import libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Import dataset

In [2]:
```
data=pd.read_csv(r"C:\Users\user\Desktop\vicky\C10_air\csvs_per_year\csvs_per_year\madrid_2
```

In [3]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   date     1500 non-null   object 
 1   BEN      371 non-null    float64
 2   CO       628 non-null    float64
 3   EBE      371 non-null    float64
 4   NMHC     313 non-null    float64
 5   NO       1500 non-null   float64
 6   NO_2     1500 non-null   float64
 7   O_3      876 non-null    float64
 8   PM10     749 non-null    float64
 9   PM25     375 non-null    float64
 10  SO_2     628 non-null    float64
 11  TCH      313 non-null    float64
 12  TOL      371 non-null    float64
 13  station  1500 non-null   int64  
dtypes: float64(12), int64(1), object(1)
memory usage: 164.2+ KB
```

In [4]:
```
data.head()
```

Out[4]:

	date	BEN	CO	EBE	NMHC	NO	NO_2	O_3	PM10	PM25	SO_2	TCH	TOL	station
0	2011-11-01 01:00:00	NaN	1.0	NaN	NaN	154.0	84.0	NaN	NaN	NaN	6.0	NaN	NaN	28079004
1	2011-11-01 01:00:00	2.5	0.4	3.5	0.26	68.0	92.0	3.0	40.0	24.0	9.0	1.54	8.7	28079008
2	2011-11-01 01:00:00	2.9	NaN	3.8	NaN	96.0	99.0	NaN	NaN	NaN	NaN	NaN	7.2	28079011
3	2011-11-01 01:00:00	NaN	0.6	NaN	NaN	60.0	83.0	2.0	NaN	NaN	NaN	NaN	NaN	28079016
4	2011-11-01 01:00:00	NaN	NaN	NaN	NaN	44.0	62.0	3.0	NaN	NaN	3.0	NaN	NaN	28079017

In [5]:
```
data.shape
```

Out[5]: (1500, 14)

In [6]:
```
data.index
```

Out[6]:
```
RangeIndex(start=0, stop=1500, step=1)
```

In [7]:
```
data.columns
```

Out[7]:
```
Index(['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM25',
       'SO_2', 'TCH', 'TOL', 'station'],
      dtype='object')
```

In [8]:
```
data.isna()
```

Out[8]:

	date	BEN	CO	EBE	NMHC	NO	NO_2	O_3	PM10	PM25	SO_2	TCH	TOL	station
0	False	True	False	True	True	False	False	True	True	True	False	True	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	True	False	True	False	False	True	True	True	True	True	False	False
3	False	True	False	True	True	False	False	False	True	True	True	True	True	False
4	False	True	True	True	True	False	False	False	True	True	False	True	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1495	False	True	True	True	False	False	False	False	True	True	True	False	True	False
1496	False	True	False	True	True	False	False	False	True	True	False	True	True	False
1497	False	True	False	True	True	False	False	True	False	True	False	True	True	False
1498	False	False	True	False	True	False	False	True	False	False	False	True	False	False
1499	False	True	False	True	True	False	False	False	True	True	True	True	True	False

1500 rows × 14 columns

In [9]:
```
data.fillna(value=0)
```

Out[9]:

	date	BEN	CO	EBE	NMHC	NO	NO_2	O_3	PM10	PM25	SO_2	TCH	TOL	station
0	2011-11-01 01:00:00	0.0	1.0	0.0	0.00	154.0	84.0	0.0	0.0	0.0	6.0	0.00	0.0	28079004
1	2011-11-01 01:00:00	2.5	0.4	3.5	0.26	68.0	92.0	3.0	40.0	24.0	9.0	1.54	8.7	28079008
2	2011-11-01 01:00:00	2.9	0.0	3.8	0.00	96.0	99.0	0.0	0.0	0.0	0.0	0.00	7.2	28079011
3	2011-11-01 01:00:00	0.0	0.6	0.0	0.00	60.0	83.0	2.0	0.0	0.0	0.0	0.00	0.0	28079016
4	2011-11-01 01:00:00	0.0	0.0	0.0	0.00	44.0	62.0	3.0	0.0	0.0	3.0	0.00	0.0	28079017
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1495	2011-11-03 15:00:00	0.0	0.0	0.0	0.14	5.0	15.0	65.0	0.0	0.0	0.0	1.23	0.0	28079027
1496	2011-11-03 15:00:00	0.0	0.2	0.0	0.00	8.0	21.0	57.0	0.0	0.0	7.0	0.00	0.0	28079035
1497	2011-11-03 15:00:00	0.0	0.2	0.0	0.00	5.0	18.0	0.0	6.0	0.0	3.0	0.00	0.0	28079036
1498	2011-11-03 15:00:00	0.2	0.0	0.3	0.00	11.0	20.0	0.0	9.0	2.0	3.0	0.00	1.0	28079038
1499	2011-11-03 15:00:00	0.0	0.2	0.0	0.00	12.0	23.0	44.0	0.0	0.0	0.0	0.00	0.0	28079039

1500 rows × 14 columns

In [10]:
```
data.isna
```

Out[10]:
```
<bound method DataFrame.isna of                      date  BEN   CO  EBE  NMHC     NO  NO_
2   O_3  PM10  PM25  \
0     2011-11-01 01:00:00  NaN  1.0  NaN   NaN  154.0  84.0   NaN   NaN   NaN
1     2011-11-01 01:00:00  2.5  0.4  3.5  0.26   68.0  92.0   3.0  40.0  24.0
2     2011-11-01 01:00:00  2.9  NaN  3.8   NaN   96.0  99.0   NaN   NaN   NaN
3     2011-11-01 01:00:00  NaN  0.6  NaN   NaN   60.0  83.0   2.0   NaN   NaN
4     2011-11-01 01:00:00  NaN  NaN  NaN   NaN   44.0  62.0   3.0   NaN   NaN
...                   ...  ...  ...  ...   ...    ...   ...   ...   ...   ...
1495  2011-11-03 15:00:00  NaN  NaN  NaN  0.14    5.0  15.0  65.0   NaN   NaN
1496  2011-11-03 15:00:00  NaN  0.2  NaN   NaN    8.0  21.0  57.0   NaN   NaN
1497  2011-11-03 15:00:00  NaN  0.2  NaN   NaN    5.0  18.0   NaN   6.0   NaN
1498  2011-11-03 15:00:00  0.2  NaN  0.3   NaN   11.0  20.0   NaN   9.0   2.0
1499  2011-11-03 15:00:00  NaN  0.2  NaN   NaN   12.0  23.0  44.0   NaN   NaN

      SO_2   TCH  TOL   station
0      6.0   NaN  NaN  28079004
1      9.0  1.54  8.7  28079008
2      NaN   NaN  7.2  28079011
3      NaN   NaN  NaN  28079016
4      3.0   NaN  NaN  28079017
...    ...   ...  ...       ...
1495   NaN  1.23  NaN  28079027
1496   7.0   NaN  NaN  28079035
1497   3.0   NaN  NaN  28079036
1498   3.0   NaN  1.0  28079038
1499   NaN   NaN  NaN  28079039

[1500 rows x 14 columns]>
```

Plotting using various method

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

Out[11]:
```
<AxesSubplot:>
```

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

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

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

Out[13]:
```
<AxesSubplot:>
```

In [14]:
```
data.plot.hist()
```

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

In [15]:
```
data.plot.pie(y="BEN")
```

In [16]:
```
data.plot.scatter(x="NO_2",y='O_3')
```

Out[16]:
```
<AxesSubplot:xlabel='NO_2', ylabel='O_3'>
```

seaborn Visualize

In [17]:
```
sns.pairplot(data)
```

Out[17]:
```
<seaborn.axisgrid.PairGrid at 0x26f16a624c0>
```

In [21]:
```
sns.distplot(data['BEN'])
```

```
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 y
our code to use either `displot` (a figure-level function with similar flexibility) or `hi
stplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[21]:
```
<AxesSubplot:xlabel='BEN', ylabel='Density'>
```

In [22]:
```
sns.heatmap(data.corr())
```

Out[22]:
```
<AxesSubplot:>
```

In [24]:
```
data1=data[['BEN', 'CO', 'EBE', 'NMHC', 'NO_2','O_3',
       'PM10', 'SO_2']]
```

In [25]:
```
data2=data1.fillna(value=1)
```

In [29]:
```
x=data2[['CO','CO','O_3']]
y=data['station']
```

Linear Regression

In [30]:
```
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)
```

In [31]:
```
from sklearn.linear_model import LinearRegression
lr=LinearRegression()
lr.fit(x_train,y_train)
```

Out[31]:
```
LinearRegression()
```

In [32]:
```
print(lr.intercept_)
```

```
28079021.45566437
```

In [33]:
```
coeff=pd.DataFrame(lr.coef_,x.columns,columns=['PM10'])
coeff
```

Out[33]:

	PM10
CO	10.358583
CO	10.358583
O_3	0.079401

In [34]:
```
prediction1=lr.predict(x_train)
plt.scatter(y_train,prediction1)
```

Out[34]:
```
<matplotlib.collections.PathCollection at 0x26f2b9f40a0>
```

In [35]:
```
lr.score(x_test,y_test)
```

Out[35]:
```
0.13212363512992098
```

In [36]:
```
prediction1=lr.predict(x_test)
```

Ridge

In [37]:
```
from sklearn.linear_model import Ridge,Lasso
rr=Ridge(alpha=10)
rr.fit(x_train,y_train)
```

Out[37]:
```
Ridge(alpha=10)
```

In [38]:
```
rr.score(x_test,y_test)
```

Out[38]:
```
0.13339869456190434
```

In [39]:
```
prediction2=rr.predict(x_test)
```

Lasso

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]:
```
0.0008216546642045852
```

In [42]:
```
prediction3=la.score(x_test,y_test)
```

Elastic Net

In [43]:
```
from sklearn.linear_model import ElasticNet
en=ElasticNet()
en.fit(x_train,y_train)
```

Out[43]:
```
ElasticNet()
```

In [44]:
```
print(en.coef_)
```

```
[2.77716268 2.77716106 0.05264608]
```

In [45]:
```
print(en.intercept_)
```

```
28079032.766343605
```

In [46]:
```
prediction4=en.predict(x_test)
```

In [47]:
```
en.score(x_test,y_test)
```

Out[47]:
```
0.06982870222647541
```

Evalution Metrics for linear

In [48]:
```
from sklearn import metrics
```

In [49]:
```
print("Mean Absolute error:",metrics.mean_absolute_error(y_test,prediction1))
```

```
Mean Absolute error: 13.984375061483847
```

In [50]:
```
print("Mean Absolute square error:",metrics.mean_squared_error(y_test,prediction1))
```

```
Mean Absolute square error: 273.2135826534148
```

Evalution Metrics for Ridge

In [51]:
```
print("Mean Absolute error:",metrics.mean_absolute_error(y_test,prediction2))
```

```
Mean Absolute error: 14.011911228423317
```

In [52]:
```
print("Mean Absolute square error:",metrics.mean_squared_error(y_test,prediction2))
```

```
Mean Absolute square error: 272.81218497788257
```

Evalution for elasticnet

In [53]:
```
print("Mean Absolute error:",metrics.mean_absolute_error(y_test,prediction4))
```

```
Mean Absolute error: 14.712615808538265
```

In [54]:
```
print("Mean Absolute square error:",metrics.mean_squared_error(y_test,prediction4))
```

```
Mean Absolute square error: 292.8244655955402
```

Feature matrix

In [55]:
```
from sklearn.preprocessing import StandardScaler
```

In [56]:
```
from sklearn import utils
```

In [57]:
```
from sklearn.linear_model import LogisticRegression
```

In [58]:
```
df=pd.read_csv(r"C:\Users\user\Desktop\vicky\C10_air\csvs_per_year\csvs_per_year\madrid_201
```

In [59]:
```
df.columns
```

Out[59]:
```
Index(['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM25',
       'SO_2', 'TCH', 'TOL', 'station'],
      dtype='object')
```

In [67]:
```
new_df=df.fillna({'BEN':1,'CO':2,'EBE':4})
new_df
```

Out[67]:

	date	BEN	CO	EBE	NMHC	NO	NO_2	O_3	PM10	PM25	SO_2	TCH	TOL	station
0	2011-11-01 01:00:00	1.0	1.0	4.0	NaN	154.0	84.0	NaN	NaN	NaN	6.0	NaN	NaN	28079004
1	2011-11-01 01:00:00	2.5	0.4	3.5	0.26	68.0	92.0	3.0	40.0	24.0	9.0	1.54	8.7	28079008
2	2011-11-01 01:00:00	2.9	2.0	3.8	NaN	96.0	99.0	NaN	NaN	NaN	NaN	NaN	7.2	28079011
3	2011-11-01 01:00:00	1.0	0.6	4.0	NaN	60.0	83.0	2.0	NaN	NaN	NaN	NaN	NaN	28079016
4	2011-11-01 01:00:00	1.0	2.0	4.0	NaN	44.0	62.0	3.0	NaN	NaN	3.0	NaN	NaN	28079017
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
209923	2011-09-01 00:00:00	1.0	0.2	4.0	NaN	5.0	19.0	44.0	NaN	NaN	NaN	NaN	NaN	28079056
209924	2011-09-01 00:00:00	1.0	0.1	4.0	NaN	6.0	29.0	NaN	11.0	NaN	7.0	NaN	NaN	28079057
209925	2011-09-01 00:00:00	1.0	2.0	4.0	0.23	1.0	21.0	28.0	NaN	NaN	NaN	1.44	NaN	28079058
209926	2011-09-01 00:00:00	1.0	2.0	4.0	NaN	3.0	15.0	48.0	NaN	NaN	NaN	NaN	NaN	28079059
209927	2011-09-01 00:00:00	1.0	2.0	4.0	NaN	4.0	33.0	38.0	13.0	NaN	NaN	NaN	NaN	28079060

209928 rows × 14 columns

In [70]:
```
feature_matrix = new_df[['CO','EBE']]
target_vector = new_df['station']
```

In [71]:
```
feature_matrix.shape
```

Out[71]: (209928, 2)

In [72]:
```
target_vector.shape
```

Out[72]: (209928,)

In [73]:
```
from sklearn.preprocessing import StandardScaler
```

In [74]:
```
fs = StandardScaler().fit_transform(feature_matrix)
```

In [75]:
```
logr=LogisticRegression()
```

In [76]:
```
logr.fit(fs,target_vector)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:763: Converge
nceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html (https://scikit-learn.org/s
table/modules/preprocessing.html)
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression (http
s://scikit-learn.org/stable/modules/linear_model.html#logistic-regression)
  n_iter_i = _check_optimize_result(
```

Out[76]:
```
LogisticRegression()
```

In [77]:
```
observation =[[3,90,5]]
```

In [78]:
```
prediction5 =logr.predict(observation)
print(prediction5)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-78-67807927e3dd> in <module>
----> 1 prediction5 =logr.predict(observation)
      2 print(prediction5)

C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\_base.py in predict(self,
X)
    307             Predicted class label per sample.
    308         """
--> 309         scores = self.decision_function(X)
    310         if len(scores.shape) == 1:
    311             indices = (scores > 0).astype(int)

C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\_base.py in decision_funct
ion(self, X)
    286         n_features = self.coef_.shape[1]
    287         if X.shape[1] != n_features:
--> 288             raise ValueError("X has %d features per sample; expecting %d"
    289                              % (X.shape[1], n_features))
    290

ValueError: X has 3 features per sample; expecting 2
```

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

In [ ]:
```
logr.predict_proba(observation)[0][1]
```

import pickle

In [ ]:
```
import pickle
```

In [ ]:
```
filename1="prediction1"
```

In [ ]:
```
filename2="prediction2"
```

In [ ]:
```
filename3="prediction3"
```

In [ ]:
```
filename4="prediction4"
```

In [ ]:
```
filename5="prediction5"
```

In [ ]:
```
pickle.dump(lr,open(filename1,'wb'))
```

In [ ]:
```
pickle.dump(lr,open(filename2,'wb'))
```

In [ ]:
```
pickle.dump(lr,open(filename3,'wb'))
```

In [ ]:
```
pickle.dump(lr,open(filename4,'wb'))
```

In [ ]:
```
pickle.dump(lr,open(filename5,'wb'))
```

In [ ]:
```

```