20104016

DEENA

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_2006.csv")
df
```

Out[2]:

	date	BEN	CO	EBE	MXY	NMHC	NO_2	NOx	OXY	O_3	
0	2006-02-01 01:00:00	NaN	1.84	NaN	NaN	NaN	155.100006	490.100006	NaN	4.880000	97.
1	2006-02-01 01:00:00	1.68	1.01	2.38	6.36	0.32	94.339996	229.699997	3.04	7.100000	25.
2	2006-02-01 01:00:00	NaN	1.25	NaN	NaN	NaN	66.800003	192.000000	NaN	4.430000	34.
3	2006-02-01 01:00:00	NaN	1.68	NaN	NaN	NaN	103.000000	407.799988	NaN	4.830000	28.
4	2006-02-01 01:00:00	NaN	1.31	NaN	NaN	NaN	105.400002	269.200012	NaN	6.990000	54.
...	...	...	...	...	...	...	...	...	...	...	
230563	2006-05-01 00:00:00	5.88	0.83	6.23	NaN	0.20	112.500000	218.000000	NaN	24.389999	93.
230564	2006-05-01 00:00:00	0.76	0.32	0.48	1.09	0.08	51.900002	54.820000	0.61	48.410000	29.
230565	2006-05-01 00:00:00	0.96	NaN	0.69	NaN	0.19	135.100006	179.199997	NaN	11.460000	64.
230566	2006-05-01 00:00:00	0.50	NaN	0.67	NaN	0.10	82.599998	105.599998	NaN	NaN	94.
230567	2006-05-01 00:00:00	1.95	0.74	1.99	4.00	0.24	107.300003	160.199997	2.01	17.730000	52.

230568 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: 24758 entries, 5 to 230567
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     24758 non-null  object 
 1   BEN      24758 non-null  float64
 2   CO       24758 non-null  float64
 3   EBE      24758 non-null  float64
 4   MXY      24758 non-null  float64
 5   NMHC     24758 non-null  float64
 6   NO_2     24758 non-null  float64
 7   NOx      24758 non-null  float64
 8   OXY      24758 non-null  float64
 9   O_3      24758 non-null  float64
 10  PM10     24758 non-null  float64
 11  PM25     24758 non-null  float64
 12  PXY      24758 non-null  float64
 13  SO_2     24758 non-null  float64
 14  TCH      24758 non-null  float64
 15  TOL      24758 non-null  float64
 16  station  24758 non-null  int64  
dtypes: float64(15), int64(1), object(1)
memory usage: 3.4+ MB
```

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

Out[6]:

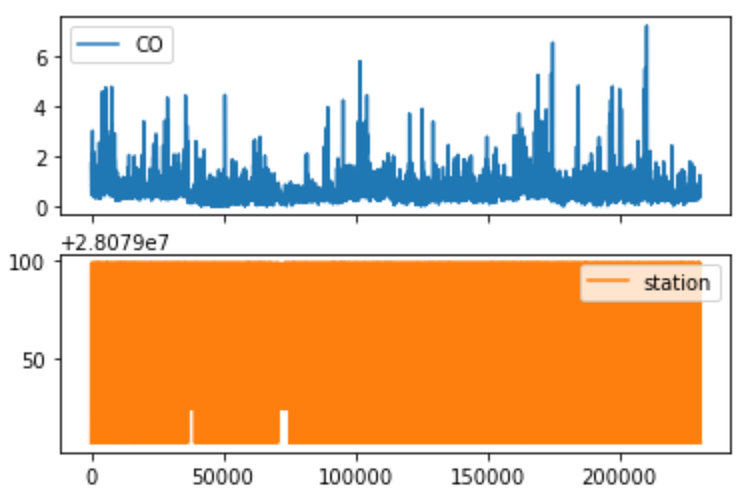
	CO	station
5	1.69	28079006
22	0.79	28079024
25	1.47	28079099
31	0.85	28079006
48	0.79	28079024
...	...	...
230538	0.40	28079024
230541	0.94	28079099
230547	1.06	28079006
230564	0.32	28079024
230567	0.74	28079099

24758 rows × 2 columns

Line chart

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

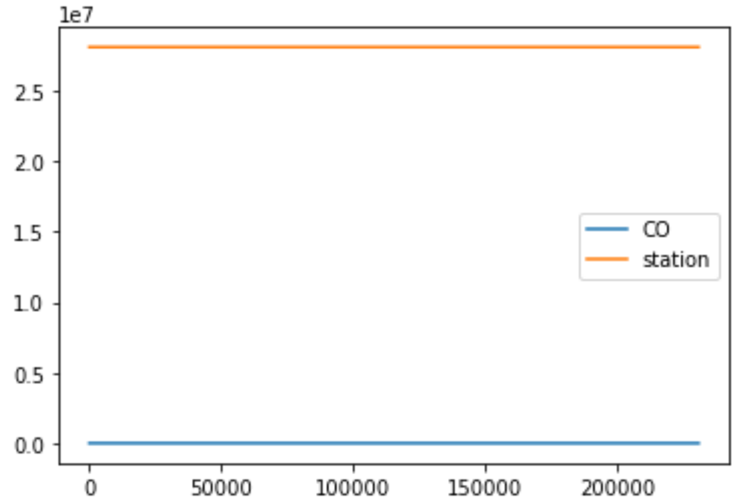
Out[7]:
```
array([<AxesSubplot:>, <AxesSubplot:>], dtype=object)
```



Line chart

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

Out[8]: <AxesSubplot:>

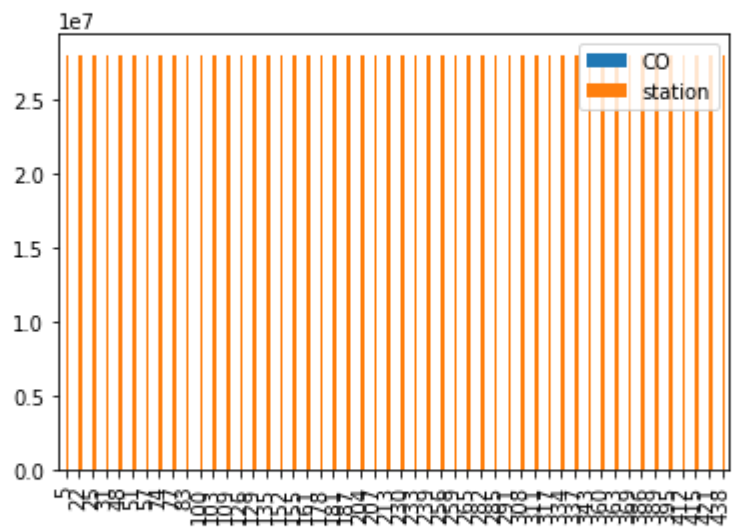


Bar chart

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

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

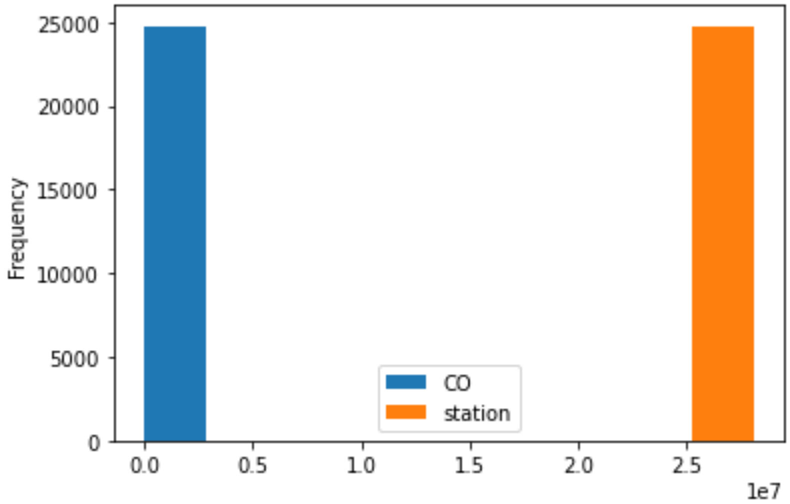
Out[10]: <AxesSubplot:>



Histogram

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

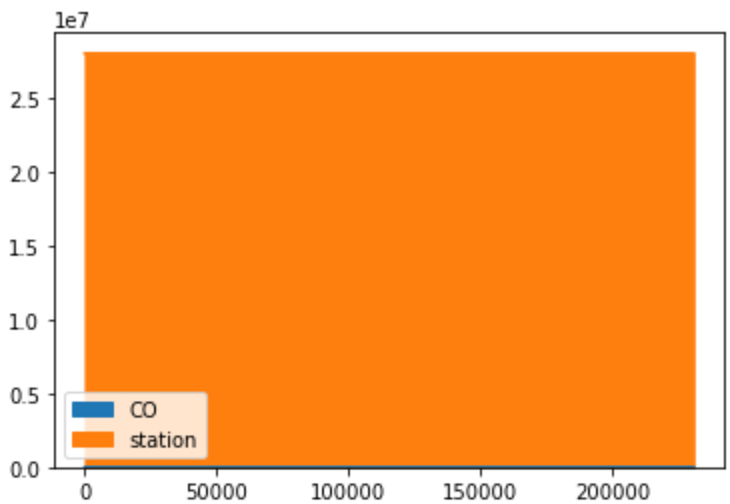
Out[11]: <AxesSubplot:ylabel='Frequency'>



Area chart

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

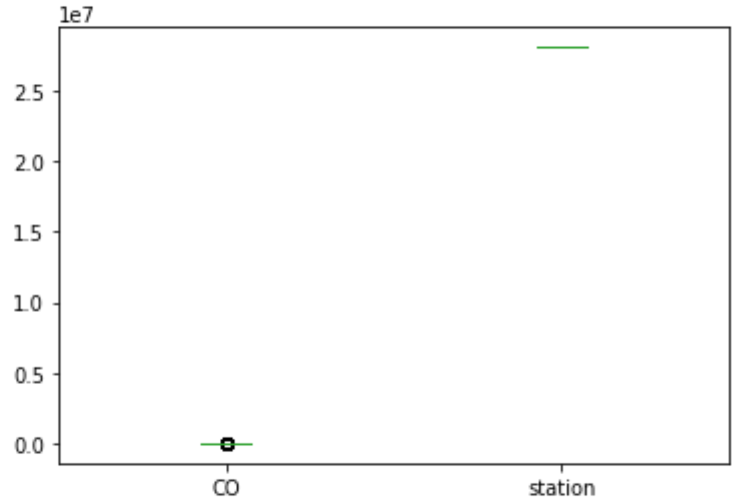
Out[12]: <AxesSubplot:>



Box chart

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

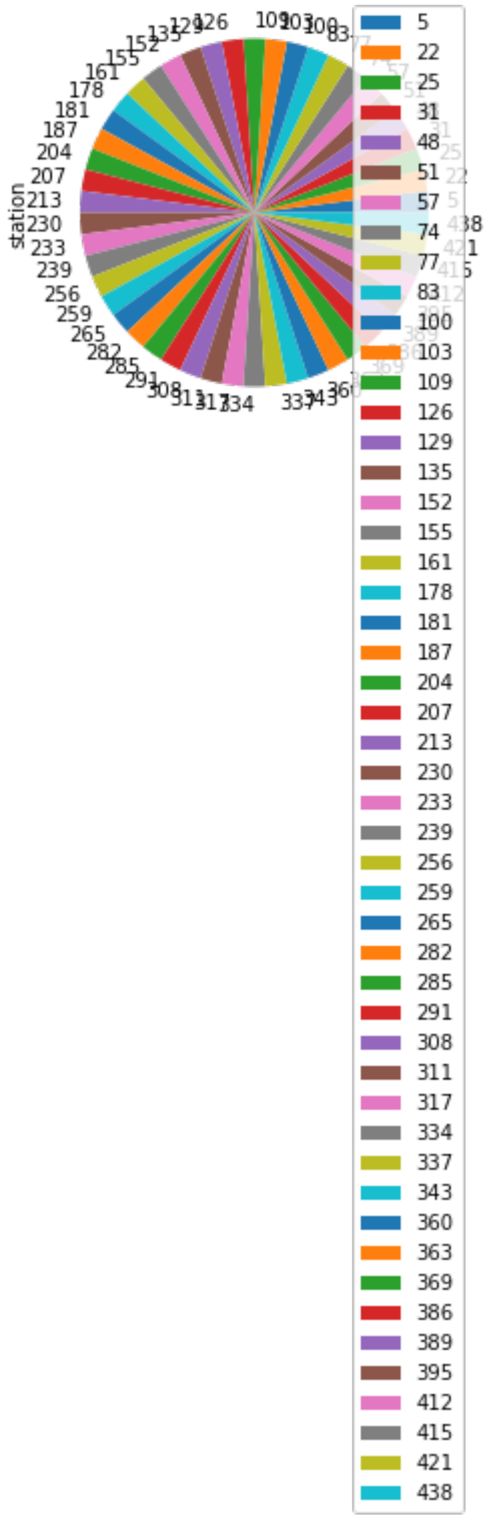
Out[13]: <AxesSubplot:>



Pie chart

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

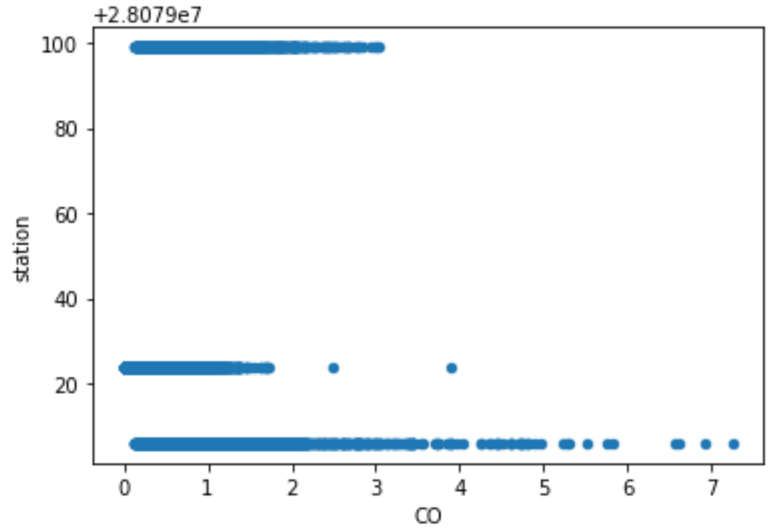
Out[14]: `<AxesSubplot:ylabel='station'>`



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: 24758 entries, 5 to 230567
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   date     24758 non-null  object
 1   BEN      24758 non-null  float64
 2   CO       24758 non-null  float64
 3   EBE      24758 non-null  float64
 4   MXY      24758 non-null  float64
 5   NMHC     24758 non-null  float64
 6   NO_2     24758 non-null  float64
 7   NOx      24758 non-null  float64
 8   OXY      24758 non-null  float64
 9   O_3      24758 non-null  float64
 10  PM10     24758 non-null  float64
 11  PM25     24758 non-null  float64
 12  PXY      24758 non-null  float64
 13  SO_2     24758 non-null  float64
 14  TCH      24758 non-null  float64
```

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

Out[17]:

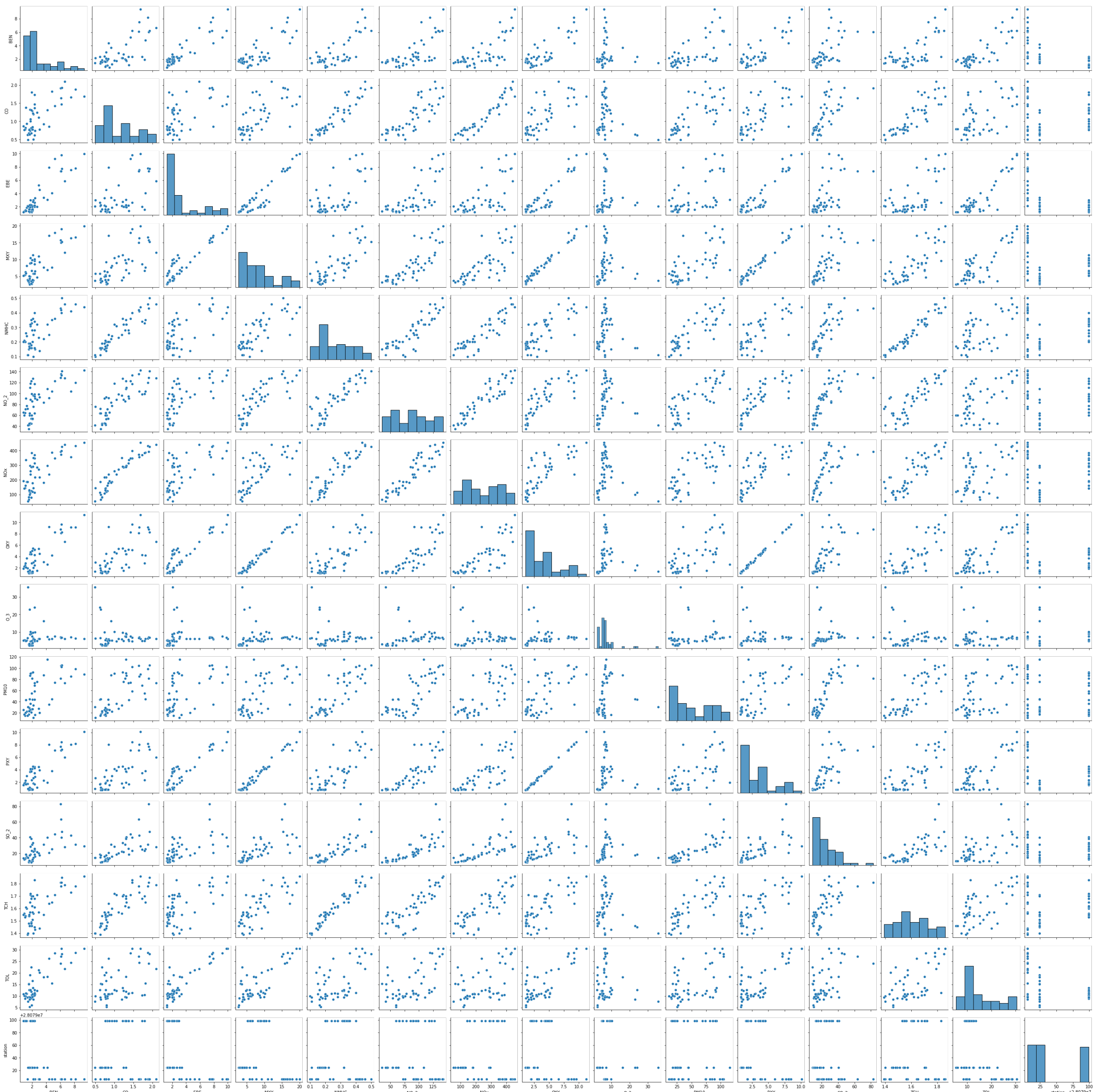
	BEN	CO	EBE	MXY	NMHC	NO_2	
count	24758.000000	24758.000000	24758.000000	24758.000000	24758.000000	24758.000000	247
mean	1.350624	0.600713	1.824534	3.835034	0.176546	58.333481	1
std	1.541636	0.419048	1.868939	4.069036	0.126683	40.529382	1
min	0.110000	0.000000	0.170000	0.150000	0.000000	1.680000	
25%	0.450000	0.360000	0.810000	1.060000	0.100000	28.450001	
50%	0.850000	0.500000	1.130000	2.500000	0.150000	52.959999	
75%	1.680000	0.720000	2.160000	5.090000	0.220000	79.347498	1
max	45.430000	7.250000	57.799999	66.900002	2.020000	461.299988	16

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])
```

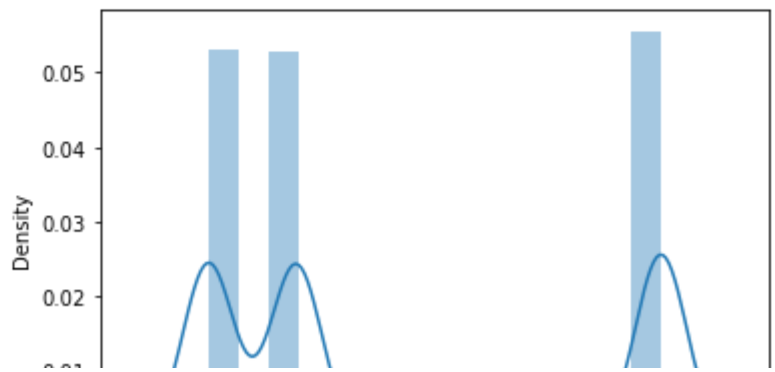
Out[19]: <seaborn.axisgrid.PairGrid at 0x20335382b80>



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

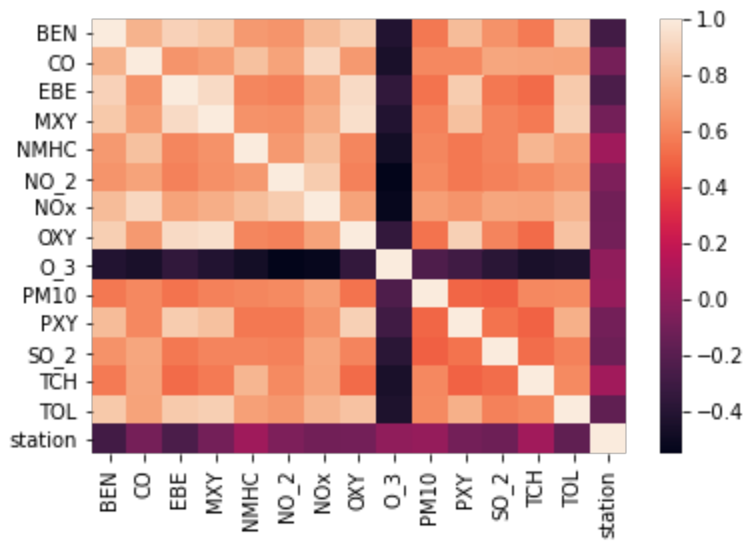
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function 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]: 28079013.414891884

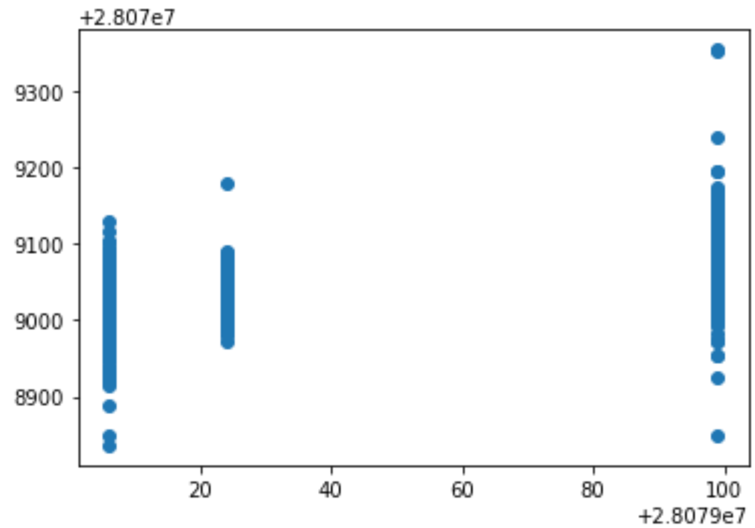
In [26]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

Out[26]:

	Co-efficient
BEN	-18.561713
CO	-13.066381
EBE	-21.247476
MXY	2.635001
NMHC	130.910680
NO_2	-0.008165
NOx	-0.009275
OXY	18.118339
O_3	-0.056919
PM10	0.143264
PXY	6.165699
SO_2	-0.653803
TCH	23.167821
TOL	-0.699416

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

Out[27]: <matplotlib.collections.PathCollection at 0x20343cfab80>



ACCURACY

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

Out[28]: 0.4082591915094467

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

Out[29]: 0.38629705567566075

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]: 0.40819618490318543

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

Out[33]: 0.3856182327593829

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]:
```
0.06110017645134225
```

Accuracy(Lasso)

In [36]:
```
la.score(x_test,y_test)
```
Out[36]:
```
0.060409984845479325
```

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([-8.25695289e+00,  0.00000000e+00, -8.43038927e+00,  3.14630738e+00,
        4.20031356e-01,  0.00000000e+00,  1.95168770e-03,  3.63985159e+00,
       -1.15134089e-01,  3.03009505e-01,  2.87657806e+00, -4.44426535e-01,
        6.20824138e-01, -1.16271554e+00])
```

In [39]:
```
en.intercept_
```
Out[39]:
```
28079051.087767527
```

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

In [41]:
```
en.score(x_test,y_test)
```
Out[41]:
```
0.24396391772353465
```

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)))
```
```
32.078270309298276
1237.8837464625722
35.18357211061112
```

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]: (24758, 14)

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

Out[46]: (24758,)

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]: 0.8741416915744405

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

Out[54]: 3.5557727473608076e-15

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

Out[55]: array([[3.55577275e-15, 7.80743173e-29, 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="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]:
```
0.875072129255626
```

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

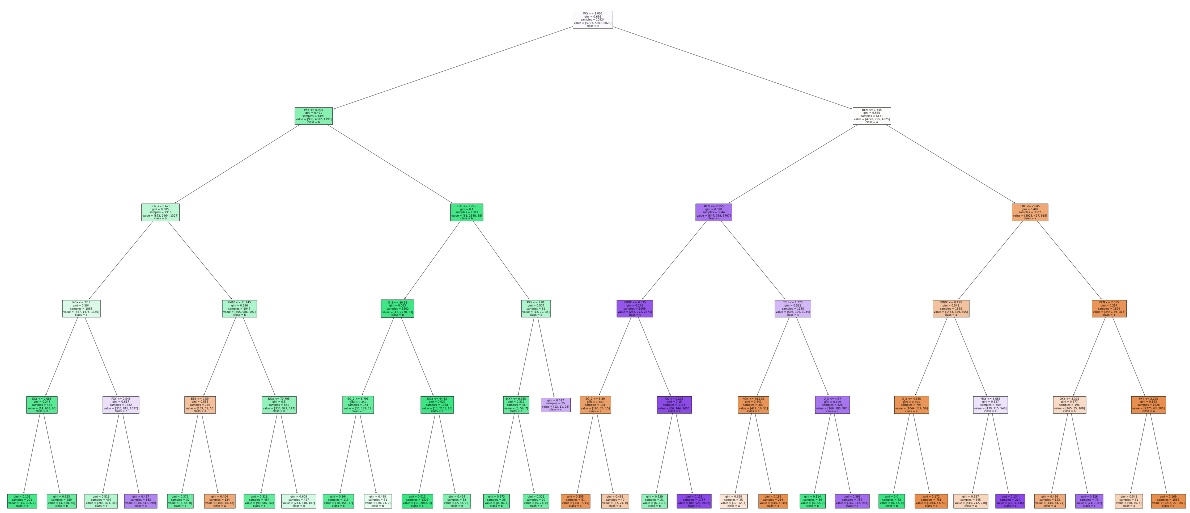
In [62]:
```
from sklearn.tree import plot_tree

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

Out[62]:
```
[Text(2222.7000000000003, 1993.2, 'OXY <= 1.005\ngini = 0.666\nsamples = 1092
4\nvalue = [5703, 5607, 6020]\nclass = c'),
 Text(1171.8000000000002, 1630.8000000000002, 'PXY <= 0.995\ngini = 0.491\nsa
mples = 4493\nvalue = [933, 4812, 1395]\nclass = b'),
 Text(595.2, 1268.4, 'BEN <= 0.625\ngini = 0.605\nsamples = 2910\nvalue = [87
2, 2464, 1327]\nclass = b'),
 Text(297.6, 906.0, 'NOx <= 22.9\ngini = 0.594\nsamples = 1863\nvalue = [367,
1478, 1130]\nclass = b'),
 Text(148.8, 543.5999999999999, 'OXY <= 0.685\ngini = 0.244\nsamples = 481\nv
alue = [14, 663, 93]\nclass = b'),
 Text(74.4, 181.19999999999982, 'gini = 0.102\nsamples = 201\nvalue = [10, 30
3, 7]\nclass = b'),
 Text(223.20000000000002, 181.19999999999982, 'gini = 0.323\nsamples = 280\nv
alue = [4, 360, 86]\nclass = b'),
 Text(446.40000000000003, 543.5999999999999, 'PXY <= 0.545\ngini = 0.617\nsam
ples = 1382\nvalue = [353, 815, 1037]\nclass = c'),
 Text(372.0, 181.19999999999982, 'gini = 0.516\nsamples = 499\nvalue = [283,
474, 38]\nclass = b'),
 Text(520.8000000000001, 181.19999999999982, 'gini = 0.437\nsamples = 883\nva
lue = [70, 341, 999]\nclass = c'),
 Text(892.8000000000001, 906.0, 'PM10 <= 11.145\ngini = 0.556\nsamples = 104
7\nvalue = [505, 986, 197]\nclass = b'),
 Text(744.0, 543.5999999999999, 'EBE <= 0.59\ngini = 0.553\nsamples = 166\nva
lue = [169, 59, 50]\nclass = a'),
 Text(669.6, 181.19999999999982, 'gini = 0.372\nsamples = 32\nvalue = [5, 45,
8]\nclass = b'),
 Text(818.4000000000001, 181.19999999999982, 'gini = 0.404\nsamples = 134\nva
lue = [164, 14, 42]\nclass = a'),
 Text(1041.6000000000001, 543.5999999999999, 'NOx <= 70.795\ngini = 0.5\nsamp
les = 881\nvalue = [336, 927, 147]\nclass = b'),
 Text(967.2, 181.19999999999982, 'gini = 0.316\nsamples = 454\nvalue = [93, 5
87, 40]\nclass = b'),
 Text(1116.0, 181.19999999999982, 'gini = 0.609\nsamples = 427\nvalue = [243,
340, 107]\nclass = b'),
 Text(1748.4, 1268.4, 'TOL <= 2.275\ngini = 0.1\nsamples = 1583\nvalue = [61,
2348, 68]\nclass = b'),
 Text(1488.0, 906.0, 'O_3 <= 30.36\ngini = 0.063\nsamples = 1502\nvalue = [4
3, 2278, 33]\nclass = b'),
 Text(1339.2, 543.5999999999999, 'SO_2 <= 8.795\ngini = 0.352\nsamples = 144\
nvalue = [30, 177, 17]\nclass = b'),
 Text(1264.8000000000002, 181.19999999999982, 'gini = 0.264\nsamples = 113\nv
alue = [10, 154, 17]\nclass = b'),
 Text(1413.6000000000001, 181.19999999999982, 'gini = 0.498\nsamples = 31\nva
lue = [20, 23, 0]\nclass = b'),
 Text(1636.8000000000002, 543.5999999999999, 'NOx <= 44.32\ngini = 0.027\nsam
ples = 1358\nvalue = [13, 2101, 16]\nclass = b'),
 Text(1562.4, 181.19999999999982, 'gini = 0.013\nsamples = 1325\nvalue = [11,
2063, 3]\nclass = b'),
 Text(1711.2, 181.19999999999982, 'gini = 0.424\nsamples = 33\nvalue = [2, 3
8, 13]\nclass = b'),
 Text(2008.8000000000002, 906.0, 'PXY <= 1.01\ngini = 0.574\nsamples = 81\nva
```

```
lue = [18, 70, 35]\nclass = b'),
 Text(1934.4, 543.5999999999999, 'MXY <= 0.985\ngini = 0.312\nsamples = 46\nv
alue = [6, 59, 7]\nclass = b'),
 Text(1860.0000000000002, 181.19999999999982, 'gini = 0.273\nsamples = 26\nva
lue = [0, 36, 7]\nclass = b'),
 Text(2008.8000000000002, 181.19999999999982, 'gini = 0.328\nsamples = 20\nva
lue = [6, 23, 0]\nclass = b'),
 Text(2083.2000000000003, 543.5999999999999, 'gini = 0.597\nsamples = 35\nval
ue = [12, 11, 28]\nclass = c'),
 Text(3273.6000000000004, 1630.8000000000002, 'BEN <= 1.345\ngini = 0.569\nsa
mples = 6431\nvalue = [4770, 795, 4625]\nclass = a'),
 Text(2678.4, 1268.4, 'BEN <= 0.955\ngini = 0.398\nsamples = 3044\nvalue = [8
47, 368, 3707]\nclass = c'),
 Text(2380.8, 906.0, 'NMHC <= 0.075\ngini = 0.246\nsamples = 1909\nvalue = [2
54, 172, 2673]\nclass = c'),
 Text(2232.0, 543.5999999999999, 'SO_2 <= 8.39\ngini = 0.343\nsamples = 131\n
value = [168, 28, 15]\nclass = a'),
 Text(2157.6000000000004, 181.19999999999982, 'gini = 0.255\nsamples = 91\nva
lue = [131, 7, 15]\nclass = a'),
 Text(2306.4, 181.19999999999982, 'gini = 0.462\nsamples = 40\nvalue = [37, 2
1, 0]\nclass = a'),
 Text(2529.6000000000004, 543.5999999999999, 'CO <= 0.205\ngini = 0.15\nsampl
es = 1778\nvalue = [86, 144, 2658]\nclass = c'),
 Text(2455.2000000000003, 181.19999999999982, 'gini = 0.529\nsamples = 21\nva
lue = [6, 21, 6]\nclass = b'),
 Text(2604.0, 181.19999999999982, 'gini = 0.135\nsamples = 1757\nvalue = [80,
123, 2652]\nclass = c'),
 Text(2976.0, 906.0, 'TCH <= 1.325\ngini = 0.561\nsamples = 1135\nvalue = [59
3, 196, 1034]\nclass = c'),
 Text(2827.2000000000003, 543.5999999999999, 'NOx <= 48.105\ngini = 0.241\nsa
mples = 309\nvalue = [427, 16, 51]\nclass = a'),
 Text(2752.8, 181.19999999999982, 'gini = 0.628\nsamples = 25\nvalue = [17, 1
2, 7]\nclass = a'),
 Text(2901.6000000000004, 181.19999999999982, 'gini = 0.189\nsamples = 284\nv
alue = [410, 4, 44]\nclass = a'),
 Text(3124.8, 543.5999999999999, 'O_3 <= 4.67\ngini = 0.419\nsamples = 826\nv
alue = [166, 180, 983]\nclass = c'),
 Text(3050.4, 181.19999999999982, 'gini = 0.114\nsamples = 39\nvalue = [4, 6
2, 0]\nclass = b'),
 Text(3199.2000000000003, 181.19999999999982, 'gini = 0.369\nsamples = 787\nv
alue = [162, 118, 983]\nclass = c'),
 Text(3868.8, 1268.4, 'EBE <= 2.845\ngini = 0.409\nsamples = 3387\nvalue = [3
923, 427, 918]\nclass = a'),
 Text(3571.2000000000003, 906.0, 'NMHC <= 0.195\ngini = 0.542\nsamples = 156
3\nvalue = [1483, 329, 605]\nclass = a'),
 Text(3422.4, 543.5999999999999, 'O_3 <= 4.635\ngini = 0.253\nsamples = 794\n
value = [1044, 114, 59]\nclass = a'),
 Text(3348.0000000000005, 181.19999999999982, 'gini = 0.0\nsamples = 43\nvalu
e = [0, 67, 0]\nclass = b'),
 Text(3496.8, 181.19999999999982, 'gini = 0.172\nsamples = 751\nvalue = [104
4, 47, 59]\nclass = a'),
 Text(3720.0000000000005, 543.5999999999999, 'MXY <= 5.685\ngini = 0.627\nsam
ples = 769\nvalue = [439, 215, 546]\nclass = c'),
 Text(3645.6000000000004, 181.19999999999982, 'gini = 0.627\nsamples = 545\nv
alue = [416, 212, 210]\nclass = a'),
 Text(3794.4, 181.19999999999982, 'gini = 0.134\nsamples = 224\nvalue = [23,
```

```
3, 336]\nclass = c'),
 Text(4166.400000000001, 906.0, 'BEN <= 2.065\ngini = 0.254\nsamples = 1824\n
value = [2440, 98, 313]\nclass = a'),
 Text(4017.6000000000004, 543.5999999999999, 'OXY <= 3.395\ngini = 0.577\nsam
ples = 196\nvalue = [165, 35, 108]\nclass = a'),
 Text(3943.2000000000003, 181.19999999999982, 'gini = 0.436\nsamples = 123\nv
alue = [144, 34, 21]\nclass = a'),
 Text(4092.0000000000005, 181.19999999999982, 'gini = 0.326\nsamples = 73\nva
lue = [21, 1, 87]\nclass = c'),
 Text(4315.200000000001, 543.5999999999999, 'PXY <= 2.295\ngini = 0.193\nsamp
les = 1628\nvalue = [2275, 63, 205]\nclass = a'),
 Text(4240.8, 181.19999999999982, 'gini = 0.541\nsamples = 61\nvalue = [60, 3
6, 8]\nclass = a'),
 Text(4389.6, 181.19999999999982, 'gini = 0.169\nsamples = 1567\nvalue = [221
```



Conclusion

Accuracy

Linear Regression :0.38629705567566075

Ridge Regression : 0.06110017645134225

Lasso Regression : 0.060409984845479325

ElasticNet Regression : 0.24396391772353465

Logistic Regression : 0.8741416915744405

Random Forest :0.875072129255626

Random Forest is suitable for this dataset