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_2005.csv")
df
```

Out[2]:

	date	BEN	CO	EBE	MXY	NMHC	NO_2	NOx	OXY	O_3	PM10
0	2005-11-01 01:00:00	NaN	0.77	NaN	NaN	NaN	57.130001	128.699997	NaN	14.720000	14.91
1	2005-11-01 01:00:00	1.52	0.65	1.49	4.57	0.25	86.559998	181.699997	1.27	11.680000	30.93
2	2005-11-01 01:00:00	NaN	0.40	NaN	NaN	NaN	46.119999	53.000000	NaN	30.469999	14.60
3	2005-11-01 01:00:00	NaN	0.42	NaN	NaN	NaN	37.220001	52.009998	NaN	21.379999	15.16
4	2005-11-01 01:00:00	NaN	0.57	NaN	NaN	NaN	32.160000	36.680000	NaN	33.410000	5.00
...	...	...	...	...	...	...	...	...	...	...	...
236995	2006-01-01 00:00:00	1.08	0.36	1.01	NaN	0.11	21.990000	23.610001	NaN	43.349998	5.00
236996	2006-01-01 00:00:00	0.39	0.54	1.00	1.00	0.11	2.200000	4.220000	1.00	69.639999	4.95
236997	2006-01-01 00:00:00	0.19	NaN	0.26	NaN	0.08	26.730000	30.809999	NaN	43.840000	4.31
236998	2006-01-01 00:00:00	0.14	NaN	1.00	NaN	0.06	13.770000	17.770000	NaN	NaN	5.00
236999	2006-01-01 00:00:00	0.50	0.40	0.73	1.84	0.13	20.940001	26.950001	1.49	48.259998	5.67

237000 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: 20070 entries, 5 to 236999
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     20070 non-null  object 
 1   BEN      20070 non-null  float64
 2   CO       20070 non-null  float64
 3   EBE      20070 non-null  float64
 4   MXY      20070 non-null  float64
 5   NMHC     20070 non-null  float64
 6   NO_2     20070 non-null  float64
 7   NOx      20070 non-null  float64
 8   OXY      20070 non-null  float64
 9   O_3      20070 non-null  float64
 10  PM10     20070 non-null  float64
 11  PM25     20070 non-null  float64
 12  PXY      20070 non-null  float64
 13  SO_2     20070 non-null  float64
 14  TCH      20070 non-null  float64
 15  TOL      20070 non-null  float64
 16  station  20070 non-null  int64  
dtypes: float64(15), int64(1), object(1)
memory usage: 2.8+ MB
```

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

Out[6]:

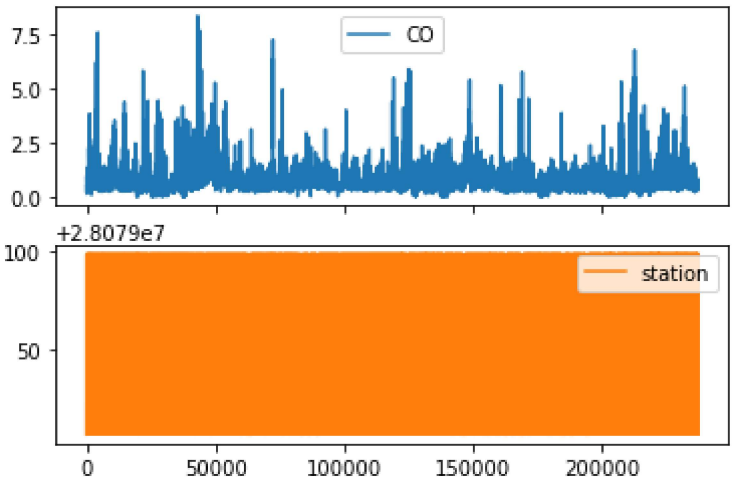
	CO	station
5	0.88	28079006
22	0.22	28079024
25	0.49	28079099
31	0.84	28079006
48	0.20	28079024
...	...	...
236970	0.39	28079024
236973	0.45	28079099
236979	0.38	28079006
236996	0.54	28079024
236999	0.40	28079099

20070 rows × 2 columns

Line chart

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

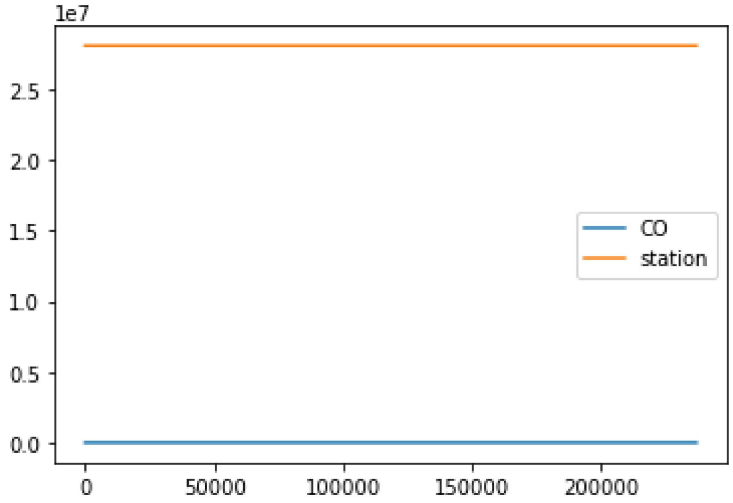
Out[7]:
```
array([<AxesSubplot:>, <AxesSubplot:>], dtype=object)
```



Line chart

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

Out[8]: <AxesSubplot:>

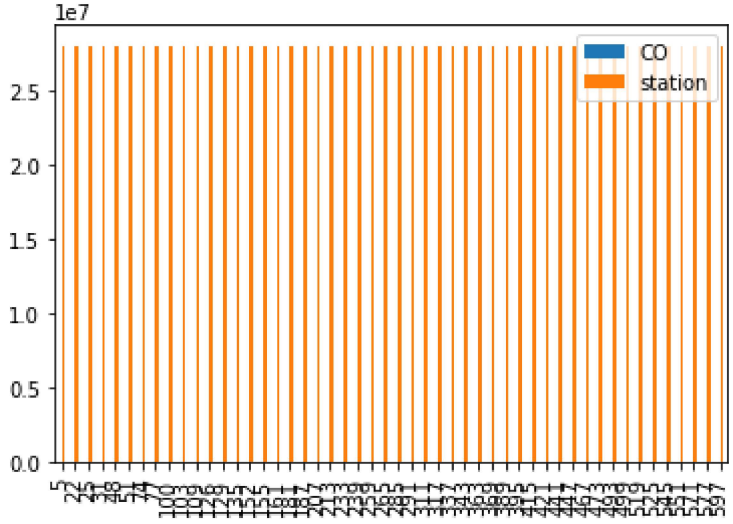


Bar chart

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

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

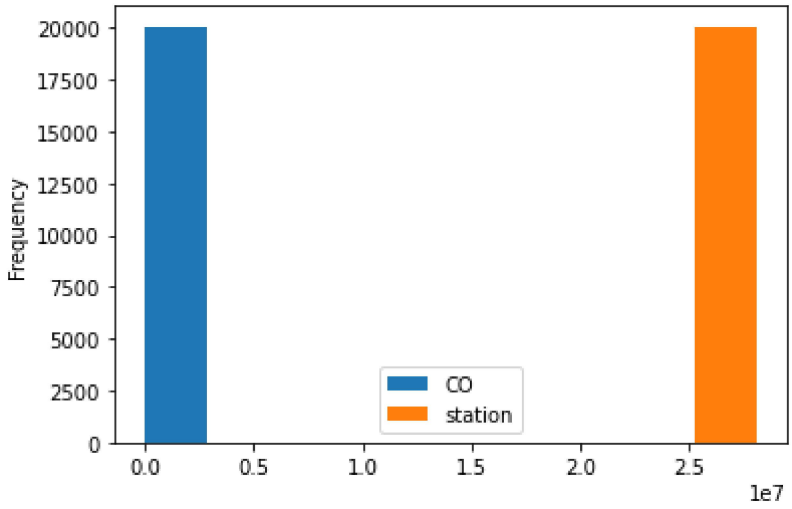
Out[10]: <AxesSubplot:>



Histogram

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

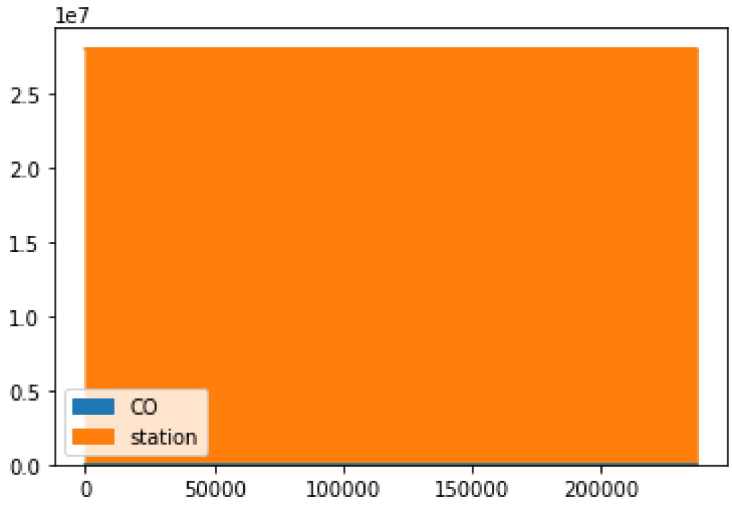
Out[11]:
```
<AxesSubplot:ylabel='Frequency'>
```



Area chart

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

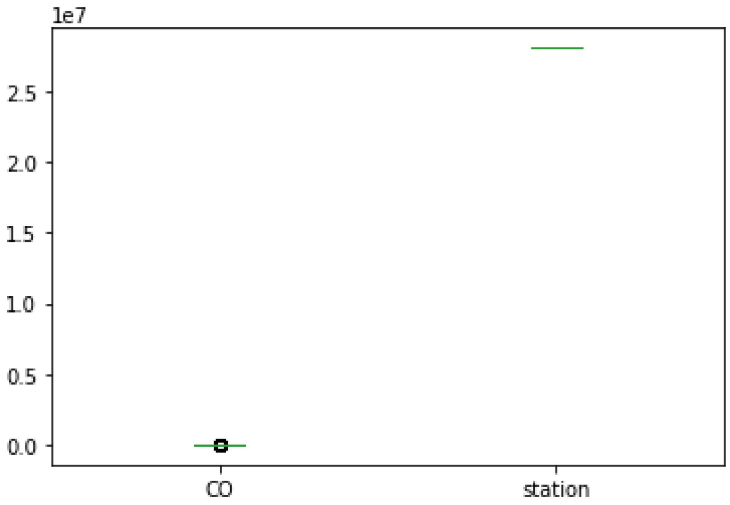
Out[12]:
```
<AxesSubplot:>
```



Box chart

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

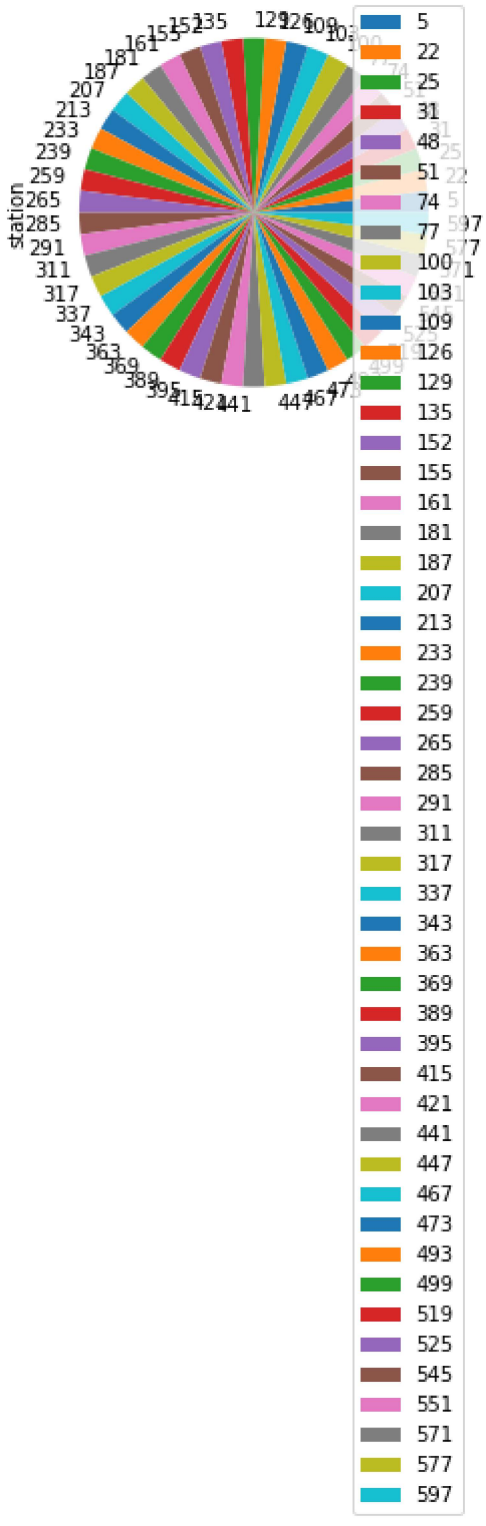
Out[13]: <AxesSubplot:>



Pie chart

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

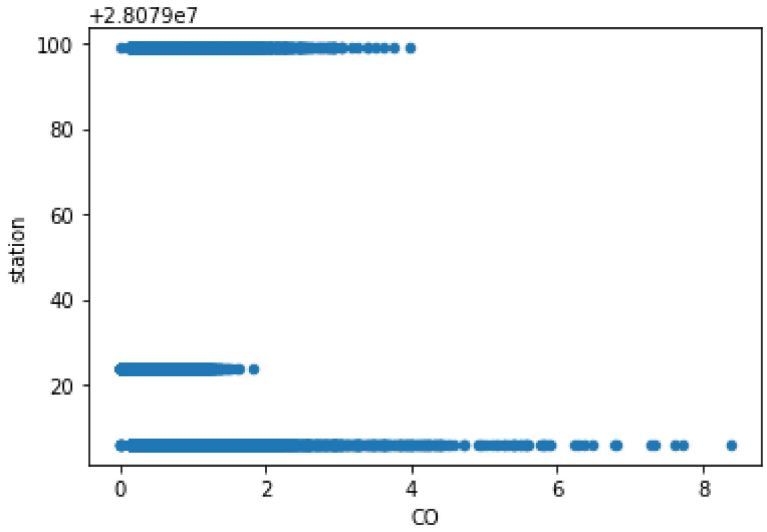
Out[14]: <AxesSubplot:ylabel='station'>



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: 20070 entries, 5 to 236999
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   date     20070 non-null  object
 1   BEN      20070 non-null  float64
 2   CO       20070 non-null  float64
 3   EBE      20070 non-null  float64
 4   MXY      20070 non-null  float64
 5   NMHC     20070 non-null  float64
 6   NO_2     20070 non-null  float64
 7   NOx      20070 non-null  float64
 8   OXY      20070 non-null  float64
 9   O_3      20070 non-null  float64
 10  PM10     20070 non-null  float64
 11  PM25     20070 non-null  float64
 12  PXY      20070 non-null  float64
 13  SO_2     20070 non-null  float64
```

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

Out[17]:

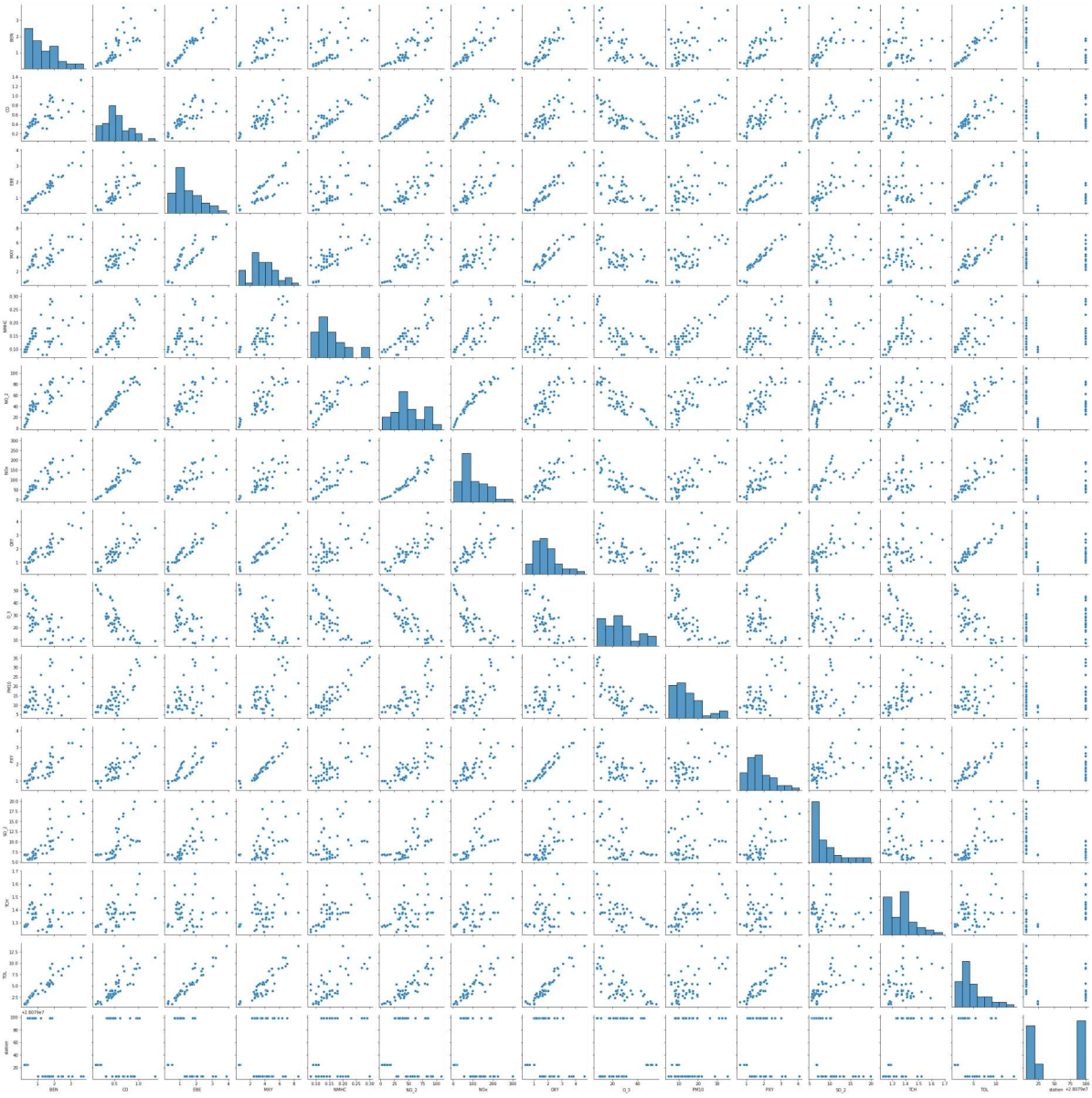
	BEN	CO	EBE	MXY	NMHC	NO_2	
count	20070.000000	20070.000000	20070.000000	20070.000000	20070.000000	20070.000000	200
mean	1.923656	0.720657	2.345423	5.457855	0.179282	66.226924	1
std	2.019061	0.549723	2.379219	5.495147	0.152783	40.568197	1
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.690000	0.400000	0.950000	1.930000	0.090000	36.602499	
50%	1.260000	0.580000	1.480000	3.800000	0.150000	60.525000	1
75%	2.510000	0.880000	2.950000	7.210000	0.220000	89.317499	1
max	26.570000	8.380000	29.870001	71.050003	1.880000	419.500000	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])
```

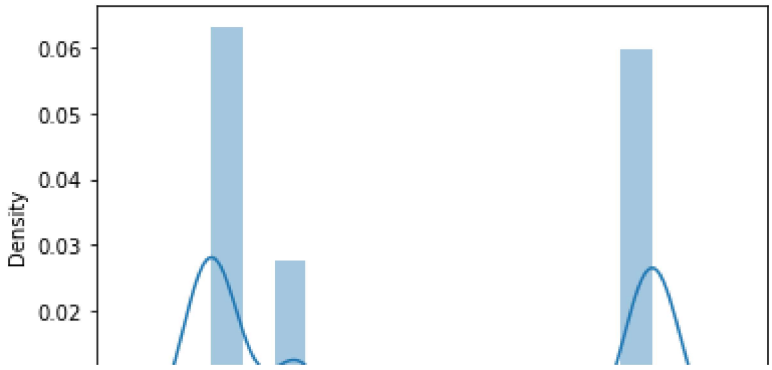
Out[19]:
```
<seaborn.axisgrid.PairGrid at 0x2082e3966a0>
```



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

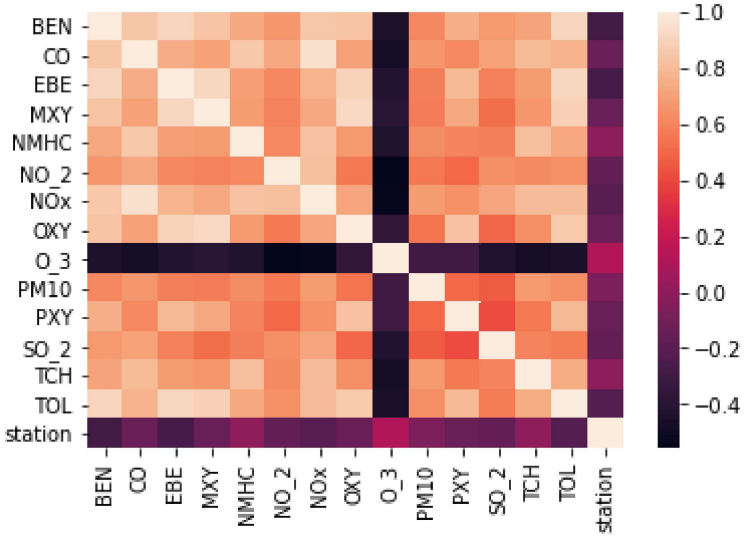
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: F
utureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-le
vel function with similar flexibility) or `histplot` (an axes-level functio
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]:
```
28078951.340651475
```

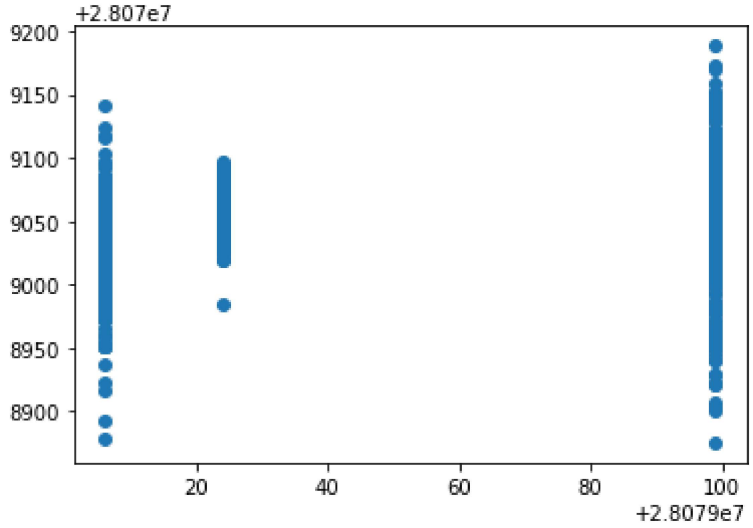
In [26]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

Out[26]:

	Co-efficient
BEN	-9.073117
CO	38.363937
EBE	-13.065268
MXY	3.619681
NMHC	72.283042
NO_2	0.110035
NOx	-0.262719
OXY	3.493396
O_3	0.017493
PM10	0.042364
PXY	2.700083
SO_2	0.099695
TCH	69.665731
TOL	-0.734365

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

Out[27]: <matplotlib.collections.PathCollection at 0x2083ccbfbe0>



ACCURACY

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

Out[28]: 0.2954070727955861

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

Out[29]: 0.30728337241705395

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]: 0.2950435635585983

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

Out[33]:
```
0.30706789915310695
```

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]:
```
0.06876368909814079
```

Accuracy(Lasso)

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

Out[36]:
```
0.056197271596408194
```

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([-5.48769594e+00,  1.49348773e+00, -7.26456706e+00,  2.58318581e+00,
        8.56280155e-01, -6.32470073e-02, -4.12763154e-03,  2.08093158e+00,
       -1.72264671e-02,  2.33084779e-01,  1.48494115e+00,  6.36231114e-02,
        1.55362219e+00, -8.81835449e-01])
```

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

Out[39]:
```
28079050.652584136
```

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

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

Out[41]:
```
0.166646715567356
```

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)))
```

```
37.21307437373933
1565.7311150288197
39.56932037613004
```

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]:
```
(20070, 14)
```

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

Out[46]:
```
(20070,)
```

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]:
```
0.879023418036871
```

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

Out[54]:
```
0.9998967601812779
```

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

Out[55]:
```
array([[9.99896760e-01, 3.21124597e-30, 1.03239819e-04]])
```

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]:
```
0.8646165156170202
```

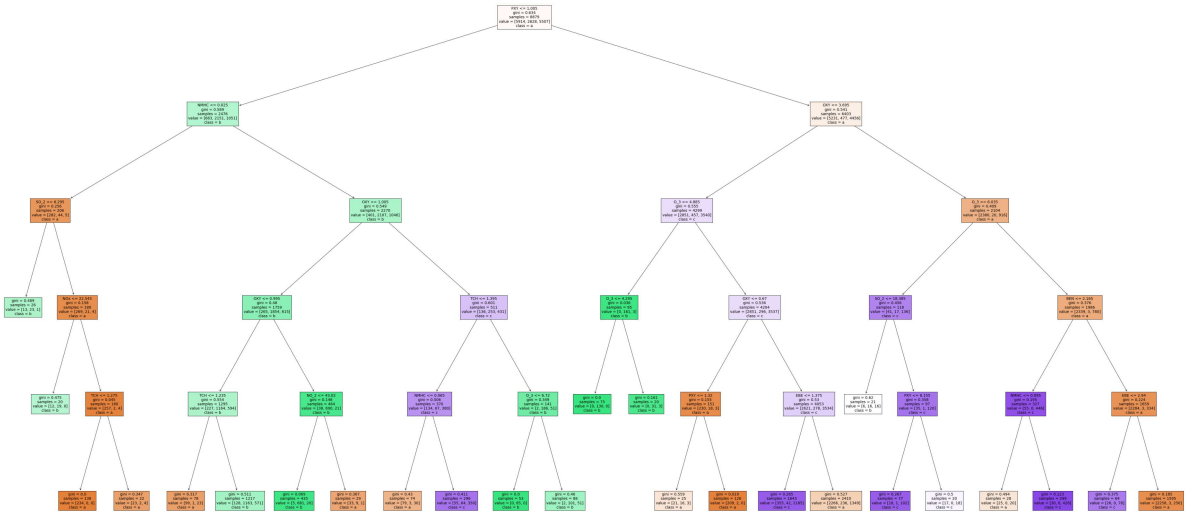
In [61]:
```
rfc_best=grid_search.best_estimator_
```

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(1978.3636363636363, 1993.2, 'PXY <= 1.005\ngini = 0.634\nsamples = 8879
\nvalue = [5914, 2628, 5507]\nclass = a'),
 Text(811.6363636363636, 1630.8000000000002, 'NMHC <= 0.025\ngini = 0.589\nsa
mples = 2476\nvalue = [683, 2151, 1051]\nclass = b'),
 Text(202.9090909090909, 1268.4, 'SO_2 <= 6.295\ngini = 0.256\nsamples = 206
\nvalue = [282, 44, 5]\nclass = a'),
 Text(101.45454545454545, 906.0, 'gini = 0.489\nsamples = 26\nvalue = [13, 2
3, 1]\nclass = b'),
 Text(304.3636363636364, 906.0, 'NOx <= 22.545\ngini = 0.158\nsamples = 180\n
value = [269, 21, 4]\nclass = a'),
 Text(202.9090909090909, 543.5999999999999, 'gini = 0.475\nsamples = 20\nvalu
e = [12, 19, 0]\nclass = b'),
 Text(405.8181818181818, 543.5999999999999, 'TCH <= 1.275\ngini = 0.045\nsamp
les = 160\nvalue = [257, 2, 4]\nclass = a'),
 Text(304.3636363636364, 181.19999999999982, 'gini = 0.0\nsamples = 138\nvalu
e = [234, 0, 0]\nclass = a'),
 Text(507.27272727272725, 181.19999999999982, 'gini = 0.347\nsamples = 22\nva
lue = [23, 2, 4]\nclass = a'),
 Text(1420.3636363636363, 1268.4, 'OXY <= 1.005\ngini = 0.549\nsamples = 2270
\nvalue = [401, 2107, 1046]\nclass = b'),
 Text(1014.5454545454545, 906.0, 'OXY <= 0.995\ngini = 0.48\nsamples = 1759\n
value = [265, 1854, 615]\nclass = b'),
 Text(811.6363636363636, 543.5999999999999, 'TCH <= 1.235\ngini = 0.554\nsamp
les = 1295\nvalue = [227, 1164, 594]\nclass = b'),
 Text(710.1818181818181, 181.19999999999982, 'gini = 0.317\nsamples = 78\nval
ue = [99, 1, 23]\nclass = a'),
 Text(913.0909090909091, 181.19999999999982, 'gini = 0.511\nsamples = 1217\nv
alue = [128, 1163, 571]\nclass = b'),
 Text(1217.4545454545455, 543.5999999999999, 'NO_2 <= 43.02\ngini = 0.148\nsa
mples = 464\nvalue = [38, 690, 21]\nclass = b'),
 Text(1116.0, 181.19999999999982, 'gini = 0.069\nsamples = 435\nvalue = [5, 6
81, 20]\nclass = b'),
 Text(1318.909090909091, 181.19999999999982, 'gini = 0.367\nsamples = 29\nval
ue = [33, 9, 1]\nclass = a'),
 Text(1826.1818181818182, 906.0, 'TCH <= 1.395\ngini = 0.601\nsamples = 511\n
value = [136, 253, 431]\nclass = c'),
 Text(1623.2727272727273, 543.5999999999999, 'NMHC <= 0.065\ngini = 0.506\nsa
mples = 370\nvalue = [134, 67, 380]\nclass = c'),
 Text(1521.8181818181818, 181.19999999999982, 'gini = 0.43\nsamples = 74\nval
ue = [79, 3, 30]\nclass = a'),
 Text(1724.7272727272727, 181.19999999999982, 'gini = 0.411\nsamples = 296\nv
alue = [55, 64, 350]\nclass = c'),
 Text(2029.090909090909, 543.5999999999999, 'O_3 <= 6.72\ngini = 0.349\nsampl
es = 141\nvalue = [2, 186, 51]\nclass = b'),
 Text(1927.6363636363635, 181.19999999999982, 'gini = 0.0\nsamples = 53\nvalu
e = [0, 85, 0]\nclass = b'),
 Text(2130.5454545454545, 181.19999999999982, 'gini = 0.46\nsamples = 88\nval
ue = [2, 101, 51]\nclass = b'),
 Text(3145.090909090909, 1630.8000000000002, 'OXY <= 3.695\ngini = 0.541\nsam
ples = 6403\nvalue = [5231, 477, 4456]\nclass = a'),
 Text(2587.090909090909, 1268.4, 'O_3 <= 4.885\ngini = 0.555\nsamples = 4299
\nvalue = [2851, 457, 3540]\nclass = c'),
 Text(2333.4545454545455, 906.0, 'O_3 <= 4.295\ngini = 0.036\nsamples = 95\nv
alue = [0, 161, 3]\nclass = b'),
 Text(2232.0, 543.5999999999999, 'gini = 0.0\nsamples = 75\nvalue = [0, 130,
0]\nclass = b'),
 Text(2434.909090909091, 543.5999999999999, 'gini = 0.161\nsamples = 20\nvalu
```

```
e = [0, 31, 3]\nclass = b'),
 Text(2840.7272727272725, 906.0, 'OXY <= 0.67\ngini = 0.536\nsamples = 4204\n
value = [2851, 296, 3537]\nclass = c'),
 Text(2637.818181818182, 543.5999999999999, 'PXY <= 1.32\ngini = 0.155\nsampl
es = 151\nvalue = [230, 18, 3]\nclass = a'),
 Text(2536.3636363636365, 181.19999999999982, 'gini = 0.559\nsamples = 25\nva
lue = [21, 16, 3]\nclass = a'),
 Text(2739.272727272727, 181.19999999999982, 'gini = 0.019\nsamples = 126\nva
lue = [209, 2, 0]\nclass = a'),
 Text(3043.6363636363635, 543.5999999999999, 'EBE <= 1.375\ngini = 0.53\nsamp
les = 4053\nvalue = [2621, 278, 3534]\nclass = c'),
 Text(2942.181818181818, 181.19999999999982, 'gini = 0.265\nsamples = 1643\nv
alue = [355, 42, 2185]\nclass = c'),
 Text(3145.090909090909, 181.19999999999982, 'gini = 0.527\nsamples = 2410\nv
alue = [2266, 236, 1349]\nclass = a'),
 Text(3703.090909090909, 1268.4, 'O_3 <= 6.035\ngini = 0.409\nsamples = 2104
\nvalue = [2380, 20, 916]\nclass = a'),
 Text(3348.0, 906.0, 'SO_2 <= 18.385\ngini = 0.456\nsamples = 118\nvalue = [4
1, 17, 136]\nclass = c'),
 Text(3246.5454545454545, 543.5999999999999, 'gini = 0.62\nsamples = 21\nvalu
e = [6, 16, 16]\nclass = b'),
 Text(3449.4545454545455, 543.5999999999999, 'PXY <= 8.155\ngini = 0.358\nsam
ples = 97\nvalue = [35, 1, 120]\nclass = c'),
 Text(3348.0, 181.19999999999982, 'gini = 0.267\nsamples = 77\nvalue = [18,
1, 102]\nclass = c'),
 Text(3550.909090909091, 181.19999999999982, 'gini = 0.5\nsamples = 20\nvalue
= [17, 0, 18]\nclass = c'),
 Text(4058.181818181818, 906.0, 'BEN <= 2.165\ngini = 0.376\nsamples = 1986\n
value = [2339, 3, 780]\nclass = a'),
 Text(3855.272727272727, 543.5999999999999, 'NMHC <= 0.095\ngini = 0.195\nsam
ples = 327\nvalue = [55, 0, 446]\nclass = c'),
 Text(3753.818181818182, 181.19999999999982, 'gini = 0.494\nsamples = 28\nval
ue = [25, 0, 20]\nclass = a'),
 Text(3956.7272727272725, 181.19999999999982, 'gini = 0.123\nsamples = 299\nv
alue = [30, 0, 426]\nclass = c'),
 Text(4261.090909090909, 543.5999999999999, 'EBE <= 2.94\ngini = 0.224\nsampl
es = 1659\nvalue = [2284, 3, 334]\nclass = a'),
 Text(4159.636363636364, 181.19999999999982, 'gini = 0.375\nsamples = 64\nval
ue = [26, 0, 78]\nclass = c'),
 Text(4362.545454545454, 181.19999999999982, 'gini = 0.185\nsamples = 1595\nv
alue = [2258, 3, 256]\nclass = a')]
```



Conclusion

Accuracy

Linear Regression:0.30728337241705395

Ridge Regression:0.30706789915310695

Lasso Regression:0.06876368909814079

ElasticNet Regression:0.166646715567356

Logistic Regression:0.879023418036871

Random Forest:0.8646165156170202

Logistic Regression is suitable for this dataset