MiddleSchool    G-06         A      IT
16    NaN         NaN       KuwaIT  MiddleSchool    G-07         B      IT
17      M          KW          NaN  MiddleSchool    G-07         A     NaN
18      F          KW       KuwaIT  MiddleSchool    G-07         A      IT
19    NaN          KW       KuwaIT  MiddleSchool    G-07         B      IT
20      F          KW          NaN  MiddleSchool    G-07         A      IT
21      F          KW       KuwaIT  MiddleSchool    G-07         B      IT
22      M          KW       KuwaIT  MiddleSchool    G-07         A      IT
23    NaN          KW       KuwaIT  MiddleSchool    G-07         A      IT
24      M          KW       KuwaIT  MiddleSchool    G-07         B     NaN
25      M          KW          NaN  MiddleSchool    G-07         A      IT
26    NaN          KW       KuwaIT  MiddleSchool    G-07         B      IT
27      M          KW       KuwaIT  MiddleSchool    G-08         A  Arabic

   Semester Relation   cns   dsa  oops  os
0         F   Father   NaN  16.0     2  20
1         F   Father  20.0  20.0     3  25
2         F   Father  10.0   7.0     0  30
3         F   Father   NaN  25.0     5  35
4         F   Father  40.0  50.0    12  50
5         F   Father  42.0  30.0    13  70
6         F   Father  35.0  12.0     0  17
7         F      NaN   NaN   NaN    15  22
8         F   Father  12.0  21.0    16  50
9         F   Father   NaN  80.0    25  70
10        F   Father  50.0  88.0    30  80
11        F   Father  19.0   6.0    19  12
12        F   Father   5.0   1.0     0  11
13        F   Father  20.0  14.0    12  19
14        F      NaN   NaN  70.0    44  60
15        F   Father  30.0  40.0    22  66
16        F   Father  36.0  30.0    20  80
17        F   Father   NaN  13.0    35  90
18        F      Mum  69.0  15.0    36  96
19        F      Mum  70.0  50.0    40  99
20        F   Father   NaN  60.0    33  90
21        F   Father  10.0  12.0     4  80
22        F   Father  15.0  21.0     2  90
23        F   Father   2.0   0.0     2  50
24        F   Father   0.0   2.0     3  70
25        F   Father   8.0   7.0    30  40
26        F   Father  19.0  19.0    25  40
27        F   Father  25.0  15.0    12  33
```

In [21]:

```
#drop the whole row which having NULL value
df.dropna(inplace=True)
print(df.isnull().sum())
df.shape
#these changes not reflect with your dataset , only change in curr data frame
#as you again read dataset, NULL are there as before
```

```
gender          0
NationalITy     0
PlaceofBirth    0
StageID         0
GradeID         0
SectionID       0
Topic           0
Semester        0
Relation        0
cns             0
dsa             0
oops            0
os              0
dtype: int64
```

Out[21]:

```
(9, 13)
```

In [23]:

```
import pandas as pd
#Dataset CSV
url = "edudata.csv"
df = pd.read_csv(url)
```

In [26]:

```
import numpy as np
```

In [27]:

```
#imputation by mean
df["cns"]=df["cns"].replace(np.NAN,df["cns"].mean())

print(df["cns"])
```

```
0     25.571429
1     20.000000
2     10.000000
3     25.571429
4     40.000000
5     42.000000
6     35.000000
7     25.571429
8     12.000000
9     25.571429
10    50.000000
11    19.000000
12     5.000000
13    20.000000
14    25.571429
15    30.000000
16    36.000000
17    25.571429
18    69.000000
19    70.000000
20    25.571429
21    10.000000
22    15.000000
23     2.000000
24     0.000000
25     8.000000
26    19.000000
27    25.000000
Name: cns, dtype: float64
```

In [28]:

```
import pandas as pd
import numpy as np
#Dataset CSV
url = "edudata.csv"
df = pd.read_csv(url)
```

In [29]:

```
#imputation by median

df["cns"]=df["cns"].replace(np.NAN,df["cns"].median())
print(df["cns"])
```

```
0     20.0
1     20.0
2     10.0
3     20.0
4     40.0
5     42.0
6     35.0
7     20.0
8     12.0
9     20.0
10    50.0
11    19.0
12     5.0
13    20.0
14    20.0
15    30.0
16    36.0
17    20.0
18    69.0
19    70.0
20    20.0
21    10.0
22    15.0
23     2.0
24     0.0
25     8.0
26    19.0
27    25.0
Name: cns, dtype: float64
```

In [30]:

```
#Dataset CSV
url = "edudata.csv"
df = pd.read_csv(url)
```

In [31]:

```
#imputation by median

import statistics
df["cns"]=df["cns"].replace(np.NAN,statistics.mode(df["cns"]))
print(df["cns"])
```

```
0     20.0
1     20.0
2     10.0
3     20.0
4     40.0
5     42.0
6     35.0
7     20.0
8     12.0
9     20.0
10    50.0
11    19.0
12     5.0
13    20.0
14    20.0
15    30.0
16    36.0
17    20.0
18    69.0
19    70.0
20    20.0
21    10.0
22    15.0
23     2.0
24     0.0
25     8.0
26    19.0
27    25.0
Name: cns, dtype: float64
```

In [33]:

```
#Dataset CSV
url = "edudata.csv"
df = pd.read_csv(url)
```

In [34]:

```
#imputation by interpolation -linear
df["cns"]=df["cns"].interpolate(method='linear',limit_direction='forward',axis=0)
print(df["cns"])
```

```
0      NaN
1     20.0
2     10.0
3     25.0
4     40.0
5     42.0
6     35.0
7     23.5
8     12.0
9     31.0
10    50.0
11    19.0
12     5.0
13    20.0
14    25.0
15    30.0
16    36.0
17    52.5
18    69.0
19    70.0
20    40.0
21    10.0
22    15.0
23     2.0
24     0.0
25     8.0
26    19.0
27    25.0
Name: cns, dtype: float64
```

In [36]:

```
#Dataset CSV
url = "edudata.csv"
df = pd.read_csv(url)

print(df.isnull().sum())
df.shape
```

```
gender          6
NationalITy     1
PlaceofBirth    5
StageID         2
GradeID         1
SectionID       0
Topic           4
Semester        0
Relation        2
cns             7
dsa             1
oops            0
os              0
dtype: int64
```

Out[36]:

```
(28, 13)
```

In [37]:

```
#replace categorical variable with random value
df["gender"]=df["gender"].fillna('unknow')
print(df["gender"])
```

```
0     unknow
1          M
2          M
3          M
4     unknow
5          F
6          M
7          M
8          F
9          F
10         M
11         M
12         M
13         M
14         F
15         F
16    unknow
17         M
18         F
19    unknow
20         F
21         F
22         M
23    unknow
24         M
25         M
26    unknow
27         M
Name: gender, dtype: object
```

In [38]:

```
#Dataset CSV
url = "edudata.csv"
df = pd.read_csv(url)

#replace categorical variable with previous value
df["gender"]=df["gender"].fillna(method='ffill')
print(df["gender"])
```

```
0     NaN
1       M
2       M
3       M
4       M
5       F
6       M
7       M
8       F
9       F
10      M
11      M
12      M
13      M
14      F
15      F
16      F
17      M
18      F
19      F
20      F
21      F
22      M
23      M
24      M
25      M
26      M
27      M
Name: gender, dtype: object
```

In [39]:

```
#Dataset CSV
url = "edudata.csv"
df = pd.read_csv(url)

#create the inconsistent data(as any no is not the value for gender)
df["gender"]=df["gender"].fillna(100)
print(df["gender"])
```

```
0     100
1       M
2       M
3       M
4     100
5       F
6       M
7       M
8       F
9       F
10      M
11      M
12      M
13      M
14      F
15      F
16    100
17      M
18      F
19    100
20      F
21      F
22      M
23    100
24      M
25      M
26    100
27      M
Name: gender, dtype: object
```

In [40]:

```
#so we replace the inconsistent data with NULL value
cnt=0;
for row in df["gender"]:
    try:
        int(row)
        df.loc[cnt,"gender"]=np.nan
    except ValueError:
        pass
    cnt+=1
```

In [41]:

```
#so the value with 100 replace by NULL
print(df["gender"])
```

```
0     NaN
1       M
2       M
3       M
4     NaN
5       F
6       M
7       M
8       F
9       F
10      M
11      M
12      M
13      M
14      F
15      F
16    NaN
17      M
18      F
19    NaN
20      F
21      F
22      M
23    NaN
24      M
25      M
26    NaN
27      M
Name: gender, dtype: object
```

```
In [ ]:
```