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(r"C:\Users\user\Downloads\C10_air\csvs_per_year\csvs(Dataset)\madrid_2008.
df
```

Out[2]:

	date	BEN	CO	EBE	MXY	NMHC	NO_2	NOx	OXY	O_3	
0	2008-06-01 01:00:00	NaN	0.47	NaN	NaN	NaN	83.089996	120.699997	NaN	16.990000	16
1	2008-06-01 01:00:00	NaN	0.59	NaN	NaN	NaN	94.820000	130.399994	NaN	17.469999	19
2	2008-06-01 01:00:00	NaN	0.55	NaN	NaN	NaN	75.919998	104.599998	NaN	13.470000	20
3	2008-06-01 01:00:00	NaN	0.36	NaN	NaN	NaN	61.029999	66.559998	NaN	23.110001	10
4	2008-06-01 01:00:00	1.68	0.80	1.70	3.01	0.30	105.199997	214.899994	1.61	12.120000	37
...	...	...	...	...	...	...	...	...	...	...	
226387	2008-11-01 00:00:00	0.48	0.30	0.57	1.00	0.31	13.050000	14.160000	0.91	57.400002	5
226388	2008-11-01 00:00:00	NaN	0.30	NaN	NaN	NaN	41.880001	48.500000	NaN	35.830002	15
226389	2008-11-01 00:00:00	0.25	NaN	0.56	NaN	0.11	83.610001	102.199997	NaN	14.130000	17
226390	2008-11-01 00:00:00	0.54	NaN	2.70	NaN	0.18	70.639999	81.860001	NaN	NaN	11
226391	2008-11-01 00:00:00	0.75	0.36	1.20	2.75	0.16	58.240002	74.239998	1.64	31.910000	12

226392 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: 25631 entries, 4 to 226391
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     25631 non-null  object 
 1   BEN      25631 non-null  float64
 2   CO       25631 non-null  float64
 3   EBE      25631 non-null  float64
 4   MXY      25631 non-null  float64
 5   NMHC     25631 non-null  float64
 6   NO_2     25631 non-null  float64
 7   NOx      25631 non-null  float64
 8   OXY      25631 non-null  float64
 9   O_3      25631 non-null  float64
 10  PM10     25631 non-null  float64
 11  PM25     25631 non-null  float64
 12  PXY      25631 non-null  float64
 13  SO_2     25631 non-null  float64
 14  TCH      25631 non-null  float64
 15  TOL      25631 non-null  float64
 16  station  25631 non-null  int64  
dtypes: float64(15), int64(1), object(1)
memory usage: 3.5+ MB
```

In [7]:

```
data=df[['PXY', 'NOx', 'OXY']]
data
```

Out[7]:

	PXY	NOx	OXY
4	1.43	214.899994	1.61
21	1.00	22.180000	1.00
25	1.22	86.709999	1.31
30	1.81	143.399994	2.03
47	0.38	27.389999	1.00
...	...	...	...
226362	1.84	25.020000	1.00
226366	1.98	106.199997	1.70
226371	2.10	158.399994	2.38
226387	1.86	14.160000	0.91
226391	1.98	74.239998	1.64

25631 rows × 3 columns

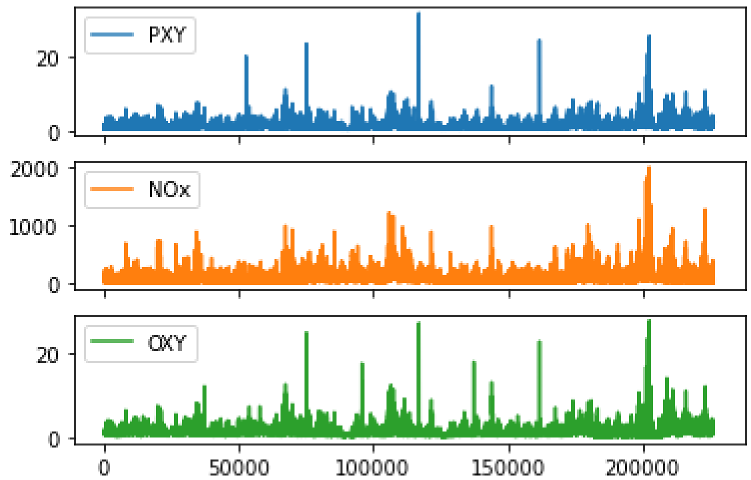
Line chart

In [8]:

```
data.plot.line(subplots=True)
```

Out[8]:

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



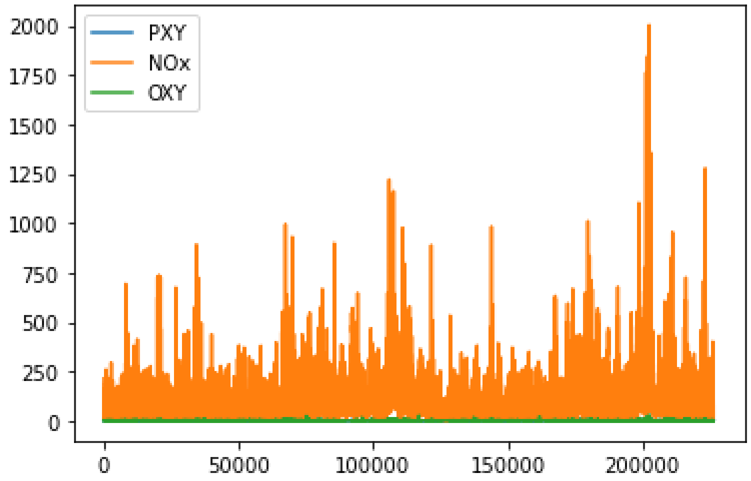
Line chart

In [9]:

```
data.plot.line()
```

Out[9]:

<AxesSubplot:>



Bar chart

In [10]:

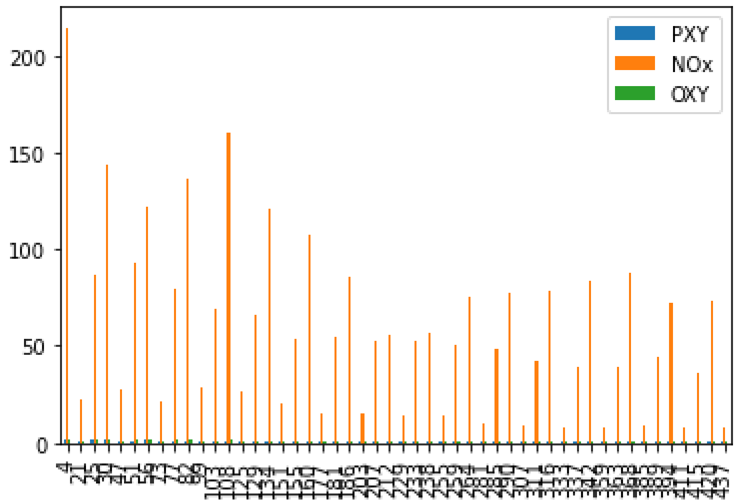
```
b=data[0:50]
```

In [11]:

```
b.plot.bar()
```

Out[11]:

<AxesSubplot:>



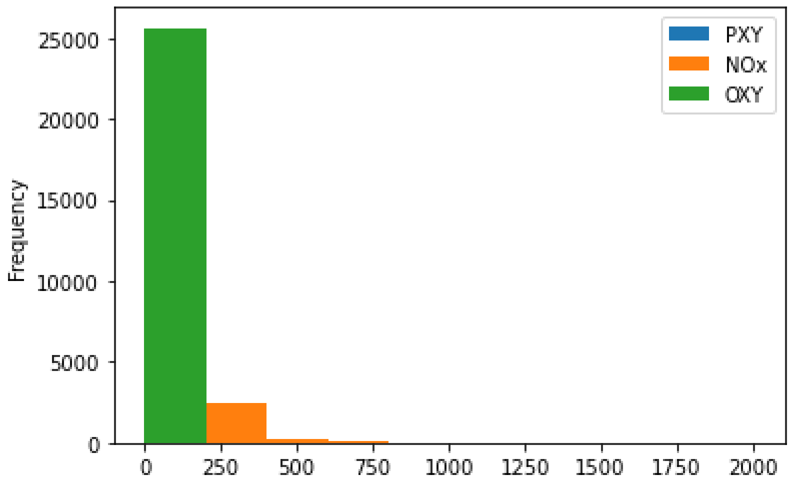
Histogram

In [12]:

```
data.plot.hist()
```

Out[12]:

```
<AxesSubplot:ylabel='Frequency'>
```



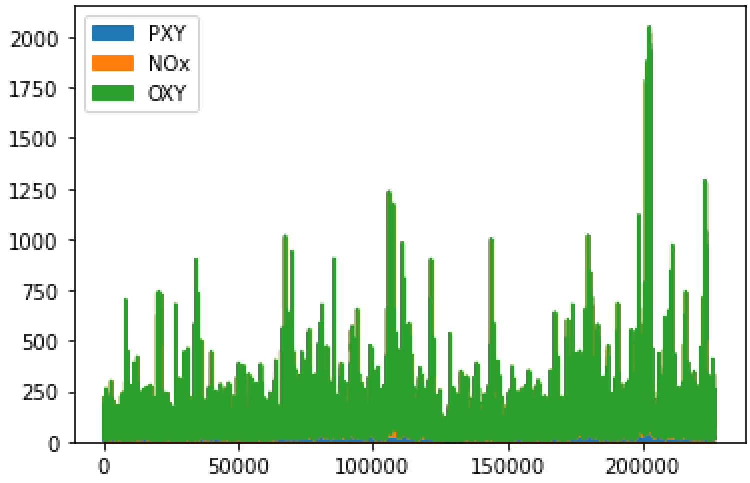
Area chart

In [13]:

```
data.plot.area()
```

Out[13]:

```
<AxesSubplot:>
```



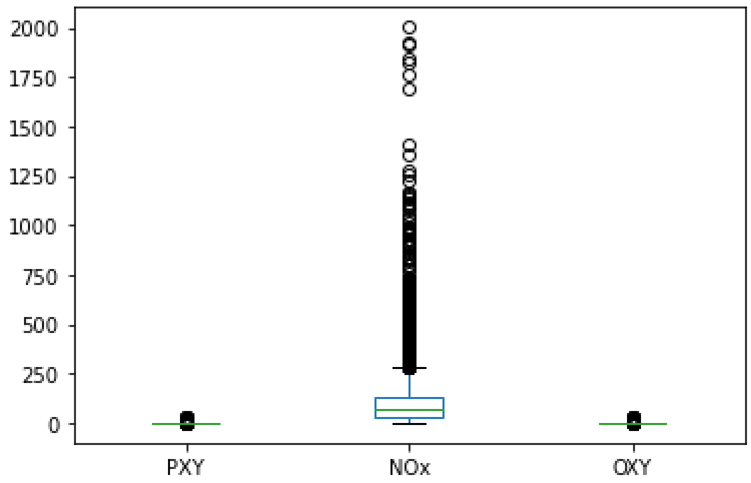
Box chart

In [14]:

```
data.plot.box()
```

Out[14]:

```
<AxesSubplot:>
```



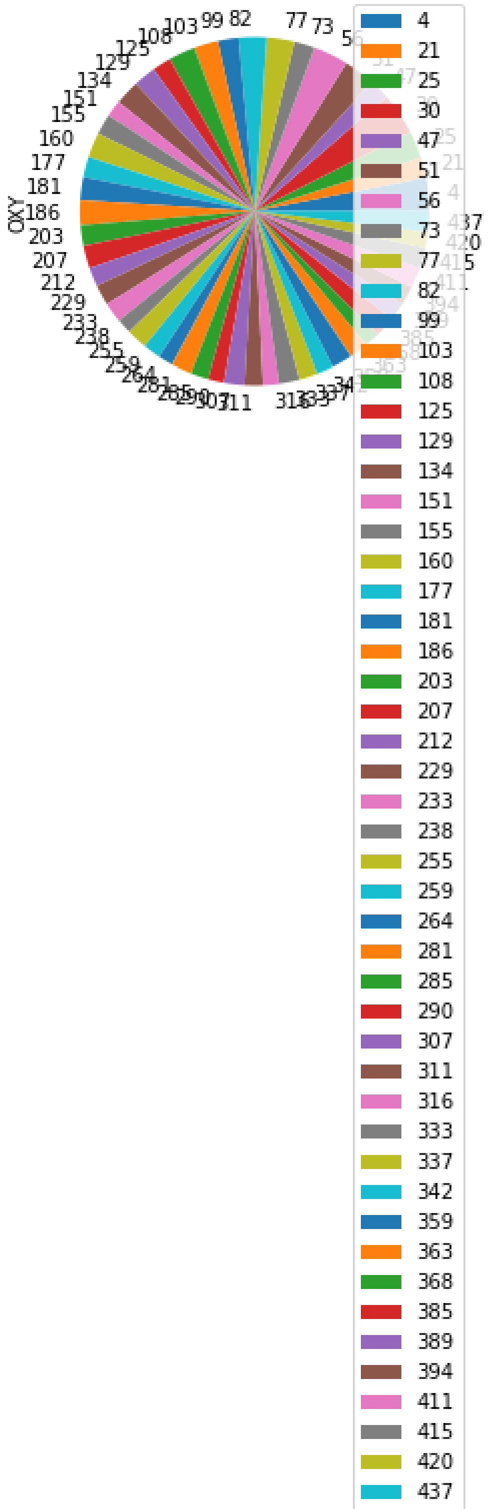
Pie chart

In [20]:

```
b.plot.pie(y='OXY' )
```

Out[20]:

```
<AxesSubplot:ylabel='OXY'>
```



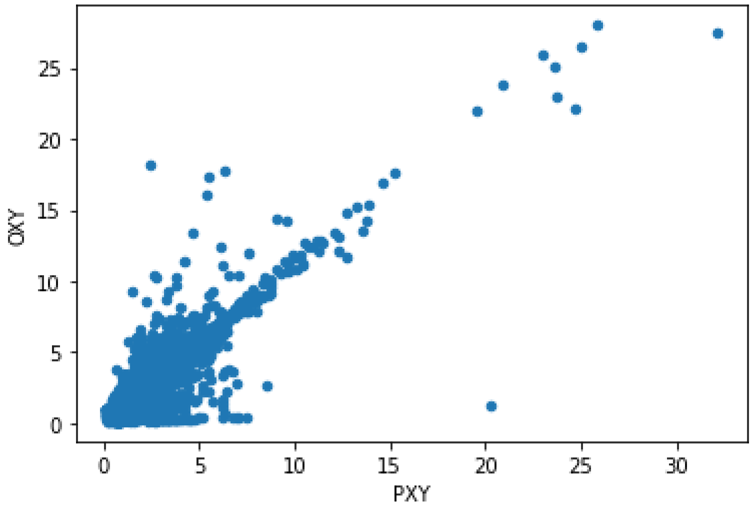
Scatter chart

In [21]:

```
data.plot.scatter(x='PXY' ,y='OXY')
```

Out[21]:

<AxesSubplot:xlabel='PXY', ylabel='OXY'>



In [22]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 25631 entries, 4 to 226391
Data columns (total 17 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     25631 non-null  object 
 1   BEN      25631 non-null  float64
 2   CO       25631 non-null  float64
 3   EBE      25631 non-null  float64
 4   MXY      25631 non-null  float64
 5   NMHC     25631 non-null  float64
 6   NO_2     25631 non-null  float64
 7   NOx      25631 non-null  float64
 8   OXY      25631 non-null  float64
 9   O_3      25631 non-null  float64
 10  PM10     25631 non-null  float64
 11  PM25     25631 non-null  float64
 12  PXY      25631 non-null  float64
 13  SO_2     25631 non-null  float64
 14  TCH      25631 non-null  float64
 15  TOL      25631 non-null  float64
 16  station  25631 non-null  int64  
dtypes: float64(15), int64(1), object(1)
memory usage: 3.5+ MB
```

In [23]:

```
df.describe()
```

Out[23]:

	BEN	CO	EBE	MXY	NMHC	NO_2
count	25631.000000	25631.000000	25631.000000	25631.000000	25631.000000	25631.000000
mean	1.090541	0.440632	1.352355	2.446045	0.213323	54.225261
std	1.146461	0.317853	1.118191	2.390023	0.123409	38.164647
min	0.100000	0.060000	0.170000	0.240000	0.000000	0.240000
25%	0.430000	0.260000	0.740000	1.000000	0.130000	25.719999
50%	0.750000	0.350000	1.000000	1.620000	0.190000	48.000000
75%	1.320000	0.510000	1.580000	3.105000	0.270000	74.924999
max	27.230000	7.030000	26.740000	55.889999	1.760000	554.900024

In [24]:

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

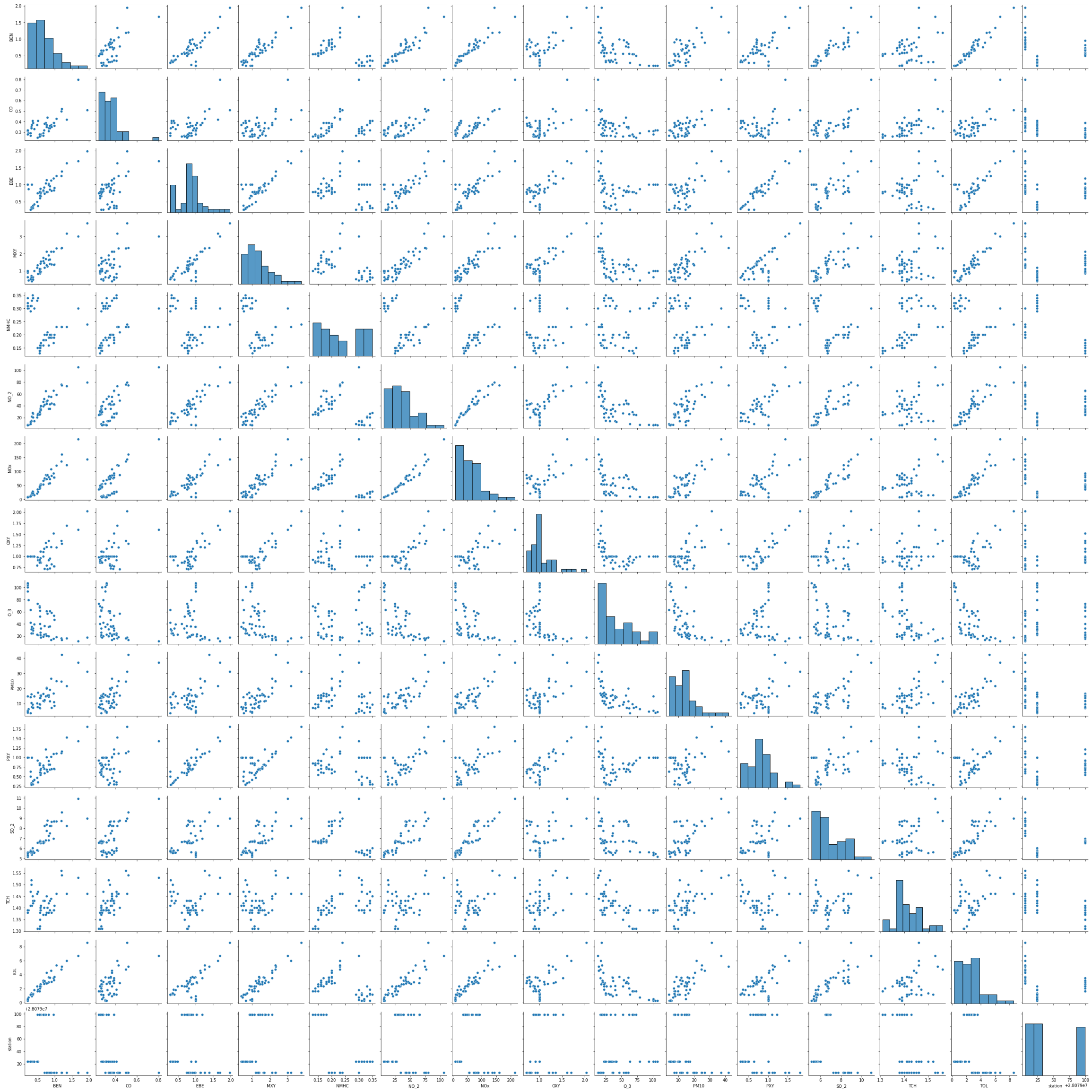
EDA AND VISUALIZATION

In [25]:

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

Out[25]:

<seaborn.axisgrid.PairGrid at 0x1fb06094610>



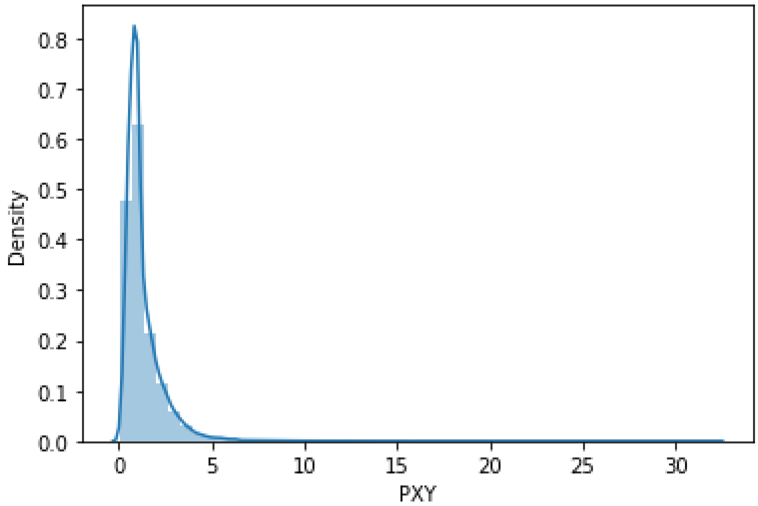
In [26]:

```
sns.distplot(df1['PXY'])
```

```
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 fun
ction for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[26]:

```
<AxesSubplot:xlabel='PXY', ylabel='Density'>
```

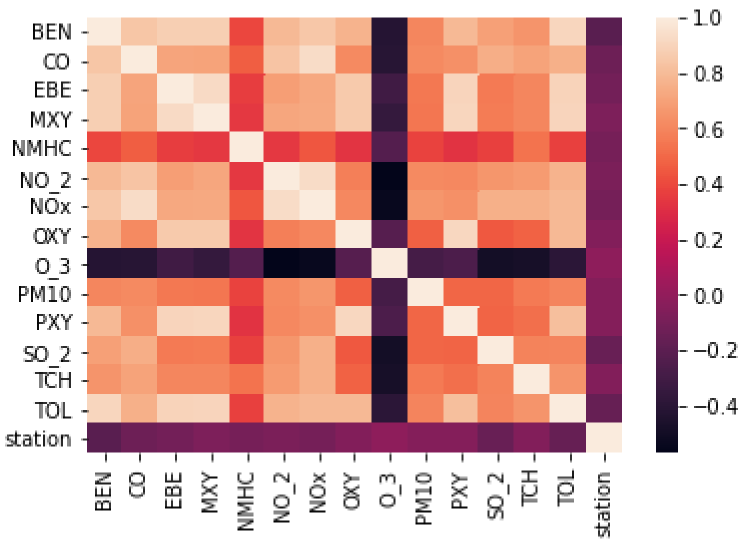


In [27]:

```
sns.heatmap(df1.corr())
```

Out[27]:

```
<AxesSubplot:>
```



TO TRAIN THE MODEL AND MODEL BULDING

In [28]:

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

In [29]:

```
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 [30]:

```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
lr.fit(x_train,y_train)
```

Out[30]:

```
LinearRegression()
```

In [31]:

```
lr.intercept_
```

Out[31]:

```
28079032.423166875
```

In [32]:

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

Out[32]:

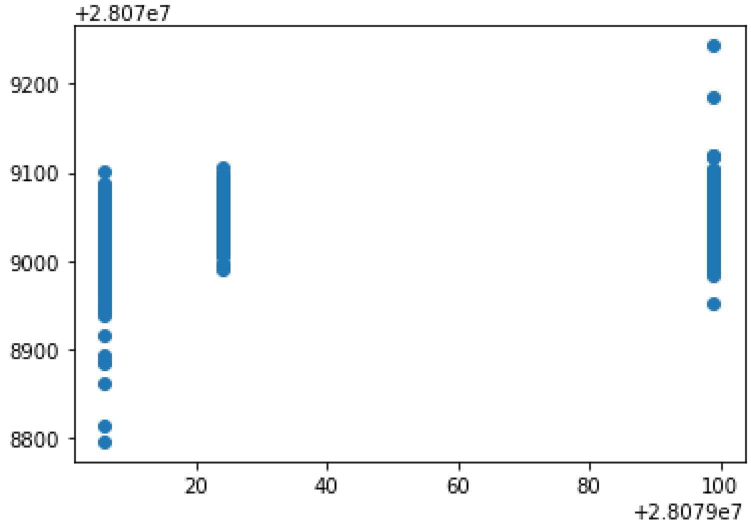
	Co-efficient
BEN	-25.858451
CO	-1.208438
EBE	-0.894502
MXY	8.047465
NMHC	-23.327643
NO_2	-0.018601
NOx	0.111604
OXY	3.930719
O_3	-0.137336
PM10	0.123197
PXY	1.299978
SO_2	-0.566373
TCH	18.890318
TOL	-1.868499

In [33]:

```
prediction =lr.predict(x_test)
plt.scatter(y_test,prediction)
```

Out[33]:

```
<matplotlib.collections.PathCollection at 0x1fb1718a730>
```



ACCURACY

In [34]:

```
lr.score(x_test,y_test)
```

Out[34]:

```
0.148686311607905466
```

In [69]:

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

Out[69]:

```
0.14103322180464228
```

Ridge and Lasso

In [36]:

```
from sklearn.linear_model import Ridge,Lasso
```

In [37]:

```
rr=Ridge(alpha=10)
rr.fit(x_train,y_train)
```

Out[37]:

```
Ridge(alpha=10)
```

Accuracy(Ridge)

In [38]:

```
rr.score(x_test,y_test)
```

Out[38]:

```
0.148523011699787
```

In [39]:

```
rr.score(x_train,y_train)
```

Out[39]:

```
0.14101087723979566
```

In [40]:

```
la=Lasso(alpha=10)
la.fit(x_train,y_train)
```

Out[40]:

```
Lasso(alpha=10)
```

In [41]:

```
la.score(x_test,y_test)
```

Out[41]:

```
0.04308170961539315
```

Accuracy(Lasso)

In [42]:

```
la.score(x_train,y_train)
```

Out[42]:

```
0.03996585184917556
```

Accuracy(Elastic Net)

In [43]:

```
from sklearn.linear_model import ElasticNet
en=ElasticNet()
en.fit(x_train,y_train)
```

Out[43]:

```
ElasticNet()
```

In [44]:

```
en.coef_
```

Out[44]:

```
array([-4.66290453, -0.        ,  0.        ,  3.32013269, -0.        ,
        0.06823102,  0.01644622,  1.42817297, -0.15557479,  0.12350788,
        1.50587257, -0.92489628,  0.        , -2.42759071])
```

In [45]:

```
en.intercept_
```

Out[45]:

```
28079057.10459771
```

In [46]:

```
prediction=en.predict(x_test)
```

In [47]:

```
en.score(x_test,y_test)
```

Out[47]:

```
0.09802778444746496
```

Evaluation Metrics

In [48]:

```
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)))
```

```
35.86636811176167
1497.9868610608958
38.70383522418542
```

Logistic Regression

In [49]:

```
from sklearn.linear_model import LogisticRegression
```

In [50]:

```
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 [51]:

```
feature_matrix.shape
```

Out[51]:

```
(25631, 14)
```

In [52]:

```
target_vector.shape
```

Out[52]:

```
(25631,)
```

In [53]:

```
from sklearn.preprocessing import StandardScaler
```

In [54]:

```
fs=StandardScaler().fit_transform(feature_matrix)
```

In [55]:

```
logr=LogisticRegression(max_iter=10000)
logr.fit(fs,target_vector)
```

Out[55]:

```
LogisticRegression(max_iter=10000)
```

In [56]:

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

In [57]:

```
prediction=logr.predict(observation)
print(prediction)
```

```
[28079099]
```

In [58]:

```
logr.classes_
```

Out[58]:

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

In [59]:

```
logr.score(fs,target_vector)
```

Out[59]:

```
0.794194530061254
```

In [60]:

```
logr.predict_proba(observation)[0][0]
```

Out[60]:

```
8.321803242555043e-09
```

In [61]:

```
logr.predict_proba(observation)
```

Out[61]:

```
array([[8.32180324e-09, 1.19114634e-13, 9.99999992e-01]])
```

Random Forest

In [62]:

```
from sklearn.ensemble import RandomForestClassifier
```

In [63]:

```
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[63]:

```
RandomForestClassifier()
```

In [64]:

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

In [65]:

```
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[65]:

```
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 [66]:

```
grid_search.best_score_
```

Out[66]:

```
0.852906715271194
```

In [67]:

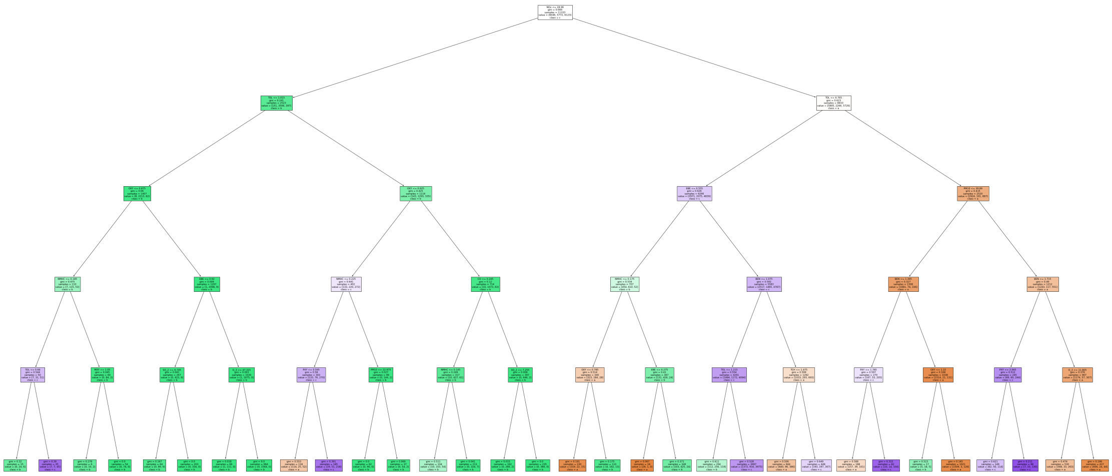
```
rfc_best=grid_search.best_estimator_
```

In [68]:

```
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'],f
```

```
[568, 33, 263]\nclass = a'),
 Text(4394.25, 181.19999999999982, 'gini = 0.186\nsamples = 437\nvalue =
[606, 24, 44]\nclass = a')]
```



Conclusion

Accuracy

Linear Regression:0.14103322180464228

Ridge Regression:0.14101087723979566

Lasso Regression:0.03996585184917556

ElasticNet Regression:0.09802778444746496

Logistic Regression:0.794194530061254

Random Forest:0.852906715271194

Random Forest is suitable for this dataset