**About Data:**

Used dataset is collected from the UCI Machine Learning repository named *Air Quality Data Set* [(Click Here)](#).

Variables are described given below:

0. Date (DD/MM/YYYY)
1. Time (HH.MM.SS)
2. True hourly averaged concentration CO in mg/m^3 (reference analyzer)
3. PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
4. True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
5. True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
6. PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
7. True hourly averaged NOx concentration in ppb (reference analyzer)
8. PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
9. True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
10. PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
11. PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
12. Temperature in Â°C
13. Relative Humidity (%)
14. AH Absolute Humidity

Primarily data cleaning part is done through remove the missing value and remove the date and time column. Then variables are normalized before the analysis. Also, data has portioned into train data (90%) and test data (10%).

```
# Spliting dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.1, shuffle = False)
print(f'X_train: {X_train.shape}\ny_train: {y_train.shape}')
print(f'X_test: {X_test.shape}\ny_test: {y_test.shape}')
```

```
X_train: (8421, 13)
y_train: (8421,)
X_test: (936, 13)
y_test: (936,)
```

Response variable is temperature (T) and others variables are explanatory variable. The result of OLS model summary is given below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.995
Model:                            OLS   Adj. R-squared:                  0.995
Method:                 Least Squares   F-statistic:                 1.473e+05
Date:                Sat, 19 Mar 2022   Prob (F-statistic):               0.00
Time:                        19:34:18   Log-Likelihood:                -21543.
No. Observations:                8421   AIC:                         4.311e+04
Df Residuals:                    8408   BIC:                         4.320e+04
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0248      0.035      0.713      0.476      -0.043       0.093
x1             0.1059      0.046      2.324      0.020       0.017       0.195
x2            -0.9070      0.160     -5.684      0.000      -1.220      -0.594
x3            -1.1148      0.039    -28.517      0.000      -1.191      -1.038
x4           -30.9288      1.205    -25.663      0.000     -33.291     -28.566
```

```
x5           1.1309      0.354      3.192      0.001       0.436       1.825
x6           1.7293      0.088     19.629      0.000       1.557       1.902
x7          -1.9554      0.097    -20.182      0.000      -2.145      -1.766
x8          -1.4780      0.076    -19.460      0.000      -1.627      -1.329
x9          11.1949      0.095    118.261      0.000      11.009      11.380
x10         -1.8427      0.117    -15.794      0.000      -2.071      -1.614
x11        -16.2445      0.134   -121.633      0.000     -16.506     -15.983
x12         82.6555      0.960     86.063      0.000      80.773      84.538
==============================================================================
Omnibus:                      503.249   Durbin-Watson:                   0.198
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              906.874
Skew:                           0.453   Prob(JB):                    1.19e-197
Kurtosis:                       4.328   Cond. No.                         115.
==============================================================================
```

**Analytical Result and Test Accuracy:**

```python
# Analytical result
w = np.linalg.inv(X_train.T@X_train)@X_train.T@y_train
print(w)
```

```
[ 2.47787942e-02  1.05853525e-01 -9.06984398e-01 -1.11476435e+00
 -3.09288246e+01  1.13088589e+00  1.72930194e+00 -1.95543557e+00
 -1.47800172e+00  1.11949172e+01 -1.84266982e+00 -1.62444642e+01
  8.26555109e+01]
```

```python
print(f'R_square value: {model.rsquared}')
print(f'Adj R_square value: {model.rsquared_adj}')
print(f'Testing accuracy: {r2_score(model.predict(X_test), y_test)}')
```

```
R_square value: 0.9952669998299652
Adj R_square value: 0.99526024483448
Testing accuracy: 0.89340427585182
```

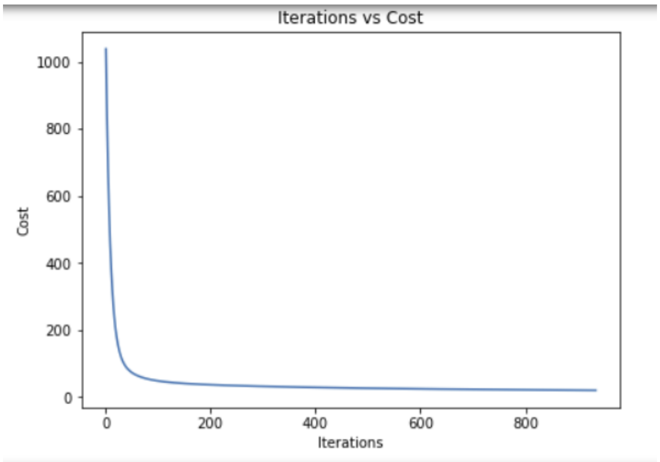
All analytical resulted coefficients are same and OLS model give approximate 90% accuracy.

From the model summary, we can conclude the following:

1) Except intercept all regression coefficients are significant. Also, from the p-value for model is less than 5% significance level, hence the OLS model is statistically significant.
2) Errors are serially uncorrelated from JB test.
3) OLS model give the 90% accuracy, which is quite good.

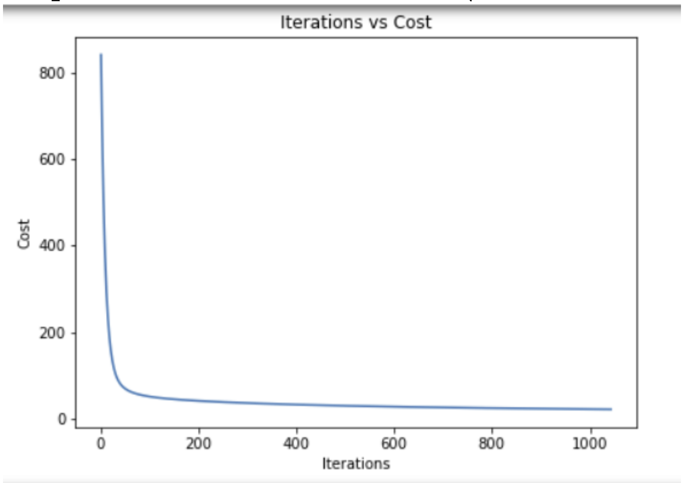
**Minimization of loss function:**

A) Batch Gradient Descent



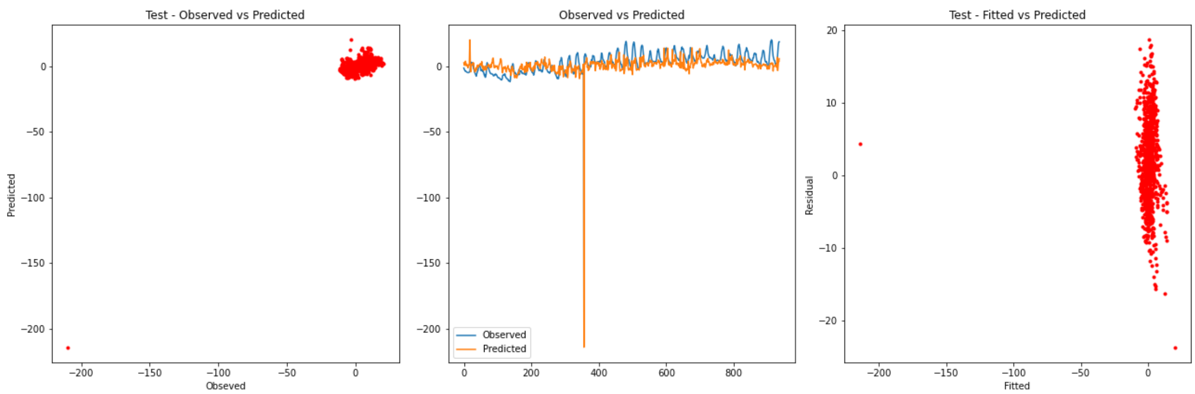

```
Iteration: 930, Cost: 20.790
Iteration: 931, Cost: 20.780
Iteration: 932, Cost: 20.770
Iteration: 933, Cost: 20.760
Iteration: 934, Cost: 20.750
Train Score: 0.9791590792254785
Test Score:  0.4270621506916217
```

B) Sequential Gradient Descent (Widrow-Hoff Algorithm)




```
Iteration: 1038, Cost: 21.225
Iteration: 1039, Cost: 21.215
Iteration: 1040, Cost: 21.205
Iteration: 1041, Cost: 21.195
Iteration: 1042, Cost: 21.185
Iteration: 1043, Cost: 21.175
Train Score: 0.9787485592224774
Test Score:  0.39555766566141337
```

From the above plots we can conclude that sequential batch gradient takes more iteration to minimizes the cost.




Residual and predicted values is correlated, which suggest there are high multicollinearity in the data.