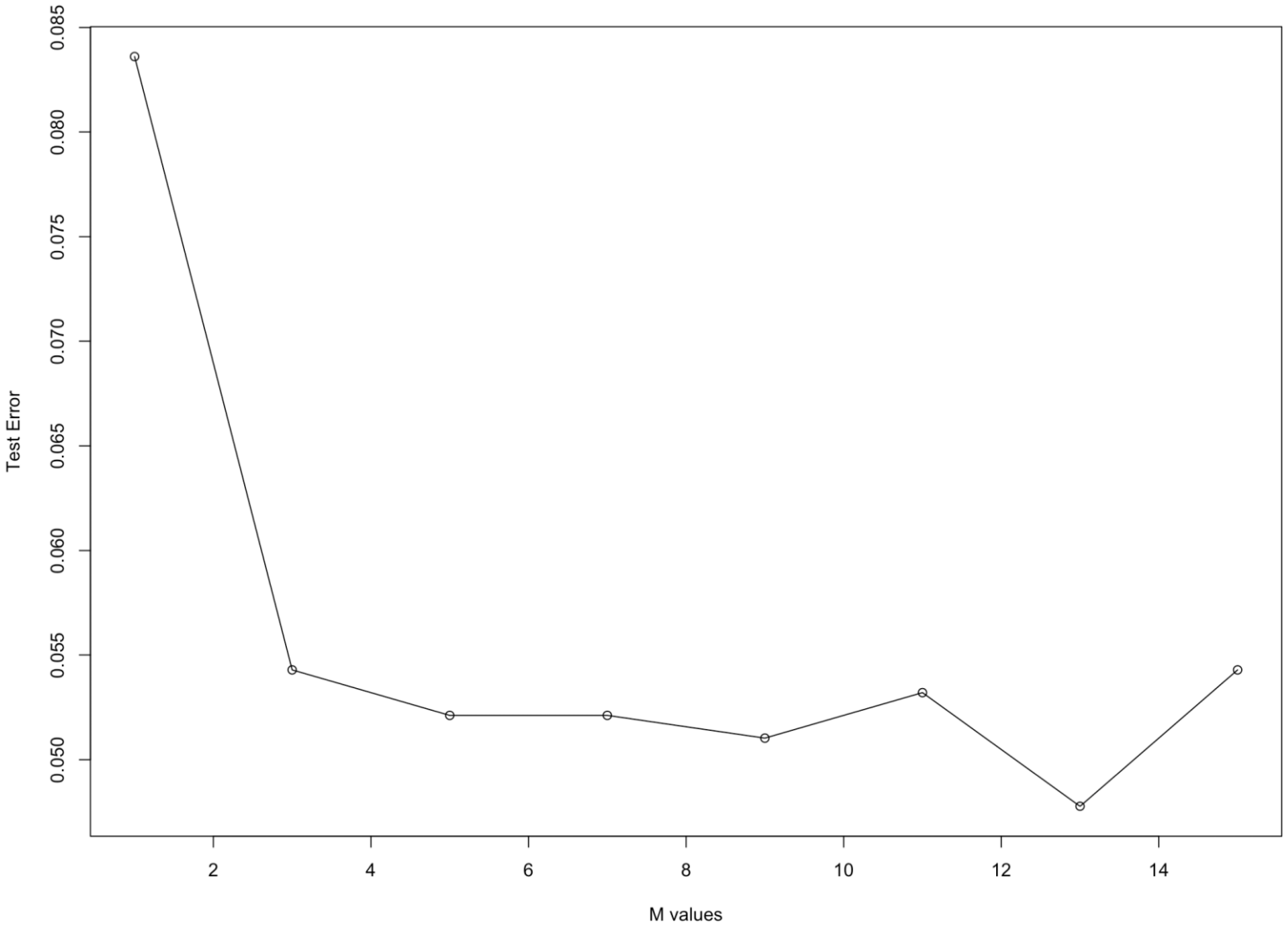
EAS 506
HW 5

By - Mohit Tripathi

Ans 1 –

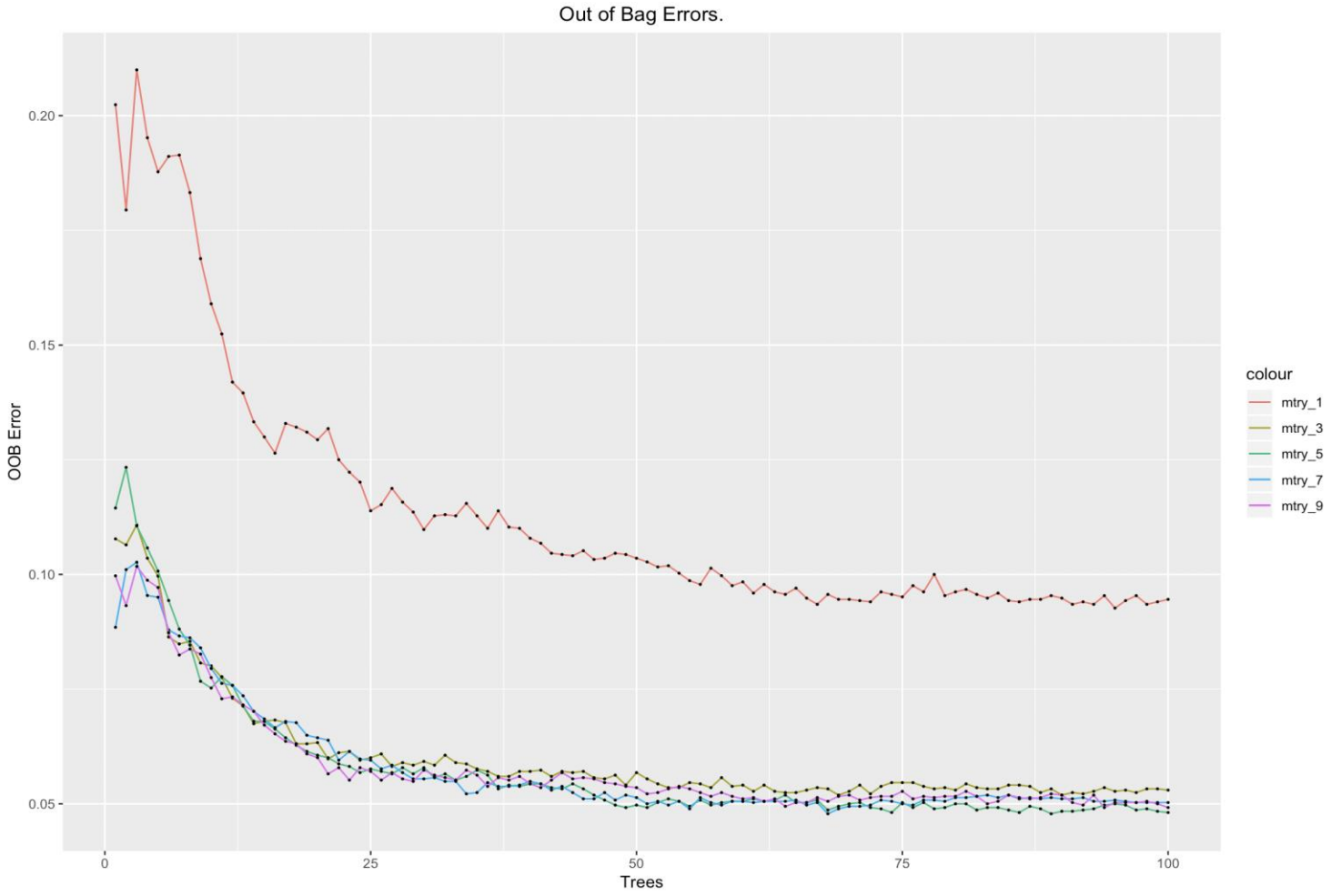
Test error list for different values of mtry-
0.06948969, 0.04343105, 0.04343105, 0.04234528, 0.04234528, 0.03800217, 0.04125950, 0.04451683

Test Error plot for different values of mtry-



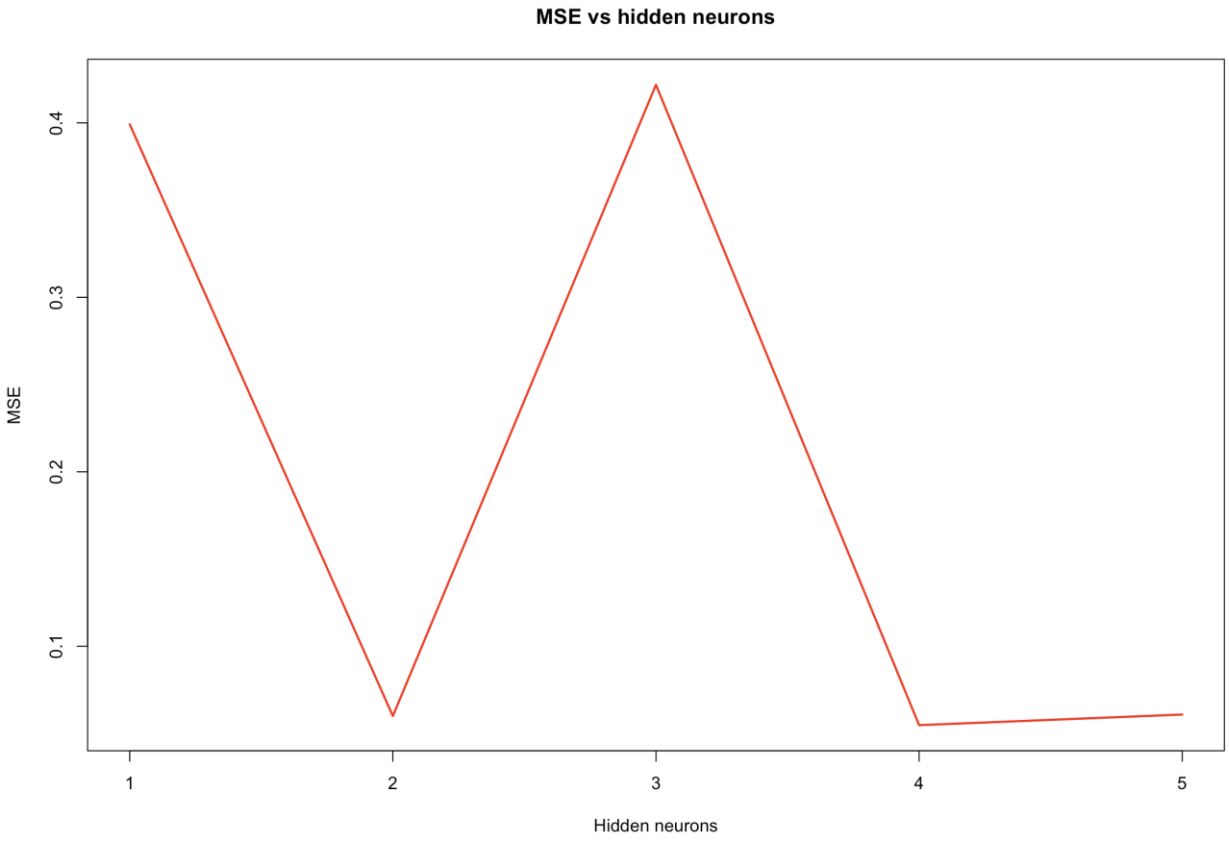
Test error is minimum for mtry = 13

OOB error for different values of mtry



Ans2-

Plotting CV error for different number of neurons used



Can observe that the minimum CV error is with 4 neurons

Test Error with 4 hidden neurons = 6.69%

Test error with logistic regression is 6.43 %. Logistic regression is giving slightly better results than Neural net in this case may be because of the less number of hidden layers are used due to computational challenges. As we will increase the hidden neurons, neural net will give better results.

In terms of Performance neural network will outperform the GAM like logistic regression, boosting, bagging. However, neural network is poor in interpretability. Their outcome is much difficult to explain in comparison to simple additive models.

Ans 3-

Test error with original data – 5.73%
Test error when the outlier has a value of 100 – 5.82%
Test error when the outlier has a value of 25 – 6.17%
Test error when the outlier has a value of 20 – 6.17%
Test error when the outlier has a value of 2 – 6.08%
Test error when the outlier has a value of .2 – 6.08%

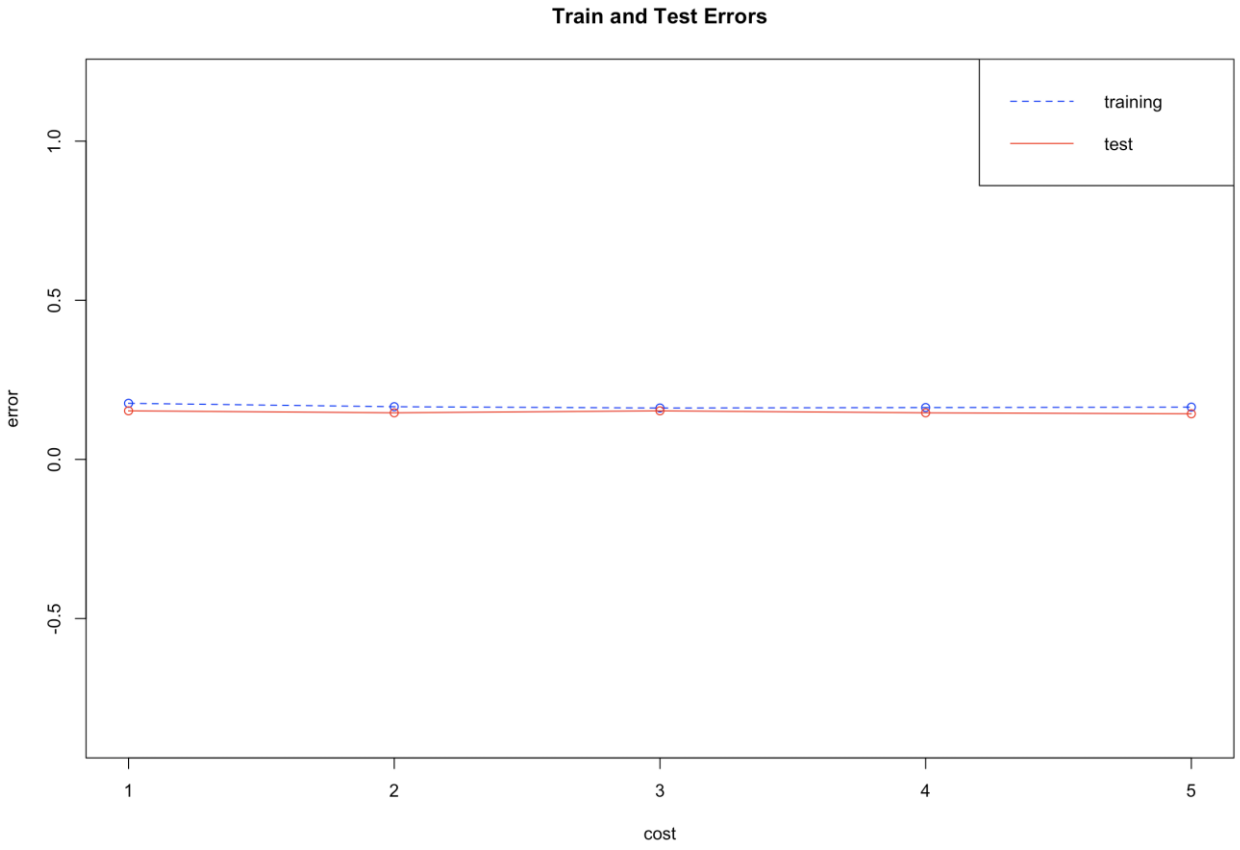
We can observe that as initially when we increased the value of a particular feature in row 1 from 0 to 100, the test error was increased to 5.82% . As we reduced the value of outlier to 25, the error increased to 6.17% . Further when the value of outlier was reduced to 2 and ,2, the test error has become constant approx.. and we can say that it's not getting affected by shrinking the outlier value to its original value.

Ans 4 –
a.)
support vector classifier (Linear Kernel)

Test Error
0.1526479751 0.1464174455 0.1526479751 0.1464174455 0.1433021807
Train Error
0.1762349800 0.1655540721 0.1615487316 0.1628838451 0.1642189586

Optimal cost is 10 where the test error is minimum.

Plots for train and test error



b.)

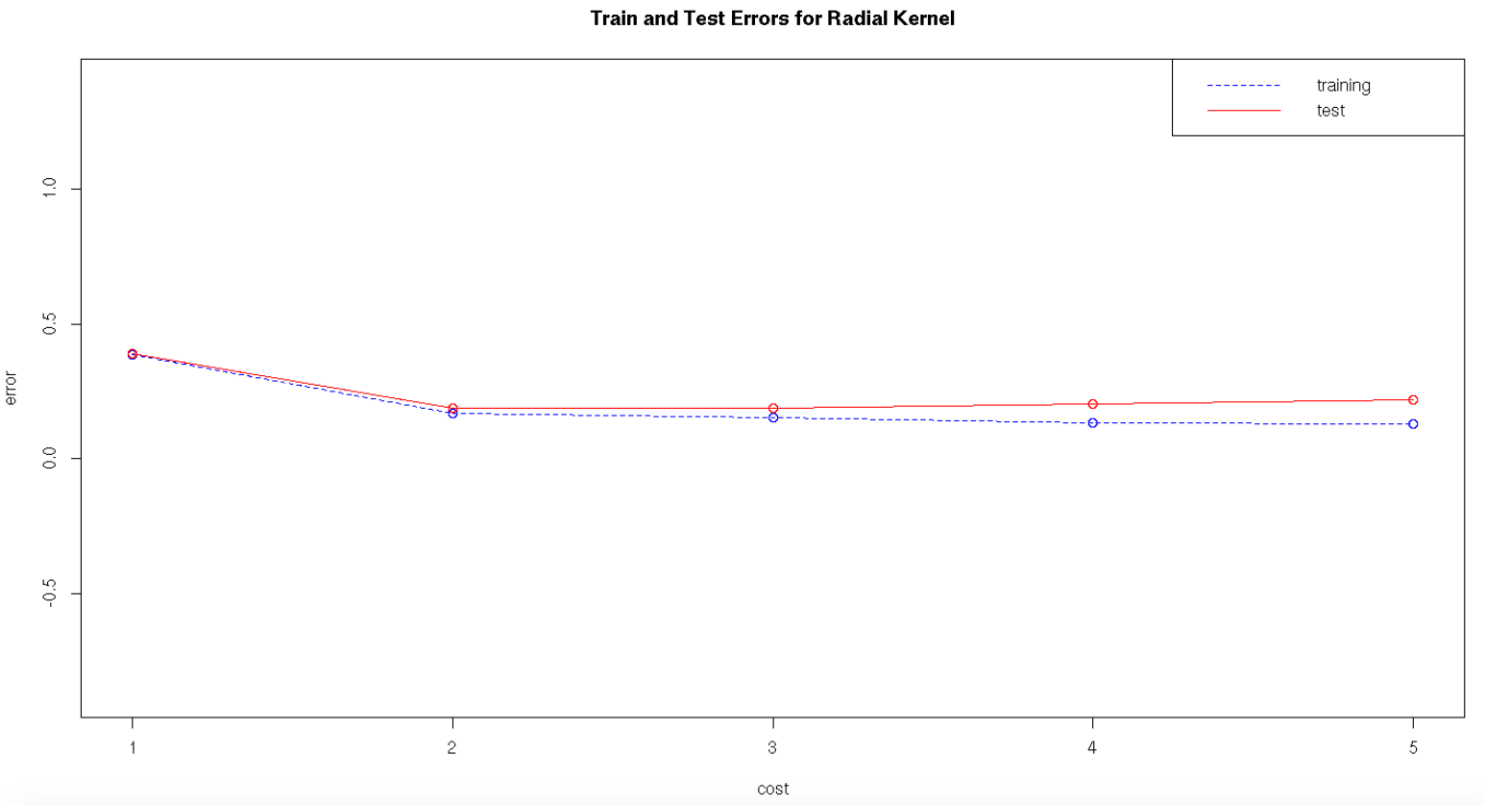
Radial Kernel

Test Error
0.3925233645, 0.1869158879, 0.1869158879, 0.2024922118, 0.2180685358

Train Error
0.3885180240, 0.1708945260, 0.1522029372, 0.1321762350, 0.1295060080

optimal cost is at cost 1

Plots for train and test error



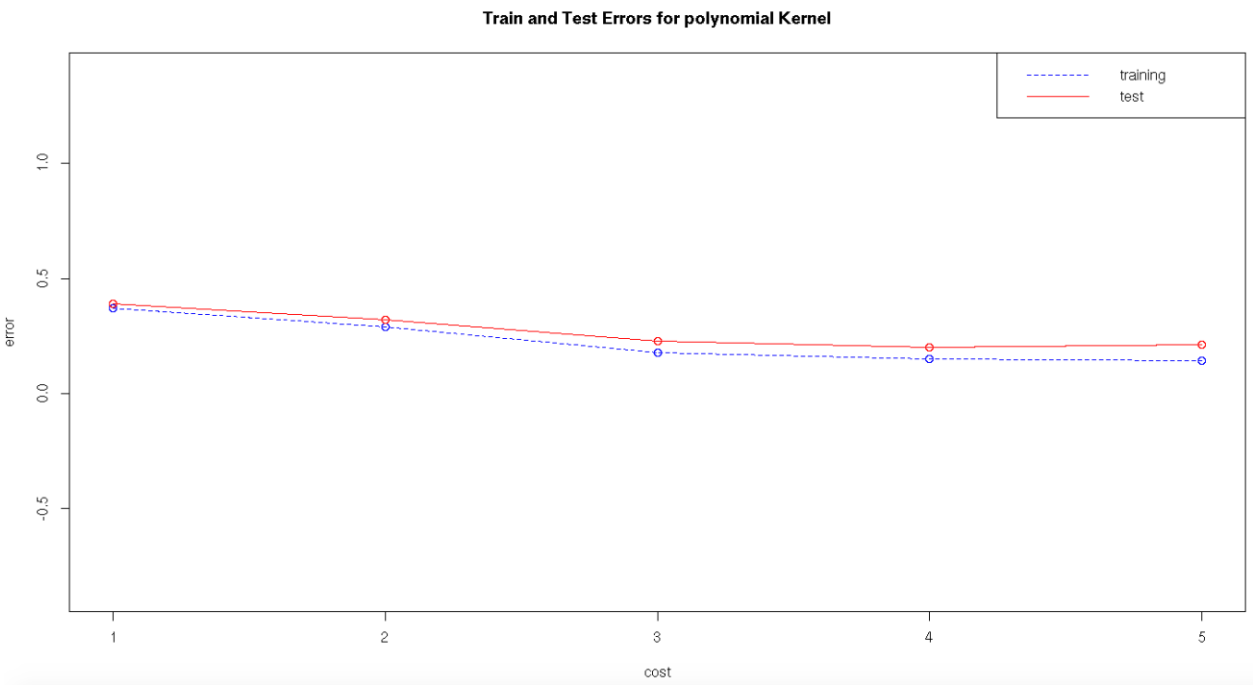
Polynomial Kernel

Test Error
0.3894080997, 0.3208722741, 0.2274143302, 0.2024922118, 0.2118380062

Train Error
0.3711615487,0.2910547397, 0.1775700935, 0.1495327103, 0.1415220294

optimal cost is at cost 5

Plots for train and test error



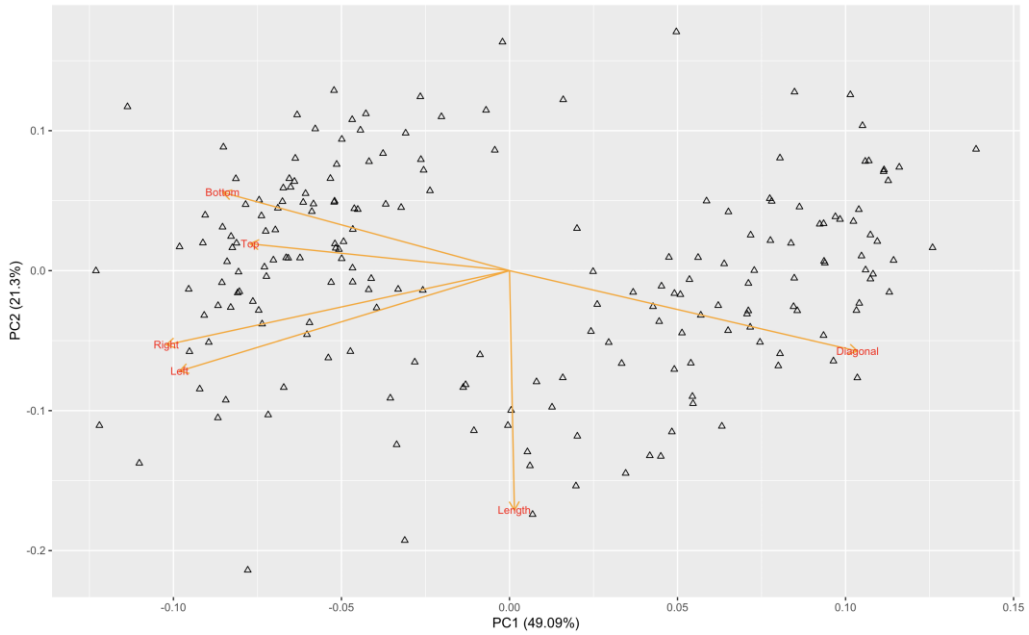
SVM with linear kernel has minimum error even at cost (cost 10) while SVM with radial and polynomial kernel (order 2) has more error at margin cost of 10, but they improvise when the cost is decreased to 1 or 5 respectively.

Linear kernel is giving the best results.

Ans5-

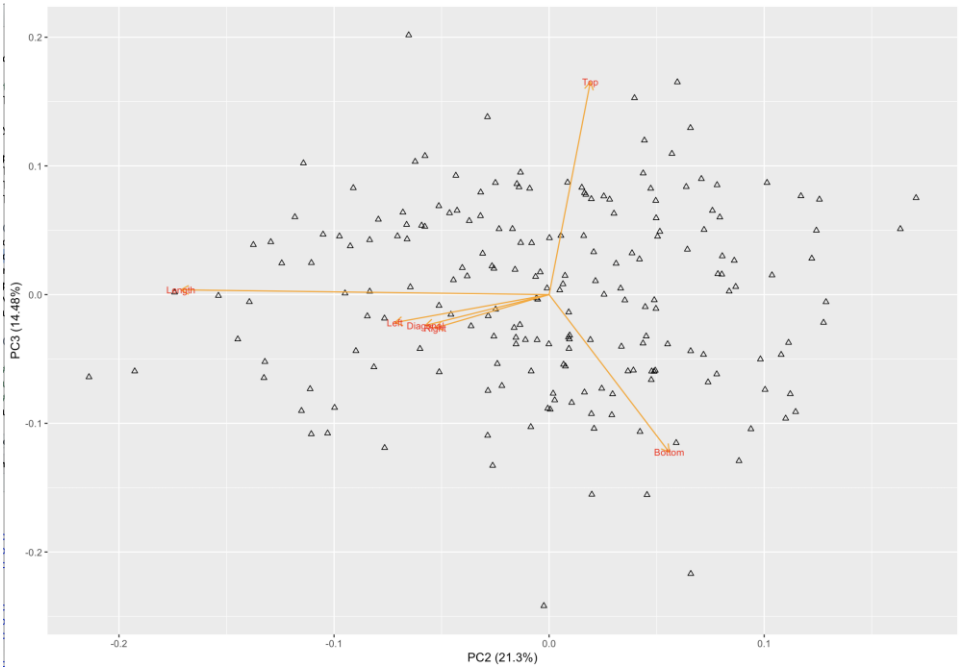
For overall notes

PCA for full data set, Plot PC1, PC2

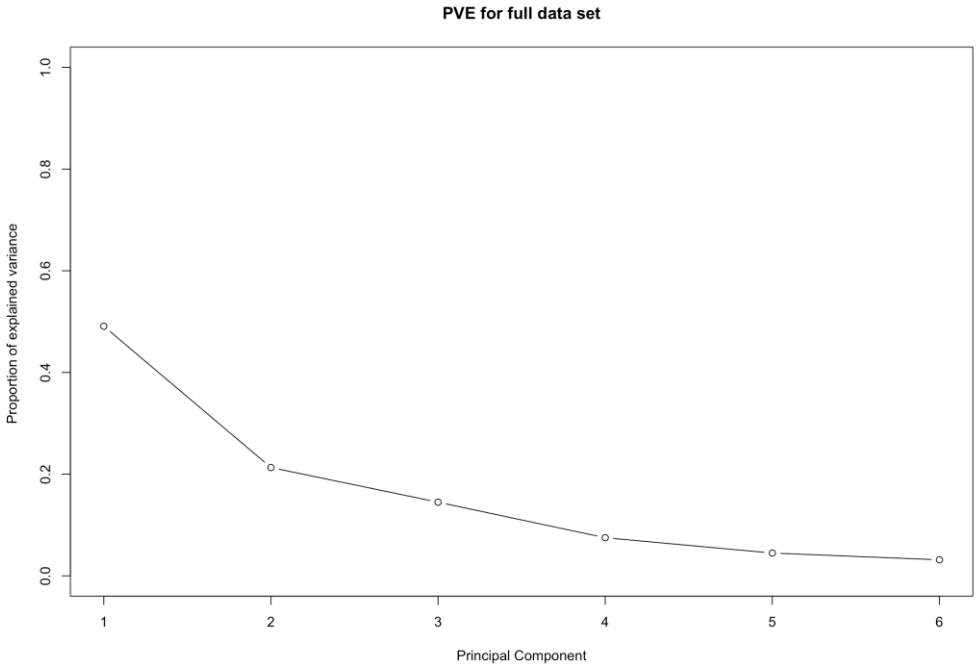


plot 2

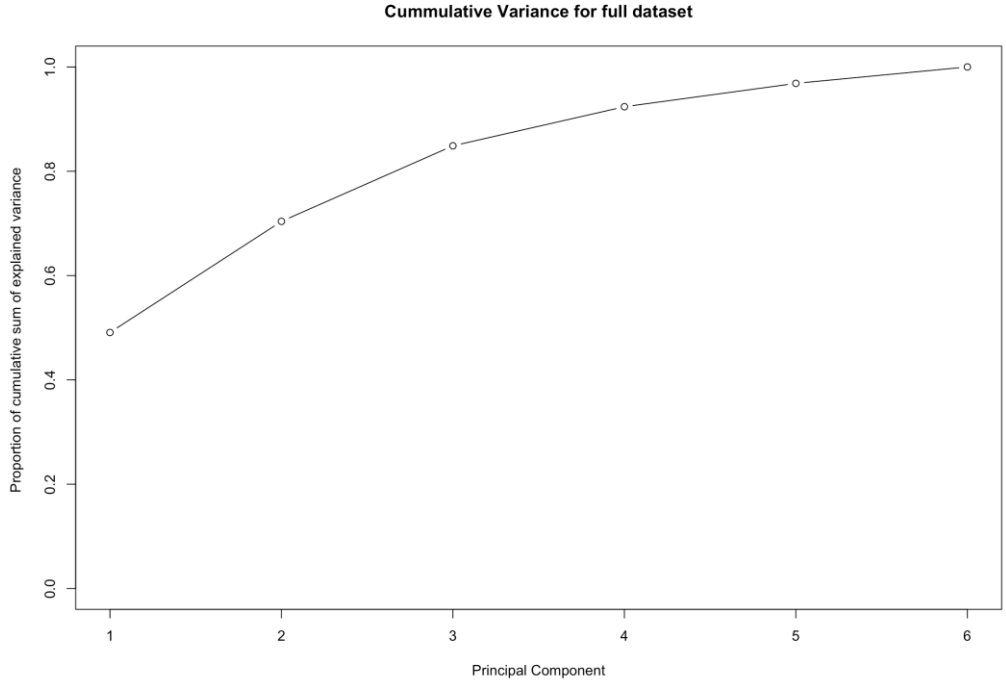
PCA for full data set, Plot PC2, PC3



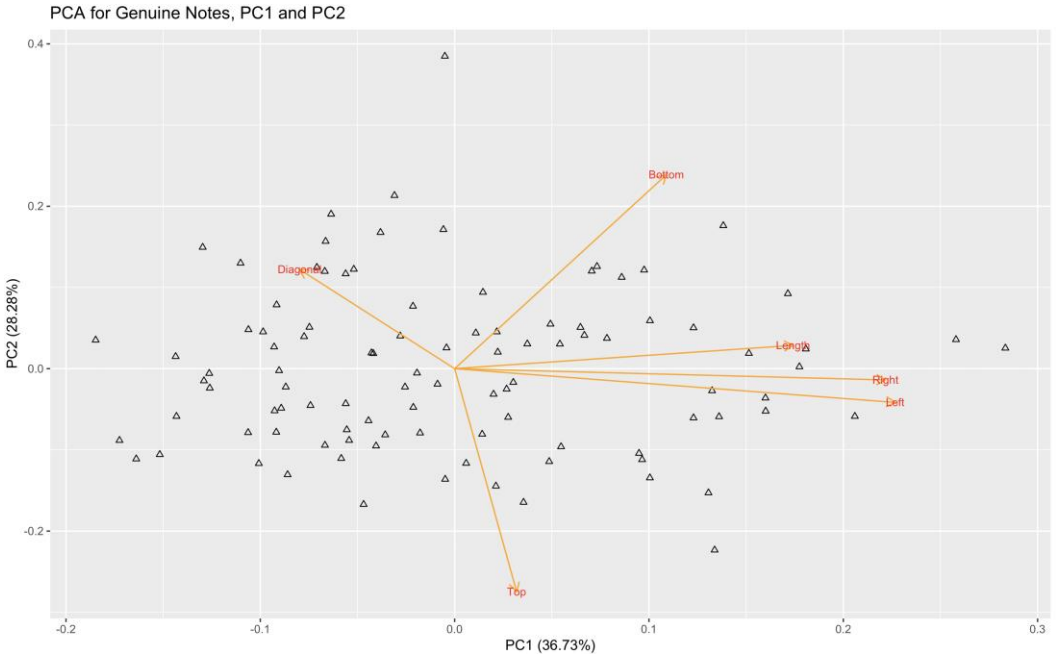
Plot of Proportion of explained variance for each principal component – We can observe that it is highest at PC1 and reducing as the number of principal components are increasing

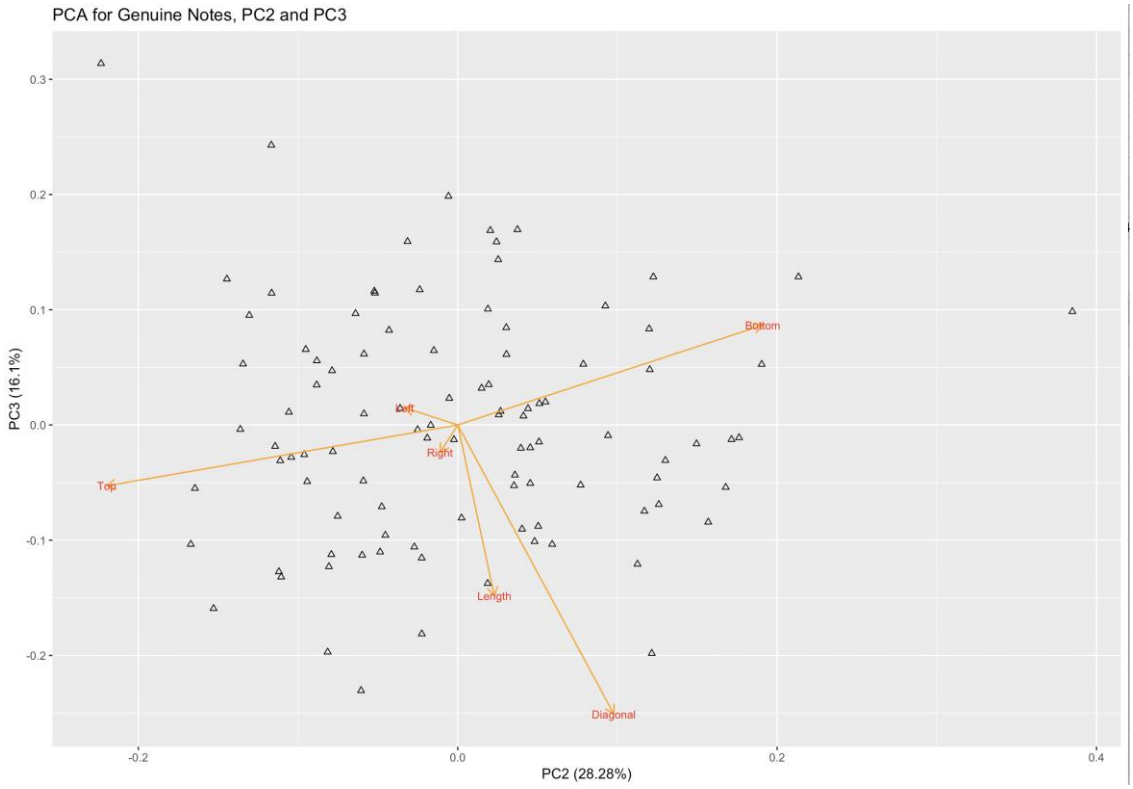


Plot of cumulative sum of Proportion of explained variance for each principal component- We can observe that the cumulative sum of variance is increasing with increase in number of principal components and it is equal to 1 at PC6.

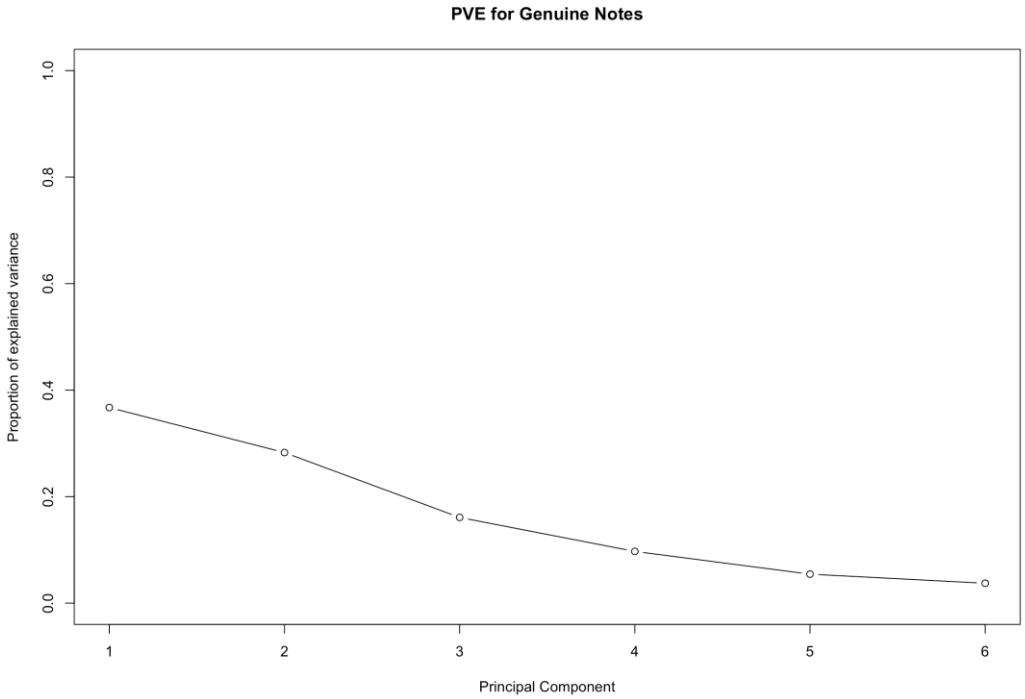


For Genuine Notes

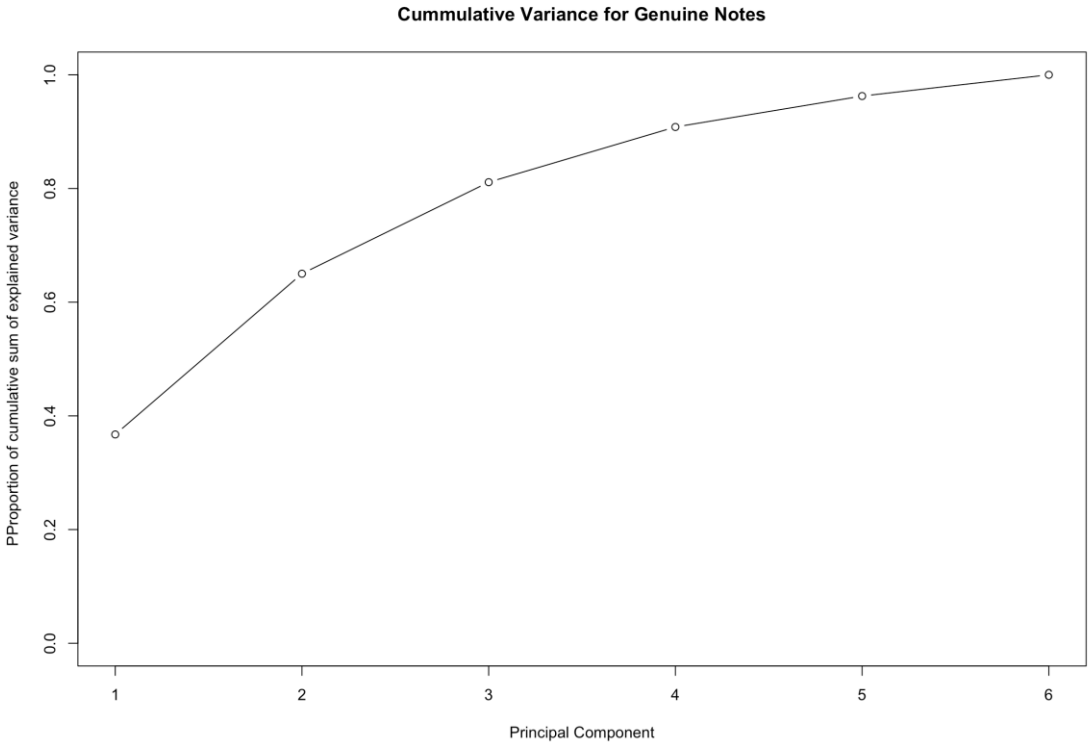




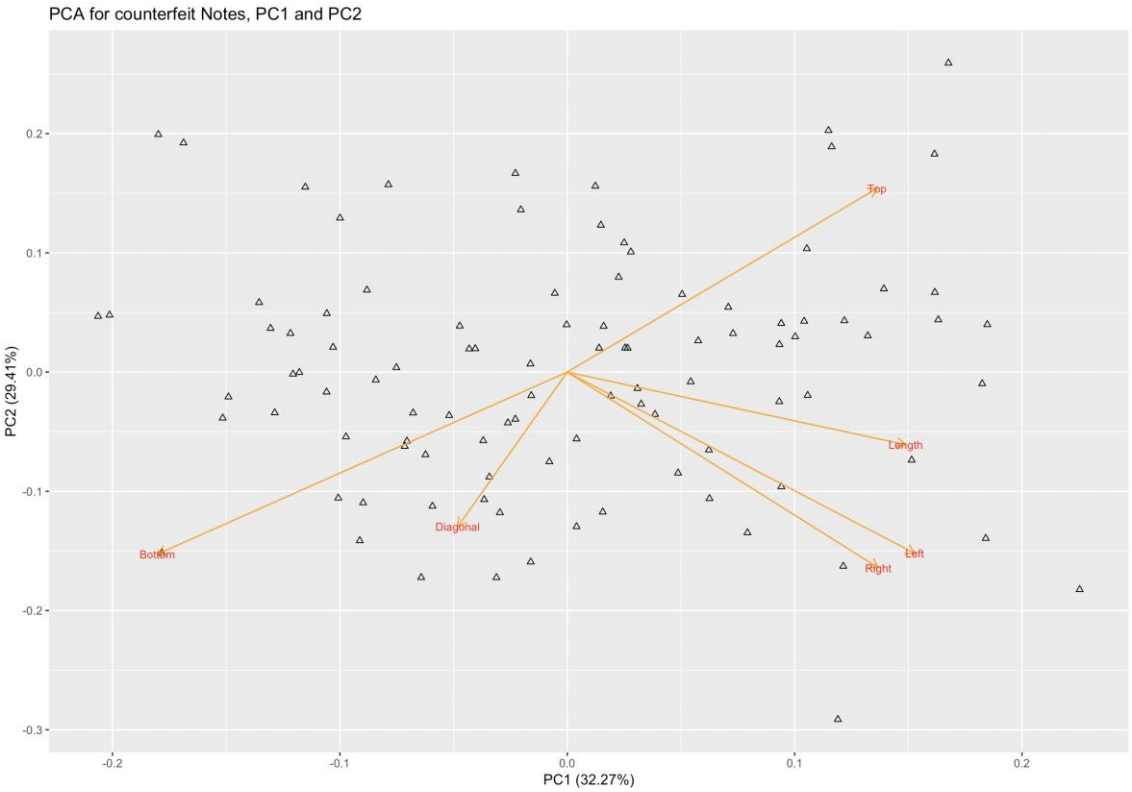
Plot of Proportion of explained variance for each principal component

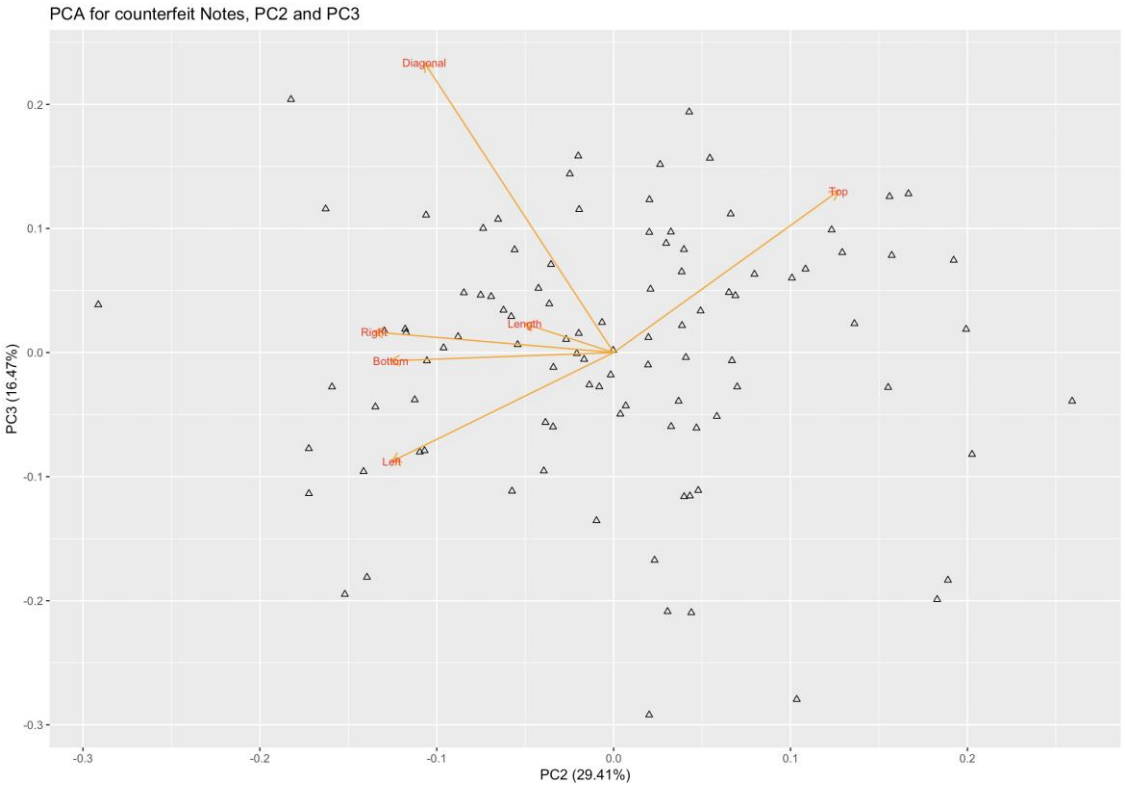


Plot of cumulative sum of Proportion of explained variance for each principal component:

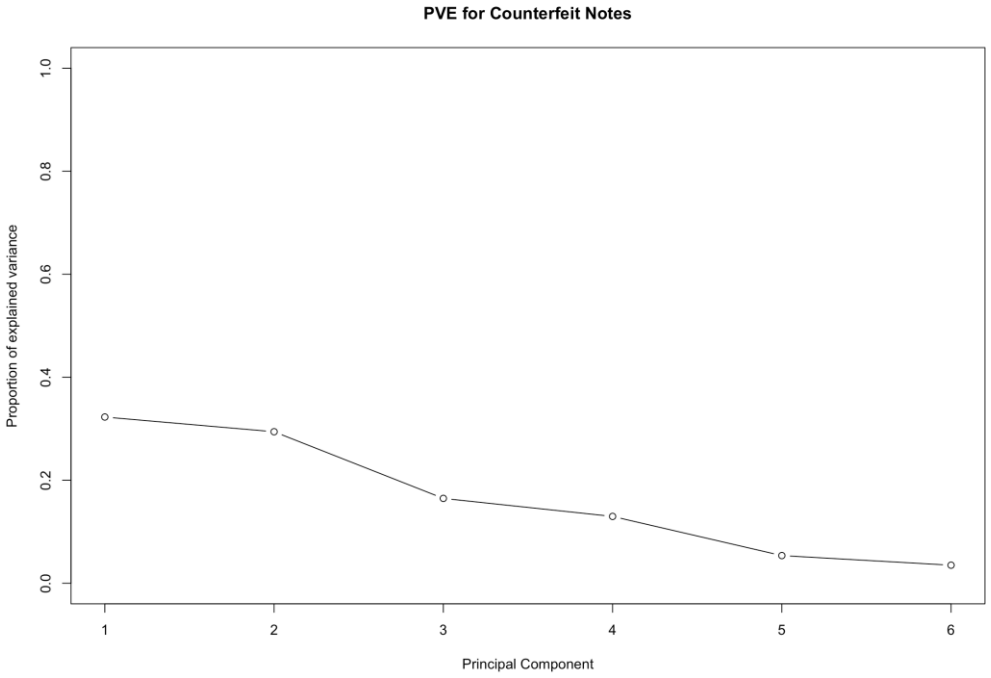


PCR for Counterfeit Notes:

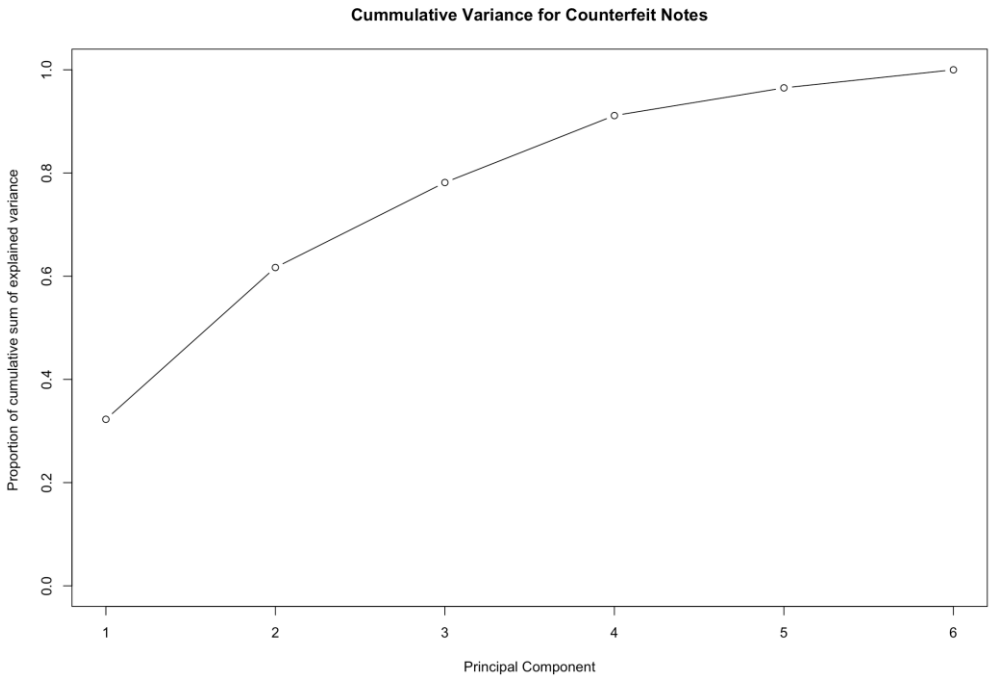




Plot of Proportion of explained variance for each principal component



Plot of cumulative sum of Proportion of explained variance for each principal component:



we can infer that the Genuine notes have more important features as the 5th component explains 10% of the variance in data.

For Fake notes we can see that only two components are dominant.