In [1]:

```
import pandas as pd
import numpy as np
```

In [3]:

```
df=pd.read_csv("C:/sameer/AirQuality.csv",encoding='cp1252')
```

```
C:\Users\samir\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:2785: DtypeWarning: Colu
mns (0) have mixed types.Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [4]:

```
df.head(5)
```

Out[4]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	date
0	150	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN	NaN	1990-02-01
1	151	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	NaN	NaN	1990-02-01
2	152	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	NaN	1990-02-01
3	150	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	NaN	1990-03-01
4	151	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	NaN	NaN	1990-03-01

In [5]:

```
df.describe()
```

Out[5]:

	so2	no2	rspm	spm	pm2_5
count	401096.000000	419509.000000	395520.000000	198355.000000	9314.000000
mean	10.829414	25.809623	108.832784	220.783480	40.791467
std	11.177187	18.503086	74.872430	151.395457	30.832525
min	0.000000	0.000000	0.000000	0.000000	3.000000
25%	5.000000	14.000000	56.000000	111.000000	24.000000
50%	8.000000	22.000000	90.000000	187.000000	32.000000
75%	13.700000	32.200000	142.000000	296.000000	46.000000
max	909.000000	876.000000	6307.033333	3380.000000	504.000000

In [6]:

```
df.shape
```

Out[6]:

```
(435742, 13)
```

In [7]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   stn_code                     291665 non-null  object 
 1   sampling_date                435739 non-null  object 
 2   state                        435742 non-null  object 
 3   location                     435739 non-null  object 
 4   agency                       286261 non-null  object 
 5   type                         430349 non-null  object 
 6   so2                          401096 non-null  float64
 7   no2                          419509 non-null  float64
 8   rspm                         395520 non-null  float64
 9   spm                          198355 non-null  float64
 10  location_monitoring_station  408251 non-null  object 
 11  pm2_5                        9314 non-null    float64
 12  date                         435735 non-null  object 
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

In [8]:

```
df.isnull().sum()
```

Out[8]:

```
stn_code                       144077
sampling_date                       3
state                               0
location                            3
agency                         149481
type                             5393
so2                             34646
no2                             16233
rspm                            40222
spm                            237387
location_monitoring_station     27491
pm2_5                          426428
date                                7
dtype: int64
```

In [9]:

```
df.count()
```

Out[9]:

```
stn_code                       291665
sampling_date                  435739
state                          435742
location                       435739
agency                         286261
type                           430349
so2                            401096
no2                            419509
rspm                           395520
spm                            198355
location_monitoring_station    408251
pm2_5                            9314
date                           435735
dtype: int64
```

In [11]:

```
df.describe()
```

Out[11]:

	so2	no2	rspm	spm	pm2_5
count	401096.000000	419509.000000	395520.000000	198355.000000	9314.000000
mean	10.829414	25.809623	108.832784	220.783480	40.791467
std	11.177187	18.503086	74.872430	151.395457	30.832525
min	0.000000	0.000000	0.000000	0.000000	3.000000
25%	5.000000	14.000000	56.000000	111.000000	24.000000
50%	8.000000	22.000000	90.000000	187.000000	32.000000
75%	13.700000	32.200000	142.000000	296.000000	46.000000
max	909.000000	876.000000	6307.033333	3380.000000	504.000000

In [12]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 435742 entries, 0 to 435741
Data columns (total 13 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   stn_code                     291665 non-null  object 
 1   sampling_date                435739 non-null  object 
 2   state                        435742 non-null  object 
 3   location                     435739 non-null  object 
 4   agency                       286261 non-null  object 
 5   type                         430349 non-null  object 
 6   so2                          401096 non-null  float64
 7   no2                          419509 non-null  float64
 8   rspm                         395520 non-null  float64
 9   spm                          198355 non-null  float64
 10  location_monitoring_station  408251 non-null  object 
 11  pm2_5                        9314 non-null    float64
 12  date                         435735 non-null  object 
dtypes: float64(5), object(8)
memory usage: 43.2+ MB
```

In [13]:

```
df = df.drop(['stn_code','agency', 'location_monitoring_station'],axis=1)
```

In [14]:

```
df.isna().sum()
```

Out[14]:

```
sampling_date         3
state                 0
location              3
type               5393
so2               34646
no2               16233
rspm              40222
spm              237387
pm2_5            426428
date                  7
dtype: int64
```

In [15]:

```
df=df.dropna(subset=['date'])
```

In [16]:

```
df.isna().sum()
```

Out[16]:

```
sampling_date         0
state                 0
location              0
type               5390
so2               34643
no2               16230
rspm              40219
spm              237380
pm2_5            426421
date                  0
dtype: int64
```

In [18]:

```
df.columns
```

Out[18]:

```
Index(['sampling_date', 'state', 'location', 'type', 'so2', 'no2', 'rspm',
       'spm', 'pm2_5', 'date'],
      dtype='object')
```

In [20]:

```
df['type'].unique()
```

Out[20]:

```
array(['Residential, Rural and other Areas', 'Industrial Area', nan,
       'Sensitive Area', 'Industrial Areas', 'Residential and others',
       'Sensitive Areas', 'Industrial', 'Residential', 'RIRUO',
       'Sensitive'], dtype=object)
```

In [21]:

```
types = {

    "Residential" : "K",
    "Residential and others" : "RO",
    "Industrial Area":"I" ,
    "Industrial Areas" : "I",
    "Industrial" : "I" ,
    "Sensitive Area": "s",
    "Sensitive Areas":"s",
    "Sensitive":"s",
    "NaN":"PRO",
    "Residential, Rural and other Areas":"MO"
 }
```

In [22]:

```
df.type = df.type.replace(types)
```

In [23]:

```
df['type'].unique()
```

Out[23]:

```
array(['MO', 'I', nan, 's', 'RO', 'K', 'RIRUO'], dtype=object)
```

In [24]:

```
df.head()
```

Out[24]:

	sampling_date	state	location	type	so2	no2	rspm	spm	pm2_5	date
0	February - M021990	Andhra Pradesh	Hyderabad	MO	4.8	17.4	NaN	NaN	NaN	1990-02-01
1	February - M021990	Andhra Pradesh	Hyderabad	I	3.1	7.0	NaN	NaN	NaN	1990-02-01
2	February - M021990	Andhra Pradesh	Hyderabad	MO	6.2	28.5	NaN	NaN	NaN	1990-02-01
3	March - M031990	Andhra Pradesh	Hyderabad	MO	6.3	14.7	NaN	NaN	NaN	1990-03-01
4	March - M031990	Andhra Pradesh	Hyderabad	I	4.7	7.5	NaN	NaN	NaN	1990-03-01

In [25]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 435735 entries, 0 to 435738
Data columns (total 10 columns):
 #   Column         Non-Null Count   Dtype  
---  ------         --------------   -----  
 0   sampling_date  435735 non-null  object 
 1   state          435735 non-null  object 
 2   location       435735 non-null  object 
 3   type           430345 non-null  object 
 4   so2            401092 non-null  float64
 5   no2            419505 non-null  float64
 6   rspm           395516 non-null  float64
 7   spm            198355 non-null  float64
 8   pm2_5          9314 non-null    float64
 9   date           435735 non-null  object 
dtypes: float64(5), object(5)
memory usage: 36.6+ MB
```

In [26]:

```
df['date']=pd.to_datetime(df['date'], errors="coerce")
df.head(5)
```

Out[26]:

	sampling_date	state	location	type	so2	no2	rspm	spm	pm2_5	date
0	February - M021990	Andhra Pradesh	Hyderabad	MO	4.8	17.4	NaN	NaN	NaN	1990-02-01
1	February - M021990	Andhra Pradesh	Hyderabad	I	3.1	7.0	NaN	NaN	NaN	1990-02-01
2	February - M021990	Andhra Pradesh	Hyderabad	MO	6.2	28.5	NaN	NaN	NaN	1990-02-01
3	March - M031990	Andhra Pradesh	Hyderabad	MO	6.3	14.7	NaN	NaN	NaN	1990-03-01
4	March - M031990	Andhra Pradesh	Hyderabad	I	4.7	7.5	NaN	NaN	NaN	1990-03-01

In [27]:

```
df['year']=df.date.dt.year
df.head()
```

Out[27]:

	sampling_date	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	February - M021990	Andhra Pradesh	Hyderabad	MO	4.8	17.4	NaN	NaN	NaN	1990-02-01	1990
1	February - M021990	Andhra Pradesh	Hyderabad	I	3.1	7.0	NaN	NaN	NaN	1990-02-01	1990
2	February - M021990	Andhra Pradesh	Hyderabad	MO	6.2	28.5	NaN	NaN	NaN	1990-02-01	1990
3	March - M031990	Andhra Pradesh	Hyderabad	MO	6.3	14.7	NaN	NaN	NaN	1990-03-01	1990
4	March - M031990	Andhra Pradesh	Hyderabad	I	4.7	7.5	NaN	NaN	NaN	1990-03-01	1990

In [28]:

```
COLS = ['so2','no2', 'rspm', 'spm', 'pm2_5']
```

In [29]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 435735 entries, 0 to 435738
Data columns (total 11 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   sampling_date  435735 non-null  object
 1   state          435735 non-null  object
 2   location       435735 non-null  object
 3   type           430345 non-null  object
 4   so2            401092 non-null  float64
 5   no2            419505 non-null  float64
 6   rspm           395516 non-null  float64
 7   spm            198355 non-null  float64
 8   pm2_5          9314 non-null    float64
 9   date           435735 non-null  datetime64[ns]
 10  year           435735 non-null  int64
dtypes: datetime64[ns](1), float64(5), int64(1), object(4)
memory usage: 39.9+ MB
```

In [30]:

```
import numpy as np
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values = np.nan, strategy='mean')
```

In [31]:

```
df[COLS] = imputer.fit_transform(df[COLS])
```

In [33]:

```
df.head()
```

Out[33]:

	sampling_date	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	February - M021990	Andhra Pradesh	Hyderabad	MO	4.8	17.4	108.833091	220.78348	40.791467	1990-02-01	1990
1	February - M021990	Andhra Pradesh	Hyderabad	I	3.1	7.0	108.833091	220.78348	40.791467	1990-02-01	1990
2	February - M021990	Andhra Pradesh	Hyderabad	MO	6.2	28.5	108.833091	220.78348	40.791467	1990-02-01	1990
3	March - M031990	Andhra Pradesh	Hyderabad	MO	6.3	14.7	108.833091	220.78348	40.791467	1990-03-01	1990
4	March - M031990	Andhra Pradesh	Hyderabad	I	4.7	7.5	108.833091	220.78348	40.791467	1990-03-01	1990

In [34]:

```
df.nunique()
```

Out[34]:

```
sampling_date    5482
state              34
location          304
type                6
so2              4198
no2              6865
rspm             6066
spm              6669
pm2_5             434
date             5067
year               29
dtype: int64
```

In [35]:

```
df.duplicated().sum()
```

Out[35]:

```
1135
```

In [36]:

```
df.drop_duplicates()
```

Out[36]:

	sampling_date	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	February - M021990	Andhra Pradesh	Hyderabad	MO	4.8	17.4	108.833091	220.78348	40.791467	1990-02-01	1990
1	February - M021990	Andhra Pradesh	Hyderabad	I	3.1	7.0	108.833091	220.78348	40.791467	1990-02-01	1990
2	February - M021990	Andhra Pradesh	Hyderabad	MO	6.2	28.5	108.833091	220.78348	40.791467	1990-02-01	1990
3	March - M031990	Andhra Pradesh	Hyderabad	MO	6.3	14.7	108.833091	220.78348	40.791467	1990-03-01	1990
4	March - M031990	Andhra Pradesh	Hyderabad	I	4.7	7.5	108.833091	220.78348	40.791467	1990-03-01	1990
...	...	...	...	...	...	...	...	...	...	...	...
435734	15-12-15	West Bengal	ULUBERIA	RIRUO	20.0	44.0	148.000000	220.78348	40.791467	2015-12-15	2015
435735	18-12-15	West Bengal	ULUBERIA	RIRUO	17.0	44.0	131.000000	220.78348	40.791467	2015-12-18	2015
435736	21-12-15	West Bengal	ULUBERIA	RIRUO	18.0	45.0	140.000000	220.78348	40.791467	2015-12-21	2015
435737	24-12-15	West Bengal	ULUBERIA	RIRUO	22.0	50.0	143.000000	220.78348	40.791467	2015-12-24	2015
435738	29-12-15	West Bengal	ULUBERIA	RIRUO	20.0	46.0	171.000000	220.78348	40.791467	2015-12-29	2015

434600 rows × 11 columns

In [37]:

```
df.head()
```

Out[37]:

	sampling_date	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	February - M021990	Andhra Pradesh	Hyderabad	MO	4.8	17.4	108.833091	220.78348	40.791467	1990-02-01	1990
1	February - M021990	Andhra Pradesh	Hyderabad	I	3.1	7.0	108.833091	220.78348	40.791467	1990-02-01	1990
2	February - M021990	Andhra Pradesh	Hyderabad	MO	6.2	28.5	108.833091	220.78348	40.791467	1990-02-01	1990
3	March - M031990	Andhra Pradesh	Hyderabad	MO	6.3	14.7	108.833091	220.78348	40.791467	1990-03-01	1990
4	March - M031990	Andhra Pradesh	Hyderabad	I	4.7	7.5	108.833091	220.78348	40.791467	1990-03-01	1990

In [38]:

```
df['type'].value_counts()
```

Out[38]:

```
MO       179013
I        148069
RO        86791
s         15010
RIRUO      1304
K           158
Name: type, dtype: int64
```

In [39]:

```
df['type'].replace({  'MO':1, 'I':2, 's':3 , 'RO':4, 'K':5, 'RIRUO':6  }, inplace=True)
```

In [40]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 435735 entries, 0 to 435738
Data columns (total 11 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   sampling_date  435735 non-null  object
 1   state          435735 non-null  object
 2   location       435735 non-null  object
 3   type           430345 non-null  float64
 4   so2            435735 non-null  float64
 5   no2            435735 non-null  float64
 6   rspm           435735 non-null  float64
 7   spm            435735 non-null  float64
 8   pm2_5          435735 non-null  float64
 9   date           435735 non-null  datetime64[ns]
 10  year           435735 non-null  int64
dtypes: datetime64[ns](1), float64(6), int64(1), object(3)
memory usage: 39.9+ MB
```

In [41]:

```
df['type']
```

Out[41]:

```
0         1.0
1         2.0
2         1.0
3         1.0
4         2.0
         ...
435734    6.0
435735    6.0
435736    6.0
435737    6.0
435738    6.0
Name: type, Length: 435735, dtype: float64
```

In [42]:

```
from sklearn.preprocessing import LabelEncoder
labelencoder = LabelEncoder()
df['state'] =labelencoder.fit_transform(df['state'])
df.head()
```

Out[42]:

	sampling_date	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	February - M021990	0	Hyderabad	1.0	4.8	17.4	108.833091	220.78348	40.791467	1990-02-01	1990
1	February - M021990	0	Hyderabad	2.0	3.1	7.0	108.833091	220.78348	40.791467	1990-02-01	1990
2	February - M021990	0	Hyderabad	1.0	6.2	28.5	108.833091	220.78348	40.791467	1990-02-01	1990
3	March - M031990	0	Hyderabad	1.0	6.3	14.7	108.833091	220.78348	40.791467	1990-03-01	1990
4	March - M031990	0	Hyderabad	2.0	4.7	7.5	108.833091	220.78348	40.791467	1990-03-01	1990

In [43]:

```
dfAndhra = df[df['state']==0]
```

In [44]:

```
dfAndhra
```

Out[44]:

	sampling_date	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	February - M021990	0	Hyderabad	1.0	4.8	17.4	108.833091	220.78348	40.791467	1990-02-01	1990
1	February - M021990	0	Hyderabad	2.0	3.1	7.0	108.833091	220.78348	40.791467	1990-02-01	1990
2	February - M021990	0	Hyderabad	1.0	6.2	28.5	108.833091	220.78348	40.791467	1990-02-01	1990
3	March - M031990	0	Hyderabad	1.0	6.3	14.7	108.833091	220.78348	40.791467	1990-03-01	1990
4	March - M031990	0	Hyderabad	2.0	4.7	7.5	108.833091	220.78348	40.791467	1990-03-01	1990
...	...	...	...	...	...	...	...	...	...	...	...
26363	13-12-15	0	Rajahmundry	2.0	7.0	13.0	71.000000	220.78348	40.791467	2015-12-13	2015
26364	16-12-15	0	Rajahmundry	2.0	7.0	18.0	77.000000	220.78348	40.791467	2015-12-16	2015
26365	19-12-15	0	Rajahmundry	2.0	8.0	23.0	64.000000	220.78348	40.791467	2015-12-19	2015
26366	22-12-15	0	Rajahmundry	2.0	7.0	19.0	61.000000	220.78348	40.791467	2015-12-22	2015
26367	25-12-15	0	Rajahmundry	2.0	6.0	17.0	71.000000	220.78348	40.791467	2015-12-25	2015

26368 rows × 11 columns

In [45]:

```
dfAndhra['location'].value_counts()
```

Out[45]:

```
Hyderabad         7764
Visakhapatnam     7108
Vijayawada        2093
Chittoor          1003
Tirupati           986
Kurnool            857
Patancheru         698
Guntur             629
Nalgonda           618
Ramagundam         554
Nellore            408
Khammam            385
Warangal           336
Ananthapur         324
Ongole             317
Kadapa             316
Srikakulam         315
Rajahmundry        311
Eluru              300
Vishakhapatnam     297
Kakinada           288
Vizianagaram       282
Sangareddy          85
Karimnagar          67
Nizamabad           27
Name: location, dtype: int64
```

In [46]:

```
from sklearn.preprocessing import OneHotEncoder
onehotencoder = OneHotEncoder(sparse=False, handle_unknown='error', drop='first')
```

In [47]:

```
pd.DataFrame(onehotencoder.fit_transform(dfAndhra[['location']]))
```

Out[47]:

	0	1	2	3	4	5	6	7	8	9	...	14	15	16	17	18	19	20	21	22	23
0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26363	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26364	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26365	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26366	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26367	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

26368 rows × 24 columns

In [48]:

```
dfAndhra['location'].value_counts()
```

Out[48]:

```
Hyderabad         7764
Visakhapatnam     7108
Vijayawada        2093
Chittoor          1003
Tirupati           986
Kurnool            857
Patancheru         698
Guntur             629
Nalgonda           618
Ramagundam         554
Nellore            408
Khammam            385
Warangal           336
Ananthapur         324
Ongole             317
Kadapa             316
Srikakulam         315
Rajahmundry        311
Eluru              300
Vishakhapatnam     297
Kakinada           288
Vizianagaram       282
Sangareddy          85
Karimnagar          67
Nizamabad           27
Name: location, dtype: int64
```

In [49]:

```
df.isnull().sum()
```

Out[49]:

```
sampling_date       0
state               0
location            0
type             5390
so2                 0
no2                 0
rspm                0
spm                 0
pm2_5               0
date                0
year                0
dtype: int64
```

In [50]:

```
df=df.fillna(df.median())
df.isnull().sum()
```

```
C:\Users\samir\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: FutureWarning: DataFrame.mean an
d DataFrame.median with numeric_only=None will include datetime64 and datetime64tz columns in a futu
re version.
  """Entry point for launching an IPython kernel.
```

Out[50]:

```
sampling_date    0
state            0
location         0
type             0
so2              0
no2              0
rspm             0
spm              0
pm2_5            0
date             0
year             0
dtype: int64
```

In [51]:

```
df.describe()
```

Out[51]:

	state	type	so2	no2	rspm	spm	pm2_5	year
count	435735.000000	435735.000000	435735.000000	435735.000000	435735.000000	435735.00000	435735.000000	435735.000000
mean	17.966833	2.035042	10.829428	25.809659	108.833091	220.78348	40.791467	2009.534123
std	9.471742	1.136631	10.723716	18.155263	71.333594	102.14629	4.507577	4.791559
min	0.000000	1.000000	0.000000	0.000000	0.000000	0.00000	3.000000	1987.000000
25%	12.000000	1.000000	5.000000	14.000000	59.000000	203.00000	40.791467	2007.000000
50%	18.000000	2.000000	9.000000	22.300000	97.666667	220.78348	40.791467	2010.000000
75%	26.000000	2.000000	13.000000	32.000000	135.000000	220.78348	40.791467	2013.000000
max	33.000000	6.000000	909.000000	876.000000	6307.033333	3380.00000	504.000000	2015.000000

In [52]:

```
df[df['so2']>100]=0
```

In [53]:

```
import pandas as pd
df=pd.read_csv("C:/sameer/heart - Copy (2).csv")
```

In [54]:

```
df.shape
```

Out[54]:

```
(303, 14)
```

In [55]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

In [57]:

```
df.dtypes
```

Out[57]:

```
age           int64
sex           int64
cp            int64
trestbps      int64
chol          int64
fbs           int64
restecg       int64
thalach       int64
exang         int64
oldpeak     float64
slope         int64
ca            int64
thal          int64
target        int64
dtype: object
```

In [58]:

```
df.nunique()
```

Out[58]:

```
age          41
sex           2
cp            4
trestbps     49
chol        152
fbs           2
restecg       3
thalach      91
exang         2
oldpeak      40
slope         3
ca            5
thal          4
target        2
dtype: int64
```

In [59]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

In [60]:

```
df['ca'].unique()
```

Out[60]:

```
array([0, 2, 1, 3, 4], dtype=int64)
```

In [61]:

```
df.ca.value_counts()
```

Out[61]:

```
0    175
1     65
2     38
3     20
4      5
Name: ca, dtype: int64
```

In [62]:

```
df.loc[df['ca']==4]
```

Out[62]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
92	52	1	2	138	223	0	1	169	0	0.0	2	4	2	1
158	58	1	1	125	220	0	1	144	0	0.4	1	4	3	1
163	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1
164	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1
251	43	1	0	132	247	1	0	143	1	0.1	1	4	3	0

In [63]:

```
df['ca'].unique()
```

Out[63]:

```
array([0, 2, 1, 3, 4], dtype=int64)
```

In [64]:

```
df.isna().sum()
```

Out[64]:

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

In [65]:

```
df=df.fillna(df.median())
df.isnull().sum()
```

Out[65]:

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

In [67]:

```
duplicates = df.duplicated(keep=False).sum()
duplicates
```

Out[67]:

```
2
```

In [68]:

```
df.describe()
```

Out[68]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.00
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.39
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.61
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.00
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.00
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.00
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.00
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.00

In [72]:

```
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
```

In [73]:

```
X = df.drop('target', axis=1)
y = df.target

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

In [75]:

```
from sklearn import svm
clf = svm.SVC(kernel='linear')
clf.fit(X_train, y_train)
y_pred = clf.predict( X_test)
```

In [77]:

```
from sklearn import metrics
accuracy = metrics.accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.7692307692307693
```

In [78]:

```
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Precision: 0.7735849056603774
Recall: 0.82
```

In [ ]:

```

```