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

In [2]:
```
df = pd.read_csv(r"C:\Users\user\Downloads\C10_air\csvs_per_year\csvs_per_year\
df
```

Out[2]:

	date	BEN	CO	EBE	NMHC	NO	NO_2	O_3	PM10	PM25	SO_2	TCH	TOL	
0	2012-09-01 01:00:00	NaN	0.2	NaN	NaN	7.0	18.0	NaN	NaN	NaN	2.0	NaN	NaN	28
1	2012-09-01 01:00:00	0.3	0.3	0.7	NaN	3.0	18.0	55.0	10.0	9.0	1.0	NaN	2.4	28
2	2012-09-01 01:00:00	0.4	NaN	0.7	NaN	2.0	10.0	NaN	NaN	NaN	NaN	NaN	1.5	28
3	2012-09-01 01:00:00	NaN	0.2	NaN	NaN	1.0	6.0	50.0	NaN	NaN	NaN	NaN	NaN	28
4	2012-09-01 01:00:00	NaN	NaN	NaN	NaN	1.0	13.0	54.0	NaN	NaN	3.0	NaN	NaN	28
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
210715	2012-03-01 00:00:00	NaN	0.6	NaN	NaN	37.0	84.0	14.0	NaN	NaN	NaN	NaN	NaN	28
210716	2012-03-01 00:00:00	NaN	0.4	NaN	NaN	5.0	76.0	NaN	17.0	NaN	7.0	NaN	NaN	28
210717	2012-03-01 00:00:00	NaN	NaN	NaN	0.34	3.0	41.0	24.0	NaN	NaN	NaN	1.34	NaN	28
210718	2012-03-01 00:00:00	NaN	NaN	NaN	NaN	2.0	44.0	36.0	NaN	NaN	NaN	NaN	NaN	28
210719	2012-03-01 00:00:00	NaN	NaN	NaN	NaN	2.0	56.0	40.0	18.0	NaN	NaN	NaN	NaN	28

210720 rows × 14 columns

In [3]:
```
df1 = df.fillna(0)
df1
```

Out[3]:

	date	BEN	CO	EBE	NMHC	NO	NO_2	O_3	PM10	PM25	SO_2	TCH	TOL	s
0	2012-09-01 01:00:00	0.0	0.2	0.0	0.00	7.0	18.0	0.0	0.0	0.0	2.0	0.00	0.0	280
1	2012-09-01 01:00:00	0.3	0.3	0.7	0.00	3.0	18.0	55.0	10.0	9.0	1.0	0.00	2.4	280
2	2012-09-01 01:00:00	0.4	0.0	0.7	0.00	2.0	10.0	0.0	0.0	0.0	0.0	0.00	1.5	280
3	2012-09-01 01:00:00	0.0	0.2	0.0	0.00	1.0	6.0	50.0	0.0	0.0	0.0	0.00	0.0	280
4	2012-09-01 01:00:00	0.0	0.0	0.0	0.00	1.0	13.0	54.0	0.0	0.0	3.0	0.00	0.0	280
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
210715	2012-03-01 00:00:00	0.0	0.6	0.0	0.00	37.0	84.0	14.0	0.0	0.0	0.0	0.00	0.0	280
210716	2012-03-01 00:00:00	0.0	0.4	0.0	0.00	5.0	76.0	0.0	17.0	0.0	7.0	0.00	0.0	280
210717	2012-03-01 00:00:00	0.0	0.0	0.0	0.34	3.0	41.0	24.0	0.0	0.0	0.0	1.34	0.0	280
210718	2012-03-01 00:00:00	0.0	0.0	0.0	0.00	2.0	44.0	36.0	0.0	0.0	0.0	0.00	0.0	280
210719	2012-03-01 00:00:00	0.0	0.0	0.0	0.00	2.0	56.0	40.0	18.0	0.0	0.0	0.00	0.0	280

210720 rows × 14 columns

In [4]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210720 entries, 0 to 210719
Data columns (total 14 columns):
 #   Column   Non-Null Count   Dtype  
---  ------   --------------   -----  
 0   date     210720 non-null  object 
 1   BEN      210720 non-null  float64
 2   CO       210720 non-null  float64
 3   EBE      210720 non-null  float64
 4   NMHC     210720 non-null  float64
 5   NO       210720 non-null  float64
 6   NO_2     210720 non-null  float64
 7   O_3      210720 non-null  float64
 8   PM10     210720 non-null  float64
 9   PM25     210720 non-null  float64
 10  SO_2     210720 non-null  float64
 11  TCH      210720 non-null  float64
 12  TOL      210720 non-null  float64
 13  station  210720 non-null  int64  
dtypes: float64(12), int64(1), object(1)
memory usage: 22.5+ MB
```

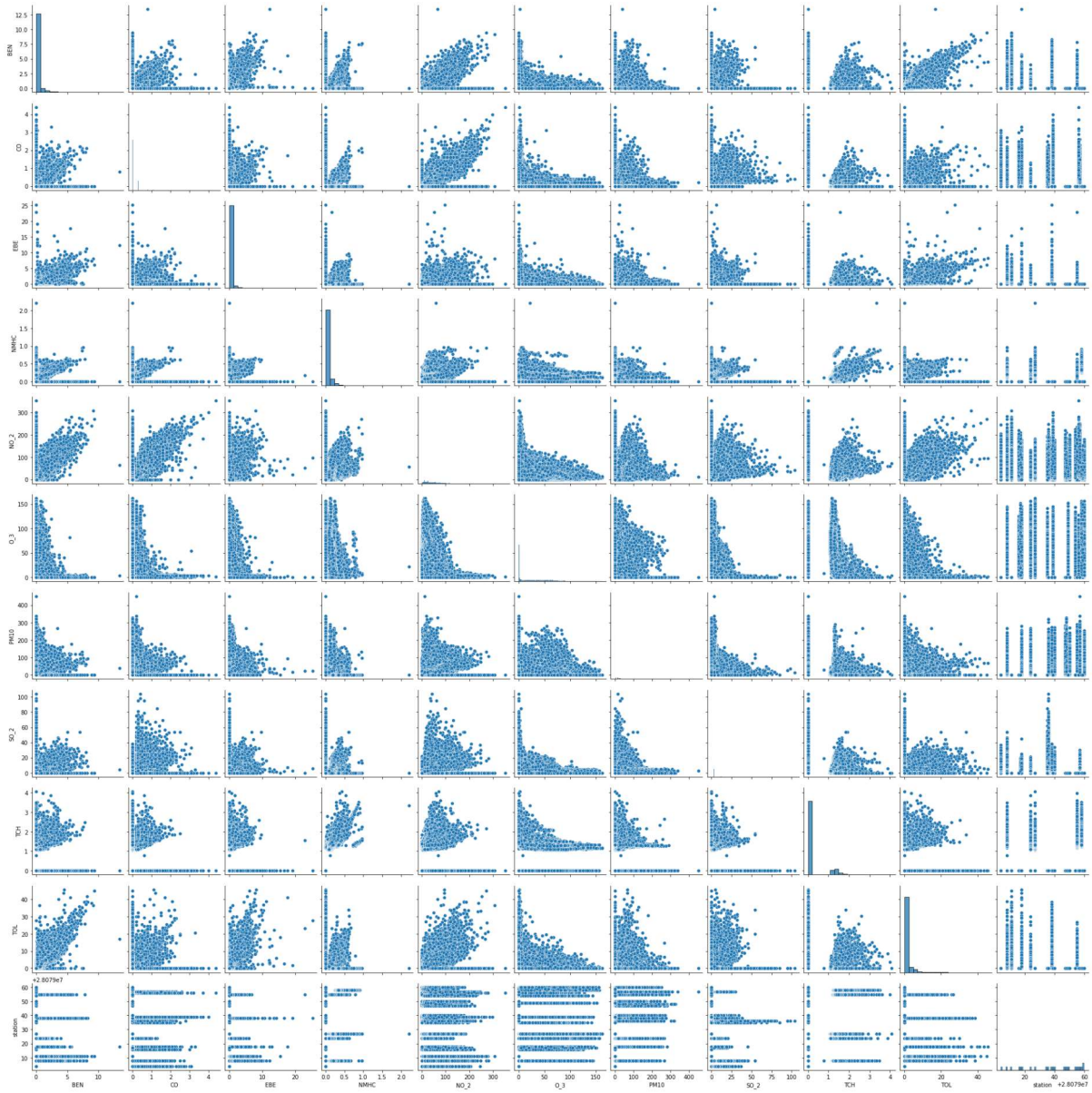
In [5]:
```
df1.columns
```

Out[5]:
```
Index(['date', 'BEN', 'CO', 'EBE', 'NMHC', 'NO', 'NO_2', 'O_3', 'PM10', 'PM2
5',
       'SO_2', 'TCH', 'TOL', 'station'],
      dtype='object')
```

In [6]:
```
df2 = df1[['BEN', 'CO', 'EBE', 'NMHC', 'NO_2', 'O_3',
       'PM10', 'SO_2', 'TCH', 'TOL', 'station']]
```

In [7]:
```
sns.pairplot(df2)
```

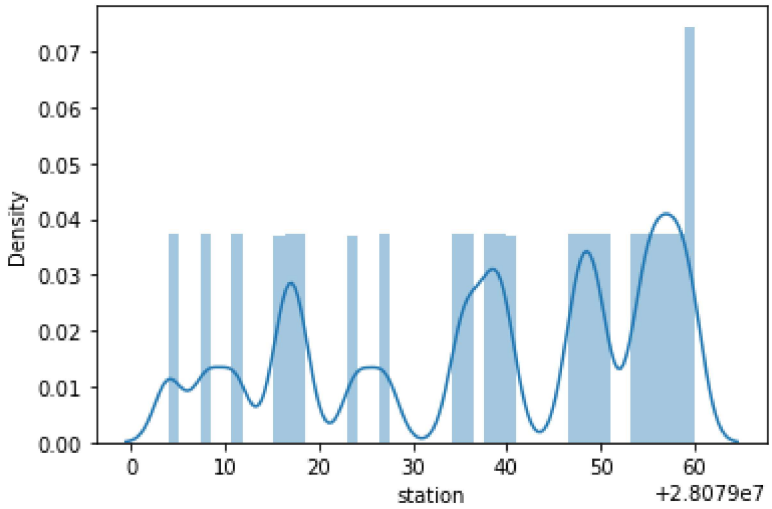
Out[7]:
```
<seaborn.axisgrid.PairGrid at 0x1b39356c7c0>
```



In [8]:
```
sns.distplot(df2['station'])
```

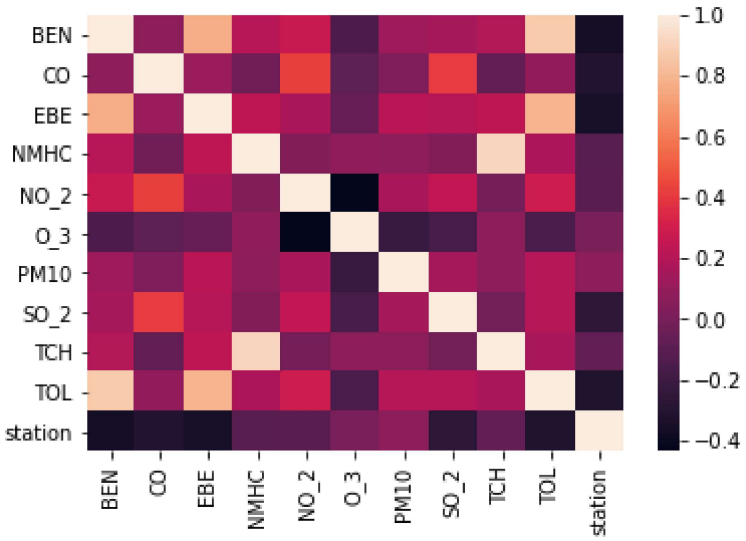
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: Fut
ureWarning: `distplot` is a deprecated function and will be removed in a futu
re version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function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

Out[8]: <AxesSubplot:xlabel='station', ylabel='Density'>



In [9]:
```
sns.heatmap(df2.corr())
```

Out[9]: <AxesSubplot:>



Linear Regression

In [10]:
```
x = df2[['BEN', 'CO', 'EBE', 'NMHC', 'NO_2', 'O_3',
       'PM10', 'SO_2', 'TCH']]
y = df2['TOL']
```

In [11]:
```
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.30)
```

In [12]:
```
from sklearn.linear_model import LinearRegression

lr = LinearRegression()
lr.fit(x_train,y_train)
```

Out[12]:
```
LinearRegression()
```

In [13]:
```
print(lr.intercept_)
```

```
-0.14997184840010114
```

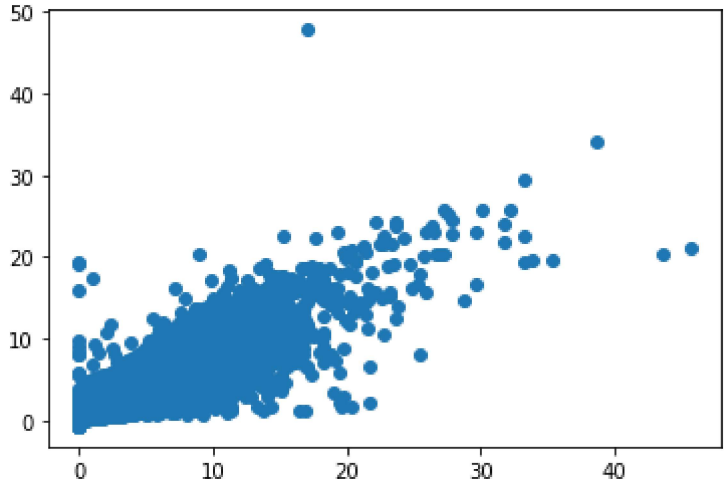
In [14]:
```
coeff = pd.DataFrame(lr.coef_,x.columns,columns=['Co-efficient'])
coeff
```

Out[14]:

	Co-efficient
BEN	2.519797
CO	-0.321403
EBE	1.122704
NMHC	-0.645097
NO_2	0.004734
O_3	-0.000088
PM10	0.005698
SO_2	0.023465
TCH	-0.071214

In [15]:
```
prediction = lr.predict(x_test)
plt.scatter(y_test,prediction)
```

Out[15]:
```
<matplotlib.collections.PathCollection at 0x1b3a31502b0>
```



In [16]:
```
print(lr.score(x_test,y_test))
```

```
0.8096303763473245
```

In [17]:
```
lr.score(x_train,y_train)
```

Out[17]:
```
0.8038041251750647
```

Ridge and Lasso

In [18]:
```
from sklearn.linear_model import Ridge,Lasso
```

In [19]:
```
rr = Ridge(alpha=10)
rr.fit(x_train,y_train)
rr.score(x_train,y_train)
```

Out[19]:
```
0.8038037485901953
```

In [20]:
```
rr.score(x_test,y_test)
```

Out[20]:
```
0.8096250636845652
```

Lasso Regression

In [21]:
```
ls = Lasso(alpha=10)
ls.fit(x_train,y_train)
ls.score(x_train,y_train)
```

Out[21]:
```
0.05748860555150703
```

In [22]:
```
ls.score(x_test,y_test)
```

Out[22]:
```
0.06126149795278113
```

ElacticNET regression

In [23]:
```
from sklearn.linear_model import ElasticNet
es = ElasticNet()
es.fit(x_train,y_train)
```

Out[23]:
```
ElasticNet()
```

In [24]:
```
print(es.coef_)
```

```
[ 5.12759826e-01 -0.00000000e+00  3.81971997e-01  0.00000000e+00
  1.51187513e-02 -3.34048862e-04  1.31388355e-02  2.23265435e-02
  0.00000000e+00]
```

In [25]:
```
print(es.intercept_)
```

```
-0.12831485667759546
```

In [26]:
```
print(es.score(x_test,y_test))
```

```
0.41574702273645137
```

In [27]:
```
print(es.score(x_train,y_train))
```

```
0.40987952955773743
```

LogisticRegression

In [28]:
```
from sklearn.linear_model import LogisticRegression
```

In [29]:
```
feature_matrix = df2.iloc[:,0:15]
target_vector = df2.iloc[:,-1]
```

In [30]:
```
feature_matrix.shape
```

Out[30]:
```
(210720, 11)
```

In [31]:
```
from sklearn.preprocessing import StandardScaler
```

In [32]:
```
fs = StandardScaler().fit_transform(feature_matrix)
```

In [33]:
```
logs = LogisticRegression()
logs.fit(fs,target_vector)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:
763: ConvergenceWarning: 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://sciki
t-learn.org/stable/modules/preprocessing.html)
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regres
sion (https://scikit-learn.org/stable/modules/linear_model.html#logistic-regr
ession)
  n_iter_i = _check_optimize_result(
```

Out[33]:
```
LogisticRegression()
```

In [34]:
```
observation = [[1.4,1.5,1.6,2.7,2.3,3.3,2.3,4.1,2.3,4.2,1.2]]
prediction = logs.predict(observation)
```

In [35]:
```
print(prediction)
```

```
[28079059]
```

In [36]:
```
logs.classes_
```

Out[36]:
```
array([28079004, 28079008, 28079011, 28079016, 28079017, 28079018,
       28079024, 28079027, 28079035, 28079036, 28079038, 28079039,
       28079040, 28079047, 28079048, 28079049, 28079050, 28079054,
       28079055, 28079056, 28079057, 28079058, 28079059, 28079060],
      dtype=int64)
```

In [37]:
```
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test = train_test_split(feature_matrix,target_vector,t
```

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

```
0.04120792204505189
```

In [39]:
```
print(logs.score(x_train,y_train))
```

```
0.04189038941316846
```

Conclusion

linear regression is bestfit model

linear regression is best fit model for dataset madrid_2001. The score of x_train,y_train is 0.8038041251750647 and x_test and y_test score is 0.8096303763473245.

In [ ]: