In [1]:
```
import os
os.chdir('desktop')
```

In [2]:
```
import pandas as pd
import numpy as np
```

In [13]:
```
df=pd.read_csv('data.csv' , encoding='cp1252')
```

```
C:\Users\AKANKSHA\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:3146: DtypeWarning: Columns (0) have mixed types.Specify dtype option on import or set low_memory=False.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name,
```

In [14]:
```
df.head()
```

Out[14]:

	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 [15]:
```
#Data cleaning
#dropping columns that aren't required
df=df.drop(['stn_code', 'agency','sampling_date','location_monitoring_station'], axis = 1)
```

In [16]:
```
df
```

Out[16]:

	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

	state	location	type	so2	no2	rspm	spm	pm2_5	date
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 [19]:
```
#checking for null values
df.isnull().sum()
```

Out[19]:
```
state            0
location         3
type          5393
so2          34646
no2          16233
rspm         40222
spm         237387
pm2_5       426428
date             7
dtype: int64
```

In [21]:
```
#fill in null values with mean in every columns using simpleimputer
#Most reqquired columns are grouped together
COL=['so2','no2','rspm','spm','pm2_5']
```

In [23]:
```
from sklearn.impute import SimpleImputer
imputer=SimpleImputer(missing_values=np.nan,strategy='mean')
df[COL]=imputer.fit_transform(df[COL])
```

In [24]:
```
df.head()
```

Out[24]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	108.832784	220.78348	40.791467	1990-02-01
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	108.832784	220.78348	40.791467	1990-02-01
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	108.832784	220.78348	40.791467	1990-02-01
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	108.832784	220.78348	40.791467	1990-03-01
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	108.832784	220.78348	40.791467	1990-03-01

In [25]:
```
df.isnull().sum()
```

Out[25]:
```
state          0
location       3
type        5393
so2            0
no2            0
rspm           0
spm            0
pm2_5          0
date           7
dtype: int64
```

In [27]:
```
#Data intergration: Adding your own data

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

df.head(50)
```

Out[27]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	108.832784	220.78348	40.791467	1990-02-01
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	108.832784	220.78348	40.791467	1990-02-01
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	108.832784	220.78348	40.791467	1990-02-01
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	108.832784	220.78348	40.791467	1990-03-01
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	108.832784	220.78348	40.791467	1990-03-01

	state	location	type	so2	no2	rspm	spm	pm2_5	date
5	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.4	25.7	108.832784	220.78348	40.791467	1990-03-01
6	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	5.4	17.1	108.832784	220.78348	40.791467	1990-04-01
7	Andhra Pradesh	Hyderabad	Industrial Area	4.7	8.7	108.832784	220.78348	40.791467	1990-04-01
8	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.2	23.0	108.832784	220.78348	40.791467	1990-04-01
9	Andhra Pradesh	Hyderabad	Industrial Area	4.0	8.9	108.832784	220.78348	40.791467	1990-05-01
10	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	3.6	18.6	108.832784	220.78348	40.791467	1990-05-01
11	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	3.9	14.1	108.832784	133.00000	40.791467	1990-06-01
12	Andhra Pradesh	Hyderabad	Industrial Area	5.6	11.8	108.832784	82.00000	40.791467	1990-06-01
13	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	3.3	19.3	108.832784	111.00000	40.791467	1990-06-01
14	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	3.9	8.2	108.832784	118.00000	40.791467	1990-07-01
15	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	3.5	12.1	108.832784	135.00000	40.791467	1990-07-01
16	Andhra Pradesh	Hyderabad	Industrial Area	7.9	10.2	108.832784	80.00000	40.791467	1990-07-01
17	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.0	9.9	108.832784	179.00000	40.791467	1990-08-01
18	Andhra Pradesh	Hyderabad	Industrial Area	12.4	11.5	108.832784	58.00000	40.791467	1990-08-01
19	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.0	12.3	108.832784	99.00000	40.791467	1990-08-01
20	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	11.5	108.832784	270.00000	40.791467	1990-09-01
21	Andhra Pradesh	Hyderabad	Industrial Area	44.8	13.7	108.832784	97.00000	40.791467	1990-09-01
22	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	8.1	17.8	108.832784	167.00000	40.791467	1990-09-01
23	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	7.7	11.3	108.832784	145.00000	40.791467	1990-10-01
24	Andhra Pradesh	Hyderabad	Industrial Area	20.6	13.6	108.832784	75.00000	40.791467	1990-10-01
25	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	20.4	27.5	108.832784	212.00000	40.791467	1990-10-01
26	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	13.9	7.2	108.832784	93.00000	40.791467	1990-11-01
27	Andhra Pradesh	Hyderabad	Industrial Area	11.2	18.6	108.832784	61.00000	40.791467	1990-11-01
28	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	22.3	35.9	108.832784	255.00000	40.791467	1990-11-01

	state	location	type	so2	no2	rspm	spm	pm2_5	date
29	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	24.5	28.0	108.832784	197.00000	40.791467	1991-01-01
30	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	7.2	10.4	108.832784	148.00000	40.791467	1991-01-01
31	Andhra Pradesh	Hyderabad	Industrial Area	28.7	16.2	108.832784	77.00000	40.791467	1991-01-01
32	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	18.7	42.2	108.832784	125.00000	40.791467	1991-02-01
33	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	24.5	18.0	108.832784	330.00000	40.791467	1991-02-01
34	Andhra Pradesh	Hyderabad	Industrial Area	20.4	12.6	108.832784	93.00000	40.791467	1991-02-01
35	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	5.2	41.3	108.832784	287.00000	40.791467	1991-03-01
36	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	7.5	12.2	108.832784	241.00000	40.791467	1991-03-01
37	Andhra Pradesh	Hyderabad	Industrial Area	4.8	8.4	108.832784	85.00000	40.791467	1991-03-01
38	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	8.5	48.5	108.832784	220.78348	40.791467	1991-04-01
39	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	9.7	12.4	108.832784	283.00000	40.791467	1991-04-01
40	Andhra Pradesh	Hyderabad	Industrial Area	21.2	11.5	108.832784	108.00000	40.791467	1991-04-01
41	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.9	15.3	108.832784	234.00000	40.791467	1991-05-01
42	Andhra Pradesh	Hyderabad	Industrial Area	17.7	14.0	108.832784	121.00000	40.791467	1991-05-01
43	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	12.3	38.6	108.832784	219.00000	40.791467	1991-05-01
44	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	3.5	11.9	108.832784	179.00000	40.791467	1991-06-01
45	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.5	108.832784	84.00000	40.791467	1991-06-01
46	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	3.0	19.0	108.832784	154.00000	40.791467	1991-06-01
47	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	10.0	108.832784	150.00000	40.791467	1991-07-01
48	Andhra Pradesh	Hyderabad	Industrial Area	7.9	9.2	108.832784	67.00000	40.791467	1991-07-01
49	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.5	17.3	108.832784	128.00000	40.791467	1991-07-01

In [28]:
```
df['year']=df.date.dt.year
```

In [29]:
```
df.head()
```

Out[29]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	4.8	17.4	108.832784	220.78348	40.791467	1990-02-01	1990.0
1	Andhra Pradesh	Hyderabad	Industrial Area	3.1	7.0	108.832784	220.78348	40.791467	1990-02-01	1990.0
2	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.2	28.5	108.832784	220.78348	40.791467	1990-02-01	1990.0
3	Andhra Pradesh	Hyderabad	Residential, Rural and other Areas	6.3	14.7	108.832784	220.78348	40.791467	1990-03-01	1990.0
4	Andhra Pradesh	Hyderabad	Industrial Area	4.7	7.5	108.832784	220.78348	40.791467	1990-03-01	1990.0

In [30]:
```
df['year']
```

Out[30]:
```
0         1990.0
1         1990.0
2         1990.0
3         1990.0
4         1990.0
           ...
435737    2015.0
435738    2015.0
435739       NaN
435740       NaN
435741       NaN
Name: year, Length: 435742, dtype: float64
```

In [31]:
```
#data transformation : done using encoding ie process of converting categorical data to the numerical one
# 1.Replace Method
df['type'].value_counts()
```

Out[31]:
```
Residential, Rural and other Areas    179014
Industrial Area                        96091
Residential and others                 86791
Industrial Areas                       51747
Sensitive Area                          8980
Sensitive Areas                         5536
RIRUO                                   1304
Sensitive                                495
Industrial                               233
Residential                              158
Name: type, dtype: int64
```

In [32]:
```
df['type'].replace({'Residential, Rural and other Areas':1,
                    'Industrial Area':2 ,
```

```
                    'Residential and others':3,
                    'Industrial Areas':4,
                    'Sensitive Area':5,
                    'Sensitive Areas':6,
                    'RIRUO':7,
                    'Sensitive':8,
                    'Industrial':9,
                   'Residential':10 },inplace=True)
```

In [33]:
```
df.head()
```

Out[33]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	Andhra Pradesh	Hyderabad	1.0	4.8	17.4	108.832784	220.78348	40.791467	1990-02-01	1990.0
1	Andhra Pradesh	Hyderabad	2.0	3.1	7.0	108.832784	220.78348	40.791467	1990-02-01	1990.0
2	Andhra Pradesh	Hyderabad	1.0	6.2	28.5	108.832784	220.78348	40.791467	1990-02-01	1990.0
3	Andhra Pradesh	Hyderabad	1.0	6.3	14.7	108.832784	220.78348	40.791467	1990-03-01	1990.0
4	Andhra Pradesh	Hyderabad	2.0	4.7	7.5	108.832784	220.78348	40.791467	1990-03-01	1990.0

In [36]:
```
#2. Using Label Encoder
#from sklearn.preprocessing import LabelEncoder
e1=LabelEncoder()
df['type']=e1.fit_transform(df['type'])
```

In [37]:
```
df.head()
```

Out[37]:

	state	location	type	so2	no2	rspm	spm	pm2_5	date	year
0	Andhra Pradesh	Hyderabad	0	4.8	17.4	108.832784	220.78348	40.791467	1990-02-01	1990.0
1	Andhra Pradesh	Hyderabad	1	3.1	7.0	108.832784	220.78348	40.791467	1990-02-01	1990.0
2	Andhra Pradesh	Hyderabad	0	6.2	28.5	108.832784	220.78348	40.791467	1990-02-01	1990.0
3	Andhra Pradesh	Hyderabad	0	6.3	14.7	108.832784	220.78348	40.791467	1990-03-01	1990.0
4	Andhra Pradesh	Hyderabad	1	4.7	7.5	108.832784	220.78348	40.791467	1990-03-01	1990.0

In [ ]: