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_2012.csv")
```

Out[2]:

	date	BEN	CO	EBE	NMHC	NO	NO_2	O_3	PM10	PM25	SO_2	TCH	TOL
0	2011-11-01 01:00:00	NaN	1.0	NaN	NaN	154.0	84.0	NaN	NaN	NaN	6.0	NaN	NaN
1	2011-11-01 01:00:00	2.5	0.4	3.5	0.26	68.0	92.0	3.0	40.0	24.0	9.0	1.54	8.7
2	2011-11-01 01:00:00	2.9	NaN	3.8	NaN	96.0	99.0	NaN	NaN	NaN	NaN	NaN	7.2
3	2011-11-01 01:00:00	NaN	0.6	NaN	NaN	60.0	83.0	2.0	NaN	NaN	NaN	NaN	NaN
4	2011-11-01 01:00:00	NaN	NaN	NaN	NaN	44.0	62.0	3.0	NaN	NaN	3.0	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
209923	2011-09-01 00:00:00	NaN	0.2	NaN	NaN	5.0	19.0	44.0	NaN	NaN	NaN	NaN	NaN
209924	2011-09-01 00:00:00	NaN	0.1	NaN	NaN	6.0	29.0	NaN	11.0	NaN	7.0	NaN	NaN
209925	2011-09-01 00:00:00	NaN	NaN	NaN	0.23	1.0	21.0	28.0	NaN	NaN	NaN	1.44	NaN
209926	2011-09-01 00:00:00	NaN	NaN	NaN	NaN	3.0	15.0	48.0	NaN	NaN	NaN	NaN	NaN
209927	2011-09-01 00:00:00	NaN	NaN	NaN	NaN	4.0	33.0	38.0	13.0	NaN	NaN	NaN	NaN

209928 rows × 14 columns

Data Cleaning and Data Preprocessing

In [3]:

In [4]:

Out[4]:
```
Index(['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM25
',
       'SO_2', 'TCH', 'TOL', 'station'],
      dtype='object')
```

In [5]:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 16460 entries, 1 to 209910
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     16460 non-null  object 
 1   BEN      16460 non-null  float64
 2   CO       16460 non-null  float64
 3   EBE      16460 non-null  float64
 4   NMHC     16460 non-null  float64
 5   NO       16460 non-null  float64
 6   NO_2     16460 non-null  float64
 7   O_3      16460 non-null  float64
 8   PM10     16460 non-null  float64
 9   PM25     16460 non-null  float64
 10  SO_2     16460 non-null  float64
 11  TCH      16460 non-null  float64
 12  TOL      16460 non-null  float64
 13  station  16460 non-null  int64  
dtypes: float64(12), int64(1), object(1)
memory usage: 1.9+ MB
```

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

Out[6]:

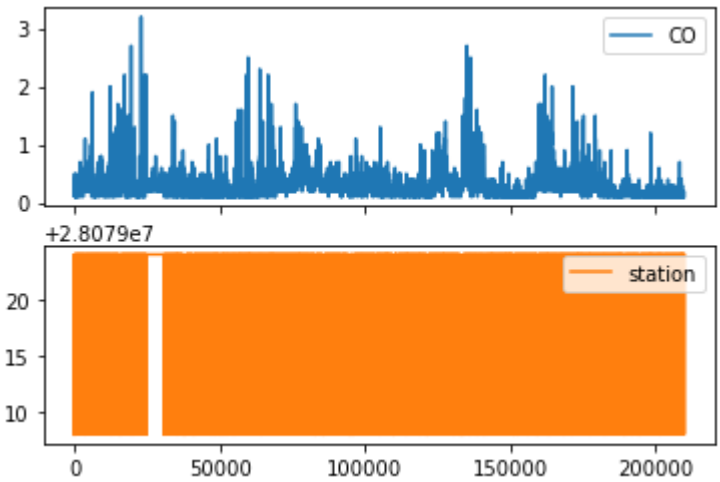
	CO	station
1	0.4	28079008
6	0.3	28079024
25	0.3	28079008
30	0.4	28079024
49	0.2	28079008
...	...	...
209862	0.1	28079024
209881	0.1	28079008
209886	0.1	28079024
209905	0.1	28079008
209910	0.1	28079024

16460 rows × 2 columns

Line chart

In [7]:
```
```

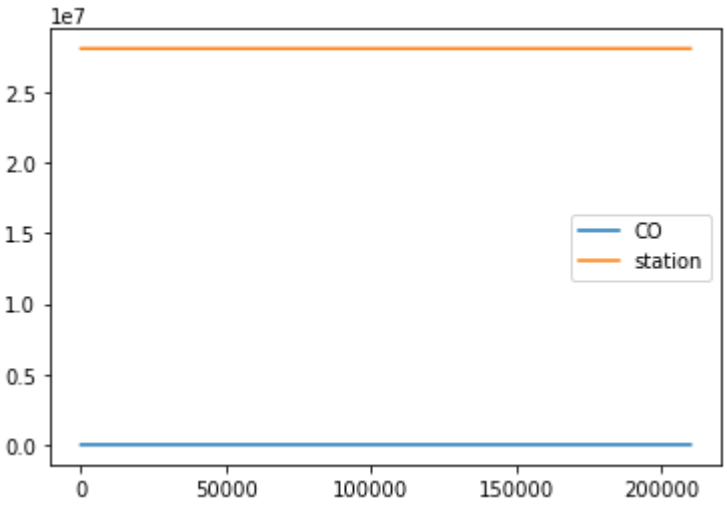
Out[7]:
```
array([<AxesSubplot:>, <AxesSubplot:>], dtype=object)
```



Line chart

In [8]:

Out[8]: <AxesSubplot:>

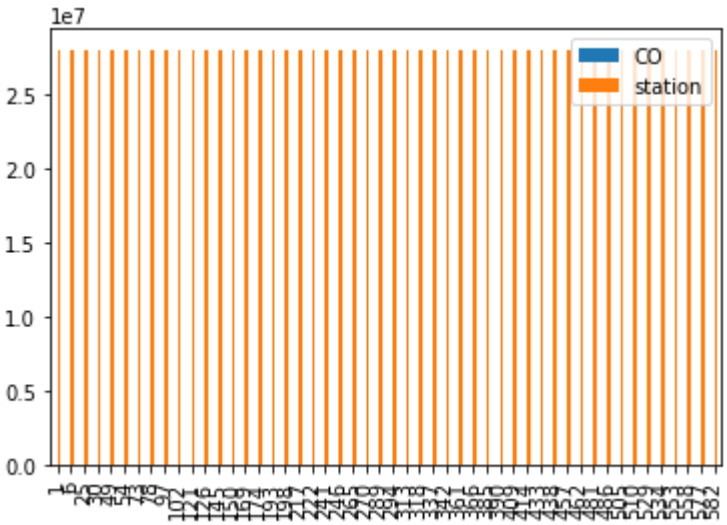


Bar chart

In [9]:

In [10]:

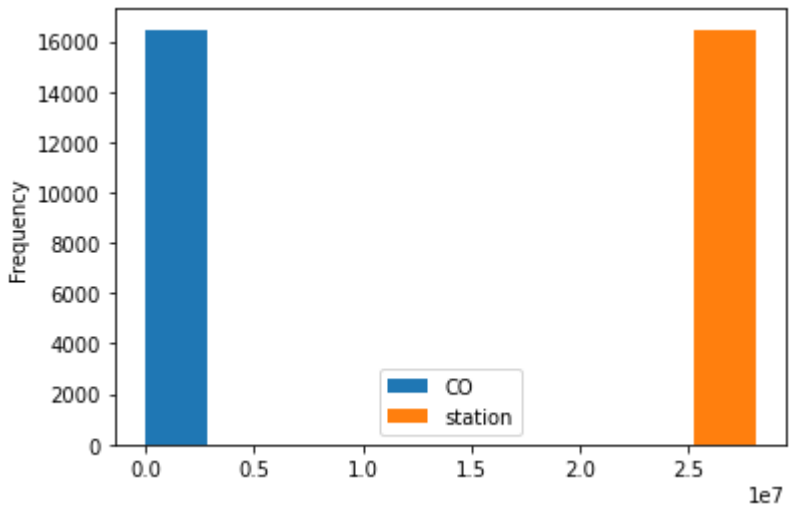
Out[10]: <AxesSubplot:>



Histogram

In [11]:

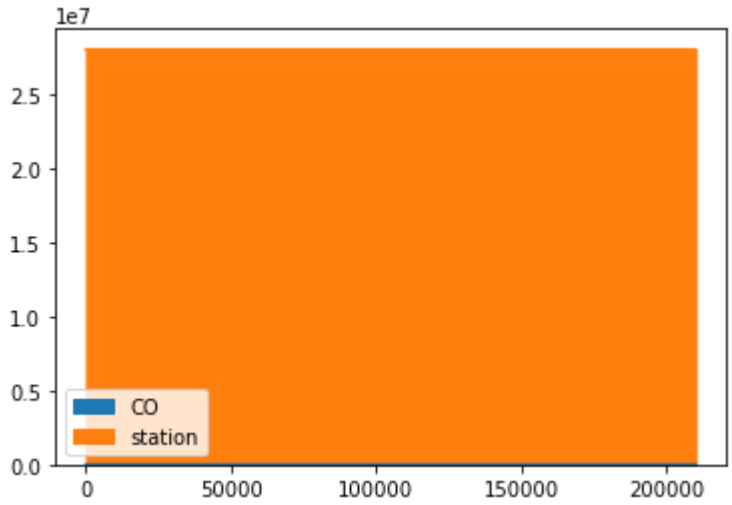
Out[11]: <AxesSubplot:ylabel='Frequency'>



Area chart

In [12]:

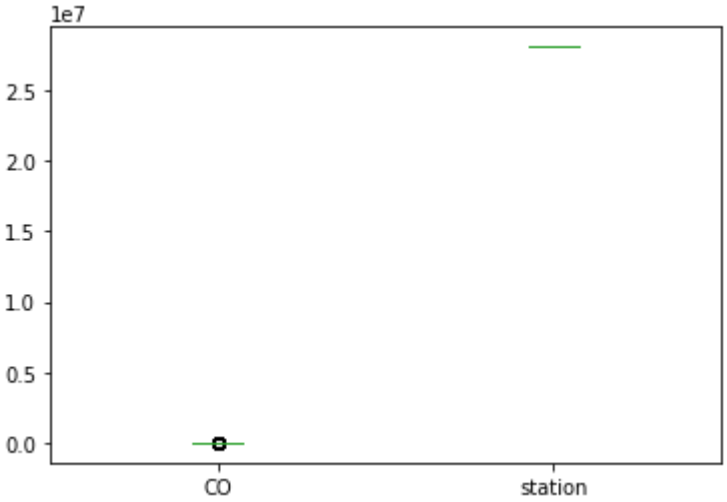
Out[12]: <AxesSubplot:>



Box chart

In [13]:

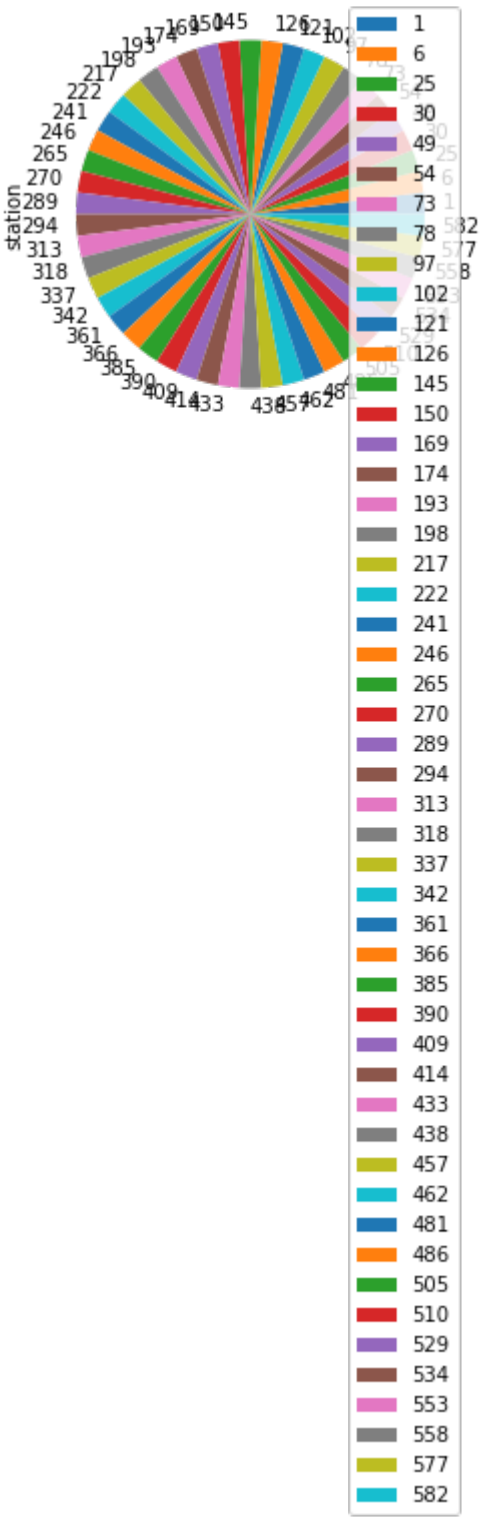
Out[13]: <AxesSubplot:>



Pie chart

In [14]:

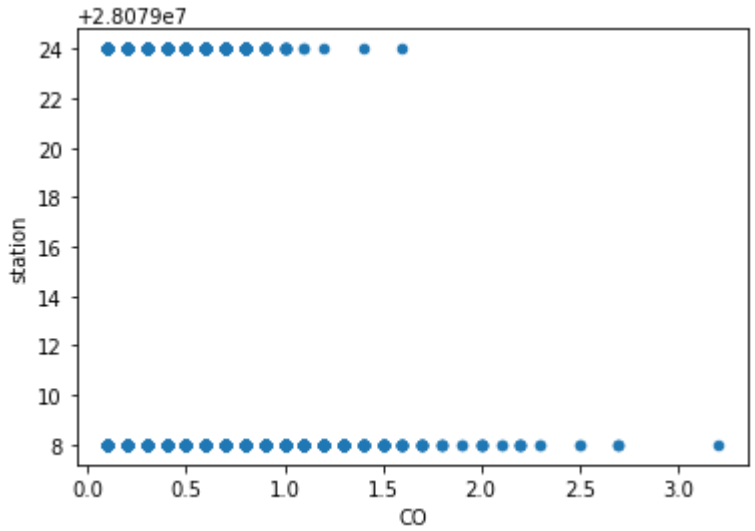
Out[14]: <AxesSubplot:ylabel='station'>



Scatter chart

In [15]:

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



In [16]:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 16460 entries, 1 to 209910
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     16460 non-null  object 
 1   BEN      16460 non-null  float64
 2   CO       16460 non-null  float64
 3   EBE      16460 non-null  float64
 4   NMHC     16460 non-null  float64
 5   NO       16460 non-null  float64
 6   NO_2     16460 non-null  float64
 7   O_3      16460 non-null  float64
 8   PM10     16460 non-null  float64
 9   PM25     16460 non-null  float64
 10  SO_2     16460 non-null  float64
 11  TCH      16460 non-null  float64
 12  TOL      16460 non-null  float64
 13  station  16460 non-null  int64
```

In [17]:

Out[17]:

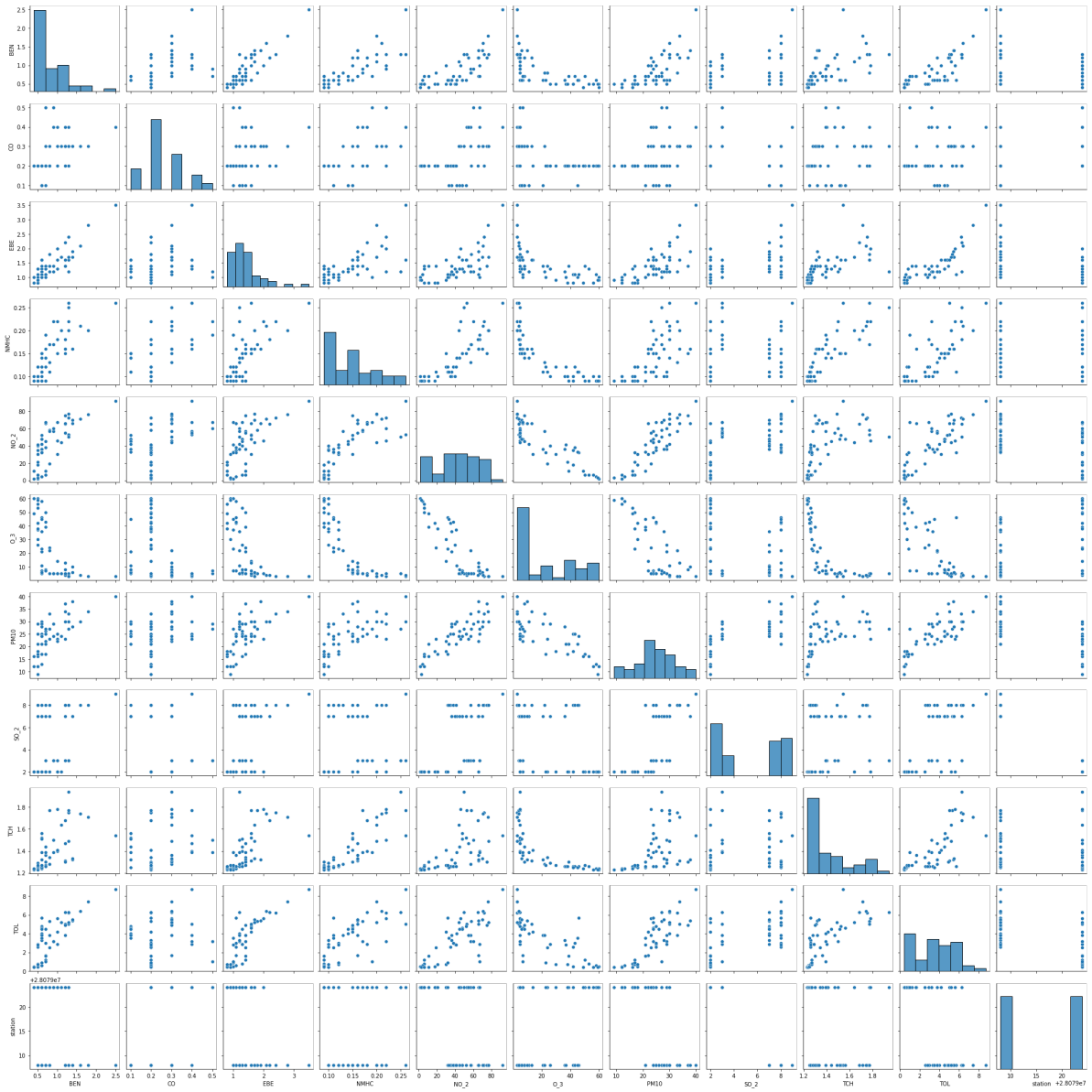
	BEN	CO	EBE	NMHC	NO	NO_2	
count	16460.000000	16460.000000	16460.000000	16460.000000	16460.000000	16460.000000	164
mean	0.900680	0.277758	1.471871	0.167043	23.671810	44.583961	
std	0.768892	0.206143	1.051004	0.075068	44.362859	31.569185	
min	0.100000	0.100000	0.200000	0.010000	1.000000	1.000000	
25%	0.500000	0.200000	0.800000	0.120000	2.000000	19.000000	
50%	0.700000	0.200000	1.200000	0.160000	7.000000	40.000000	
75%	1.100000	0.300000	1.700000	0.200000	25.000000	63.000000	
max	9.500000	3.200000	12.800000	0.840000	615.000000	289.000000	1

In [19]:
```
df1=df[['BEN', 'CO', 'EBE','NMHC', 'NO_2', 'O_3',
```

EDA AND VISUALIZATION

In [20]:

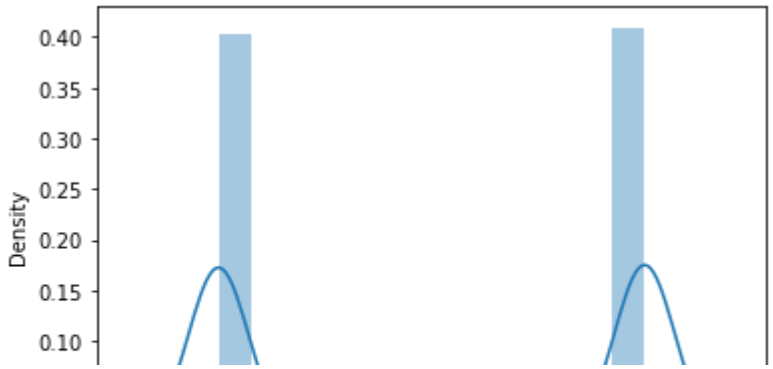
Out[20]: <seaborn.axisgrid.PairGrid at 0x2875c467610>



In [21]:

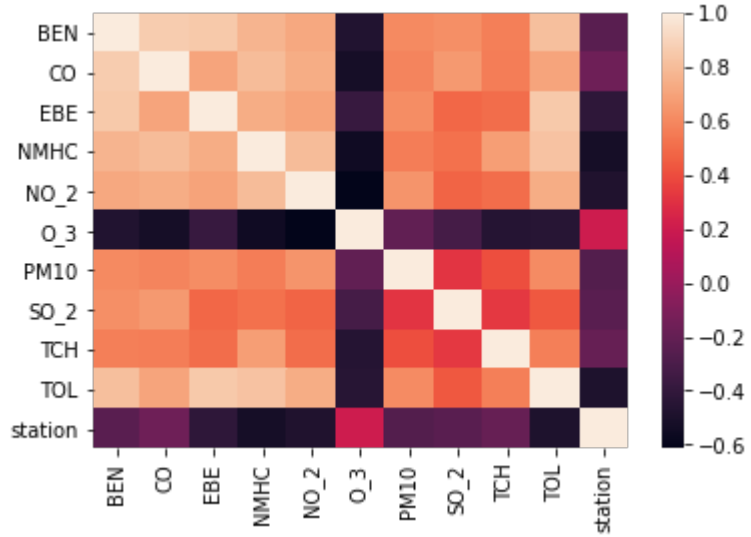
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: Fut
ureWarning: `distplot` is a deprecated function and will be removed in a futu
re version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function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

Out[21]: <AxesSubplot:xlabel='station', ylabel='Density'>



In [22]:

Out[22]: <AxesSubplot:>



TO TRAIN THE MODEL AND MODEL BULDING

In [24]:

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

In [25]:

```
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 [26]:
```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
```

Out[26]:
```
LinearRegression()
```

In [27]:
```
lr.intercept_
```

Out[27]:
```
28079017.449205846
```

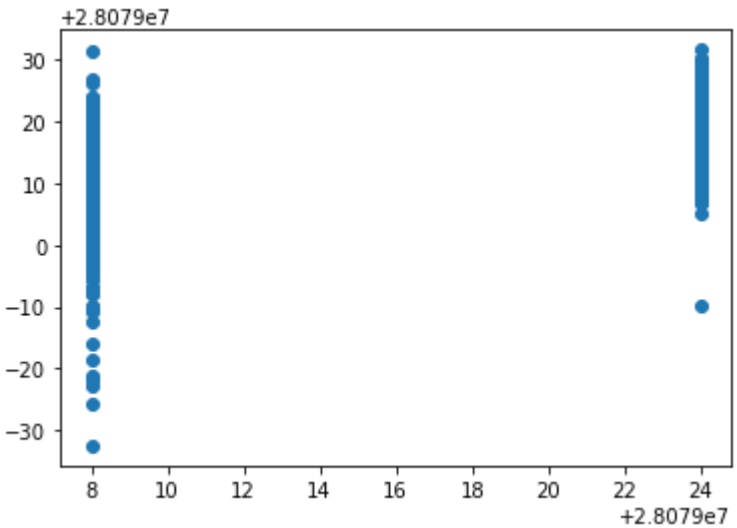
In [28]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
```

Out[28]:

	Co-efficient
BEN	3.828395
CO	32.621367
EBE	-1.974801
NMHC	-99.317516
NO_2	-0.083379
O_3	-0.016319
PM10	0.001411
SO_2	-0.498409
TCH	10.613384
TOL	-0.450018

In [29]:
```
prediction =lr.predict(x_test)
```

Out[29]:
```
<matplotlib.collections.PathCollection at 0x28765ff79a0>
```



ACCURACY

In [30]:

Out[30]: 0.6056543716808311

In [31]:

Out[31]: 0.6283925193702642

Ridge and Lasso

In [32]:

In [33]:
```
rr=Ridge(alpha=10)
```

Out[33]: Ridge(alpha=10)

Accuracy(Ridge)

In [34]:

Out[34]: 0.5832993683437642

In [36]:

Out[36]: 0.5918893311823653

In [37]:
```
la=Lasso(alpha=10)
```

Out[37]: Lasso(alpha=10)

In [39]:

Out[39]: 0.23193056355241382

Accuracy(Lasso)

In [40]:

Out[40]: 0.2395064520519974

In [41]:
```
from sklearn.linear_model import ElasticNet
en=ElasticNet()
en.fit(x_train,y_train)
```

Out[41]: ElasticNet()

In [42]:

Out[42]:
```
array([ 0.63039186,  0.        , -0.        , -0.        , -0.1319198 ,
       -0.05300076,  0.07606271, -0.        ,  0.        , -0.72625177])
```

In [43]:

Out[43]:
```
28079024.262245778
```

In [45]:
```
prediction=en.predict(x_test)
```

In [46]:

Out[46]:
```
0.3263204869423424
```

Evaluation Metrics

In [47]:
```
from sklearn import metrics
print(metrics.mean_absolute_error(y_test,prediction))
print(metrics.mean_squared_error(y_test,prediction))
```

```
5.8433173992865335
43.11503617674337
6.566204091919727
```

Logistic Regression

In [48]:

In [50]:
```
feature_matrix=df[['BEN', 'CO', 'EBE',  'NMHC', 'NO_2',  'O_3',
       'PM10', 'SO_2', 'TCH', 'TOL']]
```

In [51]:

Out[51]:
```
(16460, 10)
```

In [52]:

Out[52]:
```
(16460,)
```

In [53]:

In [54]:

In [55]:
```
logr=LogisticRegression(max_iter=10000)
```

Out[55]:
```
LogisticRegression(max_iter=10000)
```

In [62]:

In [63]:
```
prediction=logr.predict(observation)
```
```
[28079008]
```

In [58]:

Out[58]:
```
array([28079008, 28079024], dtype=int64)
```

In [59]:

Out[59]:
```
0.9237545565006076
```

In [64]:

Out[64]:
```
0.9999999999999966
```

In [66]:

Out[66]:
```
array([[1.00000000e+00, 3.47334507e-15]])
```

Random Forest

In [67]:

In [68]:
```
rfc=RandomForestClassifier()
```

Out[68]:
```
RandomForestClassifier()
```

In [69]:
```
parameters={'max_depth':[1,2,3,4,5],
            'min_samples_leaf':[5,10,15,20,25],
            'n_estimators':[10,20,30,40,50]
```

In [70]:
```
from sklearn.model_selection import GridSearchCV
grid_search =GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="ac
```

Out[70]:
```
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 [71]:

Out[71]:
```
0.9336920673494185
```

In [72]:

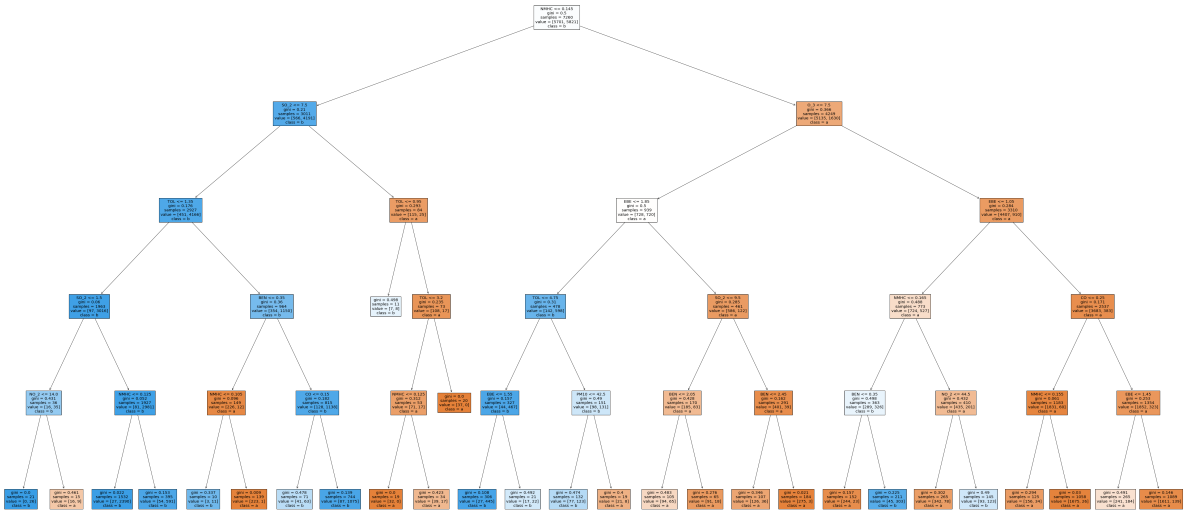
In [73]:
```
from sklearn.tree import plot_tree

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

Out[73]:
```
[Text(2103.230769230769, 1993.2, 'NMHC <= 0.145\ngini = 0.5\nsamples = 7260\n
value = [5701, 5821]\nclass = b'),
 Text(1116.0, 1630.8000000000002, 'SO_2 <= 7.5\ngini = 0.21\nsamples = 3011\n
value = [566, 4191]\nclass = b'),
 Text(686.7692307692307, 1268.4, 'TOL <= 1.35\ngini = 0.176\nsamples = 2927\n
value = [451, 4166]\nclass = b'),
 Text(343.38461538461536, 906.0, 'SO_2 <= 1.5\ngini = 0.06\nsamples = 1963\nv
alue = [97, 3016]\nclass = b'),
 Text(171.69230769230768, 543.5999999999999, 'NO_2 <= 14.0\ngini = 0.431\nsam
ples = 36\nvalue = [16, 35]\nclass = b'),
 Text(85.84615384615384, 181.19999999999982, 'gini = 0.0\nsamples = 21\nvalue
= [0, 26]\nclass = b'),
 Text(257.53846153846155, 181.19999999999982, 'gini = 0.461\nsamples = 15\nva
lue = [16, 9]\nclass = a'),
 Text(515.0769230769231, 543.5999999999999, 'NMHC <= 0.125\ngini = 0.052\nsam
ples = 1927\nvalue = [81, 2981]\nclass = b'),
 Text(429.23076923076917, 181.19999999999982, 'gini = 0.022\nsamples = 1532\n
value = [27, 2390]\nclass = b'),
 Text(600.9230769230769, 181.19999999999982, 'gini = 0.153\nsamples = 395\nva
lue = [54, 591]\nclass = b'),
 Text(1030.1538461538462, 906.0, 'BEN <= 0.35\ngini = 0.36\nsamples = 964\nva
lue = [354, 1150]\nclass = b'),
 Text(858.4615384615383, 543.5999999999999, 'NMHC <= 0.105\ngini = 0.096\nsam
ples = 149\nvalue = [226, 12]\nclass = a'),
 Text(772.6153846153845, 181.19999999999982, 'gini = 0.337\nsamples = 10\nval
ue = [3, 11]\nclass = b'),
 Text(944.3076923076923, 181.19999999999982, 'gini = 0.009\nsamples = 139\nva
lue = [223, 1]\nclass = a'),
 Text(1201.8461538461538, 543.5999999999999, 'CO <= 0.15\ngini = 0.182\nsampl
es = 815\nvalue = [128, 1138]\nclass = b'),
 Text(1116.0, 181.19999999999982, 'gini = 0.478\nsamples = 71\nvalue = [41, 6
3]\nclass = b'),
 Text(1287.6923076923076, 181.19999999999982, 'gini = 0.139\nsamples = 744\nv
alue = [87, 1075]\nclass = b'),
 Text(1545.230769230769, 1268.4, 'TOL <= 0.95\ngini = 0.293\nsamples = 84\nva
lue = [115, 25]\nclass = a'),
 Text(1459.3846153846152, 906.0, 'gini = 0.498\nsamples = 11\nvalue = [7, 8]\
nclass = b'),
 Text(1631.0769230769229, 906.0, 'TOL <= 3.2\ngini = 0.235\nsamples = 73\nval
ue = [108, 17]\nclass = a'),
 Text(1545.230769230769, 543.5999999999999, 'NMHC <= 0.125\ngini = 0.312\nsam
ples = 53\nvalue = [71, 17]\nclass = a'),
 Text(1459.3846153846152, 181.19999999999982, 'gini = 0.0\nsamples = 19\nvalu
e = [32, 0]\nclass = a'),
 Text(1631.0769230769229, 181.19999999999982, 'gini = 0.423\nsamples = 34\nva
lue = [39, 17]\nclass = a'),
 Text(1716.9230769230767, 543.5999999999999, 'gini = 0.0\nsamples = 20\nvalue
= [37, 0]\nclass = a'),
 Text(3090.461538461538, 1630.8000000000002, 'O_3 <= 7.5\ngini = 0.366\nsampl
es = 4249\nvalue = [5135, 1630]\nclass = a'),
 Text(2403.6923076923076, 1268.4, 'EBE <= 1.85\ngini = 0.5\nsamples = 939\nva
lue = [728, 720]\nclass = a'),
```

```
 Text(2060.3076923076924, 906.0, 'TOL <= 4.75\ngini = 0.31\nsamples = 478\nva
lue = [142, 598]\nclass = b'),
 Text(1888.6153846153845, 543.5999999999999, 'EBE <= 1.55\ngini = 0.157\nsamp
les = 327\nvalue = [44, 467]\nclass = b'),
 Text(1802.7692307692307, 181.19999999999982, 'gini = 0.108\nsamples = 306\nv
alue = [27, 445]\nclass = b'),
 Text(1974.4615384615383, 181.19999999999982, 'gini = 0.492\nsamples = 21\nva
lue = [17, 22]\nclass = b'),
 Text(2232.0, 543.5999999999999, 'PM10 <= 42.5\ngini = 0.49\nsamples = 151\nv
alue = [98, 131]\nclass = b'),
 Text(2146.153846153846, 181.19999999999982, 'gini = 0.474\nsamples = 132\nva
lue = [77, 123]\nclass = b'),
 Text(2317.846153846154, 181.19999999999982, 'gini = 0.4\nsamples = 19\nvalue
= [21, 8]\nclass = a'),
 Text(2747.076923076923, 906.0, 'SO_2 <= 9.5\ngini = 0.285\nsamples = 461\nva
lue = [586, 122]\nclass = a'),
 Text(2575.3846153846152, 543.5999999999999, 'BEN <= 2.05\ngini = 0.428\nsamp
les = 170\nvalue = [185, 83]\nclass = a'),
 Text(2489.5384615384614, 181.19999999999982, 'gini = 0.483\nsamples = 105\nv
alue = [94, 65]\nclass = a'),
 Text(2661.230769230769, 181.19999999999982, 'gini = 0.276\nsamples = 65\nval
ue = [91, 18]\nclass = a'),
 Text(2918.7692307692305, 543.5999999999999, 'BEN <= 2.45\ngini = 0.162\nsamp
les = 291\nvalue = [401, 39]\nclass = a'),
 Text(2832.9230769230767, 181.19999999999982, 'gini = 0.346\nsamples = 107\nv
alue = [126, 36]\nclass = a'),
 Text(3004.6153846153843, 181.19999999999982, 'gini = 0.021\nsamples = 184\nv
alue = [275, 3]\nclass = a'),
 Text(3777.230769230769, 1268.4, 'EBE <= 1.05\ngini = 0.284\nsamples = 3310\n
value = [4407, 910]\nclass = a'),
 Text(3433.8461538461534, 906.0, 'NMHC <= 0.165\ngini = 0.488\nsamples = 773\
nvalue = [724, 527]\nclass = a'),
 Text(3262.1538461538457, 543.5999999999999, 'BEN <= 0.35\ngini = 0.498\nsamp
les = 363\nvalue = [289, 326]\nclass = b'),
 Text(3176.307692307692, 181.19999999999982, 'gini = 0.157\nsamples = 152\nva
lue = [244, 23]\nclass = a'),
 Text(3347.9999999999995, 181.19999999999982, 'gini = 0.225\nsamples = 211\nv
alue = [45, 303]\nclass = b'),
 Text(3605.5384615384614, 543.5999999999999, 'NO_2 <= 44.5\ngini = 0.432\nsam
ples = 410\nvalue = [435, 201]\nclass = a'),
 Text(3519.6923076923076, 181.19999999999982, 'gini = 0.302\nsamples = 265\nv
alue = [342, 78]\nclass = a'),
 Text(3691.3846153846152, 181.19999999999982, 'gini = 0.49\nsamples = 145\nva
lue = [93, 123]\nclass = b'),
 Text(4120.615384615385, 906.0, 'CO <= 0.25\ngini = 0.171\nsamples = 2537\nva
lue = [3683, 383]\nclass = a'),
 Text(3948.9230769230767, 543.5999999999999, 'NMHC <= 0.155\ngini = 0.061\nsa
mples = 1183\nvalue = [1831, 60]\nclass = a'),
 Text(3863.076923076923, 181.19999999999982, 'gini = 0.294\nsamples = 125\nva
lue = [156, 34]\nclass = a'),
 Text(4034.7692307692305, 181.19999999999982, 'gini = 0.03\nsamples = 1058\nv
alue = [1675, 26]\nclass = a'),
 Text(4292.307692307692, 543.5999999999999, 'EBE <= 1.45\ngini = 0.253\nsampl
es = 1354\nvalue = [1852, 323]\nclass = a'),
 Text(4206.461538461538, 181.19999999999982, 'gini = 0.491\nsamples = 265\nva
lue = [241, 184]\nclass = a'),
```

```
 Text(4378.153846153846, 181.19999999999982, 'gini = 0.146\nsamples = 1089\nv
alue = [1611, 139]\nclass = a')]
```



Conclusion

Accuracy

Linear Regression :0.6283925193702642

Ridge Regression :0.23193056355241382

Lasso Regression :0.2395064520519974

ElasticNet Regression : 0.3263204869423424

Logistic Regression : 0.9237545565006076

Random Forest :0.9336920673494185

Random Forest is suitable for this dataset