20104016

DEENA

Importing Libraries

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
```

Importing Datasets

In [2]:
```
df=pd.read_csv("madrid_2003.csv")
```

Out[2]:

	date	BEN	CO	EBE	MXY	NMHC	NO_2	NOx	OXY	O_3	
0	2003-03-01 01:00:00	NaN	1.72	NaN	NaN	NaN	73.900002	316.299988	NaN	10.550000	55.2
1	2003-03-01 01:00:00	NaN	1.45	NaN	NaN	0.26	72.110001	250.000000	0.73	6.720000	52.3
2	2003-03-01 01:00:00	NaN	1.57	NaN	NaN	NaN	80.559998	224.199997	NaN	21.049999	63.2
3	2003-03-01 01:00:00	NaN	2.45	NaN	NaN	NaN	78.370003	450.399994	NaN	4.220000	67.8
4	2003-03-01 01:00:00	NaN	3.26	NaN	NaN	NaN	96.250000	479.100006	NaN	8.460000	95.7
...	...	...	...	...	...	...	...	...	...	...	
243979	2003-10-01 00:00:00	0.20	0.16	2.01	3.17	0.02	31.799999	32.299999	1.68	34.049999	7.3
243980	2003-10-01 00:00:00	0.32	0.08	0.36	0.72	NaN	10.450000	14.760000	1.00	34.610001	7.4
243981	2003-10-01 00:00:00	NaN	NaN	NaN	NaN	0.07	34.639999	50.810001	NaN	32.160000	16.8
243982	2003-10-01 00:00:00	NaN	NaN	NaN	NaN	0.07	32.580002	41.020000	NaN	NaN	13.5
243983	2003-10-01 00:00:00	1.00	0.29	2.15	6.41	0.07	37.150002	56.849998	2.28	21.480000	12.3

243984 rows × 16 columns

Data Cleaning and Data Preprocessing

In [3]:

In [4]:

Out[4]:
```
Index(['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3
',
       'PM10', 'PXY', 'SO_2', 'TCH', 'TOL', 'station'],
      dtype='object')
```

In [5]:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 33010 entries, 5 to 243983
Data columns (total 16 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     33010 non-null  object 
 1   BEN      33010 non-null  float64
 2   CO       33010 non-null  float64
 3   EBE      33010 non-null  float64
 4   MXY      33010 non-null  float64
 5   NMHC     33010 non-null  float64
 6   NO_2     33010 non-null  float64
 7   NOx      33010 non-null  float64
 8   OXY      33010 non-null  float64
 9   O_3      33010 non-null  float64
 10  PM10     33010 non-null  float64
 11  PXY      33010 non-null  float64
 12  SO_2     33010 non-null  float64
 13  TCH      33010 non-null  float64
 14  TOL      33010 non-null  float64
 15  station  33010 non-null  int64  
dtypes: float64(14), int64(1), object(1)
memory usage: 4.3+ MB
```

In [6]:
```
data=df[['CO' ,'station']]
```

Out[6]:

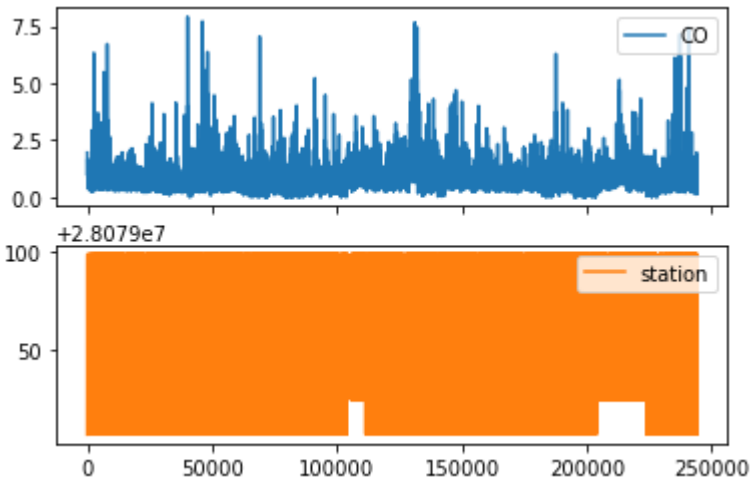
	CO	station
5	1.94	28079006
23	1.27	28079024
27	1.79	28079099
33	1.47	28079006
51	1.29	28079024
...	...	...
243955	0.41	28079099
243957	0.60	28079035
243961	0.82	28079006
243979	0.16	28079024
243983	0.29	28079099

33010 rows × 2 columns

Line chart

In [7]:

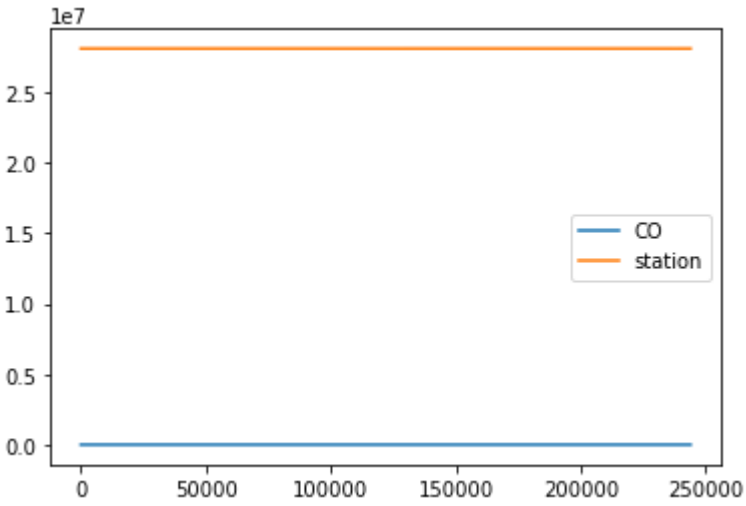
Out[7]: array([<AxesSubplot:>, <AxesSubplot:>], dtype=object)



Line chart

In [8]:

Out[8]: <AxesSubplot:>

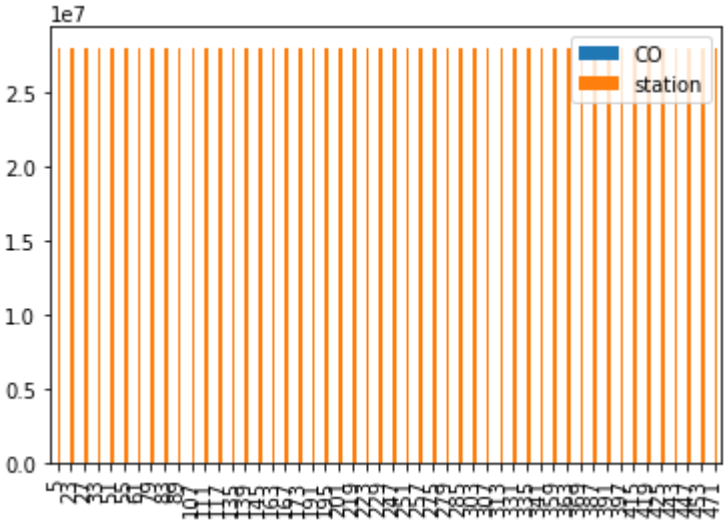


Bar chart

In [9]:

In [10]:

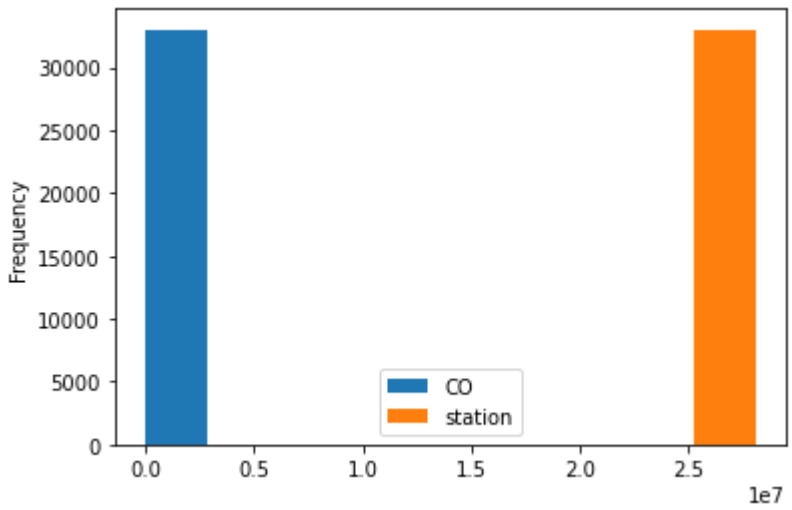
Out[10]: <AxesSubplot:>



Histogram

In [11]:

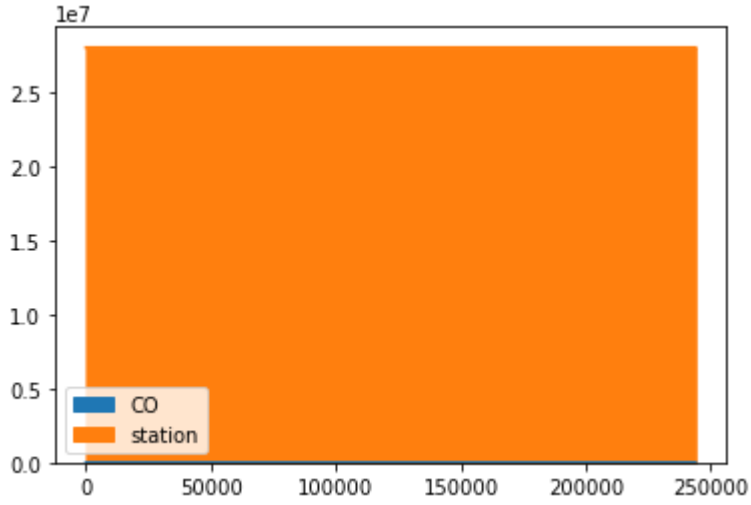
Out[11]: <AxesSubplot:ylabel='Frequency'>



Area chart

In [12]:

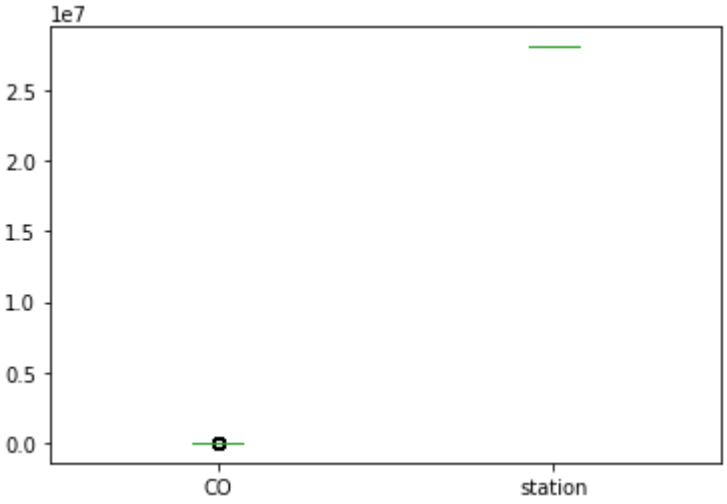
Out[12]: <AxesSubplot:>



Box chart

In [13]:

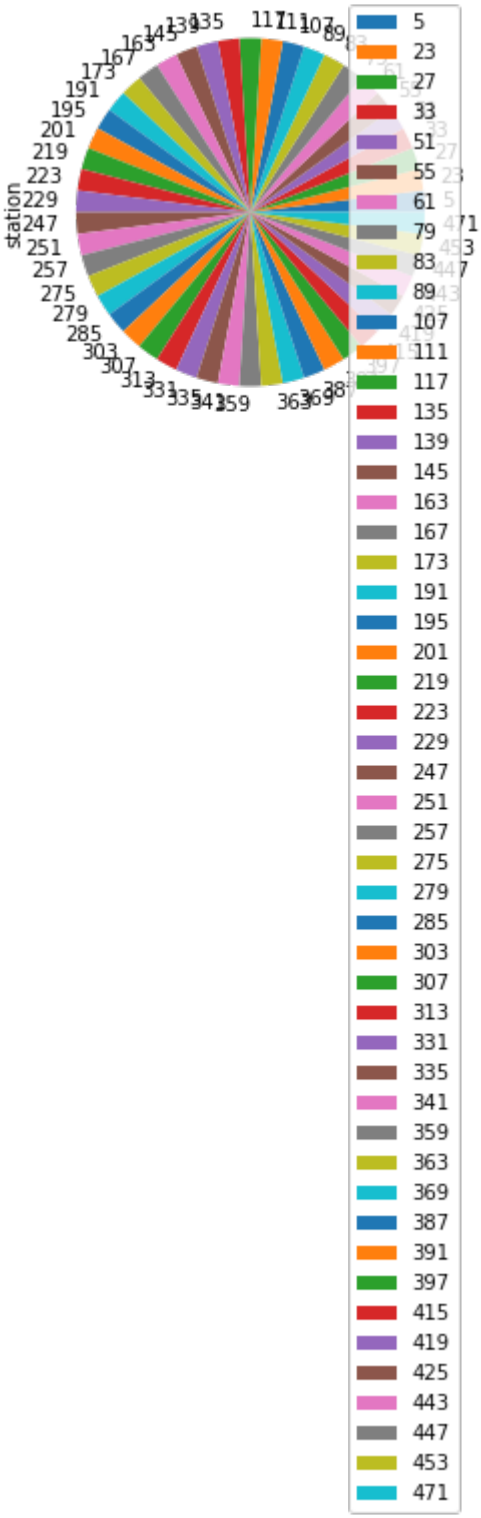
Out[13]: <AxesSubplot:>



Pie chart

In [14]:

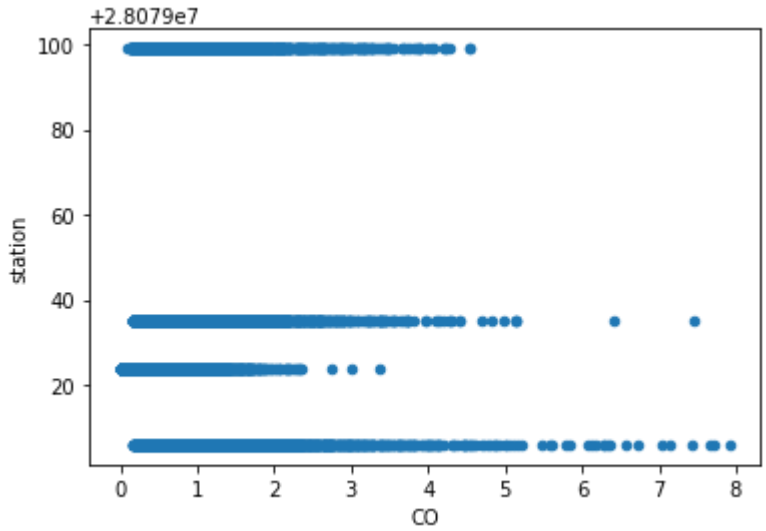
Out[14]: <AxesSubplot:ylabel='station'>



Scatter chart

In [15]:

Out[15]: <AxesSubplot:xlabel='CO', ylabel='station'>



In [16]:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 33010 entries, 5 to 243983
Data columns (total 16 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     33010 non-null  object 
 1   BEN      33010 non-null  float64
 2   CO       33010 non-null  float64
 3   EBE      33010 non-null  float64
 4   MXY      33010 non-null  float64
 5   NMHC     33010 non-null  float64
 6   NO_2     33010 non-null  float64
 7   NOx      33010 non-null  float64
 8   OXY      33010 non-null  float64
 9   O_3      33010 non-null  float64
 10  PM10     33010 non-null  float64
 11  PXY      33010 non-null  float64
 12  SO_2     33010 non-null  float64
 13  TCH      33010 non-null  float64
```

In [17]:

Out[17]:

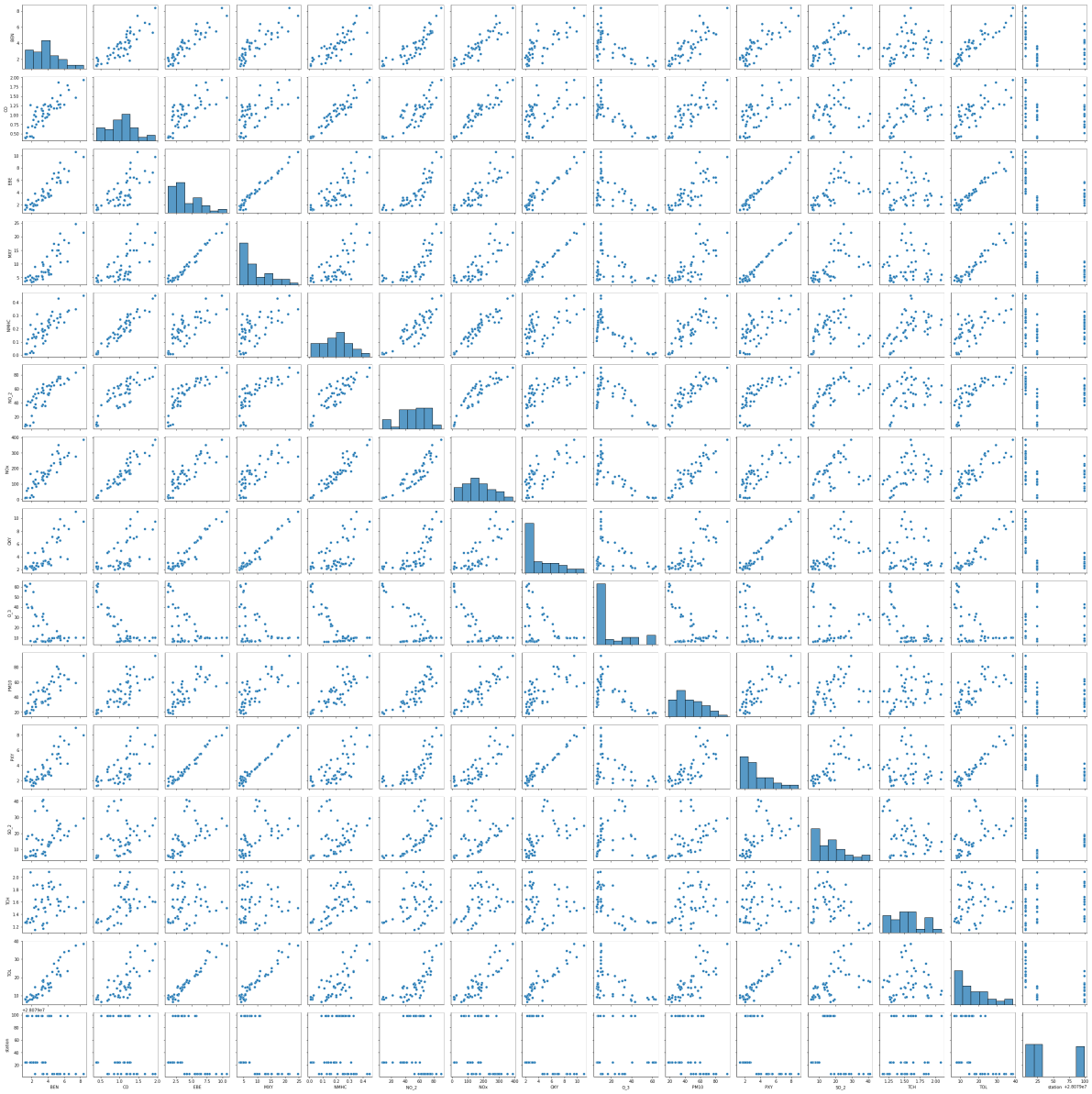
	BEN	CO	EBE	MXY	NMHC	NO_2	
count	33010.000000	33010.000000	33010.000000	33010.000000	33010.000000	33010.000000	330
mean	2.192633	0.759868	2.639726	5.838414	0.137177	57.328049	1
std	2.064160	0.545999	2.825194	6.267296	0.127863	31.811082	1
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.900000	0.430000	1.010000	1.880000	0.060000	34.529999	
50%	1.610000	0.620000	1.890000	4.070000	0.110000	55.105000	
75%	2.810000	0.930000	3.300000	7.530000	0.170000	76.160004	1
max	66.389999	7.920000	92.589996	177.600006	2.180000	342.700012	12

In [18]:
```
df1=df[['BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3',
```

EDA AND VISUALIZATION

In [19]:

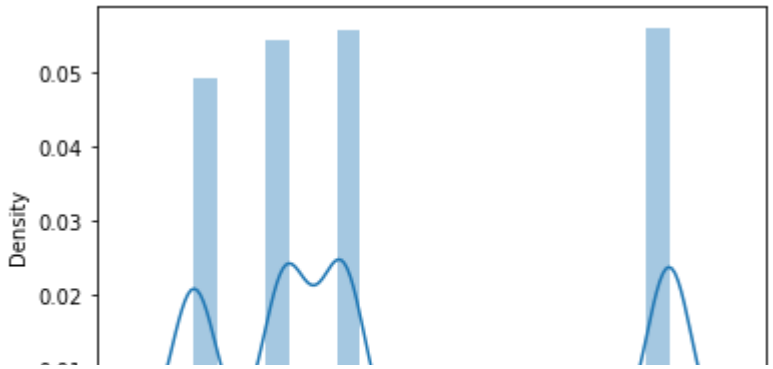
Out[19]: <seaborn.axisgrid.PairGrid at 0x26b28a419a0>



In [20]:

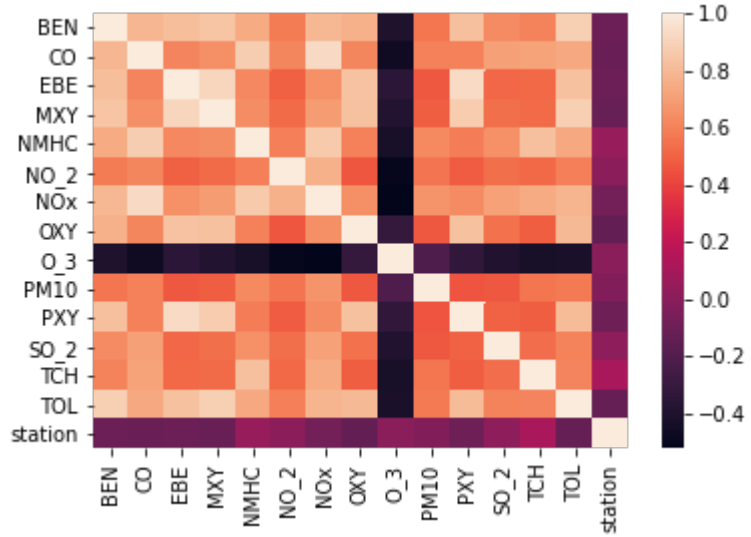
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: Fut
ureWarning: `distplot` is a deprecated function and will be removed in a futu
re version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function for hi
stograms).
  warnings.warn(msg, FutureWarning)
```

Out[20]: <AxesSubplot:xlabel='station', ylabel='Density'>



In [21]:

Out[21]: <AxesSubplot:>



TO TRAIN THE MODEL AND MODEL BULDING

In [22]:

```
x=df[['BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3',
       'PM10', 'PXY', 'SO_2', 'TCH', 'TOL']]
```

In [23]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)
```

Linear Regression

In [24]:
```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
```

Out[24]:
```
LinearRegression()
```

In [25]:
```

```

Out[25]:
```
28079002.47896372
```

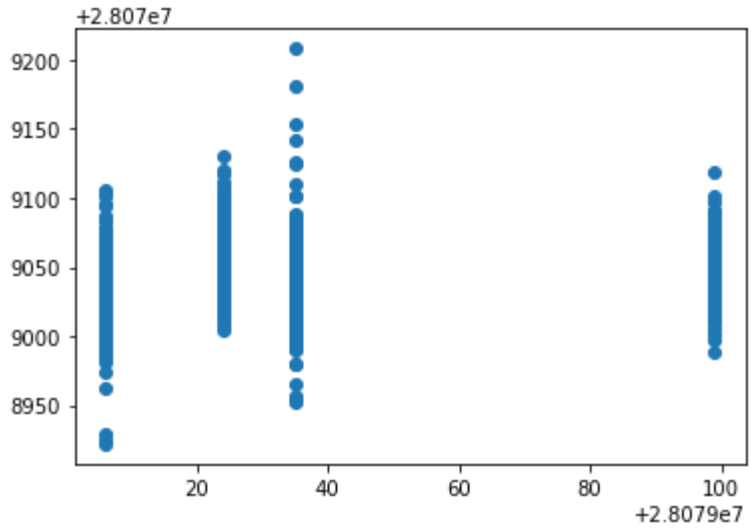
In [26]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
```

Out[26]:

	Co-efficient
BEN	1.417836
CO	-38.954765
EBE	-2.045492
MXY	0.210647
NMHC	155.356410
NO_2	0.154853
NOx	-0.067811
OXY	-1.143264
O_3	-0.021174
PM10	-0.065738
PXY	1.862734
SO_2	0.824172
TCH	35.305109
TOL	-0.855951

In [27]:
```
prediction =lr.predict(x_test)
```

Out[27]: <matplotlib.collections.PathCollection at 0x26b387636a0>



ACCURACY

In [28]:

Out[28]: 0.1780576915685128

In [29]:

Out[29]: 0.17521106566991873

Ridge and Lasso

In [30]:

In [31]:
```
rr=Ridge(alpha=10)
```

Out[31]: Ridge(alpha=10)

Accuracy(Ridge)

In [32]:

Out[32]: 0.1771586988614129

In [33]:

Out[33]: 0.17410086501313182

In [34]:
```
la=Lasso(alpha=10)
```

Out[34]:
```
Lasso(alpha=10)
```

In [35]:

Out[35]:
```
0.034744196302548325
```

Accuracy(Lasso)

In [36]:

Out[36]:
```
0.0365487599901525
```

In [37]:
```
from sklearn.linear_model import ElasticNet
en=ElasticNet()
```

Out[37]:
```
ElasticNet()
```

In [38]:

Out[38]:
```
array([ 0.        , -0.33644631,  0.00488372, -0.03607672,  0.10723709,
        0.14497582, -0.06776602, -1.09314912, -0.05070077,  0.06814352,
        0.31986254,  0.73316094,  1.59455817, -0.45370976])
```

In [39]:

Out[39]:
```
28079038.265746832
```

In [40]:

In [41]:

Out[41]:
```
0.05232787274331019
```

Evaluation Metrics

In [42]:
```
from sklearn import metrics
print(metrics.mean_absolute_error(y_test,prediction))
print(metrics.mean_squared_error(y_test,prediction))
```

```
29.097543702118223
1177.5167085582136
34.314963333190576
```

Logistic Regression

In [43]:

In [44]:
```
feature_matrix=df[['BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O
       'PM10', 'PXY', 'SO_2', 'TCH', 'TOL']]
```

In [45]:

Out[45]: (33010, 14)

In [46]:

Out[46]: (33010,)

In [47]:

In [48]:

In [49]:
```
logr=LogisticRegression(max_iter=10000)
```

Out[49]: LogisticRegression(max_iter=10000)

In [50]:

In [51]:
```
prediction=logr.predict(observation)
```

[28079035]

In [52]:

Out[52]: array([28079006, 28079024, 28079035, 28079099], dtype=int64)

In [53]:

Out[53]: 0.7584974250227204

In [54]:

Out[54]: 2.3306153265290618e-23

In [55]:

Out[55]: array([[2.33061533e-23, 1.44436075e-55, 1.00000000e+00, 6.68457491e-16]])

Random Forest

In [56]:

In [57]:
```
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[57]: RandomForestClassifier()

In [58]:
```
parameters={'max_depth':[1,2,3,4,5],
            'min_samples_leaf':[5,10,15,20,25],
            'n_estimators':[10,20,30,40,50]
```

In [59]:
```
from sklearn.model_selection import GridSearchCV
grid_search =GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="ac
```

Out[59]:
```
GridSearchCV(cv=2, estimator=RandomForestClassifier(),
             param_grid={'max_depth': [1, 2, 3, 4, 5],
                         'min_samples_leaf': [5, 10, 15, 20, 25],
                         'n_estimators': [10, 20, 30, 40, 50]},
             scoring='accuracy')
```

In [60]:

Out[60]:
```
0.7272253750995574
```

In [61]:

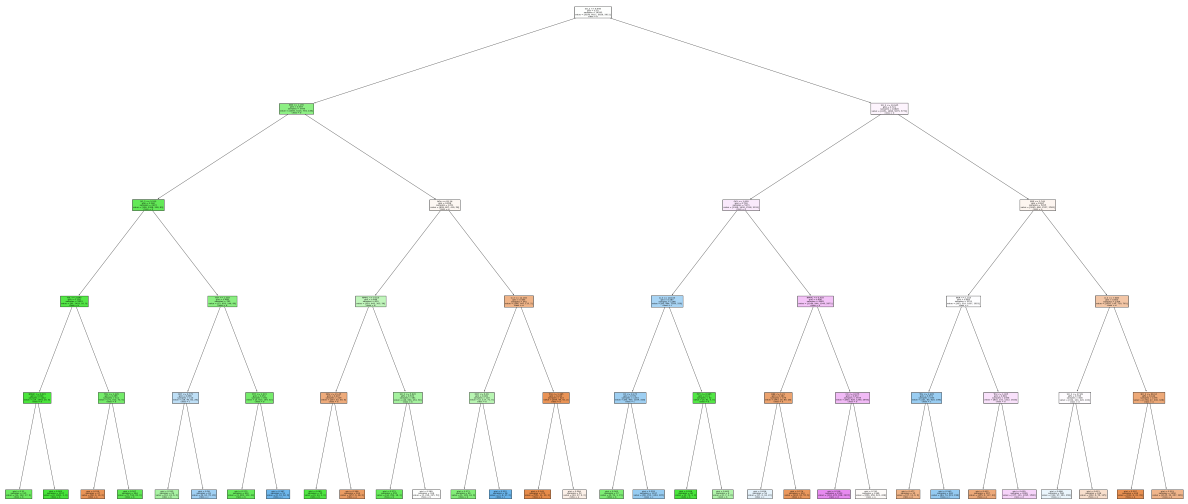
In [62]:
```
from sklearn.tree import plot_tree

plt.figure(figsize=(80,40))
```

Out[62]:
```
[Text(2232.0, 1993.2, 'SO_2 <= 6.645\ngini = 0.75\nsamples = 14592\nvalue =
[5356, 6011, 5829, 5911]\nclass = b'),
 Text(1116.0, 1630.8000000000002, 'EBE <= 1.195\ngini = 0.487\nsamples = 376
6\nvalue = [1005, 4115, 754, 138]\nclass = b'),
 Text(558.0, 1268.4, 'SO_2 <= 5.525\ngini = 0.246\nsamples = 2415\nvalue = [1
06, 3308, 338, 80]\nclass = b'),
 Text(279.0, 906.0, 'TOL <= 2.645\ngini = 0.128\nsamples = 1653\nvalue = [85,
2455, 92, 0]\nclass = b'),
 Text(139.5, 543.5999999999999, 'NMHC <= 0.015\ngini = 0.034\nsamples = 1242\
nvalue = [14, 1943, 20, 0]\nclass = b'),
 Text(69.75, 181.19999999999982, 'gini = 0.31\nsamples = 115\nvalue = [14, 14
2, 17, 0]\nclass = b'),
 Text(209.25, 181.19999999999982, 'gini = 0.003\nsamples = 1127\nvalue = [0,
1801, 3, 0]\nclass = b'),
 Text(418.5, 543.5999999999999, 'TCH <= 1.245\ngini = 0.365\nsamples = 411\nv
alue = [71, 512, 72, 0]\nclass = b'),
 Text(348.75, 181.19999999999982, 'gini = 0.229\nsamples = 47\nvalue = [66,
0, 10, 0]\nclass = a'),
 Text(488.25, 181.19999999999982, 'gini = 0.207\nsamples = 364\nvalue = [5, 5
12, 62, 0]\nclass = b'),
 Text(837.0, 906.0, 'TCH <= 1.255\ngini = 0.448\nsamples = 762\nvalue = [21,
853, 246, 80]\nclass = b'),
 Text(697.5, 543.5999999999999, 'OXY <= 0.775\ngini = 0.666\nsamples = 110\nv
alue = [18, 40, 82, 28]\nclass = c'),
 Text(627.75, 181.19999999999982, 'gini = 0.555\nsamples = 29\nvalue = [10, 2
4, 6, 0]\nclass = b'),
 Text(767.25, 181.19999999999982, 'gini = 0.58\nsamples = 81\nvalue = [8, 16,
76, 28]\nclass = c'),
 Text(976.5, 543.5999999999999, 'PXY <= 1.015\ngini = 0.352\nsamples = 652\nv
alue = [3, 813, 164, 52]\nclass = b'),
 Text(906.75, 181.19999999999982, 'gini = 0.273\nsamples = 601\nvalue = [3, 8
03, 102, 43]\nclass = b'),
 Text(1046.25, 181.19999999999982, 'gini = 0.387\nsamples = 51\nvalue = [0, 1
0, 62, 9]\nclass = c'),
 Text(1674.0, 1268.4, 'NOx <= 81.16\ngini = 0.656\nsamples = 1351\nvalue = [8
99, 807, 416, 58]\nclass = a'),
 Text(1395.0, 906.0, 'NMHC <= 0.015\ngini = 0.648\nsamples = 810\nvalue = [30
3, 642, 303, 56]\nclass = b'),
 Text(1255.5, 543.5999999999999, 'NOx <= 21.45\ngini = 0.447\nsamples = 229\n
value = [257, 52, 50, 0]\nclass = a'),
 Text(1185.75, 181.19999999999982, 'gini = 0.085\nsamples = 25\nvalue = [0, 4
3, 2, 0]\nclass = b'),
 Text(1325.25, 181.19999999999982, 'gini = 0.306\nsamples = 204\nvalue = [25
7, 9, 48, 0]\nclass = a'),
 Text(1534.5, 543.5999999999999, 'SO_2 <= 5.645\ngini = 0.533\nsamples = 581\
nvalue = [46, 590, 253, 56]\nclass = b'),
 Text(1464.75, 181.19999999999982, 'gini = 0.273\nsamples = 255\nvalue = [44,
354, 18, 3]\nclass = b'),
 Text(1604.25, 181.19999999999982, 'gini = 0.589\nsamples = 326\nvalue = [2,
236, 235, 53]\nclass = b'),
 Text(1953.0, 906.0, 'O_3 <= 12.165\ngini = 0.485\nsamples = 541\nvalue = [59
6, 165, 113, 2]\nclass = a'),
```

```
 Text(1813.5, 543.5999999999999, 'MXY <= 6.09\ngini = 0.481\nsamples = 118\nv
alue = [2, 117, 71, 0]\nclass = b'),
 Text(1743.75, 181.19999999999982, 'gini = 0.371\nsamples = 90\nvalue = [1, 1
09, 34, 0]\nclass = b'),
 Text(1883.25, 181.19999999999982, 'gini = 0.322\nsamples = 28\nvalue = [1,
8, 37, 0]\nclass = c'),
 Text(2092.5, 543.5999999999999, 'TCH <= 1.385\ngini = 0.242\nsamples = 423\n
value = [594, 48, 42, 2]\nclass = a'),
 Text(2022.75, 181.19999999999982, 'gini = 0.145\nsamples = 371\nvalue = [55
1, 11, 33, 2]\nclass = a'),
 Text(2162.25, 181.19999999999982, 'gini = 0.584\nsamples = 52\nvalue = [43,
37, 9, 0]\nclass = a'),
 Text(3348.0, 1630.8000000000002, 'SO_2 <= 10.945\ngini = 0.721\nsamples = 10
826\nvalue = [4351, 1896, 5075, 5773]\nclass = d'),
 Text(2790.0, 1268.4, 'OXY <= 1.005\ngini = 0.716\nsamples = 5371\nvalue = [1
306, 1430, 2518, 3191]\nclass = d'),
 Text(2511.0, 906.0, 'O_3 <= 104.05\ngini = 0.587\nsamples = 1286\nvalue = [6
0, 486, 1169, 320]\nclass = c'),
 Text(2371.5, 543.5999999999999, 'CO <= 0.175\ngini = 0.572\nsamples = 1234\n
value = [60, 405, 1164, 320]\nclass = c'),
 Text(2301.75, 181.19999999999982, 'gini = 0.363\nsamples = 37\nvalue = [0, 5
3, 1, 15]\nclass = b'),
 Text(2441.25, 181.19999999999982, 'gini = 0.555\nsamples = 1197\nvalue = [6
0, 352, 1163, 305]\nclass = c'),
 Text(2650.5, 543.5999999999999, 'PXY <= 0.965\ngini = 0.11\nsamples = 52\nva
lue = [0, 81, 5, 0]\nclass = b'),
 Text(2580.75, 181.19999999999982, 'gini = 0.026\nsamples = 46\nvalue = [0, 7
4, 1, 0]\nclass = b'),
 Text(2720.25, 181.19999999999982, 'gini = 0.463\nsamples = 6\nvalue = [0, 7,
4, 0]\nclass = b'),
 Text(3069.0, 906.0, 'NMHC <= 0.035\ngini = 0.696\nsamples = 4085\nvalue = [1
246, 944, 1349, 2871]\nclass = d'),
 Text(2929.5, 543.5999999999999, 'EBE <= 1.21\ngini = 0.381\nsamples = 376\nv
alue = [443, 16, 89, 28]\nclass = a'),
 Text(2859.75, 181.19999999999982, 'gini = 0.656\nsamples = 54\nvalue = [27,
0, 34, 22]\nclass = c'),
 Text(2999.25, 181.19999999999982, 'gini = 0.274\nsamples = 322\nvalue = [41
6, 16, 55, 6]\nclass = a'),
 Text(3208.5, 543.5999999999999, 'CO <= 0.675\ngini = 0.672\nsamples = 3709\n
value = [803, 928, 1260, 2843]\nclass = d'),
 Text(3138.75, 181.19999999999982, 'gini = 0.516\nsamples = 2323\nvalue = [13
4, 470, 638, 2415]\nclass = d'),
 Text(3278.25, 181.19999999999982, 'gini = 0.741\nsamples = 1386\nvalue = [66
9, 458, 622, 428]\nclass = a'),
 Text(3906.0, 1268.4, 'EBE <= 3.745\ngini = 0.697\nsamples = 5455\nvalue = [3
045, 466, 2557, 2582]\nclass = a'),
 Text(3627.0, 906.0, 'BEN <= 1.315\ngini = 0.683\nsamples = 3151\nvalue = [95
3, 316, 1837, 1821]\nclass = c'),
 Text(3487.5, 543.5999999999999, 'TOL <= 1.405\ngini = 0.573\nsamples = 655\n
value = [121, 85, 625, 198]\nclass = c'),
 Text(3417.75, 181.19999999999982, 'gini = 0.477\nsamples = 35\nvalue = [37,
11, 6, 0]\nclass = a'),
 Text(3557.25, 181.19999999999982, 'gini = 0.543\nsamples = 620\nvalue = [84,
74, 619, 198]\nclass = c'),
 Text(3766.5, 543.5999999999999, 'TCH <= 1.305\ngini = 0.681\nsamples = 2496\
nvalue = [832, 231, 1212, 1623]\nclass = d'),
```

```
 Text(3696.75, 181.19999999999982, 'gini = 0.412\nsamples = 397\nvalue = [45
5, 0, 117, 43]\nclass = a'),
 Text(3836.25, 181.19999999999982, 'gini = 0.639\nsamples = 2099\nvalue = [37
7, 231, 1095, 1580]\nclass = d'),
 Text(4185.0, 906.0, 'O_3 <= 7.885\ngini = 0.603\nsamples = 2304\nvalue = [20
92, 150, 720, 761]\nclass = a'),
 Text(4045.5, 543.5999999999999, 'SO_2 <= 27.89\ngini = 0.708\nsamples = 828\
nvalue = [334, 103, 420, 433]\nclass = d'),
 Text(3975.75, 181.19999999999982, 'gini = 0.688\nsamples = 580\nvalue = [13
2, 97, 361, 298]\nclass = c'),
 Text(4115.25, 181.19999999999982, 'gini = 0.613\nsamples = 248\nvalue = [20
2, 6, 59, 135]\nclass = a'),
 Text(4324.5, 543.5999999999999, 'NO_2 <= 66.16\ngini = 0.444\nsamples = 147
6\nvalue = [1758, 47, 300, 328]\nclass = a'),
 Text(4254.75, 181.19999999999982, 'gini = 0.207\nsamples = 426\nvalue = [63
3, 8, 33, 39]\nclass = a'),
 Text(4394.25, 181.19999999999982, 'gini = 0.519\nsamples = 1050\nvalue = [11
25, 39, 267, 289]\nclass = a')]
```



Conclusion

Accuracy

Linear Regression : 0.17521106566991873

Ridge Regression : 0.034744196302548325

Lasso Regression : 0.05232787274331019

ElasticNet Regression : 28079038.265746832

Logistic Regression : 0.7584974250227204

Logistic Regression is suitable for this dataset