# FoDS Assignment - 2

Vishesh Badjatya
2018A7PS0270H
Deotale Riddhish Anant
2018A7PS0292H
Shubhanjay Varma
2018A7PS0631H

The Assignment consisted of 4 parts:-

1. Data Pre-Processing
2. Implementing Linear Regression through Normal Equation
3. Implementing Linear Regression through Gradient Descent
4. Implementing Linear Regression through Stochastic Gradient Descent

## Data Preprocessing

In this phase, the given .txt file was input into the '.py' files and the '.ipynb' file. Lists containing all the values of age, bmi, no. of children and charges were created. All of the input parameters (parameters except charges) were normalized using the min-max normalization technique. The formula used for normalization was:-

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

After normalization, the dataset was divided. A 70-30 split was introduced with the 70 part being used for training the regression model and the 30 part became the testing dataset. In the 20-models files, this shuffling takes place 20 times for 20 different regression models to be generated.

## Normal Equation

This implementation of linear regression was done by implementing the below formula:-

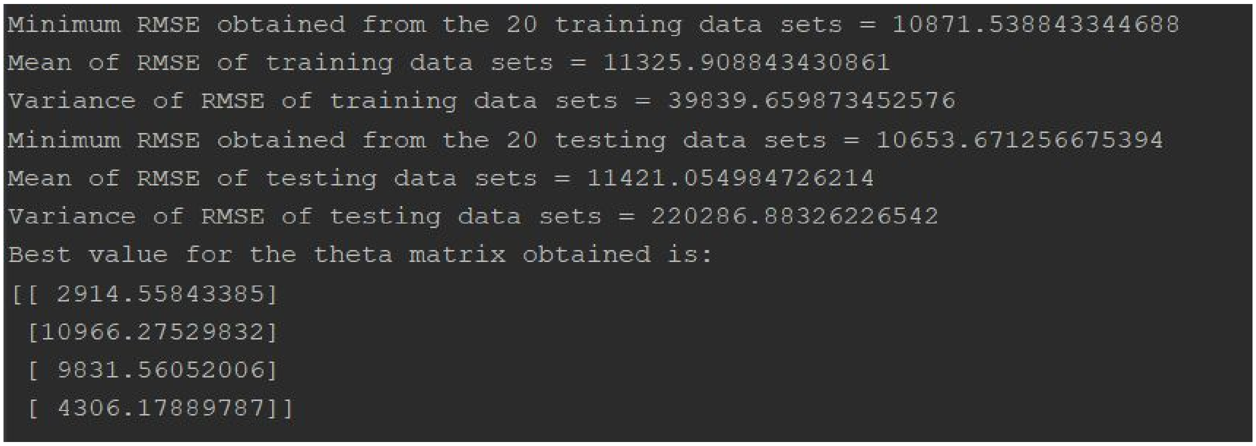
$$\boxed{\theta = \left(X^T X\right)^{-1} . \left(X^T y\right)}$$

In the formula written above, theta represents the parameters which are to be calculated. Or, in simpler terms, theta denotes the values of W0,W1,W2 and W3 in the equation:-

$$Y = W0 + W1X1 + W2X2 + W3X3$$

Where Y is the calculated value of investment. X1 is the age, X2 = bmi and X3 is the no. of children.

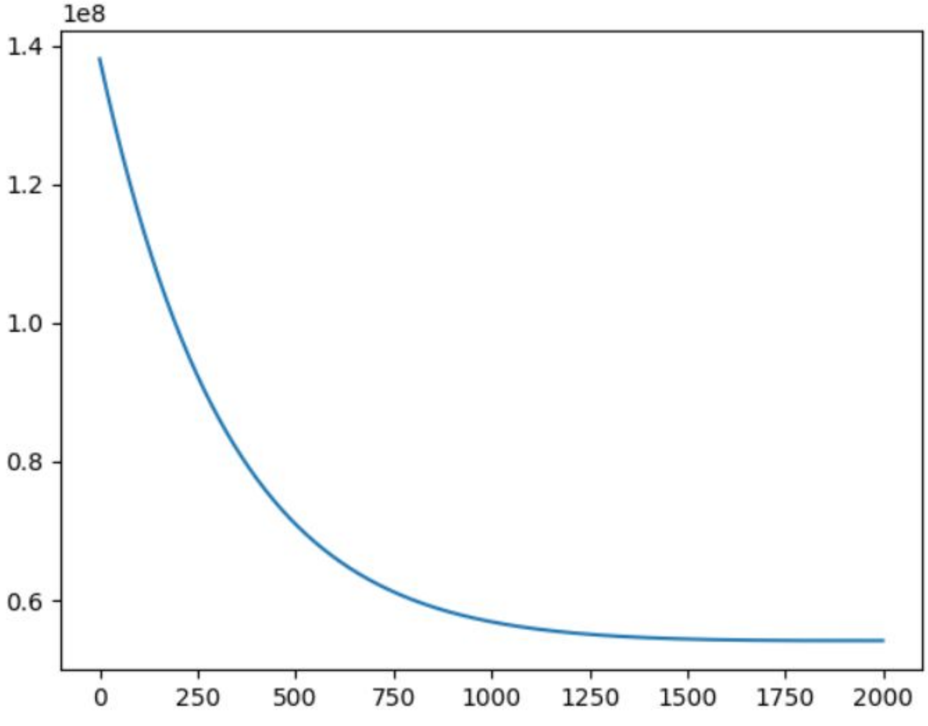
This equation was implemented with the help of the numpy library.
Apart from implementing the normal equation, the RMSE values obtained after running the regression models have also been printed. Below is a snippet of the output we get after running the normal equation part once.



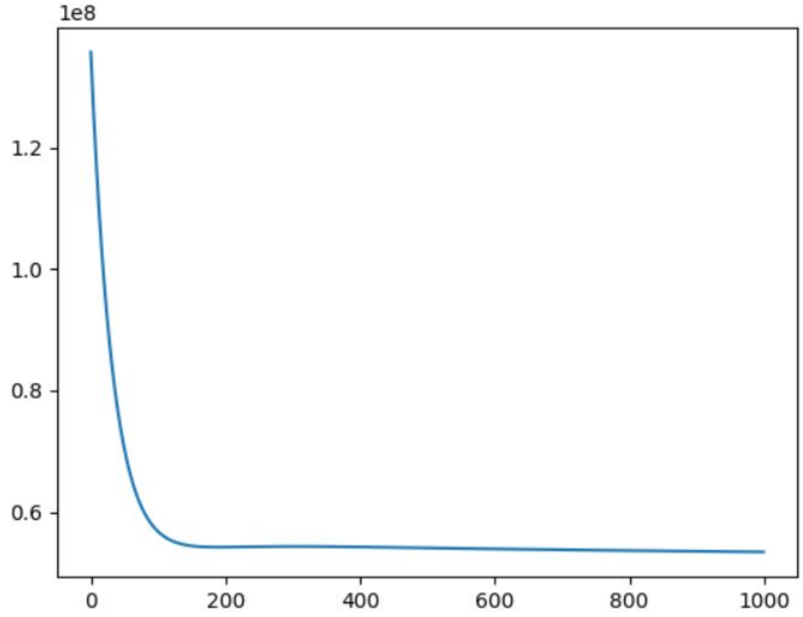
```
Minimum RMSE obtained from the 20 training data sets = 10871.538843344688
Mean of RMSE of training data sets = 11325.908843430861
Variance of RMSE of training data sets = 39839.659873452576
Minimum RMSE obtained from the 20 testing data sets = 10653.671256675394
Mean of RMSE of testing data sets = 11421.054984726214
Variance of RMSE of testing data sets = 220286.88326226542
Best value for the theta matrix obtained is:
[[ 2914.55843385]
 [10966.27529832]
 [ 9831.56052006]
 [ 4306.17889787]]
```

## Gradient Descent

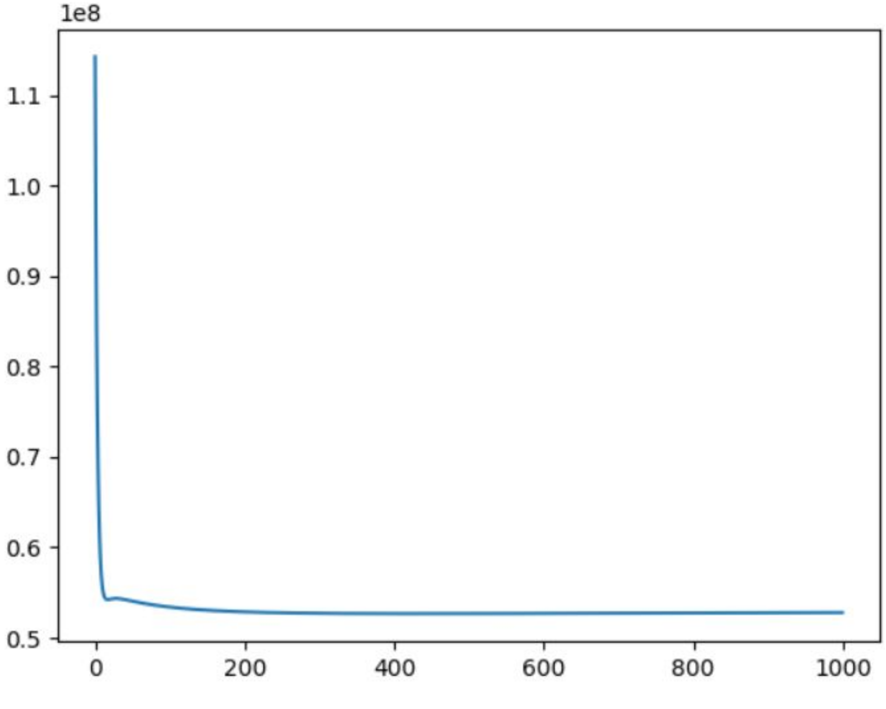
Gradient descent has been implemented with the learning rates as 0.001, 0.01 and 0.1. The theta values(w0,w1,etc.) were initialized to 1 before the gradient descent algorithm was implemented. The values of w1 and w2 tend to be among the maximum values of the 4, which suggests that the age and the bmi have the highest weight in deciding the value of the insurance charges. Plots have been made with the Y-axis as the value of the cost function and X-axis as the number of iterations that the gradient descent algorithm has run through. Only the plots giving the least RMSE value for the testing dataset have been attached in this report. Gradient descent algorithm gives the same output for the theta matrix value and the RMSE values as the normal equation method. The algorithm reaches the final value faster when the learning rate is greater. In the Jupyter Notebook, the cost function value after every 50 consecutive iterations is printed as well for better visualization. Below is a snippet of the output gotten after running the 20 GD models:-



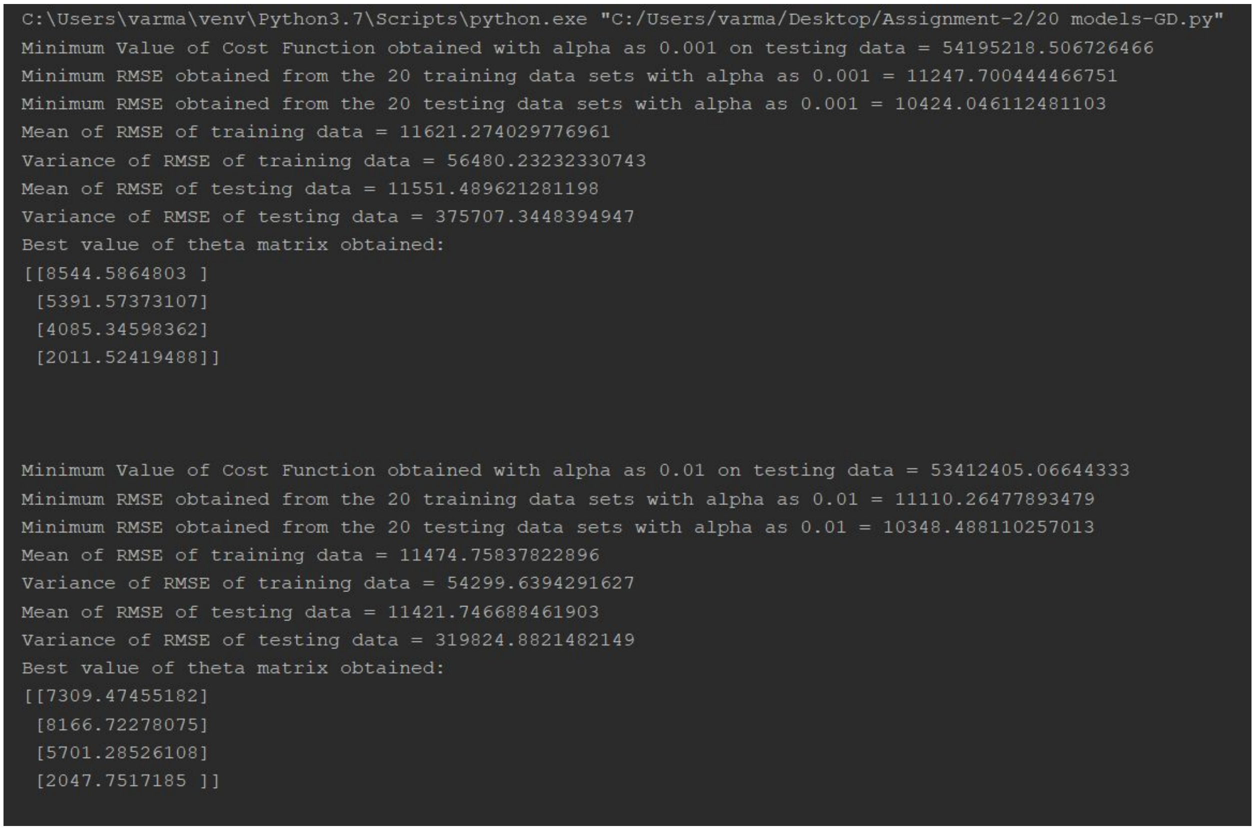
No. of Iterations
(Alpha = 0.001)



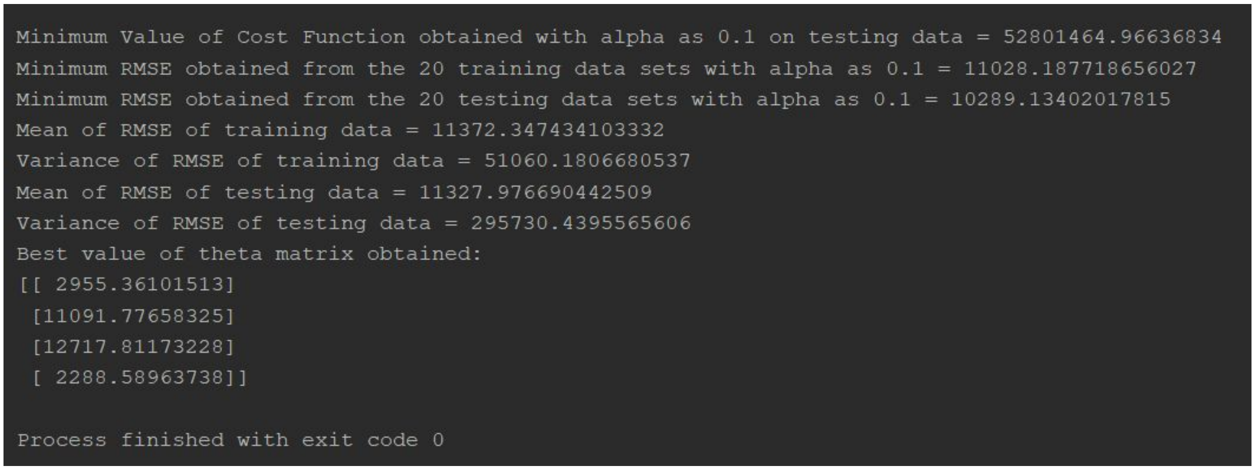
No. of iterations
(Alpha = 0.01)



No. of iterations
(Alpha = 0.1)



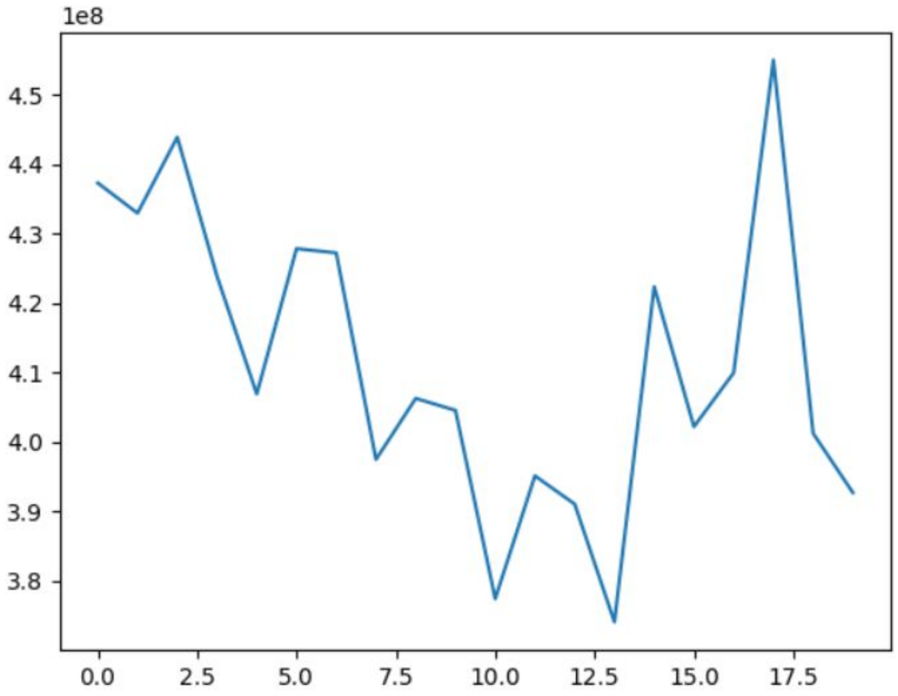
```
C:\Users\varma\venv\Python3.7\Scripts\python.exe "C:/Users/varma/Desktop/Assignment-2/20 models-GD.py"
Minimum Value of Cost Function obtained with alpha as 0.001 on testing data = 54195218.506726466
Minimum RMSE obtained from the 20 training data sets with alpha as 0.001 = 11247.700444466751
Minimum RMSE obtained from the 20 testing data sets with alpha as 0.001 = 10424.046112481103
Mean of RMSE of training data = 11621.274029776961
Variance of RMSE of training data = 56480.23232330743
Mean of RMSE of testing data = 11551.489621281198
Variance of RMSE of testing data = 375707.3448394947
Best value of theta matrix obtained:
[[8544.5864803 ]
 [5391.57373107]
 [4085.34598362]
 [2011.52419488]]



Minimum Value of Cost Function obtained with alpha as 0.01 on testing data = 53412405.06644333
Minimum RMSE obtained from the 20 training data sets with alpha as 0.01 = 11110.26477893479
Minimum RMSE obtained from the 20 testing data sets with alpha as 0.01 = 10348.488110257013
Mean of RMSE of training data = 11474.75837822896
Variance of RMSE of training data = 54299.6394291627
Mean of RMSE of testing data = 11421.746688461903
Variance of RMSE of testing data = 319824.8821482149
Best value of theta matrix obtained:
[[7309.47455182]
 [8166.72278075]
 [5701.28526108]
 [2047.7517185 ]]
```



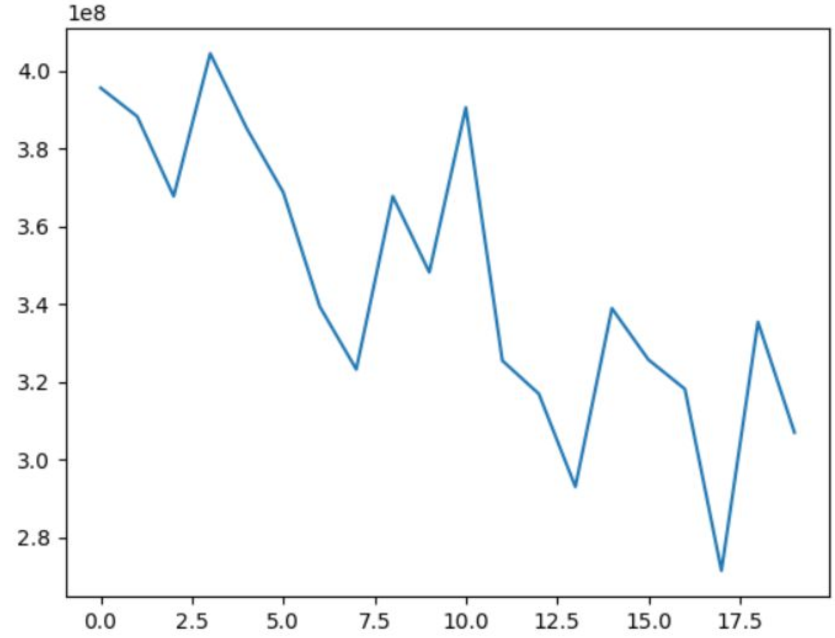
```
Minimum Value of Cost Function obtained with alpha as 0.1 on testing data = 52801464.96636834
Minimum RMSE obtained from the 20 training data sets with alpha as 0.1 = 11028.187718656027
Minimum RMSE obtained from the 20 testing data sets with alpha as 0.1 = 10289.13402017815
Mean of RMSE of training data = 11372.347434103332
Variance of RMSE of training data = 51060.1806680537
Mean of RMSE of testing data = 11327.976690442509
Variance of RMSE of testing data = 295730.4395565606
Best value of theta matrix obtained:
[[ 2955.36101513]
 [11091.77658325]
 [12717.81173228]
 [ 2288.58963738]]

Process finished with exit code 0
```

## Stochastic Gradient Descent

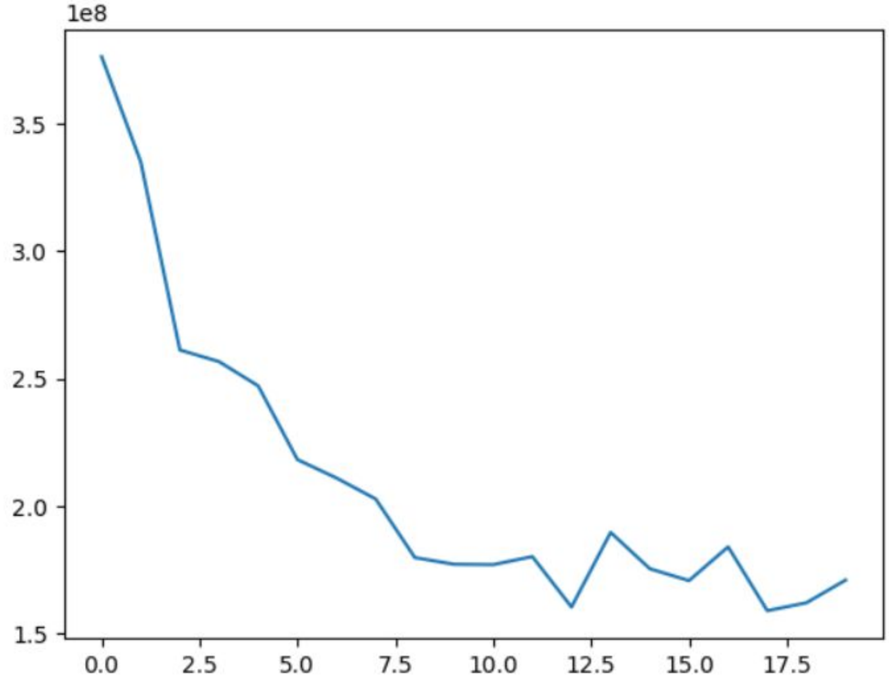
This method has only been run for 20 iterations in the '20 models-SGD.py file' and for 200 iterations in the '.ipynb' file due to it having a really high run-time. In one iteration, a random row from the input values is selected and the gradient descent algorithm is run for only that one particular datapoint. This process is repeated for the number of total rows in the input matrix. Only a single iteration of this sort proves to be computationally harder than the normal gradient descent algorithm since gradient descent simplifies calculations through the use of numpy. This method gives a greater RMSE value and the greater value of the cost function as compared to GD as well. This difference is due to less number of iterations being run on SGD. Even if SGD may give the same result as GD and the normal equation, it is not computationally feasible. Plots have been generated in the same manner as the above approach. Below is a snippet of the outputs that we get by running the SGD model:-



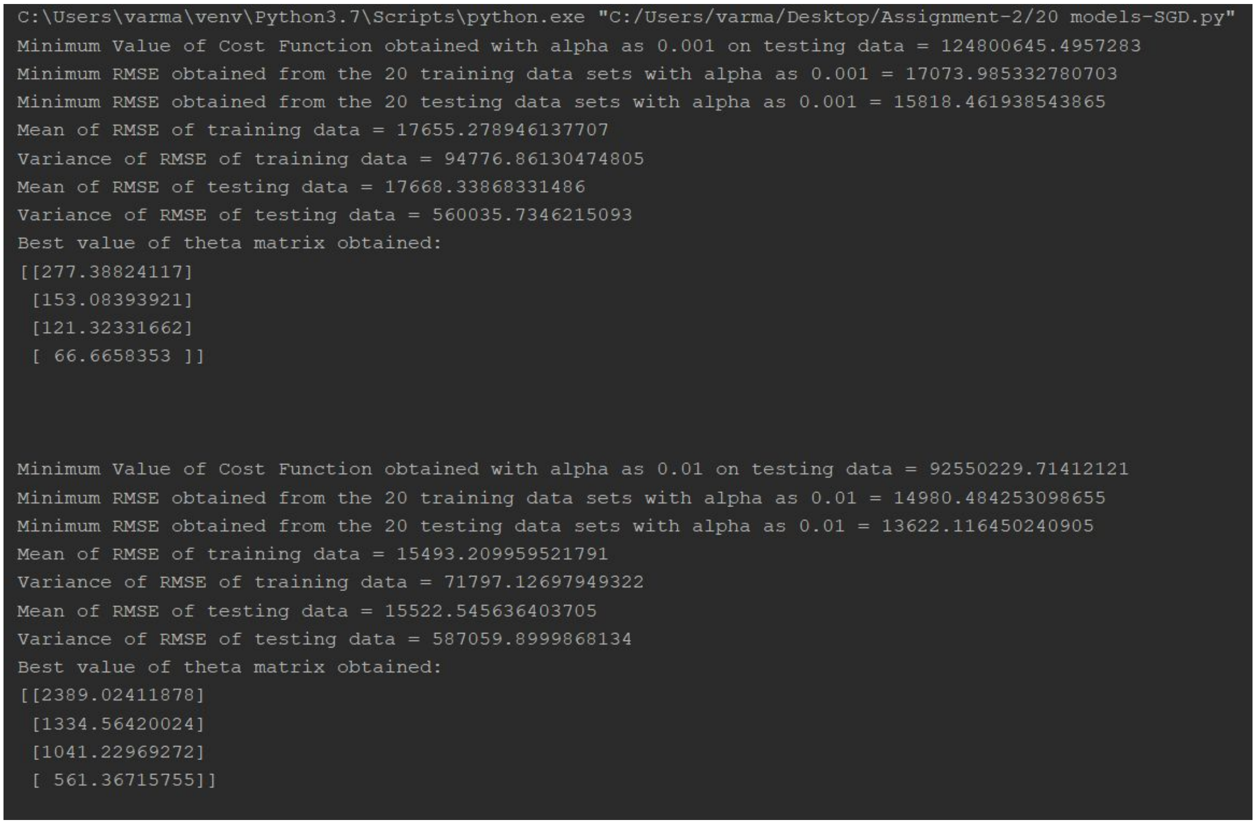
No. of iterations
(Alpha = 0.001)



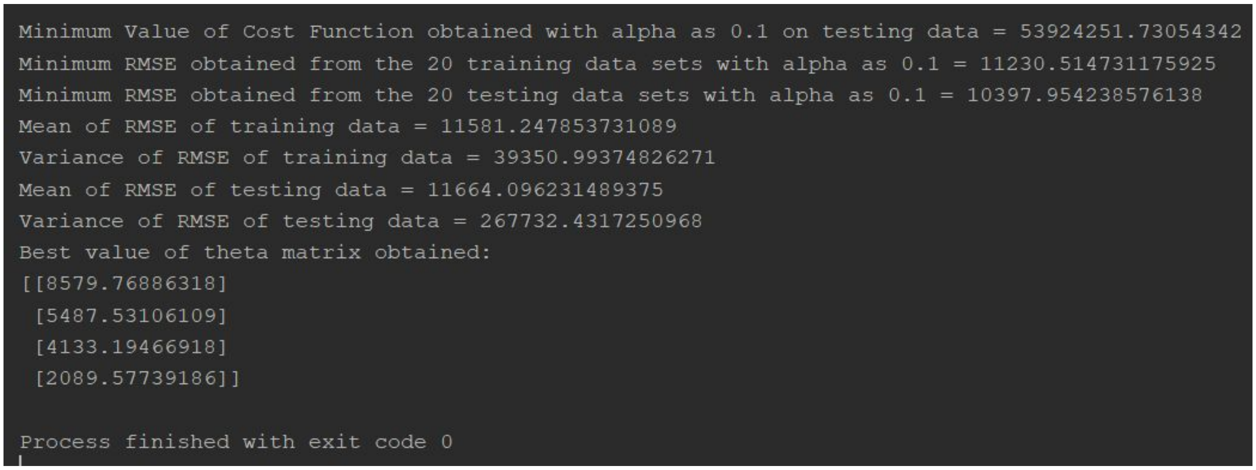
No. of iterations
(Alpha = 0.01)



No. of iterations
(Alpha = 0.1)



```
C:\Users\varma\venv\Python3.7\Scripts\python.exe "C:/Users/varma/Desktop/Assignment-2/20 models-SGD.py"
Minimum Value of Cost Function obtained with alpha as 0.001 on testing data = 124800645.4957283
Minimum RMSE obtained from the 20 training data sets with alpha as 0.001 = 17073.985332780703
Minimum RMSE obtained from the 20 testing data sets with alpha as 0.001 = 15818.461938543865
Mean of RMSE of training data = 17655.278946137707
Variance of RMSE of training data = 94776.86130474805
Mean of RMSE of testing data = 17668.33868331486
Variance of RMSE of testing data = 560035.7346215093
Best value of theta matrix obtained:
[[277.38824117]
 [153.08393921]
 [121.32331662]
 [ 66.6658353 ]]



Minimum Value of Cost Function obtained with alpha as 0.01 on testing data = 92550229.71412121
Minimum RMSE obtained from the 20 training data sets with alpha as 0.01 = 14980.484253098655
Minimum RMSE obtained from the 20 testing data sets with alpha as 0.01 = 13622.116450240905
Mean of RMSE of training data = 15493.209959521791
Variance of RMSE of training data = 71797.12697949322
Mean of RMSE of testing data = 15522.545636403705
Variance of RMSE of testing data = 587059.8999868134
Best value of theta matrix obtained:
[[2389.02411878]
 [1334.56420024]
 [1041.22969272]
 [ 561.36715755]]
```



```
Minimum Value of Cost Function obtained with alpha as 0.1 on testing data = 53924251.73054342
Minimum RMSE obtained from the 20 training data sets with alpha as 0.1 = 11230.514731175925
Minimum RMSE obtained from the 20 testing data sets with alpha as 0.1 = 10397.954238576138
Mean of RMSE of training data = 11581.247853731089
Variance of RMSE of training data = 39350.99374826271
Mean of RMSE of testing data = 11664.096231489375
Variance of RMSE of testing data = 267732.4317250968
Best value of theta matrix obtained:
[[8579.76886318]
 [5487.53106109]
 [4133.19466918]
 [2089.57739186]]

Process finished with exit code 0
```