Multi Linear Regression

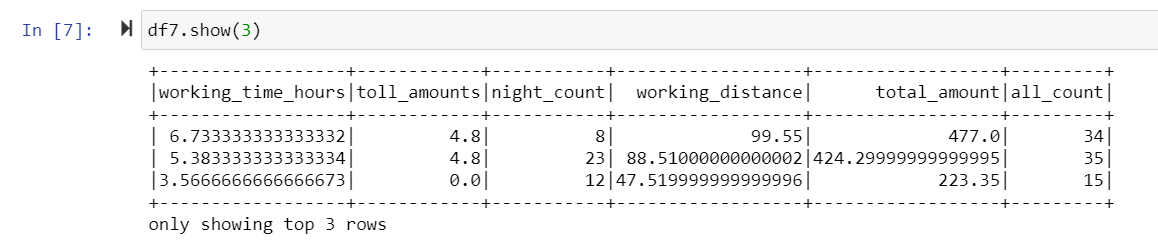
The following exercise is a multi linear regression using batch gradient descent written from scratch without any usage of any libraries in pyspark. The data set used is attached to this document. The data set comprises of the details from New York City taxi trip reports from the year 2013. We will build a data model based on Multi Linear Regression to predict the total amount of money that each driver can make per day. This is the amount that a driver gets per day (fare amount plus tip amount). We want to use the following 5 features for this linear regression model are:

* Total working time in hours that a driver worked per day (a float number)
* Total travel distance in miles that a driver drove a taxi per day (a float number)
* Total number of rides per day (an integer number)
* Total amount of toll per day (a float number - toll amount indicates the number of rides over

the NYC bridges or rides to the airport.

* Total number of night rides between 1:00 AM and 6:00 AM per day.

By using pyspark data frames, we created the data frame.

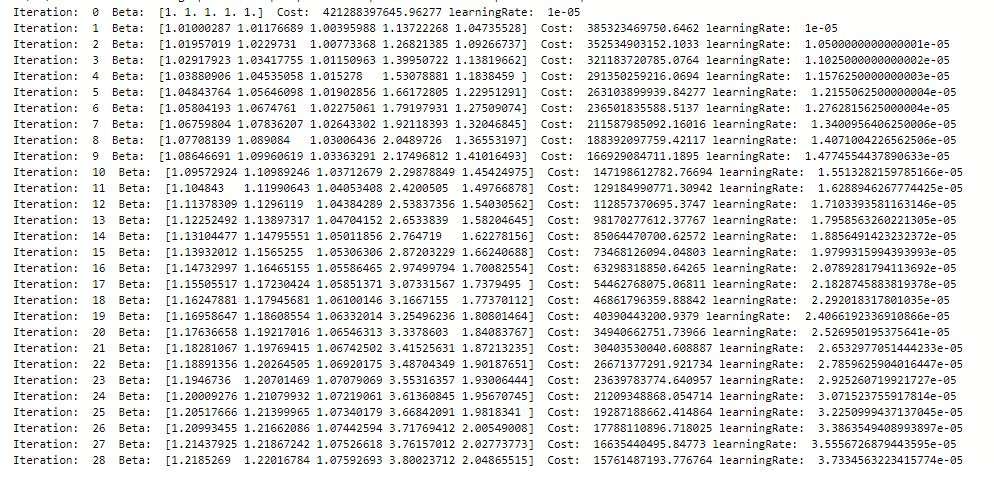


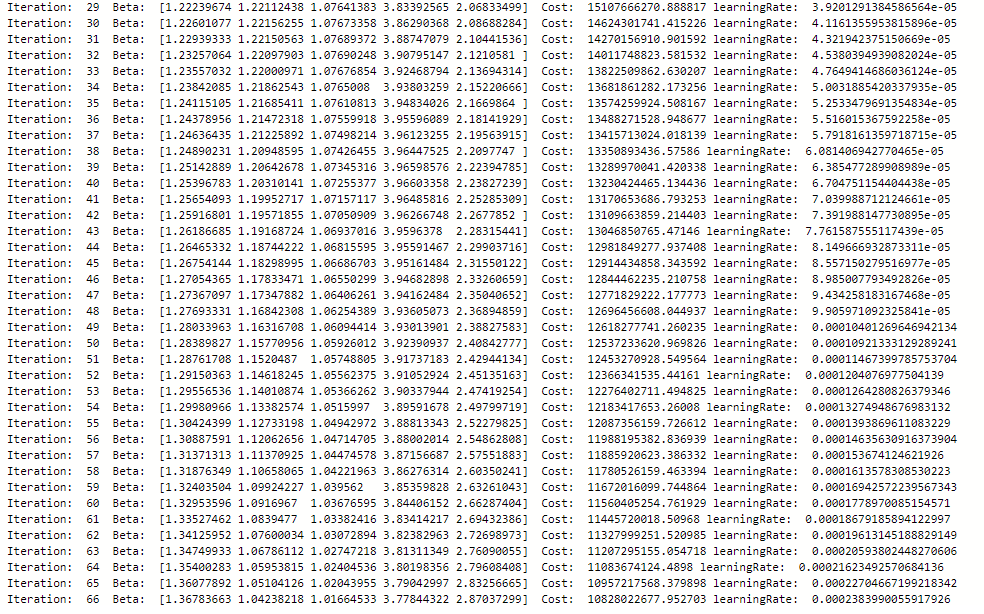
We implemented using the bold driver. Bold driver is a way to dynamically change the learning rate.

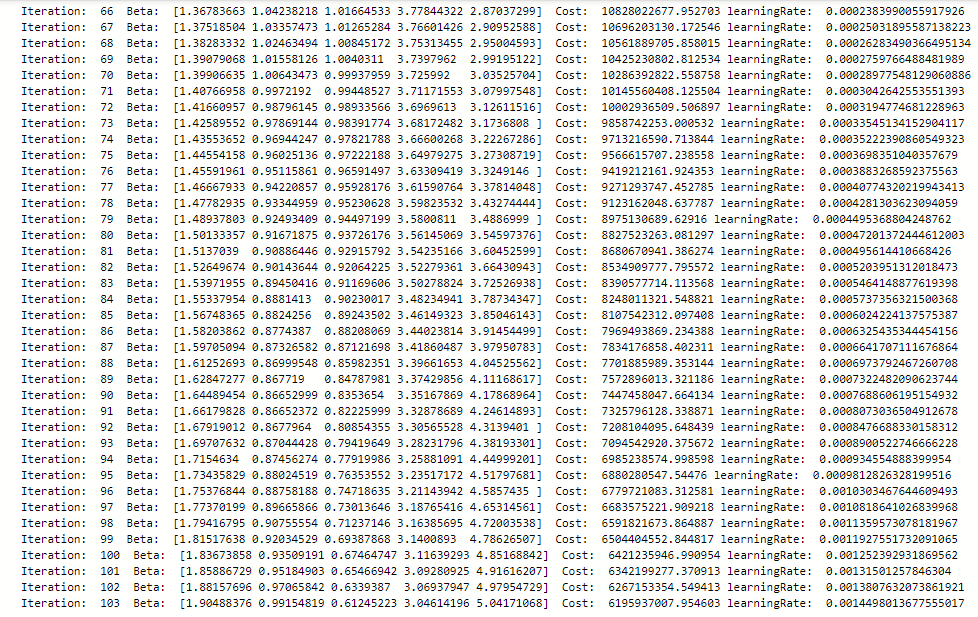
On increasing cost, learning rate is halved. While on decreasing cost, the rate is 1.05 times. This time we ran 200 iterations. The bold driver helped in getting us out of the local minima.

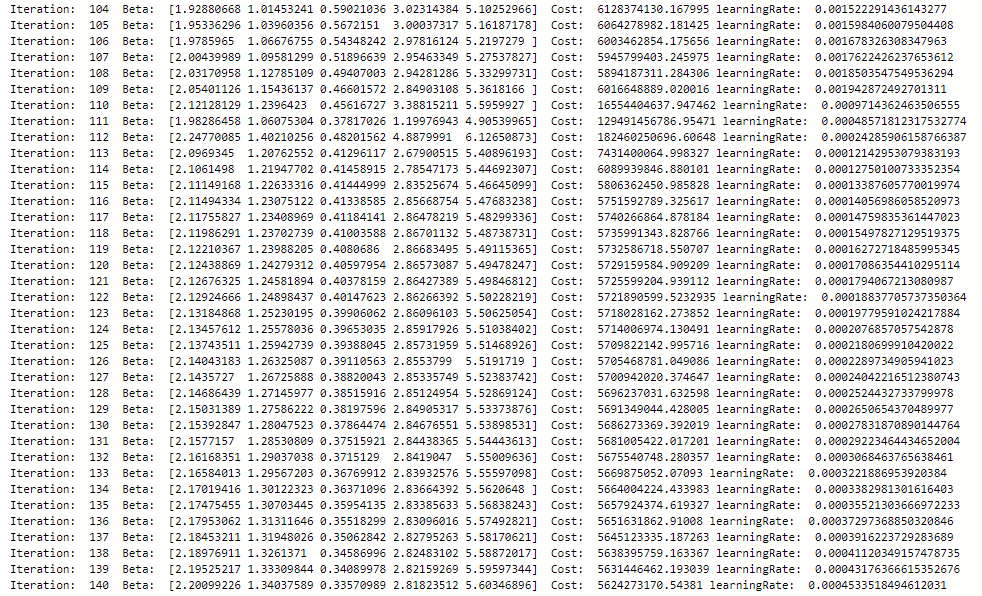
The learning rate was dynamically changed using the bold driver in this case.

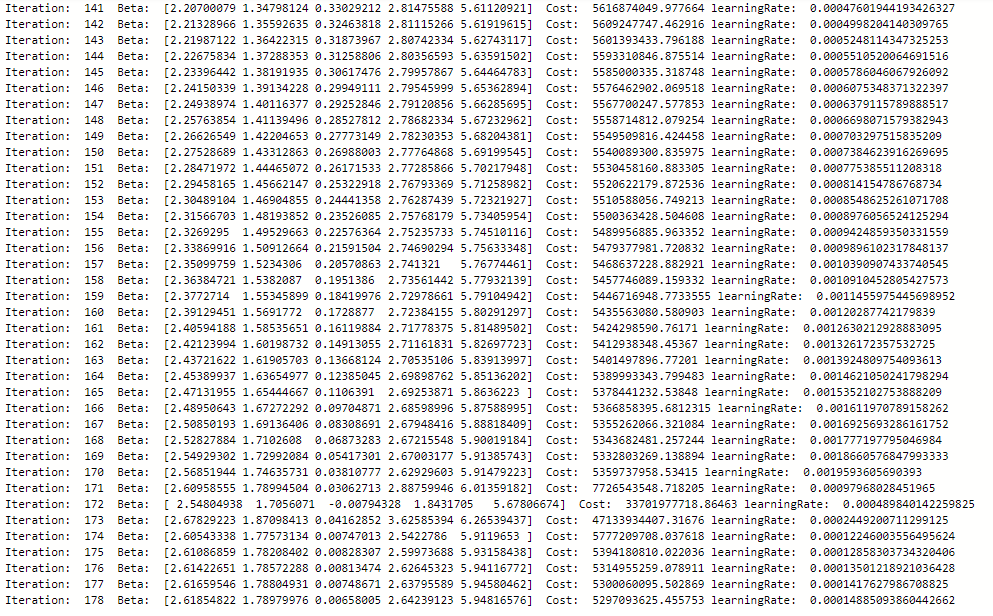
This was done on the GCP cloud services.

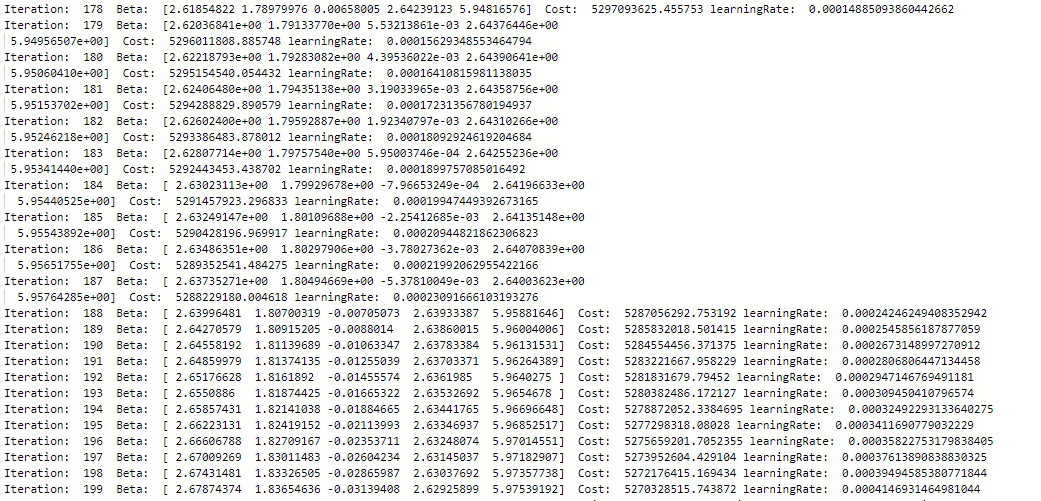




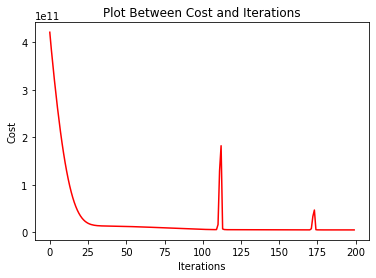








Plot showing the cost in each iteration



There were minor spikes observed. These can be due to the increasing of learning rate, and finding high cost for that specific learning rate and parameters. The cost without bold driver was 1.254e10 after 200 iterations, but 5.27e9. with the bold driver. For higher learning rate, the behaviour was absurd. The final weights can be used to predict on independent variables.