```
import numpy as np
import pandas as pd
dataset = pd.read_csv('data (1).csv')

dataset.shape

(119, 5)

dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119 entries, 0 to 118
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   COUNTY           119 non-null    object
 1   REGION           119 non-null    object
 2   AGE_GROUP        119 non-null    object
 3   TOTAL_CASE       110 non-null    float64
 4   POP_DENOMINATOR  111 non-null    float64
dtypes: float64(2), object(3)
memory usage: 4.8+ KB

dataset.describe()

         TOTAL_CASE  POP_DENOMINATOR
count    110.000000       111.000000
mean    1516.727273      4580.045045
std     2419.278446      8610.610007
min       93.000000       348.000000
25%      489.750000      1453.500000
50%      817.500000      2451.000000
75%     1509.750000      4158.000000
max    16287.000000     54039.000000

dataset = dataset.drop(['POP_DENOMINATOR'],axis=1)

dataset

        COUNTY REGION    AGE_GROUP  TOTAL_CASE
0     Anderson    ETR    0-4 YEARS      1276.0
1     Anderson    ETR   5-11 YEARS      2139.0
2     Anderson    ETR  12-18 YEARS         NaN
3      Bedford    SCR    0-4 YEARS       918.0
4      Bedford    SCR   5-11 YEARS      1223.0
..         ...    ...          ...         ...
114  Henderson    WTR    0-4 YEARS       427.0
115  Henderson    WTR   5-11 YEARS       764.0
116  Henderson    WTR  12-18 YEARS      1207.0
117      Henry    WTR    0-4 YEARS       450.0
118      Henry    WTR   5-11 YEARS       591.0
```

```
[119 rows x 4 columns]
```

```
dataset.isnull().sum()
```

```
COUNTY        0
REGION        0
AGE_GROUP     0
TOTAL_CASE    9
dtype: int64
```

```
dataset.isna().sum()
dataset.head()
```

```
     COUNTY REGION    AGE_GROUP  TOTAL_CASE
0  Anderson    ETR    0-4 YEARS      1276.0
1  Anderson    ETR   5-11 YEARS      2139.0
2  Anderson    ETR  12-18 YEARS         NaN
3   Bedford    SCR    0-4 YEARS       918.0
4   Bedford    SCR   5-11 YEARS      1223.0
```

```
#creating dummies columns
df_cat = pd.get_dummies(dataset['AGE_GROUP'],drop_first=False)
df_cat
```

```
     0-4 YEARS  12-18 YEARS  5-11 YEARS
0         True        False       False
1        False        False        True
2        False         True       False
3         True        False       False
4        False        False        True
..         ...          ...         ...
114       True        False       False
115      False        False        True
116      False         True       False
117       True        False       False
118      False        False        True

[119 rows x 3 columns]
```

```
df_cat1 = pd.get_dummies(dataset['REGION'],drop_first=False)
df_cat1
```

```
       CHR    ETR    MCR    NDR    NER    SCR    SER    UCR    WTR
0    False   True  False  False  False  False  False  False  False
1    False   True  False  False  False  False  False  False  False
2    False   True  False  False  False  False  False  False  False
3    False  False  False  False  False   True  False  False  False
4    False  False  False  False  False   True  False  False  False
..     ...    ...    ...    ...    ...    ...    ...    ...    ...
114  False  False  False  False  False  False  False  False   True
```

```
115  False  False  False  False  False  False  False  False   True
116  False  False  False  False  False  False  False  False   True
117  False  False  False  False  False  False  False  False   True
118  False  False  False  False  False  False  False  False   True

[119 rows x 9 columns]
```

```
#joining dummies columns
dataset = pd.concat([dataset,df_cat,df_cat1],axis=1)
dataset
```

```
        COUNTY REGION    AGE_GROUP  TOTAL_CASE  0-4 YEARS  12-18 YEARS
\
0     Anderson    ETR    0-4 YEARS      1276.0       True         False

1     Anderson    ETR   5-11 YEARS      2139.0      False         False

2     Anderson    ETR  12-18 YEARS         NaN      False          True

3      Bedford    SCR    0-4 YEARS       918.0       True         False

4      Bedford    SCR   5-11 YEARS      1223.0      False         False

..         ...    ...          ...         ...        ...           ...

114  Henderson    WTR    0-4 YEARS       427.0       True         False

115  Henderson    WTR   5-11 YEARS       764.0      False         False

116  Henderson    WTR  12-18 YEARS      1207.0      False          True

117      Henry    WTR    0-4 YEARS       450.0       True         False

118      Henry    WTR   5-11 YEARS       591.0      False         False


     5-11 YEARS    CHR    ETR    MCR    NDR    NER    SCR    SER
UCR    WTR
0         False  False   True  False  False  False  False  False
False  False
1          True  False   True  False  False  False  False  False
False  False
2         False  False   True  False  False  False  False  False
False  False
3         False  False  False  False  False  False   True  False
False  False
4          True  False  False  False  False  False   True  False
False  False
..          ...    ...    ...    ...    ...    ...    ...    ...    ..
.    ...
114       False  False  False  False  False  False  False  False
```

```
False   True
115       True  False  False  False  False  False  False  False
False   True
116      False  False  False  False  False  False  False  False
False   True
117      False  False  False  False  False  False  False  False
False   True
118       True  False  False  False  False  False  False  False
False   True

[119 rows x 16 columns]
```

```
dataset.drop('REGION',axis=1,inplace=True)
dataset.drop('AGE_GROUP',axis=1,inplace=True)
```

```
dataset
```

```
        COUNTY  TOTAL_CASE  0-4 YEARS  12-18 YEARS  5-11 YEARS    CHR
ETR  \
0     Anderson      1276.0       True        False       False  False
True
1     Anderson      2139.0      False        False        True  False
True
2     Anderson         NaN      False         True       False  False
True
3      Bedford       918.0       True        False       False  False
False
4      Bedford      1223.0      False        False        True  False
False
..         ...         ...        ...          ...         ...    ...
...
114  Henderson       427.0       True        False       False  False
False
115  Henderson       764.0      False        False        True  False
False
116  Henderson      1207.0      False         True       False  False
False
117      Henry       450.0       True        False       False  False
False
118      Henry       591.0      False        False        True  False
False

       MCR    NDR    NER    SCR    SER    UCR    WTR
0    False  False  False  False  False  False  False
1    False  False  False  False  False  False  False
2    False  False  False  False  False  False  False
3    False  False  False   True  False  False  False
4    False  False  False   True  False  False  False
..     ...    ...    ...    ...    ...    ...    ...
114  False  False  False  False  False  False   True
```

```
115  False  False  False  False  False  False   True
116  False  False  False  False  False  False   True
117  False  False  False  False  False  False   True
118  False  False  False  False  False  False   True

[119 rows x 14 columns]
```

```
dataset.isna().sum()
```

```
COUNTY          0
TOTAL_CASE      9
0-4 YEARS       0
12-18 YEARS     0
5-11 YEARS      0
CHR             0
ETR             0
MCR             0
NDR             0
NER             0
SCR             0
SER             0
UCR             0
WTR             0
dtype: int64
```

```
#filling missing values
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='median')
dataset['TOTAL_CASE'] = imputer.fit_transform(dataset[['TOTAL_CASE']])
```

```
dataset
```

```
        COUNTY  TOTAL_CASE  0-4 YEARS  12-18 YEARS  5-11 YEARS    CHR
ETR  \
0     Anderson      1276.0       True        False       False  False
True
1     Anderson      2139.0      False        False        True  False
True
2     Anderson       817.5      False         True       False  False
True
3      Bedford       918.0       True        False       False  False
False
4      Bedford      1223.0      False        False        True  False
False
..         ...         ...        ...          ...         ...    ...
...
114  Henderson       427.0       True        False       False  False
False
115  Henderson       764.0      False        False        True  False
False
```

```
116  Henderson      1207.0      False         True       False  False
False
117      Henry       450.0       True        False       False  False
False
118      Henry       591.0      False        False        True  False
False
```

```
       MCR    NDR    NER    SCR    SER    UCR    WTR
0    False  False  False  False  False  False  False
1    False  False  False  False  False  False  False
2    False  False  False  False  False  False  False
3    False  False  False   True  False  False  False
4    False  False  False   True  False  False  False
..     ...    ...    ...    ...    ...    ...    ...
114  False  False  False  False  False  False   True
115  False  False  False  False  False  False   True
116  False  False  False  False  False  False   True
117  False  False  False  False  False  False   True
118  False  False  False  False  False  False   True

[119 rows x 14 columns]
```

```
dataset.isna().sum()
```

```
COUNTY          0
TOTAL_CASE      0
0-4 YEARS       0
12-18 YEARS     0
5-11 YEARS      0
CHR             0
ETR             0
MCR             0
NDR             0
NER             0
SCR             0
SER             0
UCR             0
WTR             0
dtype: int64
```

```
dataset
```

```
        COUNTY  TOTAL_CASE  0-4 YEARS  12-18 YEARS  5-11 YEARS    CHR
ETR  \
0     Anderson      1276.0       True        False       False  False
True
1     Anderson      2139.0      False        False        True  False
True
2     Anderson       817.5      False         True       False  False
True
```

```
3      Bedford       918.0       True        False       False  False
False
4      Bedford      1223.0      False        False        True  False
False
..         ...         ...        ...          ...         ...    ...
...
114  Henderson       427.0       True        False       False  False
False
115  Henderson       764.0      False        False        True  False
False
116  Henderson      1207.0      False         True       False  False
False
117      Henry       450.0       True        False       False  False
False
118      Henry       591.0      False        False        True  False
False

       MCR    NDR    NER    SCR    SER    UCR    WTR
0    False  False  False  False  False  False  False
1    False  False  False  False  False  False  False
2    False  False  False  False  False  False  False
3    False  False  False   True  False  False  False
4    False  False  False   True  False  False  False
..     ...    ...    ...    ...    ...    ...    ...
114  False  False  False  False  False  False   True
115  False  False  False  False  False  False   True
116  False  False  False  False  False  False   True
117  False  False  False  False  False  False   True
118  False  False  False  False  False  False   True

[119 rows x 14 columns]
```