20104028

Heamnath N

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_2007.csv")
df
```

Out[2]:

	date	BEN	CO	EBE	MXY	NMHC	NO_2	NOx	OXY	O_3	PM10
0	2007-12-01 01:00:00	NaN	2.86	NaN	NaN	NaN	282.200012	1054.000000	NaN	4.030000	156.199997
1	2007-12-01 01:00:00	NaN	1.82	NaN	NaN	NaN	86.419998	354.600006	NaN	3.260000	80.809998
2	2007-12-01 01:00:00	NaN	1.47	NaN	NaN	NaN	94.639999	319.000000	NaN	5.310000	53.099998
3	2007-12-01 01:00:00	NaN	1.64	NaN	NaN	NaN	127.900002	476.700012	NaN	4.500000	105.300003
4	2007-12-01 01:00:00	4.64	1.86	4.26	7.98	0.57	145.100006	573.900024	3.49	52.689999	106.500000
...	...	...	...	...	...	...	...	...	...	...	...
225115	2007-03-01 00:00:00	0.30	0.45	1.00	0.30	0.26	8.690000	11.690000	1.00	42.209999	6.760000
225116	2007-03-01 00:00:00	NaN	0.16	NaN	NaN	NaN	46.820000	51.480000	NaN	22.150000	5.700000
225117	2007-03-01 00:00:00	0.24	NaN	0.20	NaN	0.09	51.259998	66.809998	NaN	18.540001	13.010000

	date	BEN	CO	EBE	MXY	NMHC	NO_2	NOx	OXY	O_3	PM10
225118	2007-03-01 00:00:00	0.11	NaN	1.00	NaN	0.05	24.240000	36.930000	NaN	NaN	6.610000
225119	2007-03-01 00:00:00	0.53	0.40	1.00	1.70	0.12	32.360001	47.860001	1.37	24.150000	10.260000

225120 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: 25443 entries, 4 to 225119
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     25443 non-null  object 
 1   BEN      25443 non-null  float64
 2   CO       25443 non-null  float64
 3   EBE      25443 non-null  float64
 4   MXY      25443 non-null  float64
 5   NMHC     25443 non-null  float64
 6   NO_2     25443 non-null  float64
 7   NOx      25443 non-null  float64
 8   OXY      25443 non-null  float64
 9   O_3      25443 non-null  float64
 10  PM10     25443 non-null  float64
 11  PM25     25443 non-null  float64
 12  PXY      25443 non-null  float64
 13  SO_2     25443 non-null  float64
 14  TCH      25443 non-null  float64
 15  TOL      25443 non-null  float64
 16  station  25443 non-null  int64  
dtypes: float64(15), int64(1), object(1)
memory usage: 3.5+ MB
```

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

Out[6]:

	CO	station
4	1.86	28079006

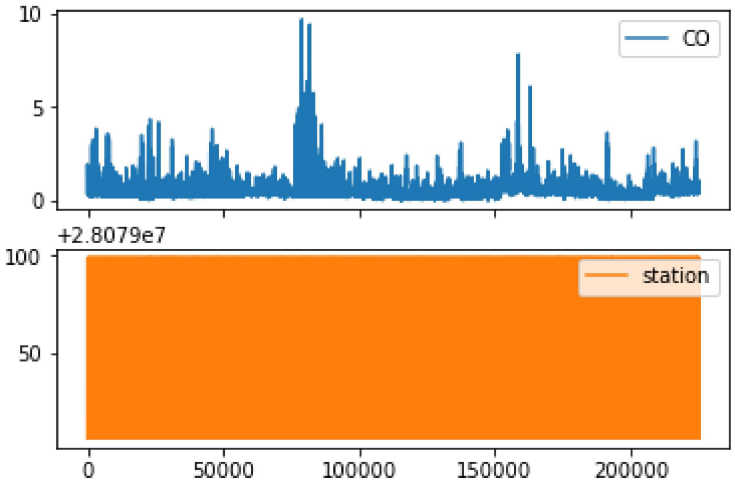
	CO	station
21	0.30	28079024
25	1.42	28079099
30	1.89	28079006
47	0.30	28079024
...	...	...
225073	0.47	28079006
225094	0.45	28079099
225098	0.41	28079006
225115	0.45	28079024
225119	0.40	28079099

25443 rows × 2 columns

Line chart

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

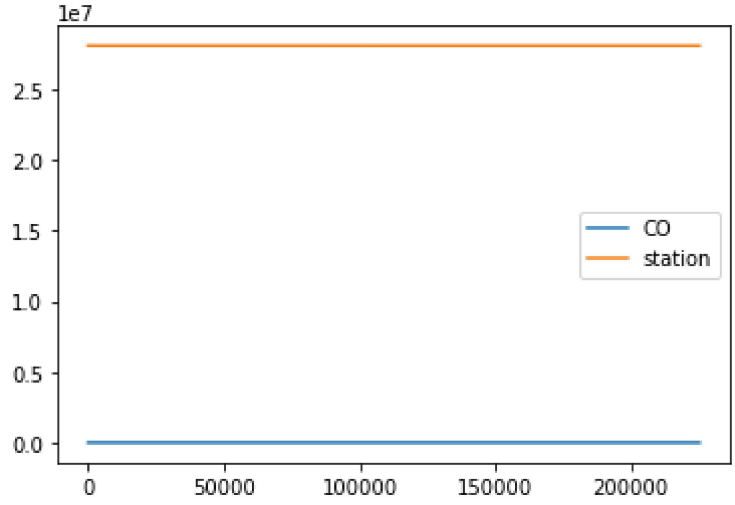
Out[7]:
```
array([<AxesSubplot:>, <AxesSubplot:>], dtype=object)
```



Line chart

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

Out[8]:
```
<AxesSubplot:>
```

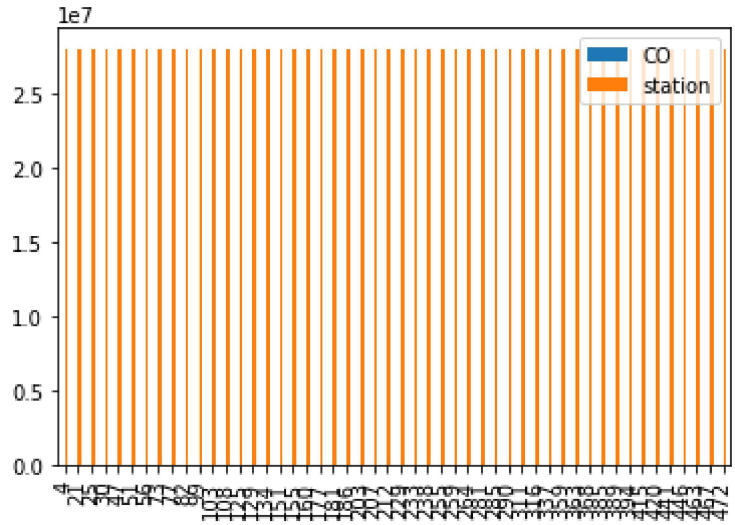


Bar chart

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

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

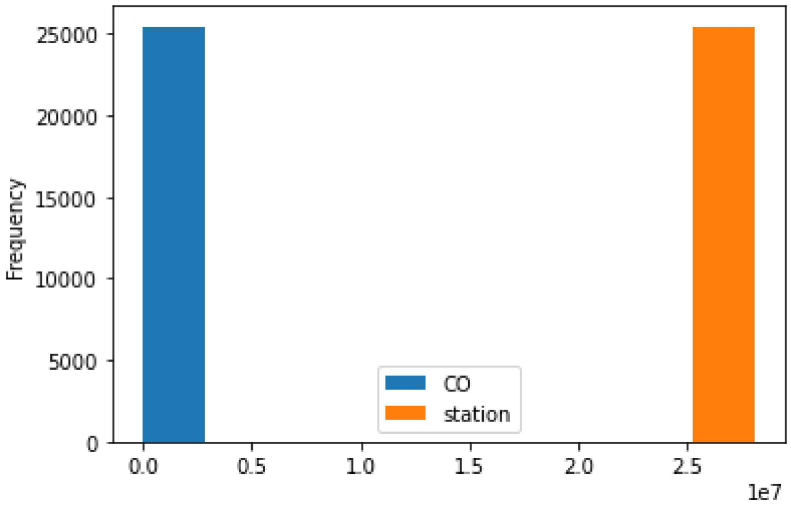
Out[10]: <AxesSubplot:>



Histogram

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

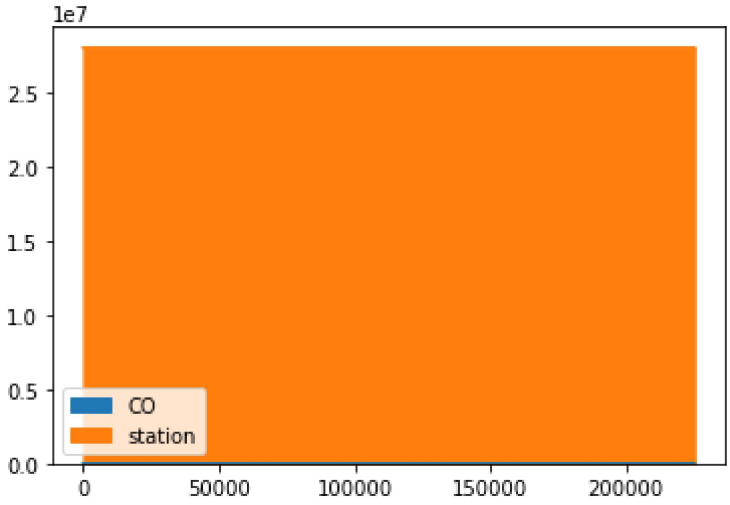
Out[11]: <AxesSubplot:ylabel='Frequency'>



Area chart

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

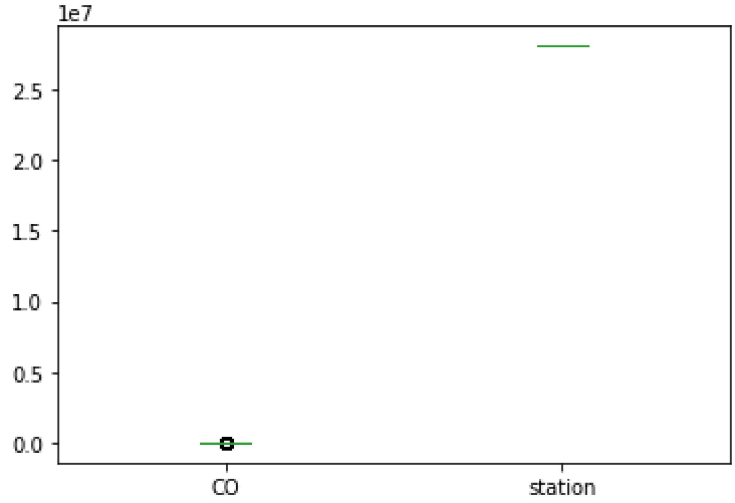
Out[12]: <AxesSubplot:>



Box chart

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

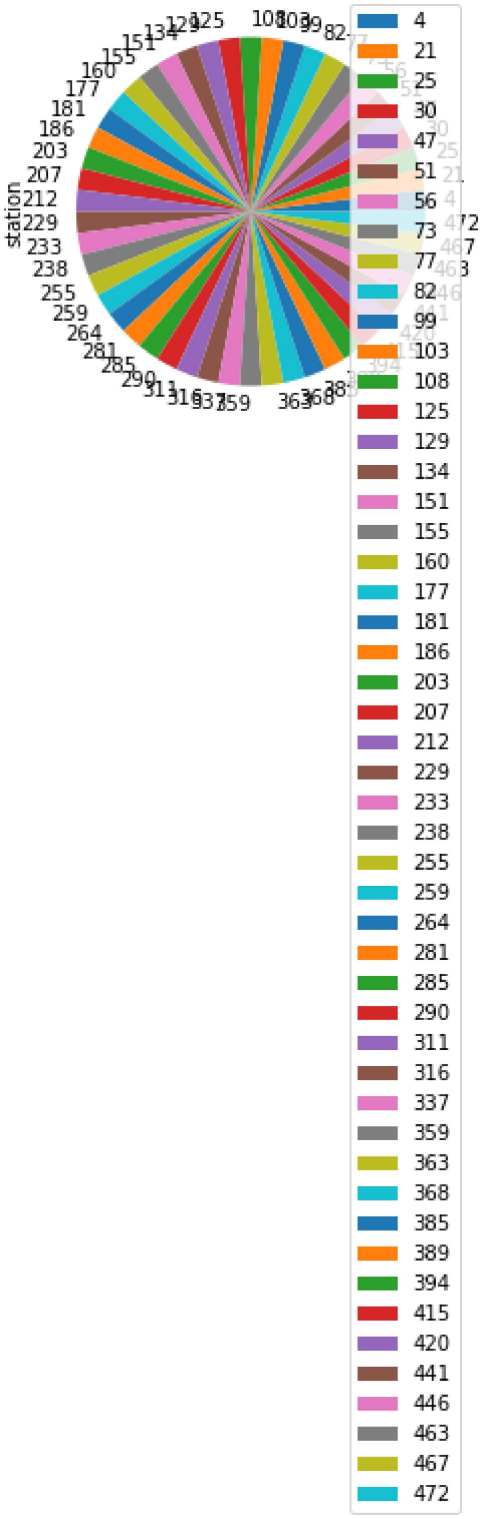
Out[13]: <AxesSubplot:>



Pie chart

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

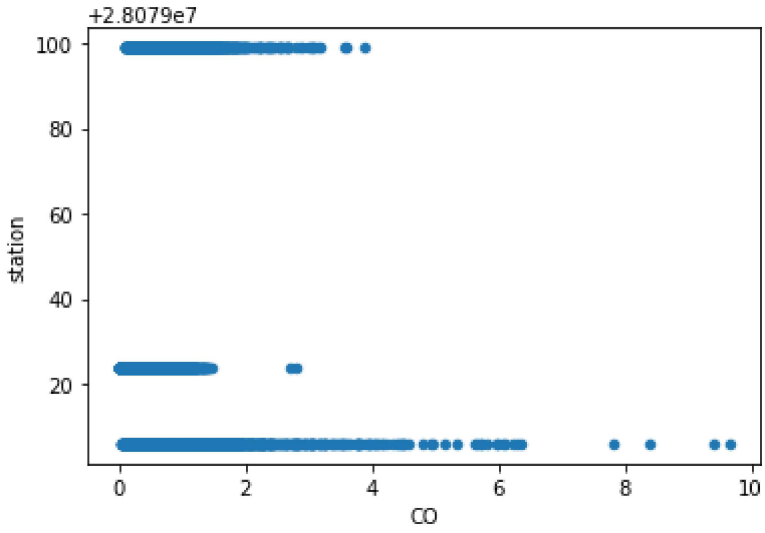
Out[14]: <AxesSubplot:ylabel='station'>



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: 25443 entries, 4 to 225119
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     25443 non-null  object 
 1   BEN      25443 non-null  float64
 2   CO       25443 non-null  float64
 3   EBE      25443 non-null  float64
 4   MXY      25443 non-null  float64
 5   NMHC     25443 non-null  float64
 6   NO_2     25443 non-null  float64
 7   NOx      25443 non-null  float64
 8   OXY      25443 non-null  float64
 9   O_3      25443 non-null  float64
 10  PM10     25443 non-null  float64
 11  PM25     25443 non-null  float64
 12  PXY      25443 non-null  float64
 13  SO_2     25443 non-null  float64
 14  TCH      25443 non-null  float64
 15  TOL      25443 non-null  float64
 16  station  25443 non-null  int64  
dtypes: float64(15), int64(1), object(1)
memory usage: 3.5+ MB
```

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

Out[17]:

	BEN	CO	EBE	MXY	NMHC	NO_2	NO
count	25443.000000	25443.000000	25443.000000	25443.000000	25443.000000	25443.000000	25443.00000
mean	1.146744	0.505120	1.394071	2.392008	0.249967	58.532683	112.74186
std	1.278733	0.423231	1.268265	2.784302	0.142627	37.755029	115.52700
min	0.130000	0.000000	0.120000	0.150000	0.000000	1.690000	1.78000
25%	0.450000	0.260000	0.780000	0.960000	0.160000	31.285001	39.91000
50%	0.770000	0.400000	1.000000	1.500000	0.220000	54.080002	82.80999

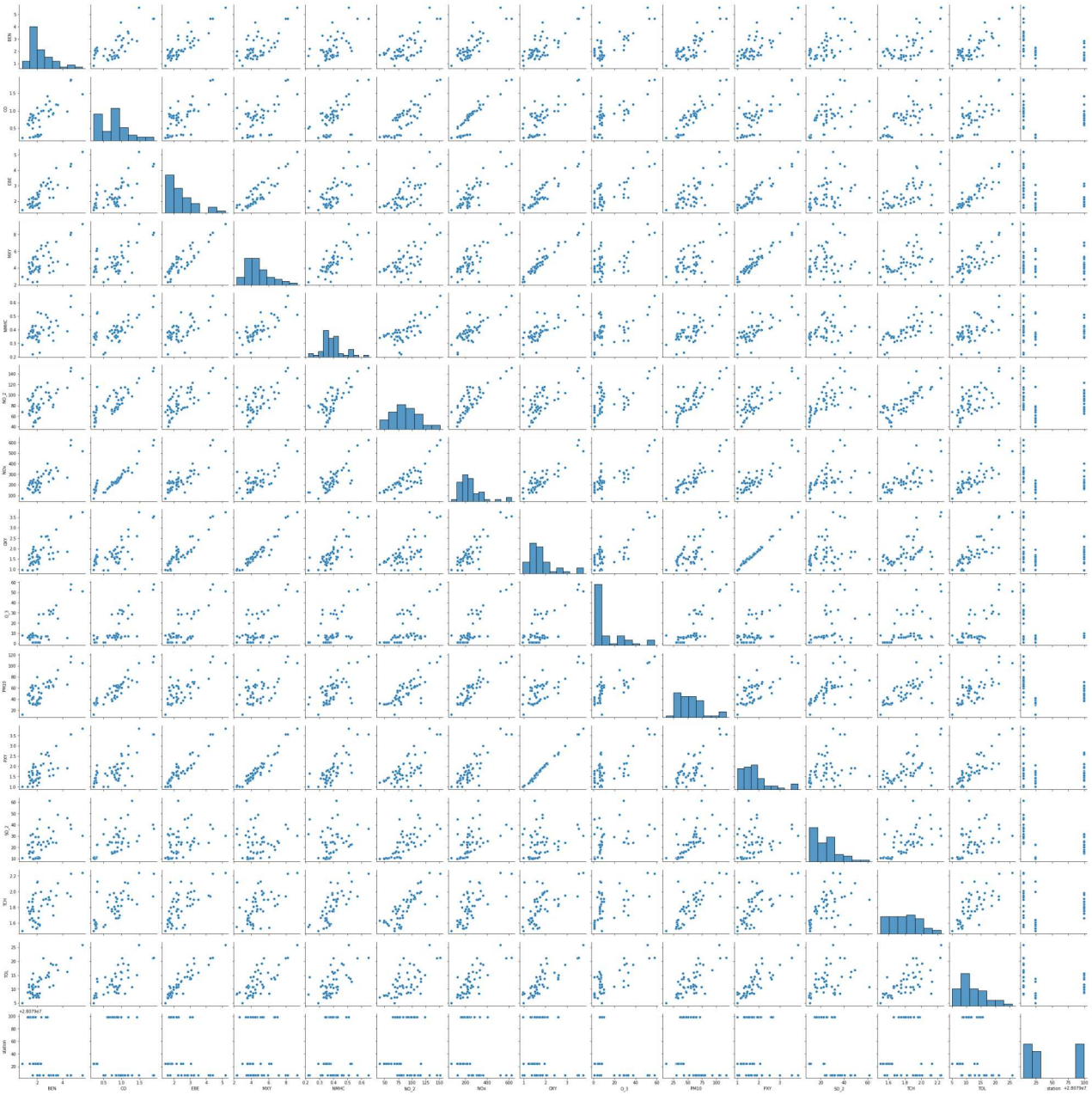
	BEN	CO	EBE	MXY	NMHC	NO_2	NO
75%	1.390000	0.640000	1.580000	2.855000	0.300000	79.230003	149.30000
max	30.139999	9.660000	31.680000	65.480003	2.570000	430.299988	1893.00000

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])
```

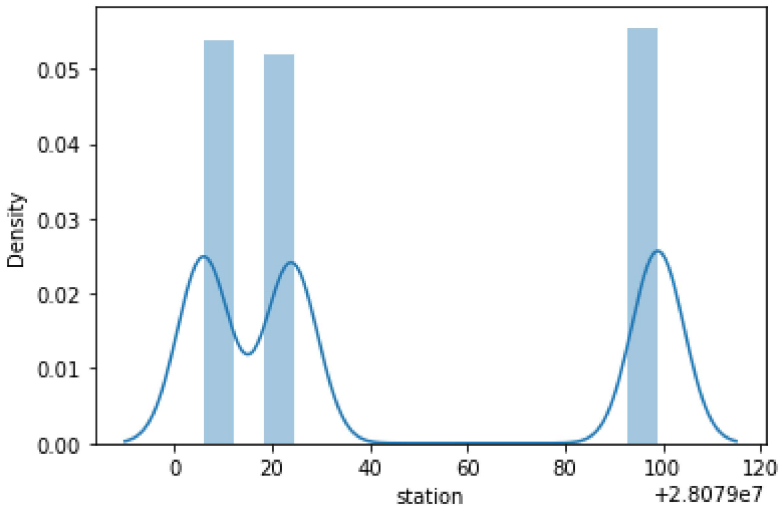
Out[19]: <seaborn.axisgrid.PairGrid at 0x21aab4b3040>



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

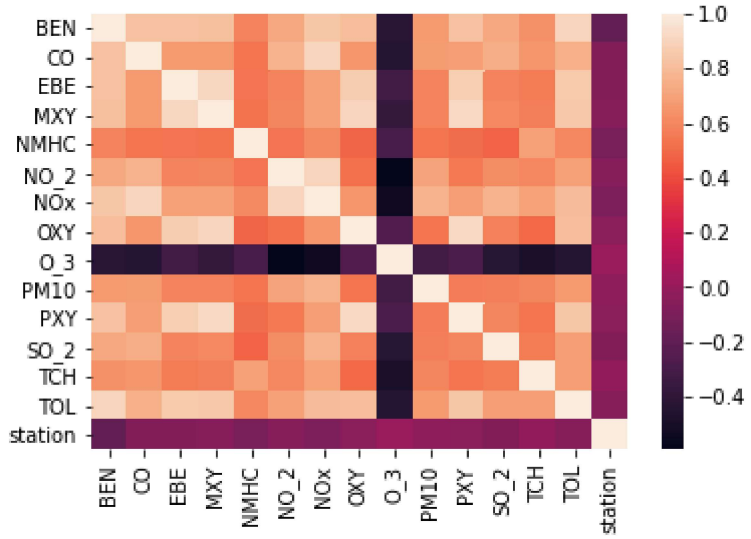
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function 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]:
```
28079008.981771596
```

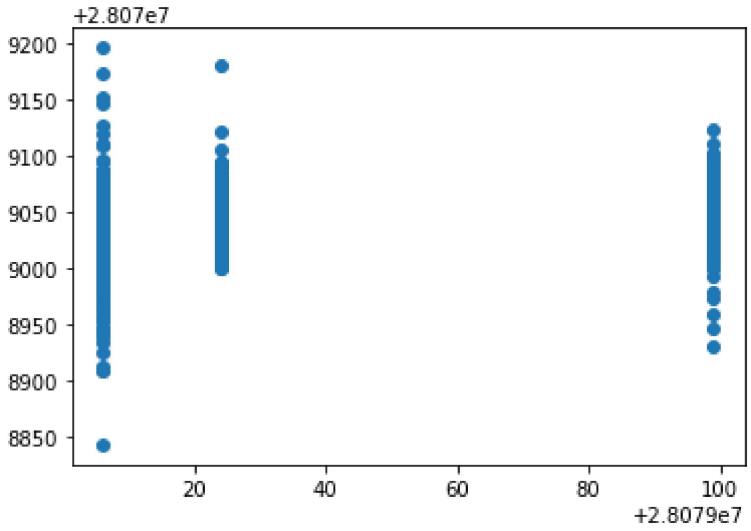
In [26]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

Out[26]:

	Co-efficient
BEN	-32.112477
CO	19.866760
EBE	-0.085819
MXY	-0.912714
NMHC	-37.697342
NO_2	0.122562
NOx	-0.047435
OXY	5.795700
O_3	-0.033518
PM10	0.122797
PXY	6.247059
SO_2	0.138420
TCH	25.915287
TOL	3.102116

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

Out[27]:
```
<matplotlib.collections.PathCollection at 0x21ab9e63730>
```



ACCURACY

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

Out[28]: 0.1640004995535138

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

Out[29]: 0.15664629403083752

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]: 0.16374855694140056

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

Out[33]: 0.15659940501416814

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]:
```
0.012371369592628745
```

Accuracy(Lasso)

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

Out[36]:
```
0.015114965823165627
```

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([-8.06135562,  0.        , -0.        ,  0.24505398, -0.        ,
        0.05818669, -0.04674434,  0.68316815, -0.06550926,  0.14444337,
        0.64612536, -0.04529324,  0.        ,  0.90520133])
```

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

Out[39]:
```
28079046.20227694
```

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

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

Out[41]:
```
0.06777801843475428
```

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)))
```

```
36.63182269054145
1542.8562100447766
39.27920836835662
```

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_3',
       'PM10', 'PXY', 'SO_2', 'TCH', 'TOL']]
target_vector=df[ 'station']
```

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

Out[45]:
```
(25443, 14)
```

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

Out[46]:
```
(25443,)
```

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)
```

```
[28079099]
```

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

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

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

Out[53]:
```
0.8146838030106512
```

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

Out[54]:
```
1.082753977181323e-19
```

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

Out[55]:
```
array([[1.08275398e-19, 1.80383815e-19, 1.00000000e+00]])
```

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="accuracy")
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]:
```
0.8216732172936553
```

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

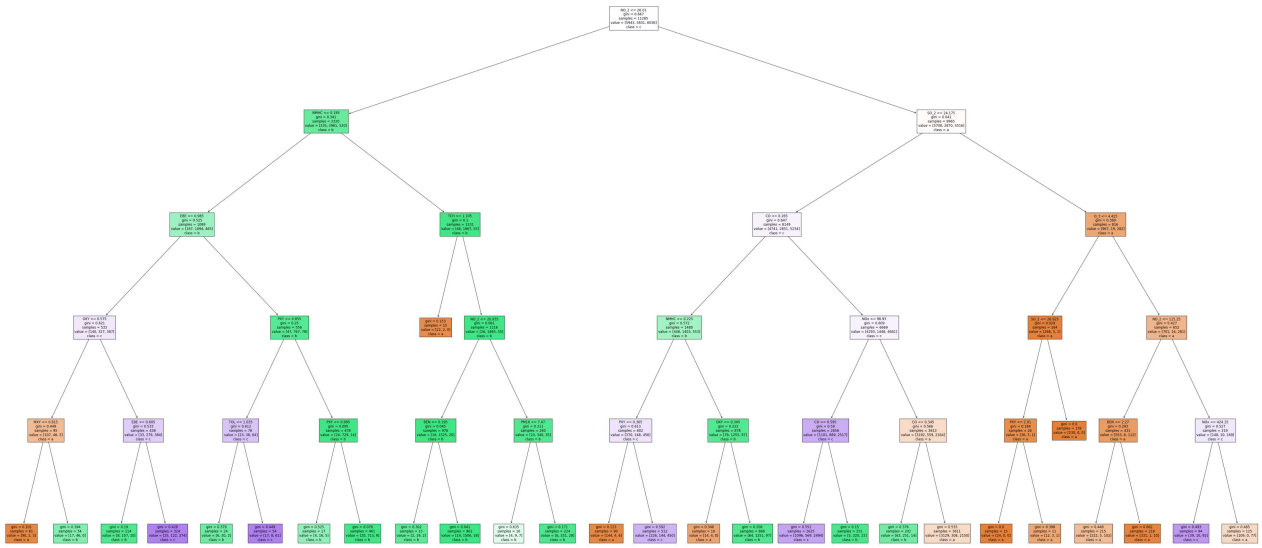
In [62]:
```
from sklearn.tree import plot_tree

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

Out[62]:
```
[Text(2242.730769230769, 1993.2, 'NO_2 <= 26.01\ngini = 0.667\nsamples = 11285\nvalue =
[5943, 5831, 6036]\nclass = c'),
 Text(1158.923076923077, 1630.8000000000002, 'NMHC <= 0.195\ngini = 0.341\nsamples = 232
0\nvalue = [235, 2961, 520]\nclass = b'),
```

```
 Text(686.7692307692307, 1268.4, 'EBE <= 0.985\ngini = 0.525\nsamples = 1089\nvalue = [1
87, 1094, 465]\nclass = b'),
 Text(343.38461538461536, 906.0, 'OXY <= 0.575\ngini = 0.621\nsamples = 533\nvalue = [14
0, 327, 387]\nclass = c'),
 Text(171.69230769230768, 543.5999999999999, 'MXY <= 0.815\ngini = 0.449\nsamples = 95\n
value = [107, 48, 3]\nclass = a'),
 Text(85.84615384615384, 181.19999999999982, 'gini = 0.101\nsamples = 61\nvalue = [90,
2, 3]\nclass = a'),
 Text(257.53846153846155, 181.19999999999982, 'gini = 0.394\nsamples = 34\nvalue = [17,
46, 0]\nclass = b'),
 Text(515.0769230769231, 543.5999999999999, 'EBE <= 0.605\ngini = 0.533\nsamples = 438\n
value = [33, 279, 384]\nclass = c'),
 Text(429.23076923076917, 181.19999999999982, 'gini = 0.19\nsamples = 114\nvalue = [8, 1
57, 10]\nclass = b'),
 Text(600.9230769230769, 181.19999999999982, 'gini = 0.428\nsamples = 324\nvalue = [25,
122, 374]\nclass = c'),
 Text(1030.1538461538462, 906.0, 'PXY <= 0.855\ngini = 0.25\nsamples = 556\nvalue = [47,
767, 78]\nclass = b'),
 Text(858.4615384615383, 543.5999999999999, 'TOL <= 1.035\ngini = 0.612\nsamples = 78\nv
alue = [23, 38, 64]\nclass = c'),
 Text(772.6153846153845, 181.19999999999982, 'gini = 0.379\nsamples = 24\nvalue = [6, 3
0, 3]\nclass = b'),
 Text(944.3076923076923, 181.19999999999982, 'gini = 0.449\nsamples = 54\nvalue = [17,
8, 61]\nclass = c'),
 Text(1201.8461538461538, 543.5999999999999, 'PXY <= 0.995\ngini = 0.095\nsamples = 478
\nvalue = [24, 729, 14]\nclass = b'),
 Text(1116.0, 181.19999999999982, 'gini = 0.525\nsamples = 17\nvalue = [4, 16, 5]\nclass
= b'),
 Text(1287.6923076923076, 181.19999999999982, 'gini = 0.076\nsamples = 461\nvalue = [20,
713, 9]\nclass = b'),
 Text(1631.0769230769229, 1268.4, 'TCH <= 1.105\ngini = 0.1\nsamples = 1231\nvalue = [4
8, 1867, 55]\nclass = b'),
 Text(1545.230769230769, 906.0, 'gini = 0.153\nsamples = 15\nvalue = [22, 2, 0]\nclass =
a'),
 Text(1716.9230769230767, 906.0, 'NO_2 <= 20.935\ngini = 0.081\nsamples = 1216\nvalue =
[26, 1865, 55]\nclass = b'),
 Text(1545.230769230769, 543.5999999999999, 'BEN <= 0.195\ngini = 0.045\nsamples = 976\n
value = [16, 1525, 20]\nclass = b'),
 Text(1459.3846153846152, 181.19999999999982, 'gini = 0.302\nsamples = 15\nvalue = [2, 1
9, 2]\nclass = b'),
 Text(1631.0769230769229, 181.19999999999982, 'gini = 0.041\nsamples = 961\nvalue = [14,
1506, 18]\nclass = b'),
 Text(1888.6153846153845, 543.5999999999999, 'PM10 <= 7.47\ngini = 0.211\nsamples = 240
\nvalue = [10, 340, 35]\nclass = b'),
 Text(1802.7692307692307, 181.19999999999982, 'gini = 0.635\nsamples = 16\nvalue = [4,
9, 7]\nclass = b'),
 Text(1974.4615384615383, 181.19999999999982, 'gini = 0.171\nsamples = 224\nvalue = [6,
331, 28]\nclass = b'),
 Text(3326.5384615384614, 1630.8000000000002, 'SO_2 <= 24.175\ngini = 0.641\nsamples = 8
965\nvalue = [5708, 2870, 5516]\nclass = a'),
 Text(2747.076923076923, 1268.4, 'CO <= 0.265\ngini = 0.647\nsamples = 8149\nvalue = [47
41, 2851, 5234]\nclass = c'),
 Text(2403.6923076923076, 906.0, 'NMHC <= 0.225\ngini = 0.572\nsamples = 1480\nvalue =
[448, 1403, 553]\nclass = b'),
 Text(2232.0, 543.5999999999999, 'PXY <= 0.365\ngini = 0.613\nsamples = 602\nvalue = [37
0, 148, 456]\nclass = c'),
 Text(2146.153846153846, 181.19999999999982, 'gini = 0.123\nsamples = 90\nvalue = [144,
4, 6]\nclass = a'),
 Text(2317.846153846154, 181.19999999999982, 'gini = 0.592\nsamples = 512\nvalue = [226,
144, 450]\nclass = c'),
 Text(2575.3846153846152, 543.5999999999999, 'OXY <= 0.345\ngini = 0.222\nsamples = 878
\nvalue = [78, 1255, 97]\nclass = b'),
 Text(2489.5384615384614, 181.19999999999982, 'gini = 0.346\nsamples = 10\nvalue = [14,
4, 0]\nclass = a'),
 Text(2661.230769230769, 181.19999999999982, 'gini = 0.208\nsamples = 868\nvalue = [64,
```

```
1251, 97]\nclass = b'),
 Text(3090.461538461538, 906.0, 'NOx <= 98.93\ngini = 0.609\nsamples = 6669\nvalue = [42
93, 1448, 4681]\nclass = c'),
 Text(2918.7692307692305, 543.5999999999999, 'CO <= 0.595\ngini = 0.59\nsamples = 2856\n
value = [1101, 889, 2517]\nclass = c'),
 Text(2832.9230769230767, 181.19999999999982, 'gini = 0.552\nsamples = 2625\nvalue = [10
96, 569, 2494]\nclass = c'),
 Text(3004.6153846153843, 181.19999999999982, 'gini = 0.15\nsamples = 231\nvalue = [5, 3
20, 23]\nclass = b'),
 Text(3262.1538461538457, 543.5999999999999, 'CO <= 0.345\ngini = 0.566\nsamples = 3813
\nvalue = [3192, 559, 2164]\nclass = a'),
 Text(3176.307692307692, 181.19999999999982, 'gini = 0.376\nsamples = 202\nvalue = [63,
251, 14]\nclass = b'),
 Text(3347.9999999999995, 181.19999999999982, 'gini = 0.535\nsamples = 3611\nvalue = [31
29, 308, 2150]\nclass = a'),
 Text(3905.9999999999995, 1268.4, 'O_3 <= 4.425\ngini = 0.369\nsamples = 816\nvalue = [9
67, 19, 282]\nclass = a'),
 Text(3691.3846153846152, 906.0, 'SO_2 <= 26.925\ngini = 0.029\nsamples = 164\nvalue =
[266, 3, 1]\nclass = a'),
 Text(3605.5384615384614, 543.5999999999999, 'PXY <= 2.01\ngini = 0.184\nsamples = 26\nv
alue = [36, 3, 1]\nclass = a'),
 Text(3519.6923076923076, 181.19999999999982, 'gini = 0.0\nsamples = 15\nvalue = [24, 0,
0]\nclass = a'),
 Text(3691.3846153846152, 181.19999999999982, 'gini = 0.398\nsamples = 11\nvalue = [12,
3, 1]\nclass = a'),
 Text(3777.230769230769, 543.5999999999999, 'gini = 0.0\nsamples = 138\nvalue = [230, 0,
0]\nclass = a'),
 Text(4120.615384615385, 906.0, 'NO_2 <= 115.25\ngini = 0.427\nsamples = 652\nvalue = [7
01, 16, 281]\nclass = a'),
 Text(3948.9230769230767, 543.5999999999999, 'BEN <= 2.27\ngini = 0.293\nsamples = 433\n
value = [553, 6, 112]\nclass = a'),
 Text(3863.076923076923, 181.19999999999982, 'gini = 0.448\nsamples = 215\nvalue = [222,
5, 102]\nclass = a'),
 Text(4034.7692307692305, 181.19999999999982, 'gini = 0.062\nsamples = 218\nvalue = [33
1, 1, 10]\nclass = a'),
 Text(4292.307692307692, 543.5999999999999, 'NOx <= 424.15\ngini = 0.527\nsamples = 219
\nvalue = [148, 10, 169]\nclass = c'),
 Text(4206.461538461538, 181.19999999999982, 'gini = 0.493\nsamples = 94\nvalue = [39, 1
0, 92]\nclass = c'),
 Text(4378.153846153846, 181.19999999999982, 'gini = 0.485\nsamples = 125\nvalue = [109,
0, 77]\nclass = a')]
```



Conclusion

Accuracy

In [63]:
```
print("Linear Regression:",lr.score(x_test,y_test))
print("Ridge Regression:",rr.score(x_test,y_test))
print("Lasso Regression",la.score(x_test,y_test))
print("ElasticNet Regression:",en.score(x_test,y_test))
print("Logistic Regression:",logr.score(fs,target_vector))
print("Random Forest:",grid_search.best_score_)
```

```
Linear Regression: 0.1640004995535138
Ridge Regression: 0.16374855694140056
Lasso Regression 0.015114965823165627
ElasticNet Regression: 0.06777801843475428
Logistic Regression: 0.8146838030106512
Random Forest: 0.8216732172936553
```

Random Forest is suitable for this dataset