```
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):
                       Non-Null Count
#
     Column
                                        Dtype
0
                       119 non-null
     COUNTY
                                        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
         110.000000
count
                           111.000000
        1516.727273
mean
                          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
       16287.000000
                         54039.000000
max
dataset = dataset.drop(['POP DENOMINATOR'],axis=1)
dataset
        COUNTY REGION
                          AGE GROUP
                                      TOTAL CASE
                          0-4 YEARS
0
      Anderson
                                          1276.0
                   ETR
1
                                          2139.0
      Anderson
                   ETR
                         5-11 YEARS
2
                        12-18 YEARS
      Anderson
                   ETR
                                             NaN
3
       Bedford
                          0-4 YEARS
                   SCR
                                           918.0
4
       Bedford
                   SCR
                         5-11 YEARS
                                          1223.0
                   . . .
                                             . . .
114
     Henderson
                  WTR
                          0-4 YEARS
                                           427.0
                         5-11 YEARS
115
     Henderson
                  WTR
                                           764.0
116
     Henderson
                  WTR 12-18 YEARS
                                          1207.0
                          0-4 YEARS
117
         Henry
                  WTR
                                           450.0
118
                         5-11 YEARS
                                           591.0
         Henry
                  WTR
```

```
[119 rows x 4 columns]
dataset.isnull().sum()
COUNTY
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
                     5-11 YEARS
1
  Anderson
               ETR
                                     2139.0
2
  Anderson
               ETR 12-18 YEARS
                                         NaN
3
                      0-4 YEARS
    Bedford
               SCR
                                      918.0
    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
                      False
115
         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
                                        False False False
     False
             True False
                          False
                                                              False
3
     False
            False False
                          False
                                 False
                                         True False False
                                                              False
4
     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
                                             True
116
  False 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 9 columns]

#joining dummies columns
dataset = pd.concat([dataset,df\_cat,df\_cat1],axis=1)

data	set = pa.cor set	icat ([ua	ataset, o	T_Cat,	ur_catij	,axis=1	)		
	COUNTY F	REGION	AGE_G	ROUP	TOTAL_CA	SE 0-4	YEARS	12-18 YE	ARS
0	Anderson	ETR	0-4 Y	EARS	1276	.0	True	Fa	lse
1	Anderson	ETR	5-11 Y	EARS	2139	.0	False	Fa	lse
2	Anderson	ETR	12-18 Y	EARS	N	aN	False	Т	rue
3	Bedford	SCR	0-4 Y	EARS	918	.0	True	Fa	lse
4	Bedford	SCR	5-11 Y	EARS	1223	.0	False	Fa	lse
114	Henderson	WTR	0-4 Y	EARS	427	.0	True	Fa	lse
115	Henderson	WTR	5-11 Y	EARS	764	.0	False	Fa	lse
116	Henderson	WTR	12-18 Y	EARS	1207	.0	False	Т	rue
117	Henry	WTR	0-4 Y	EARS	450	.0	True	Fa	lse
118	Henry	WTR	5-11 Y	EARS	591	.0	False	Fa	lse
	5-11 YEARS	CHR	ETR	MCR	NDR	NER	SCR	SER	
UCR 0	WTR False	False	True	False	False	False	False	False	
Fals 1	e False	False	True	False		False	False	False	
Fals		ratse		ratse	ratse	ratse	ratse		
2 Fals	False e False	False	True	False	False	False	False	False	
3	False	False	False	False	False	False	True	False	
4	e False True e False	False	False	False	False	False	True	False	
114		False	False	False	False	False	False	False	

```
False
       True
          True False False False False False False
115
False
       True
116
         False False False False
                                           False False False
False
       True
117
         False
                False False False
                                           False False False
False
       True
118
          True 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 \
                   1276.0
     Anderson
                               True
                                           False
                                                       False False
True
1
     Anderson
                   2139.0
                              False
                                           False
                                                       True
                                                             False
True
                      NaN
                              False
                                            True
                                                       False False
     Anderson
True
3
      Bedford
                    918.0
                               True
                                           False
                                                       False False
False
                   1223.0
      Bedford
                              False
                                           False
                                                       True False
False
114 Henderson
                    427.0
                               True
                                           False
                                                       False False
False
115
                    764.0
                              False
                                           False
                                                       True False
    Henderson
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
                         False
0
     False
          False
                  False
                               False
                                      False
                                             False
                               False
1
    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
                          False
116
     False
            False
                   False
                                 False
                                        False
                                                True
117
     False
            False False
                          False
                                 False
                                        False
                                                True
118 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
               0
NDR
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 \
                    1276.0
      Anderson
                                 True
                                             False
                                                         False False
True
      Anderson
                    2139.0
                                False
                                             False
                                                          True False
1
True
2
      Anderson
                     817.5
                                False
                                              True
                                                         False False
True
3
                     918.0
       Bedford
                                 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 H False	Henders	on	1207.0	)	False		True	False	False
117	Hen	ıry	450.0	)	True		False	False	False
False 118 False	Hen	ıry	591.6	)	False		False	True	False
1 F 2 F 3 F	False False False	NDR False False False False		SCR False False True True	False False False False	UCR False False False False	False False False False		
115 F 116 F 117 F	False False	False False False False False	False False False False False	False False False False False	False False False False False	False False False False False	True True True		
[119 r	rows x	14 col	umns]						
datase	et.isna	ı(). <mark>sum</mark>	()						
5-11 Y CHR ETR MCR NDR NER SCR SER UCR WTR	_CASE EARS YEARS	0 0 0 0 0 0 0 0							
datase	et COUN	ITV TO	TAL CASE	= 0 <sub>-</sub> 1	YEARS	12-18	VEARS	5-11 YEARS	CHR
ETR \	\ Anders		1276.0		True		False	False	False
True 1	Anders		2139.6		False		False	True	False
True 2 True	Anders		817.5		False		True	False	False

3	Bedf	ord	918.	0	True	F	alse	False	False
False	e Bedf	ord	1223.	0	False	F	alse	True	False
False						, 6, 65 6			
				•					
114 Henderson		427.0		True F		alse	False	False	
	False 115 Henderson		764.0		False F		alse	True	False
False	е								
116 False	Hender	son	1207.0		False		True	False	False
117	Не	nry	450.	0	True	F	alse	False	False
False 118 Henry		591.0		False	.se False		True	False	
False									
	MCR	NDR	NER	SCR	SER	UCR	WTR		
0 1	False False	False False	False False	False False	False False	False False	False False		
2	False	False	False	False	False	False	False		
3 4	False False	False False	False False	True True	False False	False False	False False		
114 115	False False	False False	False False	False False	False False	False False	True True		
116	False	False	False	False	False	False	True		
117 118	False False	False False	False False	False False	False False	False False	True True		
				racsc	10050	1 4 6 5 6	1140		
[119	rows x	14 col	umns]						