In [1]:

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

In [2]:

```
df=pd.read_csv("data.csv")
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12256\1381553570.py:1: DtypeW
arning: Columns (0) have mixed types. Specify dtype option on import or se
t low_memory=False.
  df=pd.read_csv("data.csv")
```

In [3]:

```
df
```

Out[3]:

	stn_code	sampling_date	state	location	agency	type	so2	no2
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5
...	...	...	...	...	...	...	...	...
435737	SAMP	24-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RIRUO	22.0	50.0
435738	SAMP	29-12-15	West Bengal	ULUBERIA	West Bengal State Pollution Control Board	RIRUO	20.0	46.0
435739	NaN	NaN	andaman-and-nicobar-islands	NaN	NaN	NaN	NaN	NaN
435740	NaN	NaN	Lakshadweep	NaN	NaN	NaN	NaN	NaN
435741	NaN	NaN	Tripura	NaN	NaN	NaN	NaN	NaN

435742 rows × 13 columns

In [4]:

```
df.transpose()
```

Out[4]:

	0	1	2	3	4	
stn_code	150.0	151.0	152.0	150.0	151.0	
sampling_date	February - M021990	February - M021990	February - M021990	March - M031990	March - M031990	M M03
state	Andhra Pradesh	Andhra Pradesh	Andhra Pradesh	Andhra Pradesh	Andhra Pradesh	A Pra
location	Hyderabad	Hyderabad	Hyderabad	Hyderabad	Hyderabad	Hyde
agency	NaN	NaN	NaN	NaN	NaN	
type	Residential, Rural and other Areas	Industrial Area	Residential, Rural and other Areas	Residential, Rural and other Areas	Industrial Area	Resid Rura other
so2	4.8	3.1	6.2	6.3	4.7	
no2	17.4	7.0	28.5	14.7	7.5	
rspm	NaN	NaN	NaN	NaN	NaN	
spm	NaN	NaN	NaN	NaN	NaN	
location_monitoring_station	NaN	NaN	NaN	NaN	NaN	
pm2_5	NaN	NaN	NaN	NaN	NaN	
date	1990-02-01	1990-02-01	1990-02-01	1990-03-01	1990-03-01	1990-

13 rows × 435742 columns

In [5]:

```
df.shape
```

Out[5]:

```
(435742, 13)
```

In [6]:

```
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 [7]:

```
df.isnull()
```

Out[7]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	l
0	False	False	False	False	True	False	False	False	True	True	
1	False	False	False	False	True	False	False	False	True	True	
2	False	False	False	False	True	False	False	False	True	True	
3	False	False	False	False	True	False	False	False	True	True	
4	False	False	False	False	True	False	False	False	True	True	
...	...	...	...	...	...	...	...	...	...	...	
435737	False	False	False	False	False	False	False	False	False	True	
435738	False	False	False	False	False	False	False	False	False	True	
435739	True	True	False	True	True	True	True	True	True	True	
435740	True	True	False	True	True	True	True	True	True	True	
435741	True	True	False	True	True	True	True	True	True	True	

435742 rows × 13 columns

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]:

```
#dataset=df.dropna() # used to drop null values
```

In [10]:

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

Out[10]:

```
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 [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

Cleaning the Data

In [12]:

```
dataset=df.drop(['stn_code', 'agency','sampling_date','location_monitoring_station'],axi
```

In [13]:

```
dataset
```

Out[13]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN	1990-02-01
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	NaN	NaN	NaN	1990-02-01
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	1990-02-01
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	1990-03-01
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	NaN	NaN	NaN	1990-03-01
...	...	...	...	...	...	...	...	...	...
435737	West Bengal	ULUBERIA	RIRUO	22.0	50.0	143.0	NaN	NaN	2015-12-24
435738	West Bengal	ULUBERIA	RIRUO	20.0	46.0	171.0	NaN	NaN	2015-12-29
435739	andaman-and-nicobar-islands	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
435740	Lakshadweep	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
435741	Tripura	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

435742 rows × 9 columns

In [14]:

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

In [15]:

```
dataset
```

Out[15]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN	1990-02-01
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	NaN	NaN	NaN	1990-02-01
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	1990-02-01
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	1990-03-01
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	NaN	NaN	NaN	1990-03-01
...	...	...	...	...	...	...	...	...	...
435734	West Bengal	ULUBERIA	RIRUO	20.0	44.0	148.0	NaN	NaN	2015-12-15
435735	West Bengal	ULUBERIA	RIRUO	17.0	44.0	131.0	NaN	NaN	2015-12-18
435736	West Bengal	ULUBERIA	RIRUO	18.0	45.0	140.0	NaN	NaN	2015-12-21
435737	West Bengal	ULUBERIA	RIRUO	22.0	50.0	143.0	NaN	NaN	2015-12-24
435738	West Bengal	ULUBERIA	RIRUO	20.0	46.0	171.0	NaN	NaN	2015-12-29

435735 rows × 9 columns

In [16]:

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

Out[16]:

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

In [17]:

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

Out[17]:

```
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 [18]:

```
types = {
    "Residential": "R",
    "Residential and others": "RO",
    "Residential, Rural and other Areas": "RRO",
    "Industrial Area": "I",
    "Industrial Areas": "I",
    "Industrial": "I",
    "Sensitive Area": "S",
    "Sensitive Areas": "S",
    "Sensitive": "S",
    "NaN": "RRO"
}
dataset.type = dataset.type.replace(types)
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12256\1220031764.py:13: Setti
ngWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  dataset.type = dataset.type.replace(types)
```

In [19]:

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

Out[19]:

```
array(['RRO', 'I', nan, 'S', 'RO', 'R', 'RIRUO'], dtype=object)
```

In [20]:

```
dataset['date'] = pd.to_datetime(dataset['date'], errors='coerce')
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12256\3924618266.py:1: Settin
gWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  dataset['date'] = pd.to_datetime(dataset['date'], errors='coerce')
```

In [21]:

```
dataset['year'] = dataset.date.dt.year
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12256\37766406.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  dataset['year'] = dataset.date.dt.year
```

In [22]:

```
dataset.head(50)
```

Out[22]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	Andhra Pradesh	Hyderabad	RRO	4.8	17.4	NaN	NaN	NaN	1990-02-01	1990
1	Andhra Pradesh	Hyderabad	I	3.1	7.0	NaN	NaN	NaN	1990-02-01	1990
2	Andhra Pradesh	Hyderabad	RRO	6.2	28.5	NaN	NaN	NaN	1990-02-01	1990
3	Andhra Pradesh	Hyderabad	RRO	6.3	14.7	NaN	NaN	NaN	1990-03-01	1990
4	Andhra Pradesh	Hyderabad	I	4.7	7.5	NaN	NaN	NaN	1990-03-01	1990
5	Andhra Pradesh	Hyderabad	RRO	6.4	25.7	NaN	NaN	NaN	1990-03-01	1990
6	Andhra Pradesh	Hyderabad	RRO	5.4	17.1	NaN	NaN	NaN	1990-04-01	1990
7	Andhra Pradesh	Hyderabad	I	4.7	8.7	NaN	NaN	NaN	1990-04-01	1990
8	Andhra Pradesh	Hyderabad	RRO	4.2	23.0	NaN	NaN	NaN	1990-04-01	1990
9	Andhra Pradesh	Hyderabad	I	4.0	8.9	NaN	NaN	NaN	1990-05-01	1990
10	Andhra Pradesh	Hyderabad	RRO	3.6	18.6	NaN	NaN	NaN	1990-05-01	1990
11	Andhra Pradesh	Hyderabad	RRO	3.9	14.1	NaN	133.0	NaN	1990-06-01	1990
12	Andhra Pradesh	Hyderabad	I	5.6	11.8	NaN	82.0	NaN	1990-06-01	1990
13	Andhra Pradesh	Hyderabad	RRO	3.3	19.3	NaN	111.0	NaN	1990-06-01	1990
14	Andhra Pradesh	Hyderabad	RRO	3.9	8.2	NaN	118.0	NaN	1990-07-01	1990
15	Andhra Pradesh	Hyderabad	RRO	3.5	12.1	NaN	135.0	NaN	1990-07-01	1990
16	Andhra Pradesh	Hyderabad	I	7.9	10.2	NaN	80.0	NaN	1990-07-01	1990
17	Andhra Pradesh	Hyderabad	RRO	4.0	9.9	NaN	179.0	NaN	1990-08-01	1990
18	Andhra Pradesh	Hyderabad	I	12.4	11.5	NaN	58.0	NaN	1990-08-01	1990
19	Andhra Pradesh	Hyderabad	RRO	4.0	12.3	NaN	99.0	NaN	1990-08-01	1990
20	Andhra Pradesh	Hyderabad	RRO	6.3	11.5	NaN	270.0	NaN	1990-09-01	1990
21	Andhra Pradesh	Hyderabad	I	44.8	13.7	NaN	97.0	NaN	1990-09-01	1990
22	Andhra Pradesh	Hyderabad	RRO	8.1	17.8	NaN	167.0	NaN	1990-09-01	1990
23	Andhra Pradesh	Hyderabad	RRO	7.7	11.3	NaN	145.0	NaN	1990-10-01	1990
24	Andhra Pradesh	Hyderabad	I	20.6	13.6	NaN	75.0	NaN	1990-10-01	1990
25	Andhra Pradesh	Hyderabad	RRO	20.4	27.5	NaN	212.0	NaN	1990-10-01	1990
26	Andhra Pradesh	Hyderabad	RRO	13.9	7.2	NaN	93.0	NaN	1990-11-01	1990
27	Andhra Pradesh	Hyderabad	I	11.2	18.6	NaN	61.0	NaN	1990-11-01	1990
28	Andhra Pradesh	Hyderabad	RRO	22.3	35.9	NaN	255.0	NaN	1990-11-01	1990
29	Andhra Pradesh	Hyderabad	RRO	24.5	28.0	NaN	197.0	NaN	1991-01-01	1991
30	Andhra Pradesh	Hyderabad	RRO	7.2	10.4	NaN	148.0	NaN	1991-01-01	1991
31	Andhra Pradesh	Hyderabad	I	28.7	16.2	NaN	77.0	NaN	1991-01-01	1991
32	Andhra Pradesh	Hyderabad	RRO	18.7	42.2	NaN	125.0	NaN	1991-02-01	1991
33	Andhra Pradesh	Hyderabad	RRO	24.5	18.0	NaN	330.0	NaN	1991-02-01	1991
34	Andhra Pradesh	Hyderabad	I	20.4	12.6	NaN	93.0	NaN	1991-02-01	1991
35	Andhra Pradesh	Hyderabad	RRO	5.2	41.3	NaN	287.0	NaN	1991-03-01	1991
36	Andhra Pradesh	Hyderabad	RRO	7.5	12.2	NaN	241.0	NaN	1991-03-01	1991

	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
37	Andhra Pradesh	Hyderabad	I	4.8	8.4	NaN	85.0	NaN	1991-03-01	1991
38	Andhra Pradesh	Hyderabad	RRO	8.5	48.5	NaN	NaN	NaN	1991-04-01	1991
39	Andhra Pradesh	Hyderabad	RRO	9.7	12.4	NaN	283.0	NaN	1991-04-01	1991
40	Andhra Pradesh	Hyderabad	I	21.2	11.5	NaN	108.0	NaN	1991-04-01	1991
41	Andhra Pradesh	Hyderabad	RRO	4.9	15.3	NaN	234.0	NaN	1991-05-01	1991
42	Andhra Pradesh	Hyderabad	I	17.7	14.0	NaN	121.0	NaN	1991-05-01	1991
43	Andhra Pradesh	Hyderabad	RRO	12.3	38.6	NaN	219.0	NaN	1991-05-01	1991
44	Andhra Pradesh	Hyderabad	RRO	3.5	11.9	NaN	179.0	NaN	1991-06-01	1991
45	Andhra Pradesh	Hyderabad	I	3.1	7.5	NaN	84.0	NaN	1991-06-01	1991
46	Andhra Pradesh	Hyderabad	RRO	3.0	19.0	NaN	154.0	NaN	1991-06-01	1991
47	Andhra Pradesh	Hyderabad	RRO	6.2	10.0	NaN	150.0	NaN	1991-07-01	1991
48	Andhra Pradesh	Hyderabad	I	7.9	9.2	NaN	67.0	NaN	1991-07-01	1991
49	Andhra Pradesh	Hyderabad	RRO	6.5	17.3	NaN	128.0	NaN	1991-07-01	1991

In [23]:

```
dataset['month'] = dataset.date.dt.month
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12256\2219524706.py:1: Settin
gWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  dataset['month'] = dataset.date.dt.month
```

In [24]:

```
dataset.head(50)
```

Out[24]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date	year	month
0	Andhra Pradesh	Hyderabad	RRO	4.8	17.4	NaN	NaN	NaN	1990-02-01	1990	2
1	Andhra Pradesh	Hyderabad	I	3.1	7.0	NaN	NaN	NaN	1990-02-01	1990	2
2	Andhra Pradesh	Hyderabad	RRO	6.2	28.5	NaN	NaN	NaN	1990-02-01	1990	2
3	Andhra Pradesh	Hyderabad	RRO	6.3	14.7	NaN	NaN	NaN	1990-03-01	1990	3
4	Andhra Pradesh	Hyderabad	I	4.7	7.5	NaN	NaN	NaN	1990-03-01	1990	3
5	Andhra Pradesh	Hyderabad	RRO	6.4	25.7	NaN	NaN	NaN	1990-03-01	1990	3
6	Andhra Pradesh	Hyderabad	RRO	5.4	17.1	NaN	NaN	NaN	1990-04-01	1990	4
7	Andhra Pradesh	Hyderabad	I	4.7	8.7	NaN	NaN	NaN	1990-04-01	1990	4
8	Andhra Pradesh	Hyderabad	RRO	4.2	23.0	NaN	NaN	NaN	1990-04-01	1990	4
9	Andhra Pradesh	Hyderabad	I	4.0	8.9	NaN	NaN	NaN	1990-05-01	1990	5
10	Andhra Pradesh	Hyderabad	RRO	3.6	18.6	NaN	NaN	NaN	1990-05-01	1990	5
11	Andhra Pradesh	Hyderabad	RRO	3.9	14.1	NaN	133.0	NaN	1990-06-01	1990	6
12	Andhra Pradesh	Hyderabad	I	5.6	11.8	NaN	82.0	NaN	1990-06-01	1990	6
13	Andhra Pradesh	Hyderabad	RRO	3.3	19.3	NaN	111.0	NaN	1990-06-01	1990	6
14	Andhra Pradesh	Hyderabad	RRO	3.9	8.2	NaN	118.0	NaN	1990-07-01	1990	7
15	Andhra Pradesh	Hyderabad	RRO	3.5	12.1	NaN	135.0	NaN	1990-07-01	1990	7
16	Andhra Pradesh	Hyderabad	I	7.9	10.2	NaN	80.0	NaN	1990-07-01	1990	7
17	Andhra Pradesh	Hyderabad	RRO	4.0	9.9	NaN	179.0	NaN	1990-08-01	1990	8
18	Andhra Pradesh	Hyderabad	I	12.4	11.5	NaN	58.0	NaN	1990-08-01	1990	8
19	Andhra Pradesh	Hyderabad	RRO	4.0	12.3	NaN	99.0	NaN	1990-08-01	1990	8
20	Andhra Pradesh	Hyderabad	RRO	6.3	11.5	NaN	270.0	NaN	1990-09-01	1990	9
21	Andhra Pradesh	Hyderabad	I	44.8	13.7	NaN	97.0	NaN	1990-09-01	1990	9
22	Andhra Pradesh	Hyderabad	RRO	8.1	17.8	NaN	167.0	NaN	1990-09-01	1990	9
23	Andhra Pradesh	Hyderabad	RRO	7.7	11.3	NaN	145.0	NaN	1990-10-01	1990	10

	state	location	type	so2	no2	rspm	spm	pm2_5	date	year	month
24	Andhra Pradesh	Hyderabad	I	20.6	13.6	NaN	75.0	NaN	1990-10-01	1990	10
25	Andhra Pradesh	Hyderabad	RRO	20.4	27.5	NaN	212.0	NaN	1990-10-01	1990	10
26	Andhra Pradesh	Hyderabad	RRO	13.9	7.2	NaN	93.0	NaN	1990-11-01	1990	11
27	Andhra Pradesh	Hyderabad	I	11.2	18.6	NaN	61.0	NaN	1990-11-01	1990	11
28	Andhra Pradesh	Hyderabad	RRO	22.3	35.9	NaN	255.0	NaN	1990-11-01	1990	11
29	Andhra Pradesh	Hyderabad	RRO	24.5	28.0	NaN	197.0	NaN	1991-01-01	1991	1
30	Andhra Pradesh	Hyderabad	RRO	7.2	10.4	NaN	148.0	NaN	1991-01-01	1991	1
31	Andhra Pradesh	Hyderabad	I	28.7	16.2	NaN	77.0	NaN	1991-01-01	1991	1
32	Andhra Pradesh	Hyderabad	RRO	18.7	42.2	NaN	125.0	NaN	1991-02-01	1991	2
33	Andhra Pradesh	Hyderabad	RRO	24.5	18.0	NaN	330.0	NaN	1991-02-01	1991	2
34	Andhra Pradesh	Hyderabad	I	20.4	12.6	NaN	93.0	NaN	1991-02-01	1991	2
35	Andhra Pradesh	Hyderabad	RRO	5.2	41.3	NaN	287.0	NaN	1991-03-01	1991	3
36	Andhra Pradesh	Hyderabad	RRO	7.5	12.2	NaN	241.0	NaN	1991-03-01	1991	3
37	Andhra Pradesh	Hyderabad	I	4.8	8.4	NaN	85.0	NaN	1991-03-01	1991	3
38	Andhra Pradesh	Hyderabad	RRO	8.5	48.5	NaN	NaN	NaN	1991-04-01	1991	4
39	Andhra Pradesh	Hyderabad	RRO	9.7	12.4	NaN	283.0	NaN	1991-04-01	1991	4
40	Andhra Pradesh	Hyderabad	I	21.2	11.5	NaN	108.0	NaN	1991-04-01	1991	4
41	Andhra Pradesh	Hyderabad	RRO	4.9	15.3	NaN	234.0	NaN	1991-05-01	1991	5
42	Andhra Pradesh	Hyderabad	I	17.7	14.0	NaN	121.0	NaN	1991-05-01	1991	5
43	Andhra Pradesh	Hyderabad	RRO	12.3	38.6	NaN	219.0	NaN	1991-05-01	1991	5
44	Andhra Pradesh	Hyderabad	RRO	3.5	11.9	NaN	179.0	NaN	1991-06-01	1991	6
45	Andhra Pradesh	Hyderabad	I	3.1	7.5	NaN	84.0	NaN	1991-06-01	1991	6
46	Andhra Pradesh	Hyderabad	RRO	3.0	19.0	NaN	154.0	NaN	1991-06-01	1991	6
47	Andhra Pradesh	Hyderabad	RRO	6.2	10.0	NaN	150.0	NaN	1991-07-01	1991	7
48	Andhra Pradesh	Hyderabad	I	7.9	9.2	NaN	67.0	NaN	1991-07-01	1991	7

In [25]:

```
dataset['hour']=dataset.date.dt.hour #Creating new hour column
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12256\421403579.py:1: Setting
WithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  dataset['hour']=dataset.date.dt.hour #Creating new hour column
```

In [26]:

```
dataset
```

Out[26]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date	year	month	ho
0	Andhra Pradesh	Hyderabad	RRO	4.8	17.4	NaN	NaN	NaN	1990-02-01	1990	2	
1	Andhra Pradesh	Hyderabad	I	3.1	7.0	NaN	NaN	NaN	1990-02-01	1990	2	
2	Andhra Pradesh	Hyderabad	RRO	6.2	28.5	NaN	NaN	NaN	1990-02-01	1990	2	
3	Andhra Pradesh	Hyderabad	RRO	6.3	14.7	NaN	NaN	NaN	1990-03-01	1990	3	
4	Andhra Pradesh	Hyderabad	I	4.7	7.5	NaN	NaN	NaN	1990-03-01	1990	3	
...	...	...	...	...	...	...	...	...	...	...	...	
435734	West Bengal	ULUBERIA	RIRUO	20.0	44.0	148.0	NaN	NaN	2015-12-15	2015	12	
435735	West Bengal	ULUBERIA	RIRUO	17.0	44.0	131.0	NaN	NaN	2015-12-18	2015	12	
435736	West Bengal	ULUBERIA	RIRUO	18.0	45.0	140.0	NaN	NaN	2015-12-21	2015	12	
435737	West Bengal	ULUBERIA	RIRUO	22.0	50.0	143.0	NaN	NaN	2015-12-24	2015	12	
435738	West Bengal	ULUBERIA	RIRUO	20.0	46.0	171.0	NaN	NaN	2015-12-29	2015	12	

435735 rows × 12 columns

In [27]:

```
#defining columns which have more impact on accracy of the computation using sklearn impo
```

In [28]:

```
from sklearn.impute import SimpleImputer
```

In [29]:

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

In [30]:

```
cols
```

Out[30]:

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

In [31]:

```
# invoking SimpleImputer to fill missing values
imputer=SimpleImputer(missing_values=np.nan,strategy='mean')
dataset[cols]=imputer.fit_transform(dataset[cols])
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12256\1356822368.py:3: Settin
gWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  dataset[cols]=imputer.fit_transform(dataset[cols])
```

In [32]:

```
imputer
```

Out[32]:

```
SimpleImputer()
```

In [33]:

```
dataset
```

Out[33]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date	ye
0	Andhra Pradesh	Hyderabad	RRO	4.8	17.4	108.833091	220.78348	40.791467	1990-02-01	19
1	Andhra Pradesh	Hyderabad	I	3.1	7.0	108.833091	220.78348	40.791467	1990-02-01	19
2	Andhra Pradesh	Hyderabad	RRO	6.2	28.5	108.833091	220.78348	40.791467	1990-02-01	19
3	Andhra Pradesh	Hyderabad	RRO	6.3	14.7	108.833091	220.78348	40.791467	1990-03-01	19
4	Andhra Pradesh	Hyderabad	I	4.7	7.5	108.833091	220.78348	40.791467	1990-03-01	19
...	...	...	...	...	...	...	...	...	...	
435734	West Bengal	ULUBERIA	RIRUO	20.0	44.0	148.000000	220.78348	40.791467	2015-12-15	20
435735	West Bengal	ULUBERIA	RIRUO	17.0	44.0	131.000000	220.78348	40.791467	2015-12-18	20
435736	West Bengal	ULUBERIA	RIRUO	18.0	45.0	140.000000	220.78348	40.791467	2015-12-21	20
435737	West Bengal	ULUBERIA	RIRUO	22.0	50.0	143.000000	220.78348	40.791467	2015-12-24	20
435738	West Bengal	ULUBERIA	RIRUO	20.0	46.0	171.000000	220.78348	40.791467	2015-12-29	20

435735 rows × 12 columns

In [34]:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 435735 entries, 0 to 435738
Data columns (total 12 columns):
 #   Column    Non-Null Count   Dtype         
---  ------    --------------   -----         
 0   state     435735 non-null  object        
 1   location  435735 non-null  object        
 2   type      430345 non-null  object        
 3   so2       435735 non-null  float64       
 4   no2       435735 non-null  float64       
 5   rspm      435735 non-null  float64       
 6   spm       435735 non-null  float64       
 7   pm2_5     435735 non-null  float64       
 8   date      435735 non-null  datetime64[ns]
 9   year      435735 non-null  int64         
 10  month     435735 non-null  int64         
 11  hour      435735 non-null  int64         
dtypes: datetime64[ns](1), float64(5), int64(3), object(3)
memory usage: 43.2+ MB
```

Data Transformation

In [35]:

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

Out[35]:

```
RRO      179013
I        148069
RO        86791
S         15010
RIRUO      1304
R           158
Name: type, dtype: int64
```

In [36]:

```
dataset['type'].replace({
    "RRO":1,
    "I":2,
    "RO":3,
    "S":4,
    "RIRUO":5,
    "R":6
})
```

Out[36]:

```
0         1.0
1         2.0
2         1.0
3         1.0
4         2.0
         ...
435734    5.0
435735    5.0
435736    5.0
435737    5.0
435738    5.0
Name: type, Length: 435735, dtype: float64
```

In [37]:

```
dataset
```

Out[37]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date	ye
0	Andhra Pradesh	Hyderabad	RRO	4.8	17.4	108.833091	220.78348	40.791467	1990-02-01	19
1	Andhra Pradesh	Hyderabad	I	3.1	7.0	108.833091	220.78348	40.791467	1990-02-01	19
2	Andhra Pradesh	Hyderabad	RRO	6.2	28.5	108.833091	220.78348	40.791467	1990-02-01	19
3	Andhra Pradesh	Hyderabad	RRO	6.3	14.7	108.833091	220.78348	40.791467	1990-03-01	19
4	Andhra Pradesh	Hyderabad	I	4.7	7.5	108.833091	220.78348	40.791467	1990-03-01	19
...	...	...	...	...	...	...	...	...	...	
435734	West Bengal	ULUBERIA	RIRUO	20.0	44.0	148.000000	220.78348	40.791467	2015-12-15	20
435735	West Bengal	ULUBERIA	RIRUO	17.0	44.0	131.000000	220.78348	40.791467	2015-12-18	20
435736	West Bengal	ULUBERIA	RIRUO	18.0	45.0	140.000000	220.78348	40.791467	2015-12-21	20
435737	West Bengal	ULUBERIA	RIRUO	22.0	50.0	143.000000	220.78348	40.791467	2015-12-24	20
435738	West Bengal	ULUBERIA	RIRUO	20.0	46.0	171.000000	220.78348	40.791467	2015-12-29	20

435735 rows × 12 columns

In [38]:

```
dataset['state'].value_counts()
```

Out[38]:

```
Maharashtra             60382
Uttar Pradesh           42816
Andhra Pradesh          26368
Punjab                  25634
Rajasthan               25589
Kerala                  24728
Himachal Pradesh        22896
West Bengal             22463
Gujarat                 21279
Tamil Nadu              20597
Madhya Pradesh          19920
Assam                   19361
Odisha                  19278
Karnataka               17118
Delhi                    8551
Chandigarh               8520
Chhattisgarh             7831
Goa                      6206
Jharkhand                5968
Mizoram                  5338
Telangana                3978
Meghalaya                3853
Puducherry               3785
Haryana                  3420
Nagaland                 2463
Bihar                    2275
Uttarakhand              1961
Jammu & Kashmir          1289
Daman & Diu               782
Dadra & Nagar Haveli      634
Uttaranchal               285
Arunachal Pradesh          90
Manipur                    76
Sikkim                      1
Name: state, dtype: int64
```

In [41]:

```
from sklearn.preprocessing import LabelEncoder
labelencoder=LabelEncoder()
dataset["state"]=labelencoder.fit_transform(dataset["state"])
df.head(70)
```

```
C:\Users\Lenovo\AppData\Local\Temp\ipykernel_12256\1170522469.py:3: Settin
gWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  dataset["state"]=labelencoder.fit_transform(dataset["state"])
```

Out[41]:

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm
0	150.0	February - M021990	0	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN
1	151.0	February - M021990	0	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN
2	152.0	February - M021990	0	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN
3	150.0	March - M031990	0	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN
4	151.0	March - M031990	0	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN
...	...	...	...	...	...	...	...	...	...
65	95.0	January - M011992	0	Hyderabad	Andhra Pradesh Pollution Control Board	RRO	6.2	28.5	108.833091
66	202.0	January - M011992	0	Hyderabad	Andhra Pradesh Pollution Control Board	I	4.7	7.5	108.833091
67	203.0	January - M011992	0	Hyderabad	Andhra Pradesh Pollution Control Board	NaN	35.8	12.5	NaN
68	232.0	January - M011992	0	Vishakhapatnam	Andhra Pradesh Pollution Control Board	NaN	52.6	89.6	NaN
69	233.0	January - M011992	0	Vishakhapatnam	Andhra Pradesh Pollution Control Board	NaN	55.8	33.8	NaN

70 rows × 13 columns

In [42]:

```
datasetMaharashtra=dataset[(dataset['state']==0)]
```

In [43]:

```
datasetMaharashtra
```

Out[43]:

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

26368 rows × 12 columns

In [44]:

```
datasetMaharashtra.value_counts()
```

Out[44]:

```
state  location   type  so2   no2   rspm        spm        pm2_5      date
year  month  hour
0      Chittoor   S     4.0    9.0   42.000000   114.00000  40.791467  2011
-03-16  2011  3      0       2
                     5.0    13.0  61.000000   220.78348  40.791467  2014
-03-02  2014  3      0       2
                                  68.000000   220.78348  40.791467  2014
-09-02  2014  9      0       2
                                  67.000000   220.78348  40.791467  2014
-02-22  2014  2      0       2
       Nalgonda   RRO   4.0    16.0  113.000000  220.78348  40.791467  2012
-08-10  2012  8      0       2

..
       Hyderabad  RRO   6.0    17.7  108.833091  220.78348  40.791467  1993
-06-01  1993  6      0       1
                      17.0   20.000000   220.78348  40.791467  2010
-07-23  2010  7      0       1
                      16.0   38.000000   220.78348  40.791467  2010
-07-10  2010  7      0       1
                      14.0   34.000000   220.78348  40.791467  2010
-11-11  2010  11     0       1
       Warangal   RRO   39.0   39.0  67.000000   220.78348  40.791467  2013
-02-14  2013  2      0       1
Length: 25539, dtype: int64
```

In [45]:

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

In [46]:

```
pd.DataFrame(onehotencoder.fit_transform(datasetMaharashtra[["location"]]))
```

Out[46]:

	0	1	2	3	4	5	6	7	8	9	...	14	15	16	17	18	19	20	21
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
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
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
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
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
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
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
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
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
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

26368 rows × 24 columns

In [ ]:

```

```