Assignment-1 Report

a. Abstract:

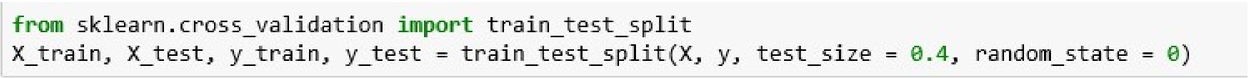
Hazardous air pollutants, also known as toxic air pollutants or air toxics, are those pollutants that are known or suspected to cause cancer or other serious health effects, such as reproductive effects or birth defects, or adverse environmental effects The AQI is an index for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you.The four air pollutants AQI are Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone which are estimated each day to know the increase and decrease of pollution yearly.

b. Introduction:

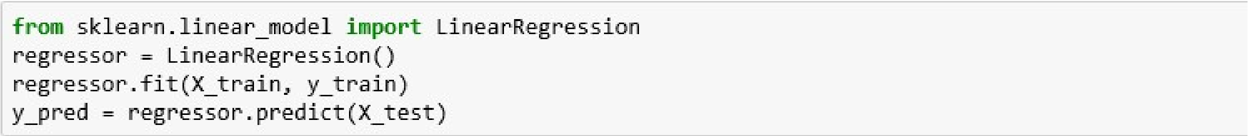
The dataset has been analyzed by taking into consideration four pollutants air quality index (AQI) yearly. There are various graph plotted using matplotlib and seaborn which represents increase in pollution state wise as well increase rate of pollutants yearly. The data has been analyzed using two machine learning algorithm which are used for prediction such as linear regression and polynomial linear regression. They are used to predict increase or decrease in a particular pollutant based on year i.e. only one pollutant has been analyzed at a time.

c. Code :

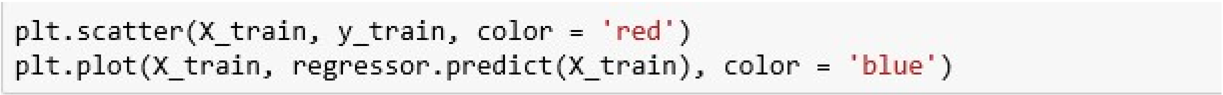
Simple Linear Regression



```
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.4, random_state = 0)
```

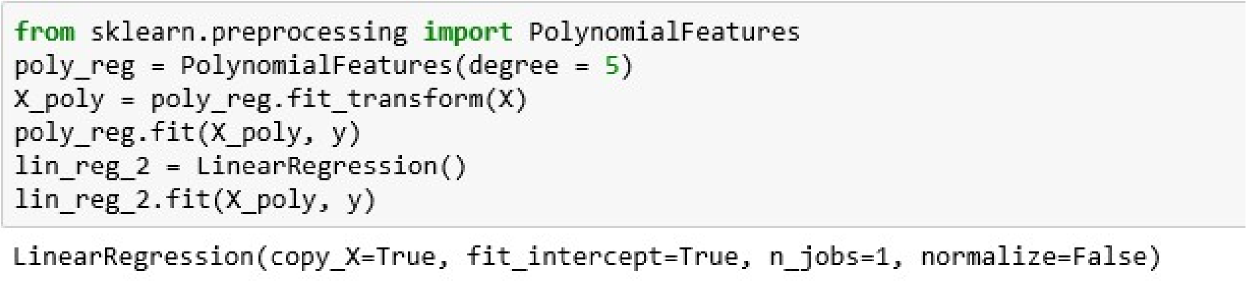


```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)
```



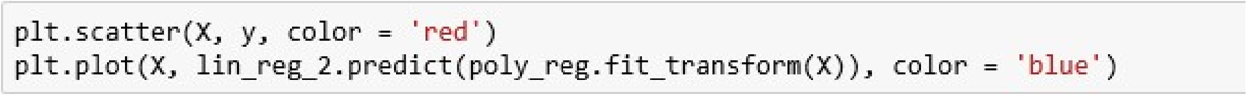
```
plt.scatter(X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
```

Polynomial Regression



```
from sklearn.preprocessing import PolynomialFeatures
poly_reg = PolynomialFeatures(degree = 5)
X_poly = poly_reg.fit_transform(X)
poly_reg.fit(X_poly, y)
lin_reg_2 = LinearRegression()
lin_reg_2.fit(X_poly, y)
```

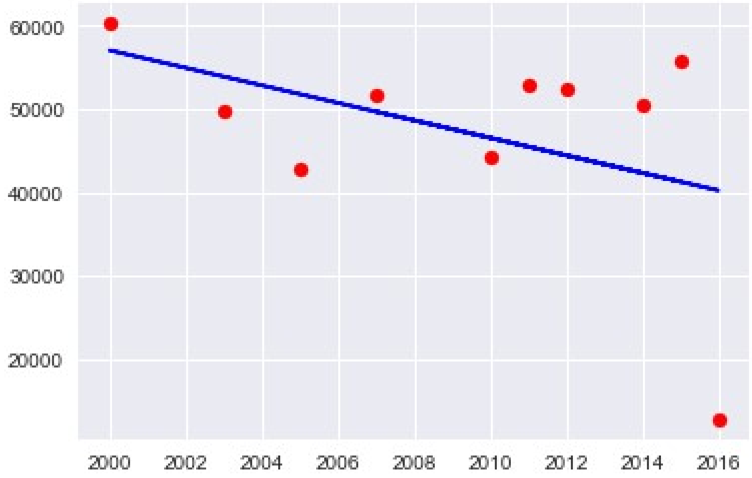
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```



```
plt.scatter(X, y, color = 'red')
plt.plot(X, lin_reg_2.predict(poly_reg.fit_transform(X)), color = 'blue')
```

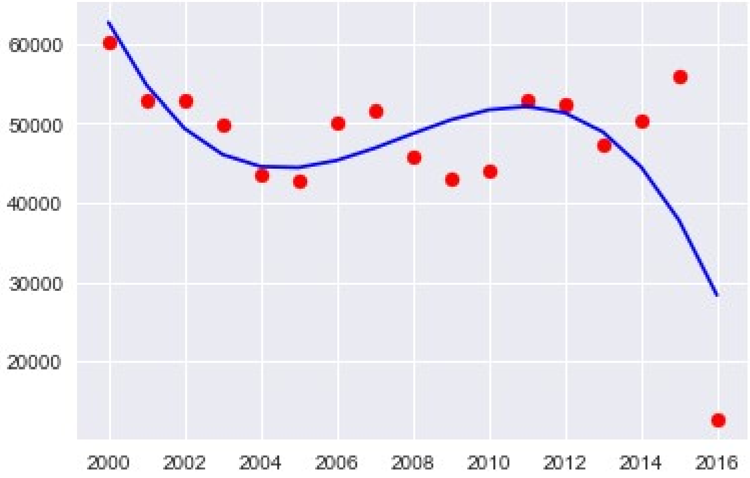
d. Results:

1. Applying simple linear regression to the data set (Year v/s CO Pollutant)



The similarity of the data predicted on test data set using this algorithm by RMSE is 4568.450373152374
While that for train dataset is 11271.521835355828.

2. Applying polynomial regression on the data set (Year v/s CO pollutant)



The similarity of the data predicted using this algorithm by RMSE is 6920.927351849981

e. Discussion:

1. It has been observed from the graph plotted that pollution in the USA has been decreased over time period so as per our prediction algorithm there is the possibility of decrease in pollution in upcoming year.
2. The prediction is only based on air quality index which tells the approximate value of pollutants over the years, further prediction can be done based on state, month or date by taking other pollutants into consideration .

f. References:

[1] https://www.kaggle.com/sogun3/uspollution
[2] https://www.kaggle.com/epa/hazardous-air-pollutants
[3] http://www.lung.org/assets/documents/healthy-air/state-of-the-air/sota-2016-full.pdf