Importing Libraries

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Importing Datasets ¶

In [2]:
```
df=pd.read_csv("madrid_2004.csv")
df
```

Out[2]:

	date	BEN	CO	EBE	MXY	NMHC	NO_2	NOx	OXY	O_3	P
0	2004-08-01 01:00:00	NaN	0.66	NaN	NaN	NaN	89.550003	118.900002	NaN	40.020000	39.99(
1	2004-08-01 01:00:00	2.66	0.54	2.99	6.08	0.18	51.799999	53.860001	3.28	51.689999	22.95(
2	2004-08-01 01:00:00	NaN	1.02	NaN	NaN	NaN	93.389999	138.600006	NaN	20.860001	49.48(
3	2004-08-01 01:00:00	NaN	0.53	NaN	NaN	NaN	87.290001	105.000000	NaN	36.730000	31.07(
4	2004-08-01 01:00:00	NaN	0.17	NaN	NaN	NaN	34.910000	35.349998	NaN	86.269997	54.08(
...	...	...	...	...	...	...	...	...	...	...	
245491	2004-06-01 00:00:00	0.75	0.21	0.85	1.55	0.07	59.580002	64.389999	0.66	33.029999	30.90(
245492	2004-06-01 00:00:00	2.49	0.75	2.44	4.57	NaN	97.139999	146.899994	2.34	7.740000	37.68
245493	2004-06-01 00:00:00	NaN	NaN	NaN	NaN	0.13	102.699997	132.600006	NaN	17.809999	22.84(
245494	2004-06-01 00:00:00	NaN	NaN	NaN	NaN	0.09	82.599998	102.599998	NaN	NaN	45.63(
245495	2004-06-01 00:00:00	3.01	0.67	2.78	5.12	0.20	92.550003	141.000000	2.60	11.460000	24.38

245496 rows × 17 columns

Data Cleaning and Data Preprocessing

In [3]:
```
df=df.dropna()
```

In [4]:
```
df.columns
```

Out[4]:
```
Index(['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_
3',
       'PM10', 'PM25', 'PXY', 'SO_2', 'TCH', 'TOL', 'station'],
      dtype='object')
```

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19397 entries, 5 to 245495
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     19397 non-null  object 
 1   BEN      19397 non-null  float64
 2   CO       19397 non-null  float64
 3   EBE      19397 non-null  float64
 4   MXY      19397 non-null  float64
 5   NMHC     19397 non-null  float64
 6   NO_2     19397 non-null  float64
 7   NOx      19397 non-null  float64
 8   OXY      19397 non-null  float64
 9   O_3      19397 non-null  float64
 10  PM10     19397 non-null  float64
 11  PM25     19397 non-null  float64
 12  PXY      19397 non-null  float64
 13  SO_2     19397 non-null  float64
 14  TCH      19397 non-null  float64
 15  TOL      19397 non-null  float64
 16  station  19397 non-null  int64  
dtypes: float64(15), int64(1), object(1)
memory usage: 2.7+ MB
```

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

Out[6]:

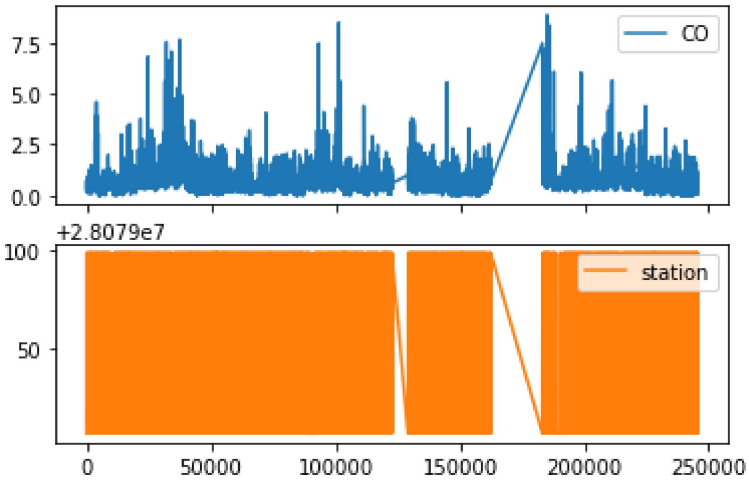
	CO	station
5	0.63	28079006
22	0.36	28079024
26	0.46	28079099
32	0.67	28079006
49	0.30	28079024
...	...	...
245463	0.08	28079024
245467	0.67	28079099
245473	1.12	28079006
245491	0.21	28079024
245495	0.67	28079099

19397 rows × 2 columns

Line chart

In [7]:
```
data.plot.line(subplots=True)
```

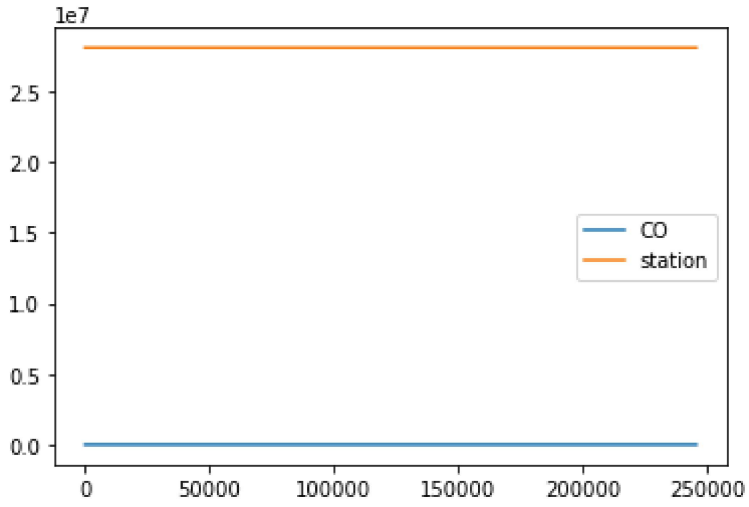
Out[7]:
```
array([<AxesSubplot:>, <AxesSubplot:>], dtype=object)
```



Line chart

In [8]:
```
data.plot.line()
```

Out[8]: <AxesSubplot:>

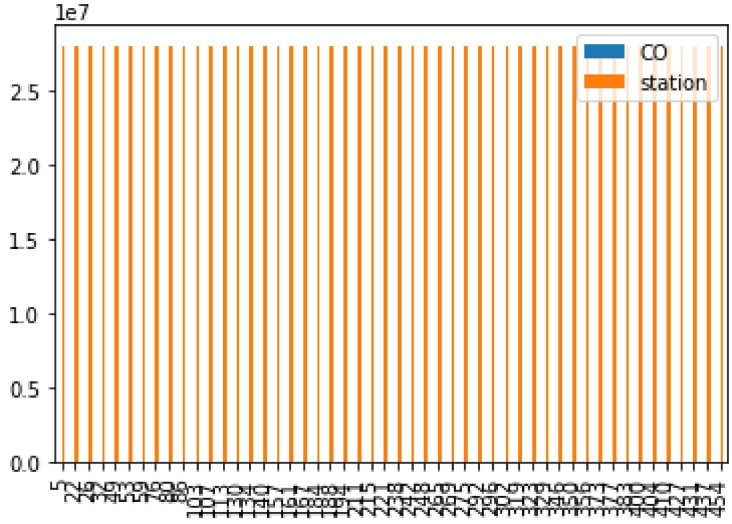


Bar chart

In [9]:
```
b=data[0:50]
```

In [10]:
```
b.plot.bar()
```

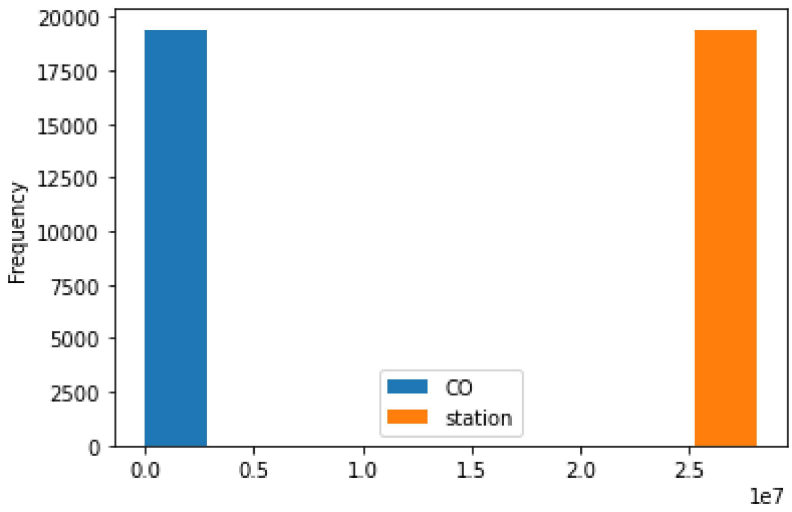
Out[10]: <AxesSubplot:>



Histogram

In [11]:
```
data.plot.hist()
```

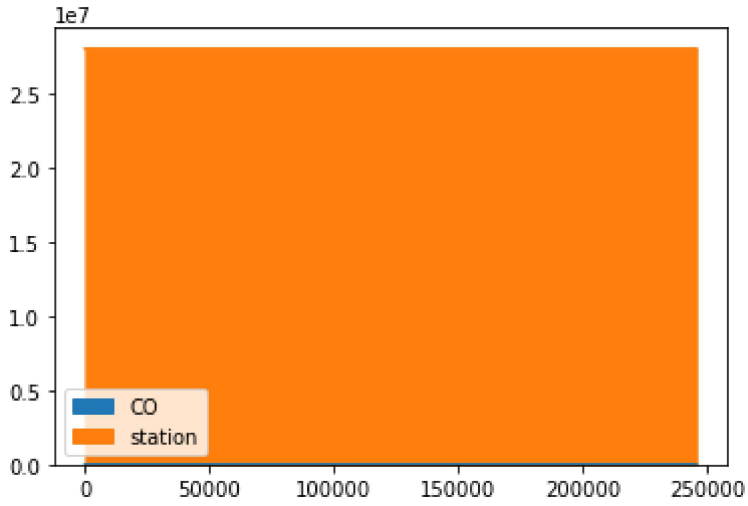
Out[11]: <AxesSubplot:ylabel='Frequency'>



Area chart

In [12]:
```
data.plot.area()
```

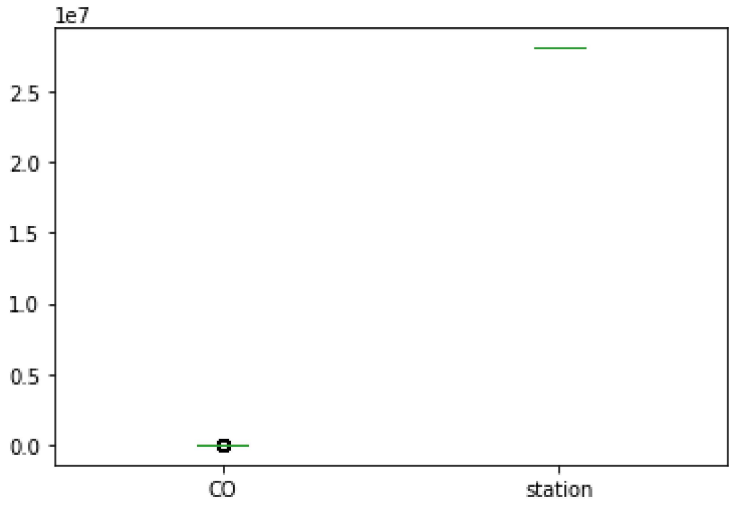
Out[12]: <AxesSubplot:>



Box chart

In [13]:
```
data.plot.box()
```

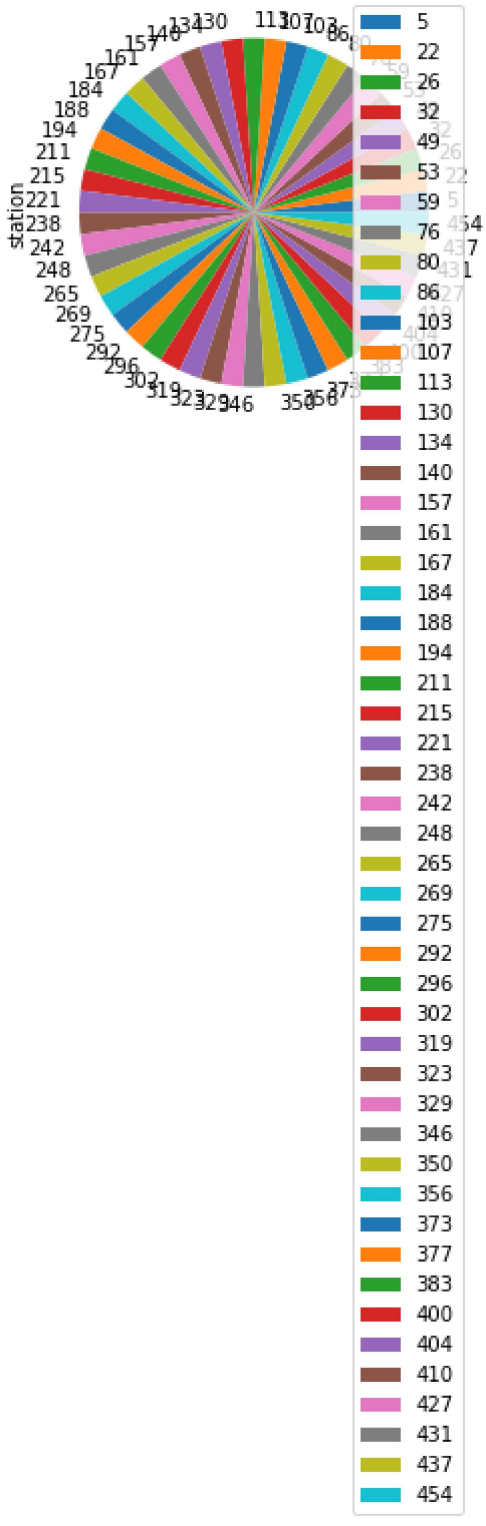
Out[13]: <AxesSubplot:>



Pie chart

In [14]:
```
b.plot.pie(y='station' )
```

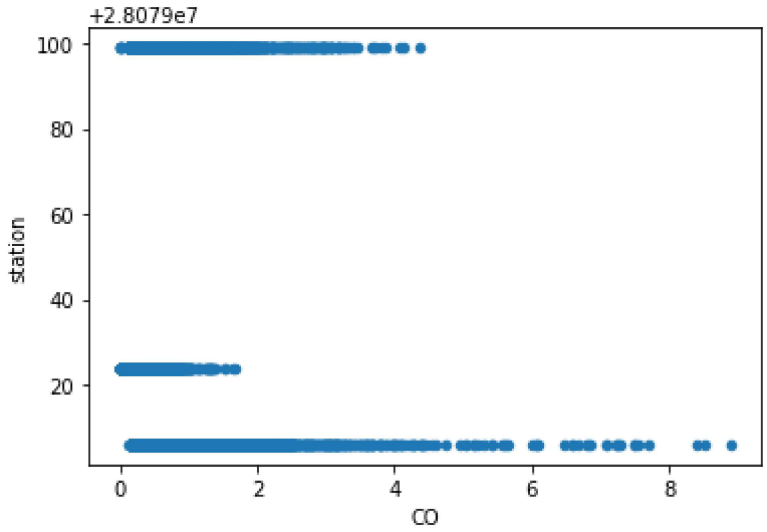
Out[14]: <AxesSubplot:ylabel='station'>



Scatter chart

In [15]:
```
data.plot.scatter(x='CO' ,y='station')
```

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



In [16]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 19397 entries, 5 to 245495
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     19397 non-null  object 
 1   BEN      19397 non-null  float64
 2   CO       19397 non-null  float64
 3   EBE      19397 non-null  float64
 4   MXY      19397 non-null  float64
 5   NMHC     19397 non-null  float64
 6   NO_2     19397 non-null  float64
 7   NOx      19397 non-null  float64
 8   OXY      19397 non-null  float64
 9   O_3      19397 non-null  float64
 10  PM10     19397 non-null  float64
 11  PM25     19397 non-null  float64
 12  PXY      19397 non-null  float64
 13  SO_2     19397 non-null  float64
 14  TCH      19397 non-null  float64
```

In [17]:
```
df.describe()
```

Out[17]:

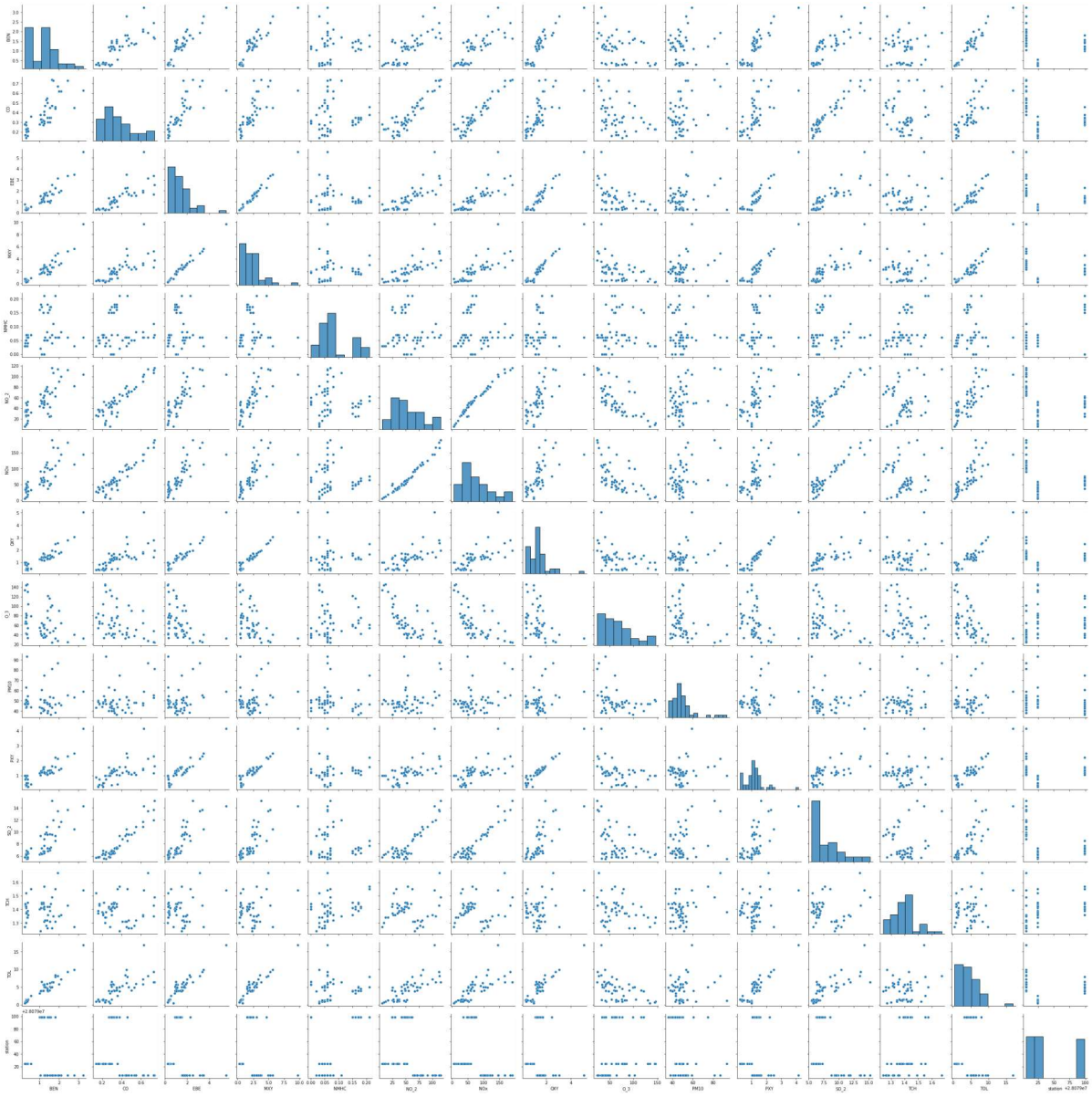
	BEN	CO	EBE	MXY	NMHC	NO_2	
count	19397.000000	19397.000000	19397.000000	19397.000000	19397.000000	19397.000000	193
mean	2.250781	0.675347	2.775913	5.424809	0.151024	62.887023	1
std	2.184724	0.591026	2.729622	5.554358	0.158603	37.952255	1
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.090000	
25%	0.870000	0.320000	1.020000	1.780000	0.060000	35.150002	
50%	1.620000	0.520000	1.970000	3.800000	0.110000	58.310001	
75%	2.910000	0.860000	3.580000	7.260000	0.200000	85.730003	1
max	34.180000	8.900000	41.880001	91.599998	4.810000	355.100006	17

In [18]:
```
df1=df[['BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3',
       'PM10', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']]
```

EDA AND VISUALIZATION

In [19]:
```
sns.pairplot(df1[0:50])
```

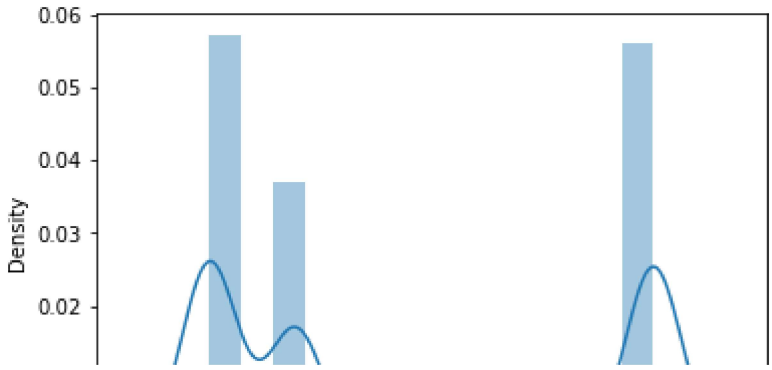
Out[19]:
```
<seaborn.axisgrid.PairGrid at 0x23df70b17f0>
```



In [20]:
```
sns.distplot(df1['station'])
```

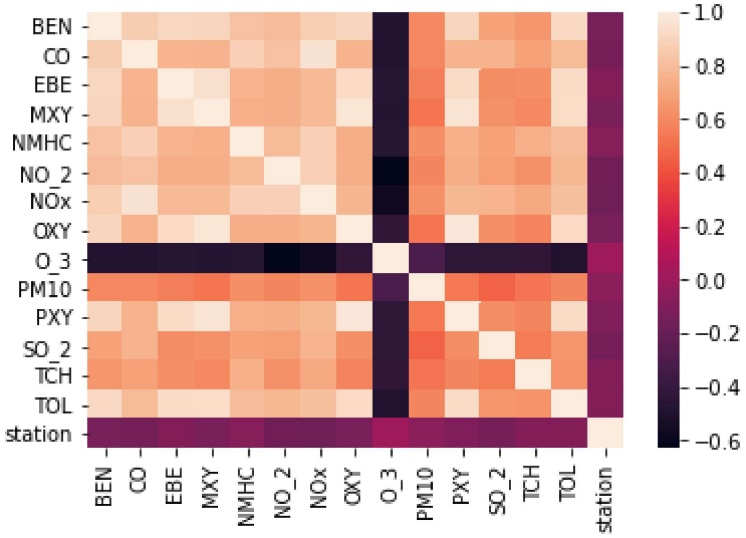
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: F
utureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-le
vel function with similar flexibility) or `histplot` (an axes-level functio
n for histograms).
  warnings.warn(msg, FutureWarning)
```

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



In [21]:
```
sns.heatmap(df1.corr())
```

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']]
y=df['station']
```

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()
lr.fit(x_train,y_train)
```

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

In [25]:
```
lr.intercept_
```

Out[25]:
```
28079074.048487846
```

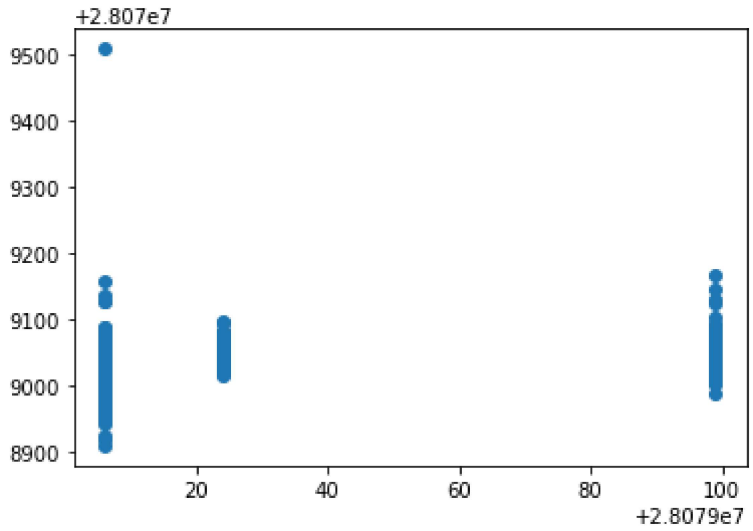
In [26]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

Out[26]:

	Co-efficient
BEN	-3.184705
CO	22.980565
EBE	3.137702
MXY	-3.517054
NMHC	108.597432
NO_2	-0.155018
NOx	-0.265446
OXY	-1.941774
O_3	-0.288851
PM10	0.081617
PXY	6.307638
SO_2	-0.176862
TCH	-5.877365
TOL	0.985554

In [27]:
```
prediction =lr.predict(x_test)
plt.scatter(y_test,prediction)
```

Out[27]: <matplotlib.collections.PathCollection at 0x23d860d7190>



ACCURACY

In [28]:
```
lr.score(x_test,y_test)
```

Out[28]: 0.10168717177729591

In [29]:
```
lr.score(x_train,y_train)
```

Out[29]: 0.1070216317851852

Ridge and Lasso

In [30]:
```
from sklearn.linear_model import Ridge,Lasso
```

In [31]:
```
rr=Ridge(alpha=10)
rr.fit(x_train,y_train)
```

Out[31]: Ridge(alpha=10)

Accuracy(Ridge)

In [32]:
```
rr.score(x_test,y_test)
```

Out[32]: 0.10553218701600331

In [33]:
```
rr.score(x_train,y_train)
```

Out[33]:
```
0.10630389078272184
```

In [34]:
```
la=Lasso(alpha=10)
la.fit(x_train,y_train)
```

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

In [35]:
```
la.score(x_train,y_train)
```

Out[35]:
```
0.05171973007134234
```

Accuracy(Lasso)

In [36]:
```
la.score(x_test,y_test)
```

Out[36]:
```
0.05686434388098016
```

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

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

In [38]:
```
en.coef_
```

Out[38]:
```
array([-0.        ,  0.42118511,  1.37824819, -1.91764273,  0.        ,
       -0.1701129 , -0.09029162, -0.        , -0.21386947,  0.12318448,
        0.56066268, -0.08432894,  0.        ,  1.15467386])
```

In [39]:
```
en.intercept_
```

Out[39]:
```
28079065.552796025
```

In [40]:
```
prediction=en.predict(x_test)
```

In [41]:
```
en.score(x_test,y_test)
```

Out[41]:
```
0.07140776525431125
```

Evaluation Metrics

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

```
38.469931908028
1655.7960308014744
40.69147368677466
```

Logistic Regression

In [43]:
```
from sklearn.linear_model import LogisticRegression
```

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

In [45]:
```
feature_matrix.shape
```

Out[45]:
```
(19397, 14)
```

In [46]:
```
target_vector.shape
```

Out[46]:
```
(19397,)
```

In [47]:
```
from sklearn.preprocessing import StandardScaler
```

In [48]:
```
fs=StandardScaler().fit_transform(feature_matrix)
```

In [49]:
```
logr=LogisticRegression(max_iter=10000)
logr.fit(fs,target_vector)
```

Out[49]:
```
LogisticRegression(max_iter=10000)
```

In [50]:
```
observation=[[1,2,3,4,5,6,7,8,9,10,11,12,13,14]]
```

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

```
[28079006]
```

In [52]:
```
logr.classes_
```

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

In [53]:
```
logr.score(fs,target_vector)
```

Out[53]:
```
0.7360416559261741
```

In [54]:
```
logr.predict_proba(observation)[0][0]
```

Out[54]:
```
0.9999978255573396
```

In [55]:
```
logr.predict_proba(observation)
```

Out[55]:
```
array([[9.99997826e-01, 7.75018107e-20, 2.17444266e-06]])
```

Random Forest

In [56]:
```
from sklearn.ensemble import RandomForestClassifier
```

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
grid_search.fit(x_train,y_train)
```

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]:
```
grid_search.best_score_
```

Out[60]:
```
0.772851274284817
```

In [61]:
```
rfc_best=grid_search.best_estimator_
```

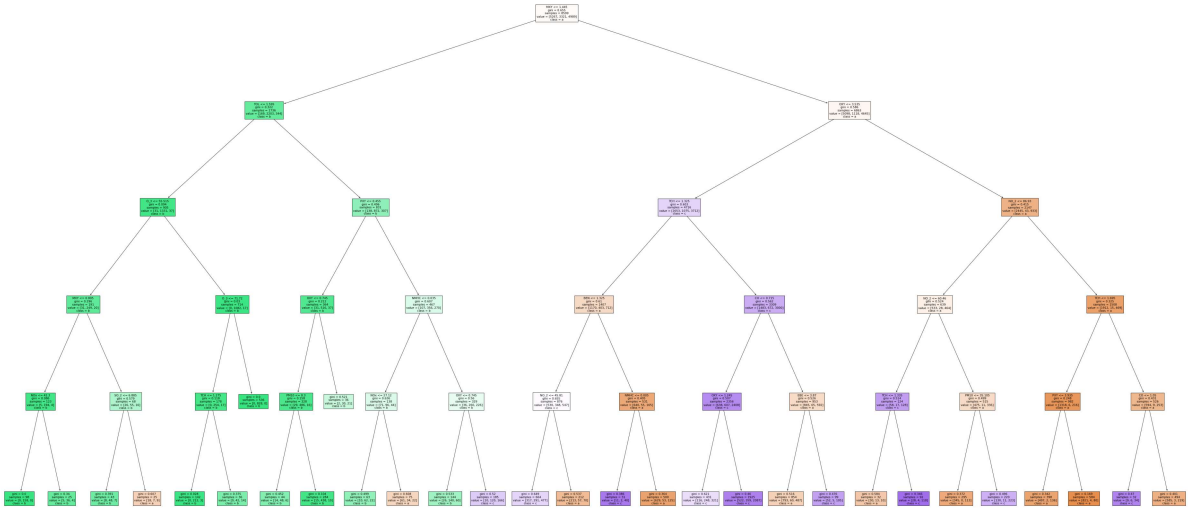
In [62]:
```
from sklearn.tree import plot_tree

plt.figure(figsize=(80,40))
plot_tree(rfc_best.estimators_[5],feature_names=x.columns,class_names=['a','b'
```

Out[62]:
```
[Text(2092.5, 1993.2, 'MXY <= 1.445\ngini = 0.655\nsamples = 8599\nvalue = [5
267, 3321, 4989]\nclass = a'),
 Text(996.4285714285713, 1630.8000000000002, 'TOL <= 1.595\ngini = 0.322\nsam
ples = 1736\nvalue = [169, 2203, 344]\nclass = b'),
 Text(597.8571428571428, 1268.4, 'O_3 <= 55.515\ngini = 0.094\nsamples = 905
\nvalue = [31, 1331, 37]\nclass = b'),
 Text(318.85714285714283, 906.0, 'MXY <= 0.805\ngini = 0.296\nsamples = 191\n
value = [31, 249, 20]\nclass = b'),
 Text(159.42857142857142, 543.5999999999999, 'NOx <= 42.3\ngini = 0.086\nsamp
les = 123\nvalue = [5, 194, 4]\nclass = b'),
 Text(79.71428571428571, 181.19999999999982, 'gini = 0.0\nsamples = 98\nvalue
= [0, 158, 0]\nclass = b'),
 Text(239.1428571428571, 181.19999999999982, 'gini = 0.34\nsamples = 25\nvalu
e = [5, 36, 4]\nclass = b'),
 Text(478.2857142857142, 543.5999999999999, 'SO_2 <= 6.885\ngini = 0.579\nsam
ples = 68\nvalue = [26, 55, 16]\nclass = b'),
 Text(398.57142857142856, 181.19999999999982, 'gini = 0.391\nsamples = 43\nva
lue = [8, 48, 7]\nclass = b'),
 Text(558.0, 181.19999999999982, 'gini = 0.607\nsamples = 25\nvalue = [18, 7,
9]\nclass = a'),
 Text(876.8571428571428, 906.0, 'O_3 <= 71.72\ngini = 0.03\nsamples = 714\nva
lue = [0, 1082, 17]\nclass = b'),
 Text(797.1428571428571, 543.5999999999999, 'TCH <= 1.275\ngini = 0.118\nsamp
les = 178\nvalue = [0, 254, 17]\nclass = b'),
 Text(717.4285714285713, 181.19999999999982, 'gini = 0.028\nsamples = 142\nva
lue = [0, 212, 3]\nclass = b'),
 Text(876.8571428571428, 181.19999999999982, 'gini = 0.375\nsamples = 36\nval
ue = [0, 42, 14]\nclass = b'),
 Text(956.5714285714284, 543.5999999999999, 'gini = 0.0\nsamples = 536\nvalue
= [0, 828, 0]\nclass = b'),
 Text(1395.0, 1268.4, 'PXY <= 0.455\ngini = 0.496\nsamples = 831\nvalue = [13
8, 872, 307]\nclass = b'),
 Text(1195.7142857142856, 906.0, 'OXY <= 0.745\ngini = 0.212\nsamples = 364\n
value = [31, 516, 37]\nclass = b'),
 Text(1116.0, 543.5999999999999, 'PM10 <= 8.3\ngini = 0.158\nsamples = 328\nv
alue = [29, 486, 16]\nclass = b'),
 Text(1036.2857142857142, 181.19999999999982, 'gini = 0.452\nsamples = 44\nva
lue = [14, 48, 6]\nclass = b'),
 Text(1195.7142857142856, 181.19999999999982, 'gini = 0.104\nsamples = 284\nv
alue = [15, 438, 10]\nclass = b'),
 Text(1275.4285714285713, 543.5999999999999, 'gini = 0.521\nsamples = 36\nval
ue = [2, 30, 21]\nclass = b'),
 Text(1594.2857142857142, 906.0, 'NMHC <= 0.035\ngini = 0.607\nsamples = 467
\nvalue = [107, 356, 270]\nclass = b'),
 Text(1434.8571428571427, 543.5999999999999, 'NOx <= 27.12\ngini = 0.636\nsam
ples = 138\nvalue = [71, 96, 44]\nclass = b'),
 Text(1355.142857142857, 181.19999999999982, 'gini = 0.499\nsamples = 63\nval
ue = [10, 62, 22]\nclass = b'),
 Text(1514.5714285714284, 181.19999999999982, 'gini = 0.608\nsamples = 75\nva
lue = [61, 34, 22]\nclass = a'),
 Text(1753.7142857142856, 543.5999999999999, 'OXY <= 0.745\ngini = 0.56\nsamp
les = 329\nvalue = [36, 260, 226]\nclass = b'),
 Text(1673.9999999999998, 181.19999999999982, 'gini = 0.533\nsamples = 144\nv
alue = [26, 140, 60]\nclass = b'),
 Text(1833.4285714285713, 181.19999999999982, 'gini = 0.52\nsamples = 185\nva
lue = [10, 120, 166]\nclass = c'),
 Text(3188.5714285714284, 1630.8000000000002, 'OXY <= 3.535\ngini = 0.586\nsa
```

mples = 6863\nvalue = [5098, 1118, 4645]\nclass = a'),
 Text(2550.8571428571427, 1268.4, 'TCH <= 1.325\ngini = 0.603\nsamples = 4716
\nvalue = [2653, 1075, 3712]\nclass = c'),
 Text(2232.0, 906.0, 'BEN <= 1.325\ngini = 0.61\nsamples = 1407\nvalue = [117
0, 403, 712]\nclass = a'),
 Text(2072.5714285714284, 543.5999999999999, 'NO_2 <= 45.81\ngini = 0.655\nsa
mples = 876\nvalue = [530, 348, 547]\nclass = c'),
 Text(1992.8571428571427, 181.19999999999982, 'gini = 0.649\nsamples = 664\nv
alue = [317, 291, 477]\nclass = c'),
 Text(2152.285714285714, 181.19999999999982, 'gini = 0.537\nsamples = 212\nva
lue = [213, 57, 70]\nclass = a'),
 Text(2391.428571428571, 543.5999999999999, 'NMHC <= 0.005\ngini = 0.405\nsam
ples = 531\nvalue = [640, 55, 165]\nclass = a'),
 Text(2311.7142857142853, 181.19999999999982, 'gini = 0.386\nsamples = 31\nva
lue = [11, 2, 40]\nclass = c'),
 Text(2471.142857142857, 181.19999999999982, 'gini = 0.364\nsamples = 500\nva
lue = [629, 53, 125]\nclass = a'),
 Text(2869.7142857142853, 906.0, 'CO <= 0.715\ngini = 0.562\nsamples = 3309\n
value = [1483, 672, 3000]\nclass = c'),
 Text(2710.285714285714, 543.5999999999999, 'OXY <= 1.245\ngini = 0.507\nsamp
les = 2356\nvalue = [638, 607, 2408]\nclass = c'),
 Text(2630.5714285714284, 181.19999999999982, 'gini = 0.621\nsamples = 431\nv
alue = [116, 248, 321]\nclass = c'),
 Text(2790.0, 181.19999999999982, 'gini = 0.46\nsamples = 1925\nvalue = [522,
359, 2087]\nclass = c'),
 Text(3029.142857142857, 543.5999999999999, 'EBE <= 3.87\ngini = 0.526\nsampl
es = 953\nvalue = [845, 65, 592]\nclass = a'),
 Text(2949.428571428571, 181.19999999999982, 'gini = 0.516\nsamples = 854\nva
lue = [793, 60, 487]\nclass = a'),
 Text(3108.8571428571427, 181.19999999999982, 'gini = 0.476\nsamples = 99\nva
lue = [52, 5, 105]\nclass = c'),
 Text(3826.2857142857138, 1268.4, 'NO_2 <= 86.93\ngini = 0.415\nsamples = 214
7\nvalue = [2445, 43, 933]\nclass = a'),
 Text(3507.428571428571, 906.0, 'NO_2 <= 60.46\ngini = 0.524\nsamples = 639\n
value = [533, 28, 464]\nclass = a'),
 Text(3347.9999999999995, 543.5999999999999, 'TCH <= 1.335\ngini = 0.514\nsam
ples = 124\nvalue = [58, 17, 128]\nclass = c'),
 Text(3268.285714285714, 181.19999999999982, 'gini = 0.584\nsamples = 32\nval
ue = [30, 13, 10]\nclass = a'),
 Text(3427.7142857142853, 181.19999999999982, 'gini = 0.346\nsamples = 92\nva
lue = [28, 4, 118]\nclass = c'),
 Text(3666.8571428571427, 543.5999999999999, 'PM10 <= 35.105\ngini = 0.499\ns
amples = 515\nvalue = [475, 11, 336]\nclass = a'),
 Text(3587.142857142857, 181.19999999999982, 'gini = 0.372\nsamples = 295\nva
lue = [345, 0, 113]\nclass = a'),
 Text(3746.5714285714284, 181.19999999999982, 'gini = 0.496\nsamples = 220\nv
alue = [130, 11, 223]\nclass = c'),
 Text(4145.142857142857, 906.0, 'TCH <= 1.695\ngini = 0.325\nsamples = 1508\n
value = [1912, 15, 469]\nclass = a'),
 Text(3985.7142857142853, 543.5999999999999, 'PXY <= 3.935\ngini = 0.248\nsam
ples = 982\nvalue = [1318, 6, 216]\nclass = a'),
 Text(3905.9999999999995, 181.19999999999982, 'gini = 0.342\nsamples = 398\nv
alue = [497, 2, 136]\nclass = a'),
 Text(4065.428571428571, 181.19999999999982, 'gini = 0.169\nsamples = 584\nva
lue = [821, 4, 80]\nclass = a'),
 Text(4304.571428571428, 543.5999999999999, 'CO <= 1.05\ngini = 0.431\nsample
s = 526\nvalue = [594, 9, 253]\nclass = a'),

```
 Text(4224.857142857142, 181.19999999999982, 'gini = 0.47\nsamples = 32\nvalu
e = [9, 6, 34]\nclass = c'),
 Text(4384.285714285714, 181.19999999999982, 'gini = 0.401\nsamples = 494\nva
lue = [585, 3, 219]\nclass = a')]
```



Conclusion

Accuracy

Linear Regression:0.1070216317851852

Ridge Regression:0.10630389078272184

Lasso Regression:0.05171973007134234

ElasticNet Regression:0.07140776525431125

Logistic Regression:0.7360416559261741

Random Forest:0.772851274284817

Random Forest is suitable for this dataset

In [ ]: