Importing Libraries

In [182]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Importing Datasets

In [183]:
```
df=pd.read_csv(r"C:\Users\user\Desktop\csvs_per_year\csvs_per_year\madrid_2005.
df
```

Out[183]:

	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 [184]:
```
df=df.dropna()
```

In [185]:
```
df.columns
```

Out[185]:
```
Index(['date', 'BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_
3',
       'PM10', 'PM25', 'PXY', 'SO_2', 'TCH', 'TOL', 'station'],
      dtype='object')
```

In [186]:
```
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 [187]:
```
data=df[['CO' ,'station']]
data
```

Out[187]:

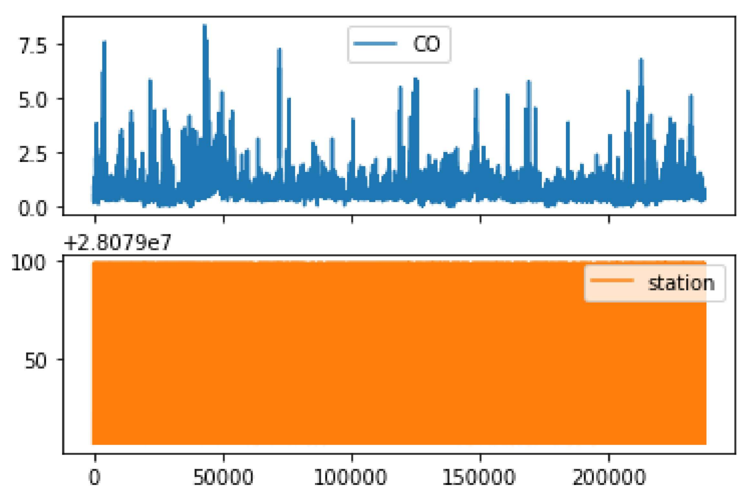
	CO	station
5	0.88	28079006
22	0.22	28079024
25	0.49	28079099
31	0.84	28079006
48	0.20	28079024
...	...	...
236970	0.39	28079024
236973	0.45	28079099
236979	0.38	28079006
236996	0.54	28079024
236999	0.40	28079099

20070 rows × 2 columns

Line chart

In [188]:
```
data.plot.line(subplots=True)
```

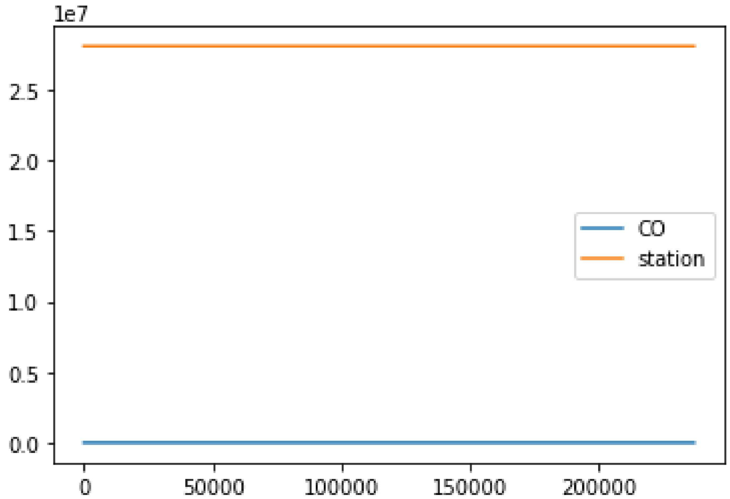
Out[188]:
```
array([<AxesSubplot:>, <AxesSubplot:>], dtype=object)
```



Line chart

In [189]:
```
data.plot.line()
```

Out[189]: <AxesSubplot:>

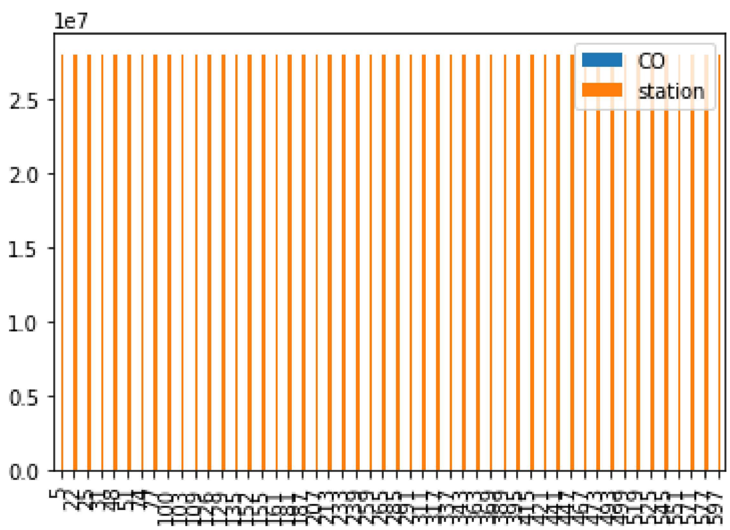


Bar chart

In [190]:
```
b=data[0:50]
```

In [191]:
```
b.plot.bar()
```

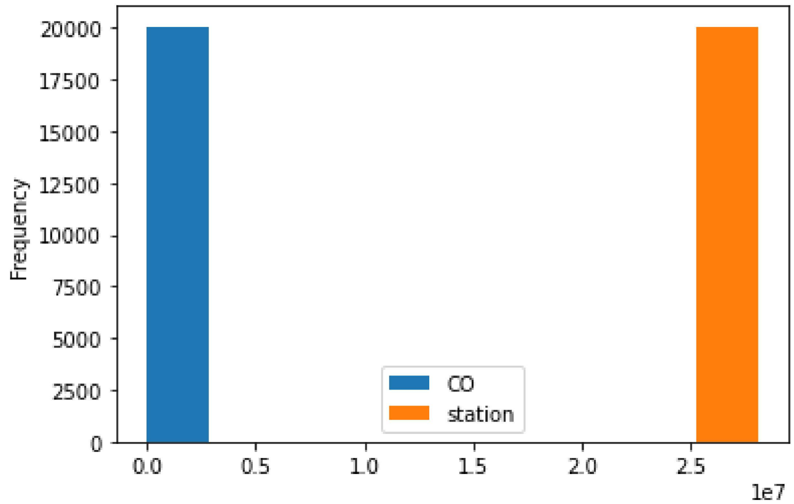
Out[191]: <AxesSubplot:>



Histogram

In [192]:
```
data.plot.hist()
```

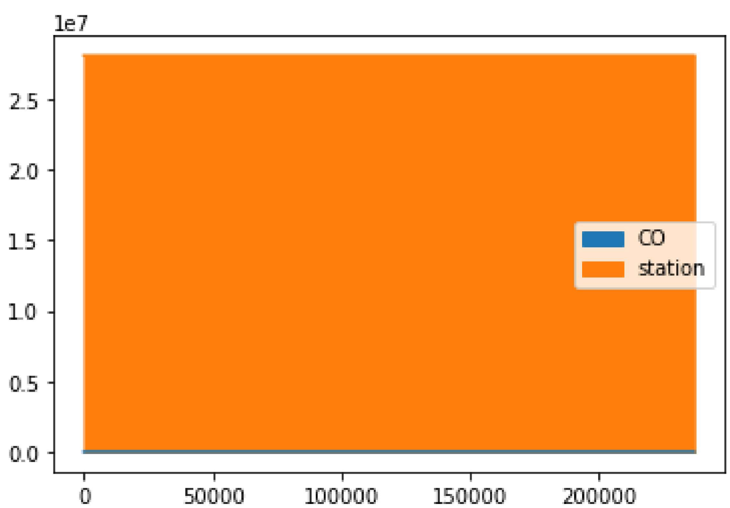
Out[192]: <AxesSubplot:ylabel='Frequency'>



Area chart

In [193]:
```
data.plot.area()
```

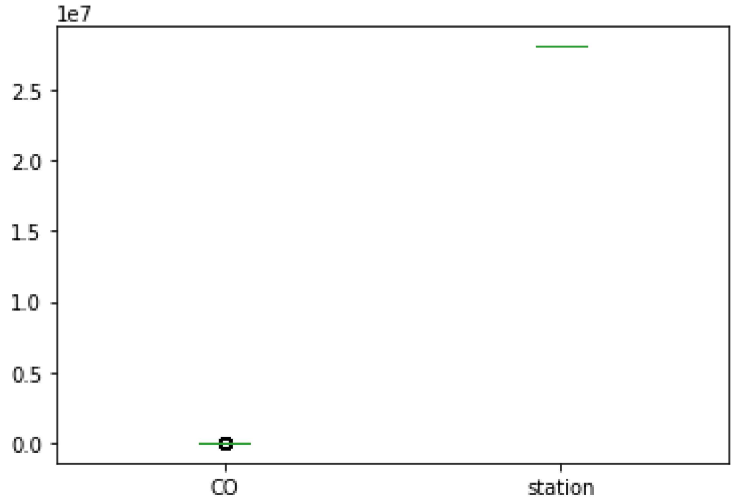
Out[193]: <AxesSubplot:>



Box chart

In [194]:
```
data.plot.box()
```

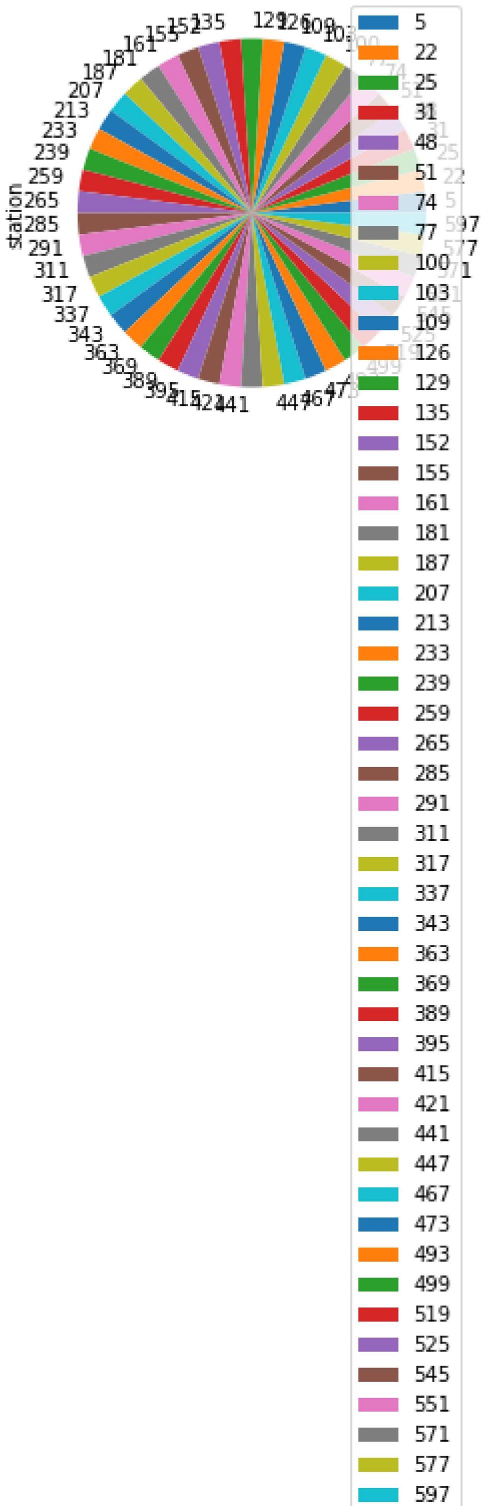
Out[194]: <AxesSubplot:>



Pie chart

In [195]:
```
b.plot.pie(y='station' )
```

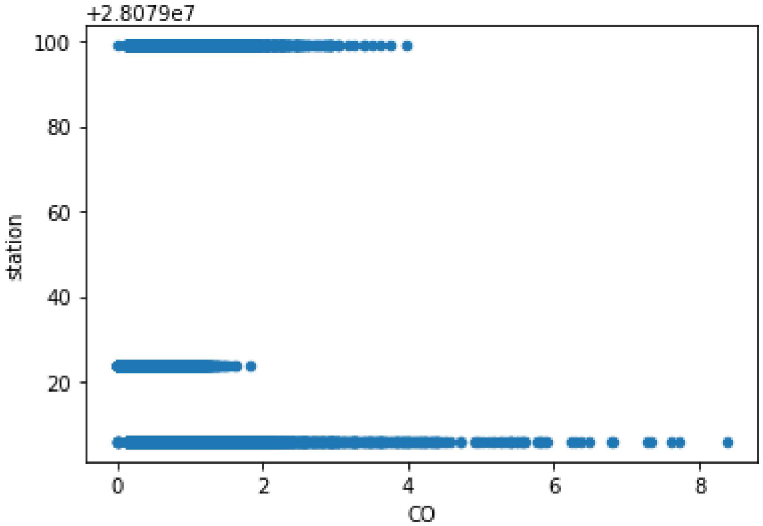
Out[195]:
```
<AxesSubplot:ylabel='station'>
```



Scatter chart

In [196]:
```
data.plot.scatter(x='CO' ,y='station')
```

Out[196]: <AxesSubplot:xlabel='CO', ylabel='station'>



In [197]:
```
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 [198]:
```
df.describe()
```

Out[198]:

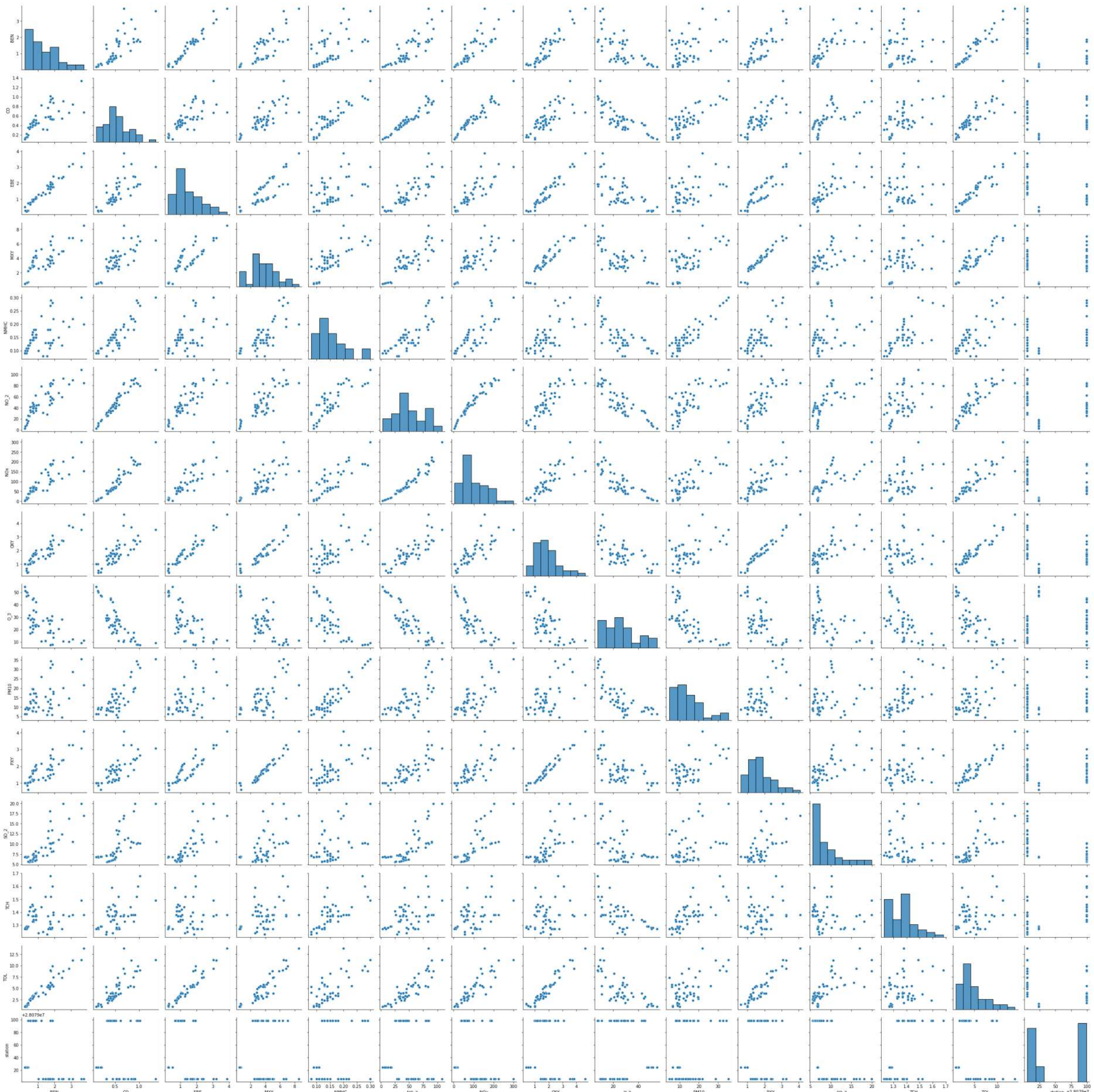
	BEN	CO	EBE	MXY	NMHC	NO_2	
count	20070.000000	20070.000000	20070.000000	20070.000000	20070.000000	20070.000000	200
mean	1.923656	0.720657	2.345423	5.457855	0.179282	66.226924	1
std	2.019061	0.549723	2.379219	5.495147	0.152783	40.568197	1
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.690000	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 [199]:
```
df1=df[['BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3',
 'PM10', 'PXY', 'SO_2', 'TCH', 'TOL', 'station']]
```

EDA AND VISUALIZATION

In [200]:
```
sns.pairplot(df1[0:50])
```

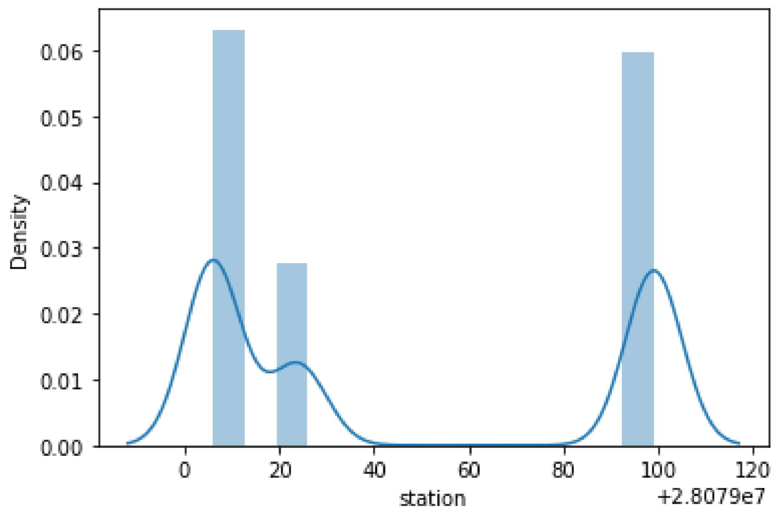
Out[200]: <seaborn.axisgrid.PairGrid at 0x1cd0a0801f0>



In [201]:
```
sns.distplot(df1['station'])
```

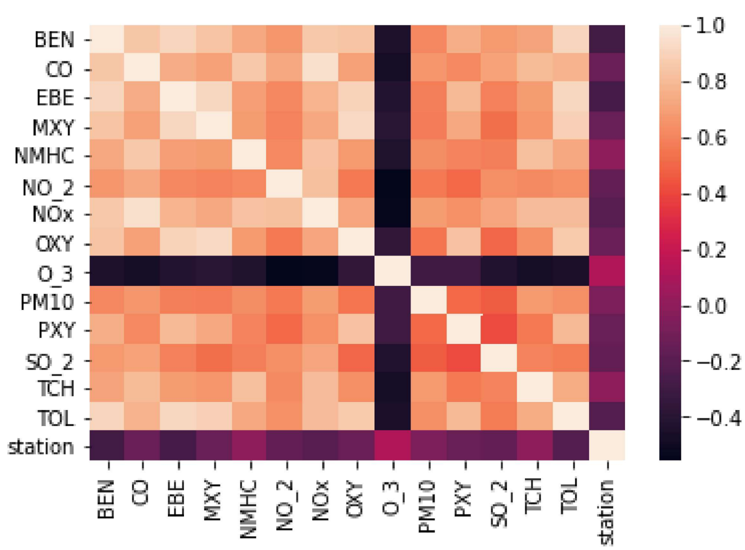
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function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

Out[201]: <AxesSubplot:xlabel='station', ylabel='Density'>



In [202]:
```
sns.heatmap(df1.corr())
```

Out[202]: <AxesSubplot:>



TO TRAIN THE MODEL AND MODEL BULDING

In [203]:
```
x=df[['BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_3',
 'PM10', 'PXY', 'SO_2', 'TCH', 'TOL']]
y=df['station']
```

In [204]:
```
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 [205]:
```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
lr.fit(x_train,y_train)
```

Out[205]:
```
LinearRegression()
```

In [206]:
```
lr.intercept_
```

Out[206]:
```
28078959.58039427
```

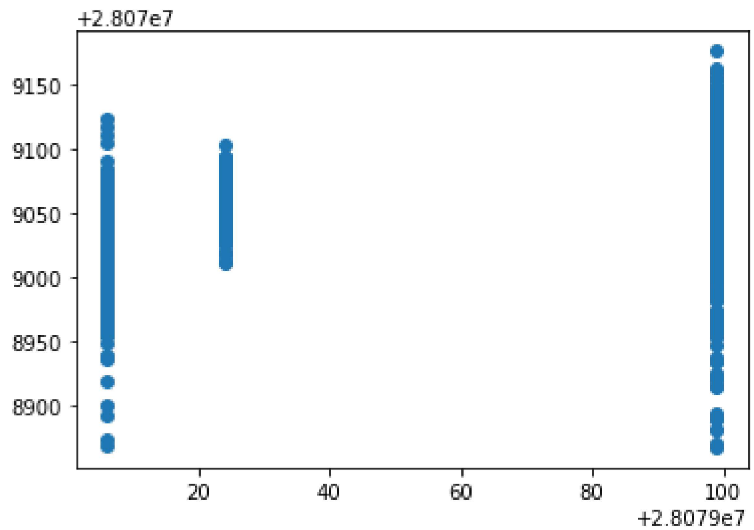
In [207]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

Out[207]:

	Co-efficient
BEN	-9.663451
CO	38.491543
EBE	-14.147655
MXY	3.821455
NMHC	74.033915
NO_2	0.119170
NOx	-0.263549
OXY	3.444876
O_3	0.016010
PM10	0.056085
PXY	2.832713
SO_2	0.225392
TCH	62.036186
TOL	-0.522836

In [208]:
```
prediction =lr.predict(x_test)
plt.scatter(y_test,prediction)
```

Out[208]: <matplotlib.collections.PathCollection at 0x1cd3df1e070>



ACCURACY

In [209]:
```
lr.score(x_test,y_test)
```

Out[209]: 0.28512114725842885

In [210]:
```
lr.score(x_train,y_train)
```

Out[210]: 0.3120143412245254

Ridge and Lasso

In [211]:
```
from sklearn.linear_model import Ridge,Lasso
```

In [212]:
```
rr=Ridge(alpha=10)
rr.fit(x_train,y_train)
```

Out[212]: Ridge(alpha=10)

Accuracy(Ridge)

In [213]:
```
rr.score(x_test,y_test)
```

Out[213]: 0.28383484258388436

In [214]:
```
rr.score(x_train,y_train)
```

Out[214]: 0.31180826883293056

In [215]:
```
la=Lasso(alpha=10)
la.fit(x_train,y_train)
```

Out[215]: Lasso(alpha=10)

In [216]:
```
la.score(x_train,y_train)
```

Out[216]: 0.0683076866149277

Accuracy(Lasso)

In [217]:
```
la.score(x_test,y_test)
```

Out[217]: 0.05686629934258647

In [218]:
```
from sklearn.linear_model import ElasticNet
en=ElasticNet()
en.fit(x_train,y_train)
```

Out[218]: ElasticNet()

In [219]:
```
en.coef_
```

Out[219]:
```
array([-5.66801685,  1.42490601, -7.75175396,  2.71663535,  0.85527075,
       -0.05134181, -0.01140651,  2.05978166, -0.01382005,  0.2530742 ,
        1.51377336,  0.14798522,  1.46534522, -0.81923336])
```

In [220]:
```
en.intercept_
```

Out[220]: 28079049.288084675

In [221]:
```
prediction=en.predict(x_test)
```

In [222]:
```
en.score(x_test,y_test)
```

Out[222]: 0.1576872672135906

Evaluation Metrics

In [223]:
```
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)))
```

```
37.16520319961406
1579.8310424576925
39.74708847774504
```

Logistic Regression

In [224]:
```
from sklearn.linear_model import LogisticRegression
```

In [225]:
```
feature_matrix=df[['BEN', 'CO', 'EBE', 'MXY', 'NMHC', 'NO_2', 'NOx', 'OXY', 'O_
 'PM10', 'PXY', 'SO_2', 'TCH', 'TOL']]
target_vector=df[ 'station']
```

In [226]:
```
feature_matrix.shape
```

Out[226]:
```
(20070, 14)
```

In [227]:
```
target_vector.shape
```

Out[227]:
```
(20070,)
```

In [228]:
```
from sklearn.preprocessing import StandardScaler
```

In [229]:
```
fs=StandardScaler().fit_transform(feature_matrix)
```

In [230]:
```
logr=LogisticRegression(max_iter=10000)
logr.fit(fs,target_vector)
```

Out[230]:
```
LogisticRegression(max_iter=10000)
```

In [231]:
```
observation=[[1,2,3,4,5,6,7,8,9,10,11,12,13,14]]
```

In [232]:
```
prediction=logr.predict(observation)
print(prediction)
```

```
[28079006]
```

In [233]:
```
logr.classes_
```

Out[233]:
```
array([28079006, 28079024, 28079099], dtype=int64)
```

In [234]:
```
logr.score(fs,target_vector)
```

Out[234]:
```
0.879023418036871
```

In [235]:
```
logr.predict_proba(observation)[0][0]
```

Out[235]:
```
0.9998967601812779
```

In [236]:
```
logr.predict_proba(observation)
```

Out[236]:
```
array([[9.99896760e-01, 3.21124597e-30, 1.03239819e-04]])
```

Random Forest

In [237]:
```
from sklearn.ensemble import RandomForestClassifier
```

In [238]:
```
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[238]:
```
RandomForestClassifier()
```

In [239]:
```
parameters={'max_depth':[1,2,3,4,5],
 'min_samples_leaf':[5,10,15,20,25],
 'n_estimators':[10,20,30,40,50]
}
```

In [240]:
```
from sklearn.model_selection import GridSearchCV
grid_search =GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="acc
grid_search.fit(x_train,y_train)
```

Out[240]:
```
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 [241]:
```
grid_search.best_score_
```

Out[241]:
```
0.8631926024854287
```

In [242]:
```
rfc_best=grid_search.best_estimator_
```

In [243]:
```
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',
```

```
 Text(1004.4, 181.19999999999982, 'gini = 0.33\nsamples = 15\nvalue = [19,
0, 5]\nclass = a'),
 Text(1227.6, 181.19999999999982, 'gini = 0.0\nsamples = 47\nvalue = [76,
0, 0]\nclass = a'),
 Text(1227.6, 906.0, 'gini = 0.439\nsamples = 16\nvalue = [23, 3, 6]\nclass
= a'),
 Text(2985.2999999999997, 1630.8000000000002, 'BEN <= 1.525\ngini = 0.58\ns
amples = 7600\nvalue = [5563, 1116, 5296]\nclass = a'),
 Text(2232.0, 1268.4, 'MXY <= 3.735\ngini = 0.518\nsamples = 3831\nvalue =
[1352, 820, 3923]\nclass = c'),
 Text(1785.6, 906.0, 'PM10 <= 33.535\ngini = 0.591\nsamples = 2604\nvalue =
[1103, 746, 2288]\nclass = c'),
 Text(1562.3999999999999, 543.5999999999999, 'NMHC <= 0.055\ngini = 0.56\ns
amples = 2005\nvalue = [945, 393, 1856]\nclass = c'),
 Text(1450.8, 181.19999999999982, 'gini = 0.187\nsamples = 369\nvalue = [52
2, 18, 41]\nclass = a'),
 Text(1674.0, 181.19999999999982, 'gini = 0.471\nsamples = 1636\nvalue = [4
23, 375, 1815]\nclass = c'),
 Text(2008.8, 543.5999999999999, 'TOL <= 2.925\ngini = 0.622\nsamples = 599
```

Conclusion

Accuracy

Linear Regression:0.312467100401984

Ridge Regression:0.3121714361922914

Lasso Regression:0.06353386104215053

ElasticNet Regression:0.17690741038500357

Logistic Regression:0.879023418036871

```
Random Forest:0.8634068653280262
```

From the above data, we can conclude that logistic regression and random forest is preferrable to other regression types

In [ ]:

```

```