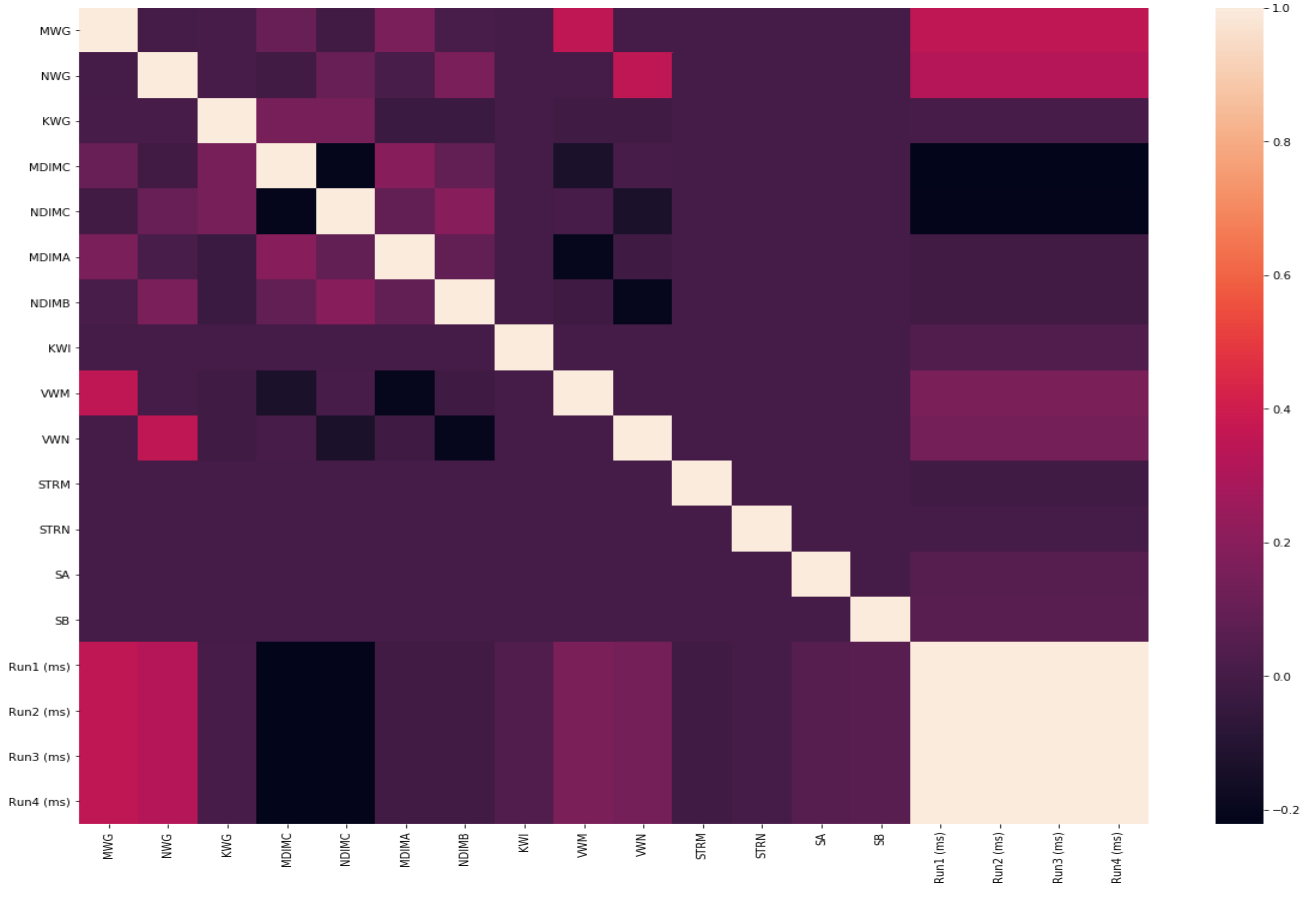
Linear Regression

In this project, I have experimented with different values of the alpha and the learning rate. The process involved exploring the data, data pre-processing, defining algorithms, and checking the model evaluation metrics i.e., root mean square error, mean square error in linear regression.

- I used different alpha values for experimentation and concluded that alpha = 0.03 is the optimal value for minimizing the cost function.
- I have generated a random column index and selected them to model my algorithm but the 'rmse' values were way worse than taking all columns as well as the best column chosen by me by checking all the experimentation.
- There was not much correlation in the data as shown by the heatmap below



Data Pre-processing

- There are no missing values in the dataset, as it is evident from data.isna() values
- I checked the data, which were in different scales. So I have used min-max scaling for scaling the data
- The values of the run were highly skewed, so I took log transform to reduce the effect of skewness
- I have taken a split of 0.7 for training and 0.3 for testing data

```
data.isna().sum()

MWG          0
NWG          0
KWG          0
MDIMC        0
NDIMC        0
MDIMA        0
NDIMB        0
KWI          0
VWM          0
VWN          0
STRM         0
STRN         0
SA           0
SB           0
Run1 (ms)    0
Run2 (ms)    0
Run3 (ms)    0
Run4 (ms)    0
dtype: int64
```

BUAN6341.004 Machine Learning Assignment1 Chitresh Kumar

Algorithms :

I have defined a cost function for rmse and mse for calculation. Then I have created a function each for linear regression and gradient descent with threshold and alpha . I am not using any loops for alpha values; rather, I am giving the values manually.I am using three parameters to control my gradient descent function alpha, it(iteration) , tol (tolerance level) to calculate our beta values and cost after each iteration.

```
print("theta: {}, \n number of iterations = {}".format(theta, count))
```

```
theta: [ 3.73807865e+00  1.51300770e+00  1.19212564e+00  2.09164277e-01
 -1.31876532e+00 -1.27008671e+00  2.19712449e-02  2.37064030e-02
 -8.58334756e-03 -9.40584144e-03 -1.01230274e-01 -1.11221872e-01
  1.15435377e-03 -1.70981305e-01 -2.57931207e-02],
 number of iterations = 602
```

I have defined the beta randomly initially .After applying the gradient descent we can getting beta values as shown above and the number of iteration was 602 when I gave the tolerance level of 0.001 and number of iteration of 2000.

Linear Regression Equation

Run = beta0 + beta1*MWG+ beta2*NWG+ beta3*KWG+ beta4*MDIMC+ beta5*NDIMC+ beta6*MDIMA+ beta7*NDIMB+ beta8*KWI+ beta9*VWM+ beta10*VWN+ beta11*STRM+ beta12*STRN+ beta13*SA+ beta14*SB

Experiment 1)

The initial theta values are

```
theta: [-0.34889445  0.98370343  0.58092283  0.07028444  0.77753268  0.58195875
  1.47179053  1.66318101 -0.26117712 -0.68867681 -0.69492326  1.94042346
  1.80541519  0.45631385 -0.57481204],
```

The optimized theta values after iteration are

```
theta: [ 3.86263268  1.49719843  1.17711662  0.1960955  -1.35461902 -1.3063491
  0.00554319  0.00614963 -0.02541197 -0.05469389 -0.14920599 -0.12796896
 -0.01583042 -0.187738   -0.0426499 ],
 number of iterations = 3000
```

Final Run equation = 3.8626 + 1.4971*MWG+ 1.1771*NWG+ 0.19609*KWG-1.3546*MDIMC-1.3063*NDIMC+ 0.00554*MDIMA+ 0.00614*NDIMB-0.0254*KWI-0.05469*VWM-0.149205*VWN-0.127968*STRM+ -0.01583*STRN-0.187738*SA-0.0426499*SB

The values which I am getting after iteration with optimal alpha and least rmse for test which is nearest value as compared to the train error

```
mse_cost(y_test, y_pred_test)
```

```
0.28295852466997395
```

```
rmse_cost(y_test, y_pred_test)
```

```
0.5319384594762574
```

Actual Values	Predicted Values
5.008065	5.155649
4.226213	3.741866
5.422624	5.279745
3.806718	1.860283
3.828043	3.977924
...	...
5.280064	5.050970
3.382694	4.263414
5.057026	5.047071
3.369449	4.472287
5.640896	4.873098

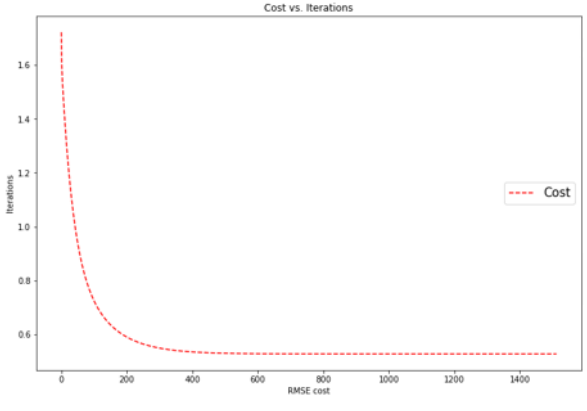
The rmse cost of the test data is around 0.5319

While our rmse cost of the train data is 0.52717.

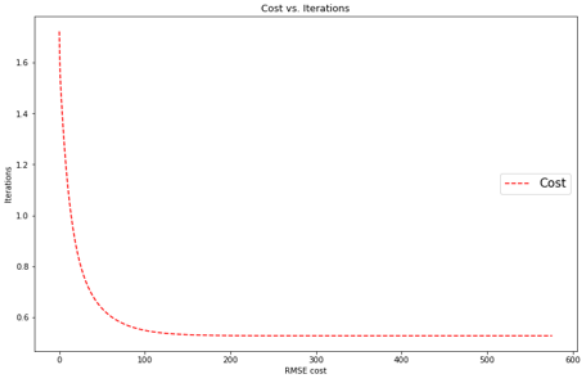
Experiment 2)

With various alpha values the plot of the rmse with iteration are as follows:

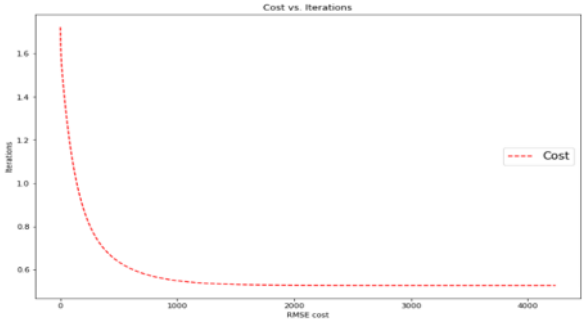
Alpha = 0.1



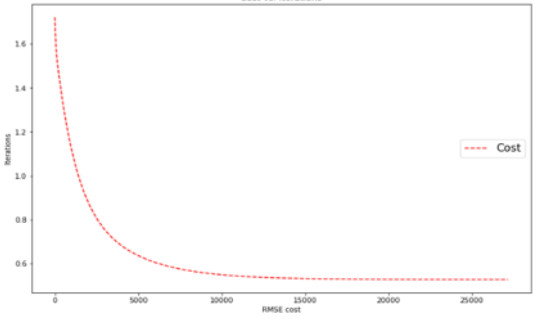
Alpha = 0.3



Alpha = 0.03



Alpha = 0.003



BUAN6341.004 Machine Learning Assignment1 Chitresh Kumar

Experiment 3)

The rmse values for randomly selected columns ('MWG', 'KWG', 'NDIMC', 'KWI', 'VWM','VWN', 'SA', 'SB')

Train: 0.6764446695893828
Test: 0.6809499473720186

The train and test values for original data is 0.5272002042389978 ,0.5319384594762574.

Experiment 4)

The rmse values for columns selected by me is 0.6587239211749101,0.6546666463611563
The columns of my choice performed better than the column selecting randomly while it performed worse th an the original data. I think the skewness of the output variable is decreasing my model performance. As we d on't have much correlation in our data. So every variable contains important variation which are not able to p rovide while selecting less columns. My choice of variables was based on the columns and its relationship with the output variable so I think it is performing slightly better than randomly selected columns.

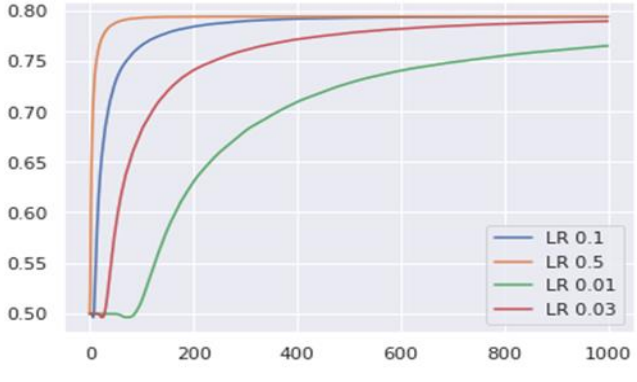
From the results we can say that MWG, NWG, MDIMC,NDIMC will have the highest impact on Run time.
I think I should have removed the outliers in the output variable. I could have used more iteration, but it was not able to run it on my system.

Logistic Regression

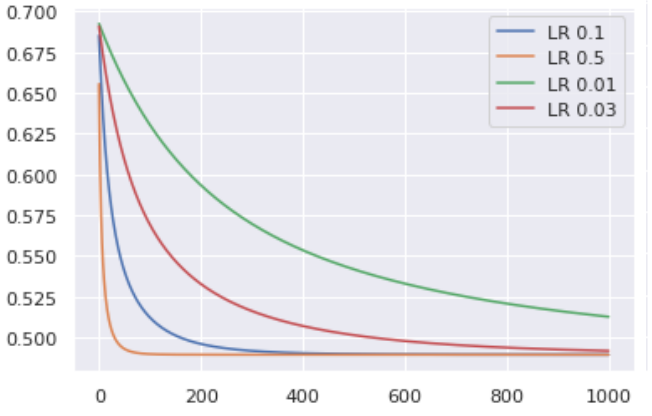
In the logistic regression,I have used the median of the run time to convert the variable into a binary variable.I have created the sigmoid function and gradient descent algorithm to run the logistic regression.
Experiment 1)
The accuracy values for the logistic regression when alpha is [0.1,0.5,0.01,0.03] and the number of iteration taken is as follows:
The accuracy of the test set is : 0.7949227373068433
The accuracy of the train set is: 0.7942999053926206
The accuracy of the test set is : 0.7949089403973509
The accuracy of the train set is: 0.7944063386944181
The accuracy of the test set is : 0.7638520971302428
The accuracy of the train set is: 0.7652377010406812
The accuracy of the test set is : 0.7898730684326711
The accuracy of the train set is: 0.7898119678334911

Experiment 2) As we can see from the values above best performing alpha value is 0.5 .We can see from accuracy the accuracy value is the highest in the alpha = 0.5.

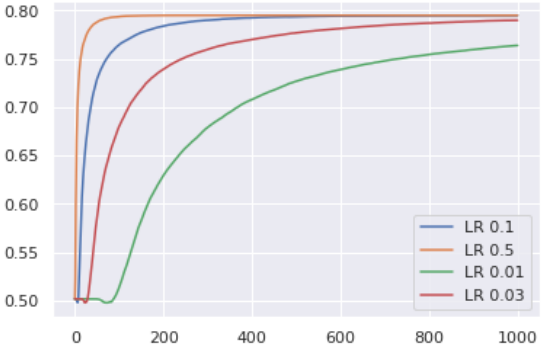
Test Accuracy vs Iteration



Cost vs Iteration



Training Accuracy vs Iteration



Experiment 3)

I have selected all the columns which had randomly selected in the linear regression experiment .
The accuracy of test data is : 0.6784871688741722
The accuracy of train data is : 0.6787665562913907
The accuracy for randomly selected data is less than the original dataset we have used originally

Experiment 4)
The selected columns similar to the linear regression experiment.
The accuracy of test data is : 0.6984478476821192
The accuracy of train data is : 0.7000758830022075

The accuracy of selected data is less than the original data and more than the randomly selected data.