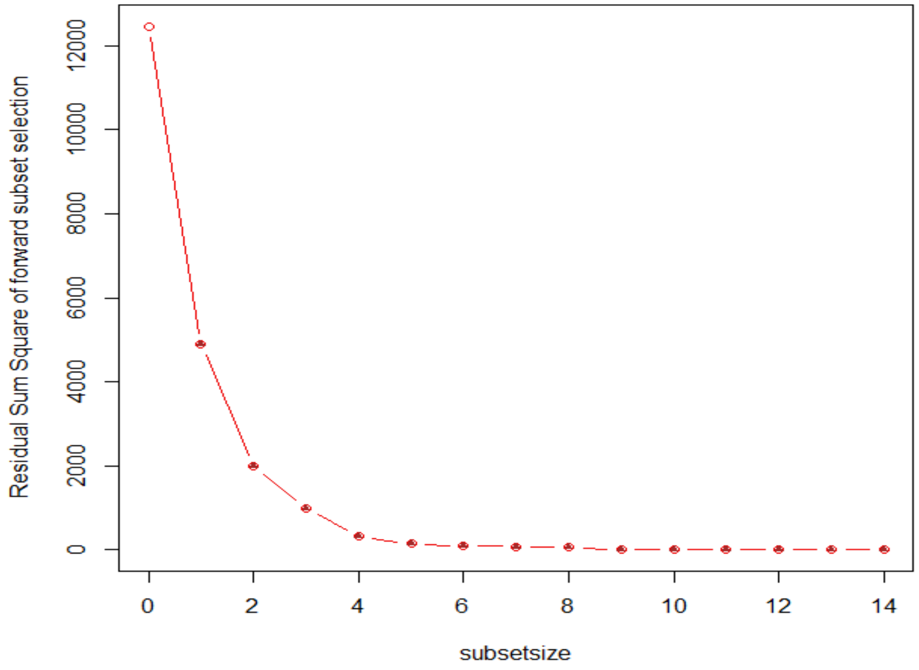
Task1:Visualizing and plotting cereal data set

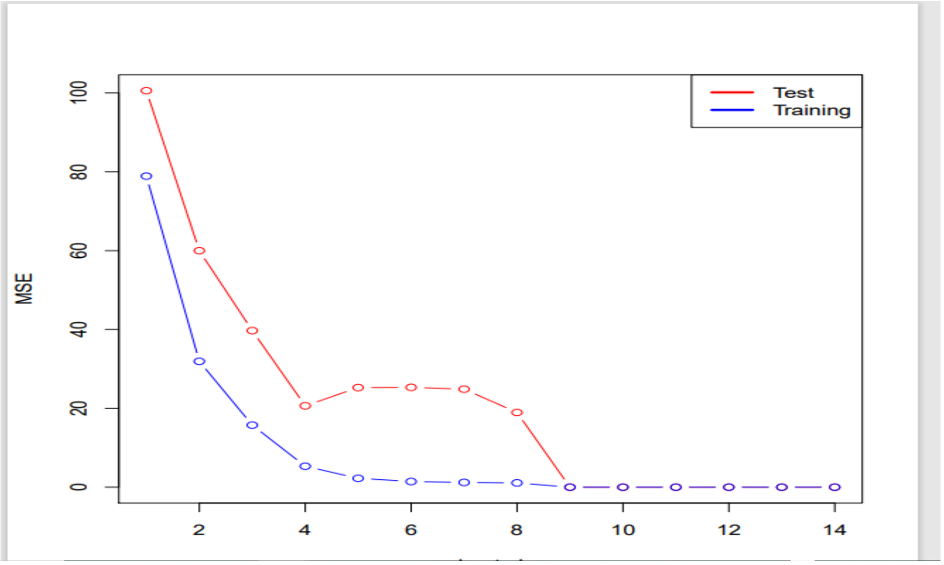
mean(model_linear_regression$residuals^2)= 6.361827e-14

Residual Sum Square Error Plot of Forward subset selection

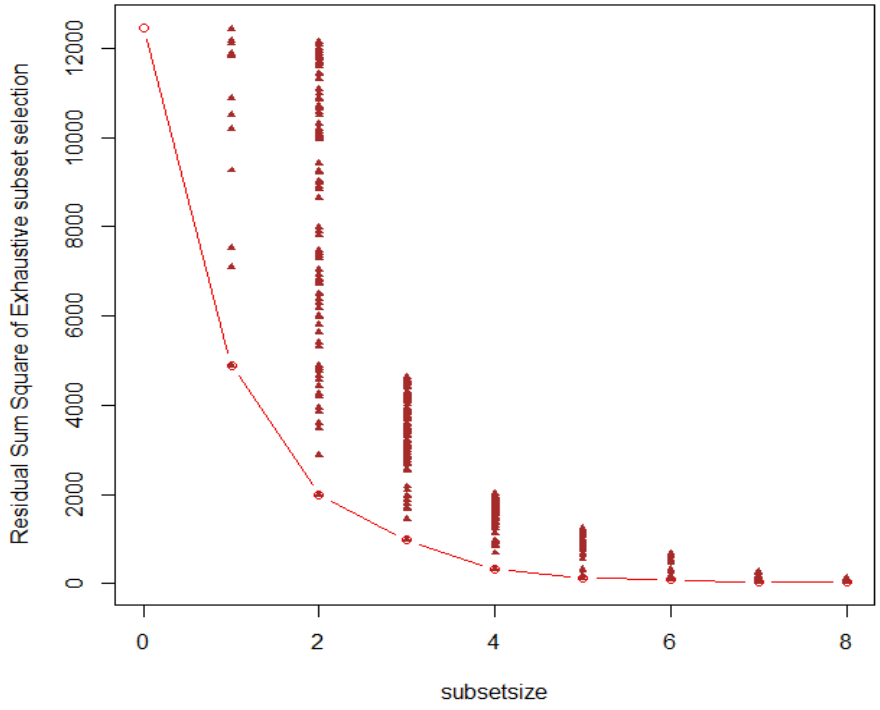


From the above plot of residual error its quite indicative that the residual error is least for a predictor value of 6.Hence there are six such variables which fits the model very well .

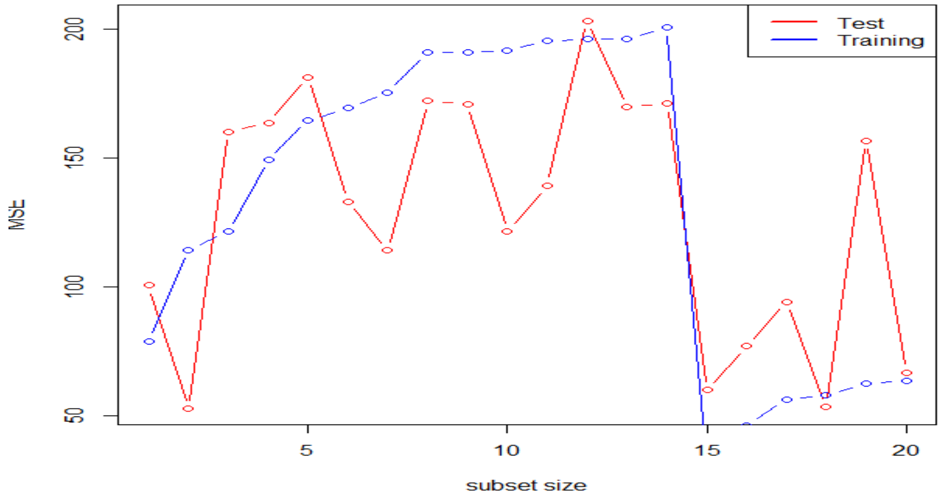
MSE of Forward Subset Selection



From the above plot it seems that though the MSE error of training is decreasing with increase in subset size however for the test set its quite unusual in the sense that it has got its peak value for a subset size of 2 and then it sharply falls at subset size=4 and from there it forms a increasing curve again at a subset size of 9 it starts stabilizing



From the above plot of residual error its quite indicative that the residual error has decreased and quite stable for a subset size of 8.

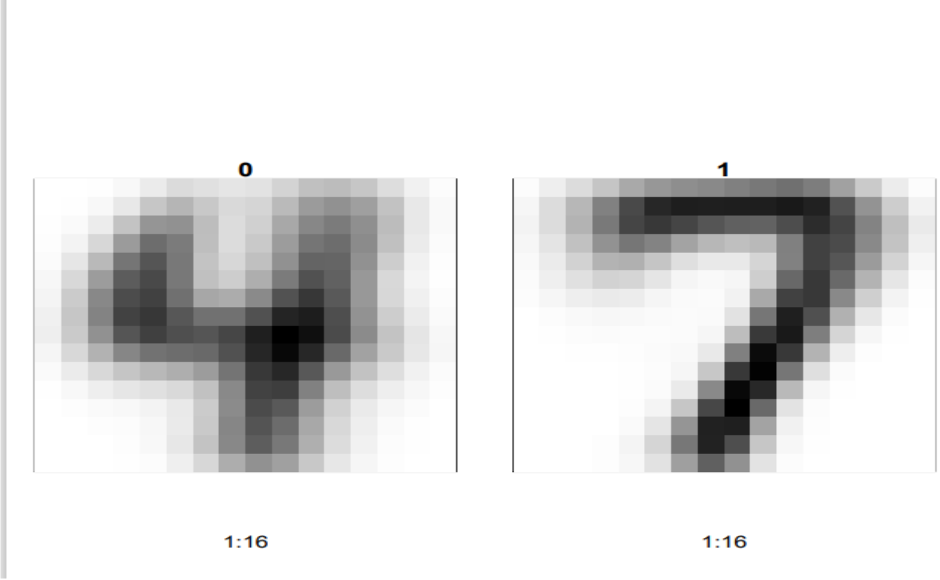


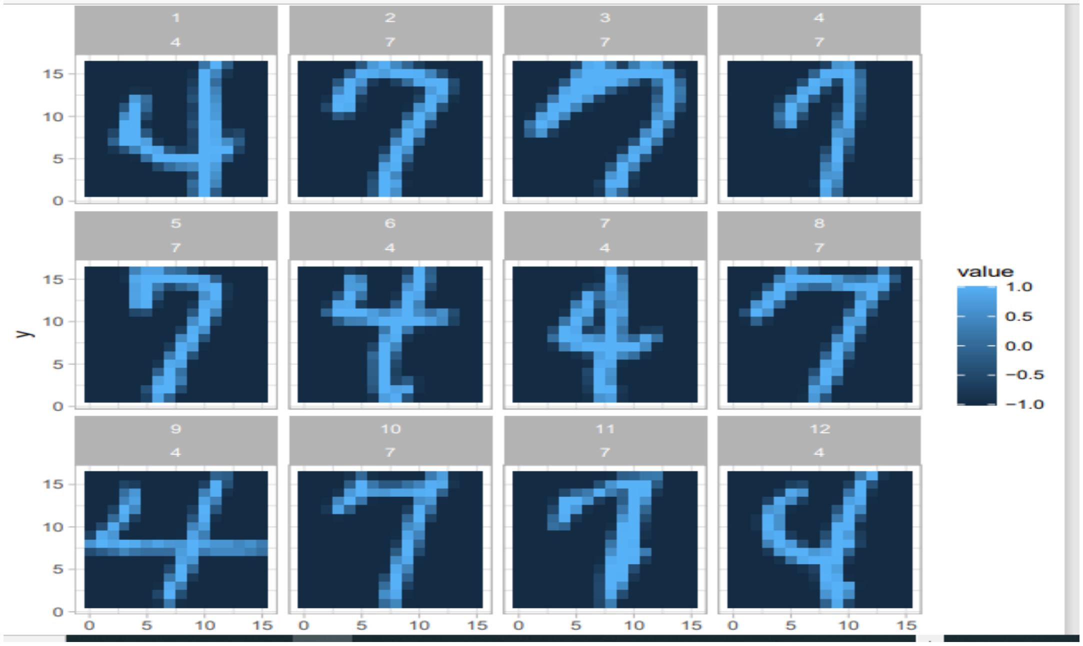
From the above plot of mean square error, it's quite indicative that the MSE for both test a has decreased and quite stable for a subset size of 20.

On comparing the three models it seems Multiple Linear Regression is reporting minimum error compared to other two models, but in real cases best subset selection provides minimum MSE. It may be due to various reasons such as the data split was not good(though other strategies were also used to split the data properly) .If we look at the residual plot and MSE of other two models that is exhaustive subset selection and forward subset selection bot of them are performing well taking under consideration that both the models are predicting the best results on the basis of certain subsets of variables or predictors which may or may not be common to them.

Task2: Visualization and cleaning of data for Handwritten Digit Recognition

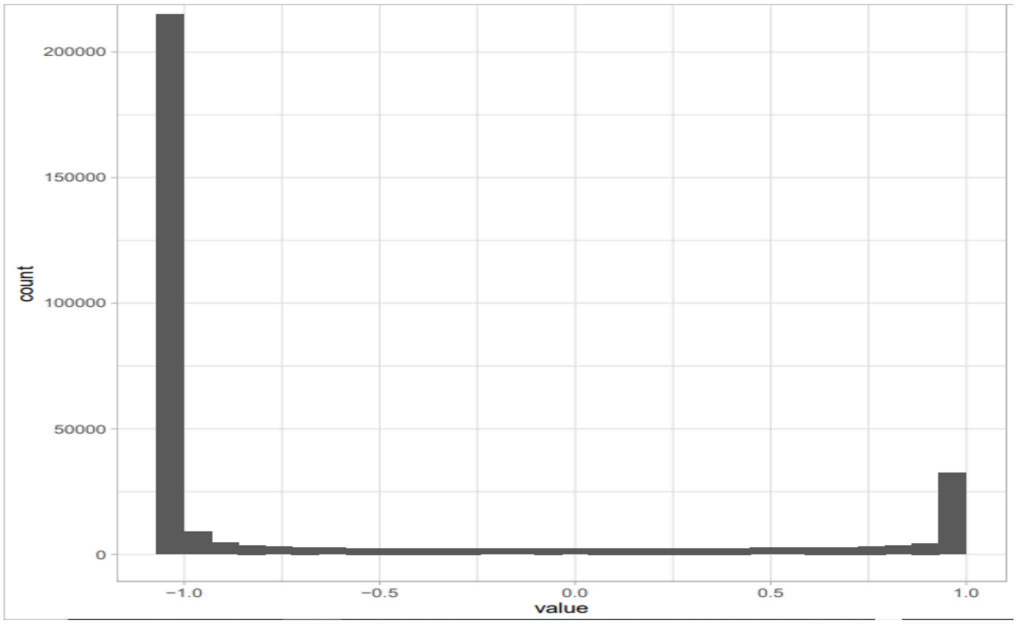
Display digits of training data set (An average of each digit)





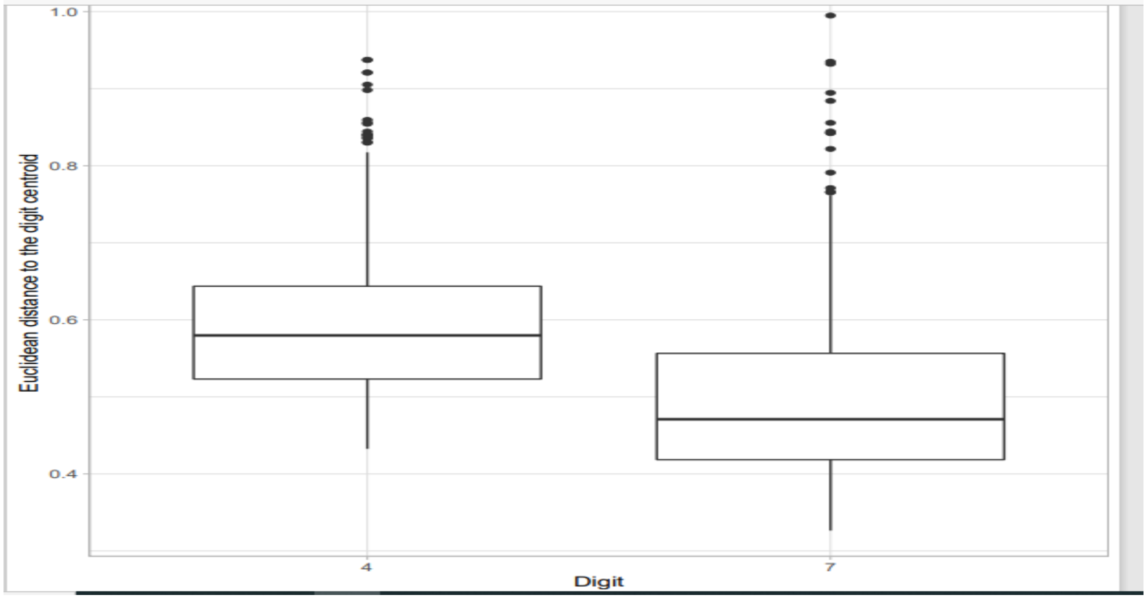
In this figure we have considered each pixel as an observation that is we have one row for each pixel in an image. This is a useful format because it lets us visualize the data along the way. For example, we can visualize the first 12 instances with a couple lines of ggplot2.

Exploring the pixel data:



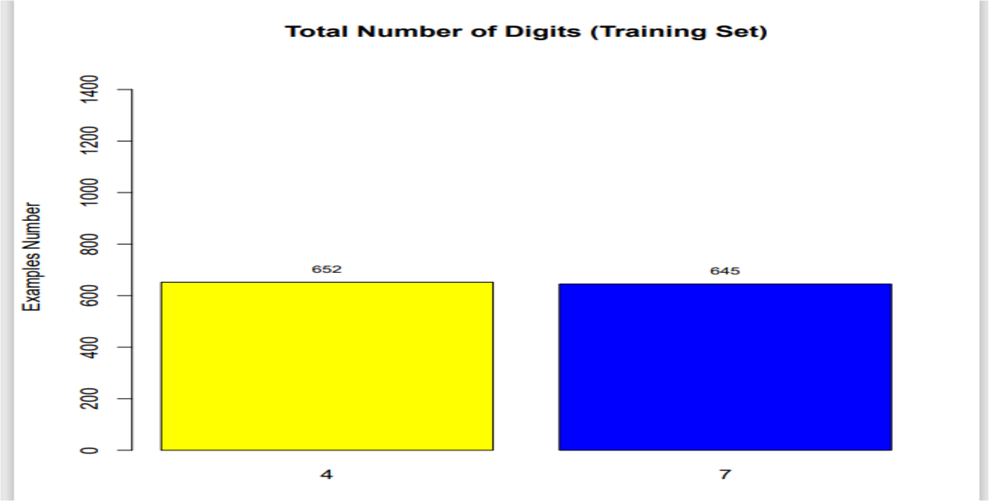
Most pixels in the dataset are completely white, along with another set of pixels that are completely dark, with relatively few in between. This gives us a hint for later feature engineering steps: if we wanted to, we could probably replace each pixel with a binary 0 or 1 with very little loss of information.

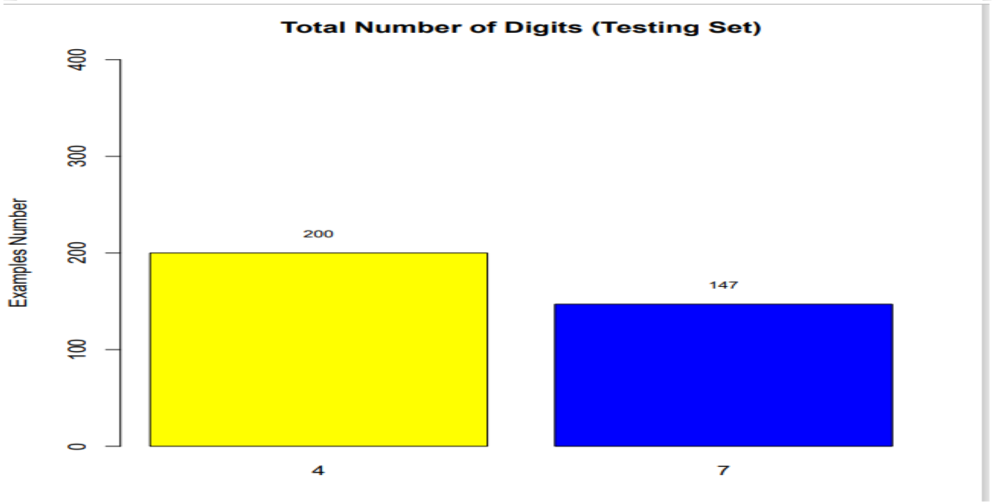
Further exploration:



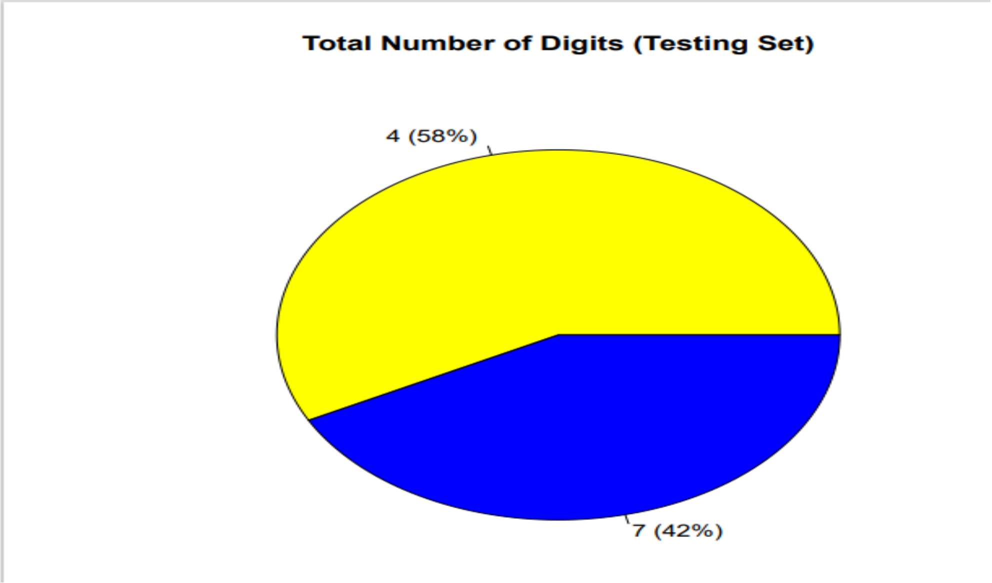
It looks like 7s have comparatively low distances to their centroid: for the most part there's not a ton of variability in how people draw that digit.

Histogram display of training and testing set of zi pcode data

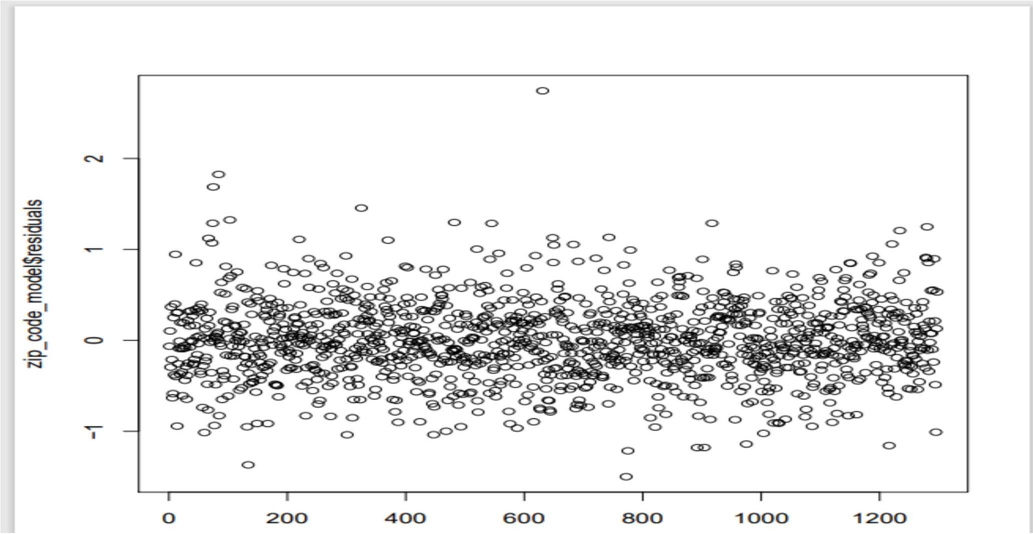




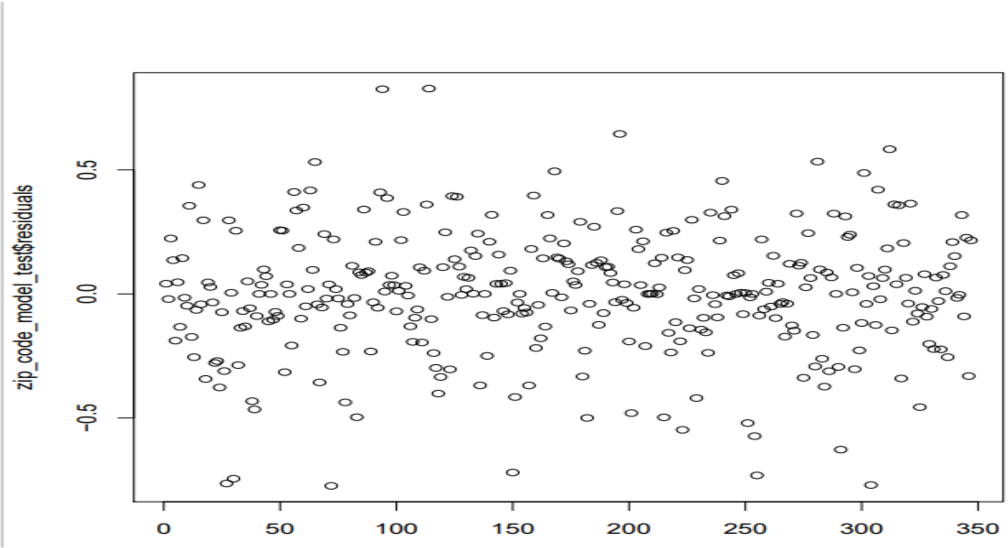
Pie Plot of test data:



Residual error plot in **training data** of Zipcode after trained with linear regression model

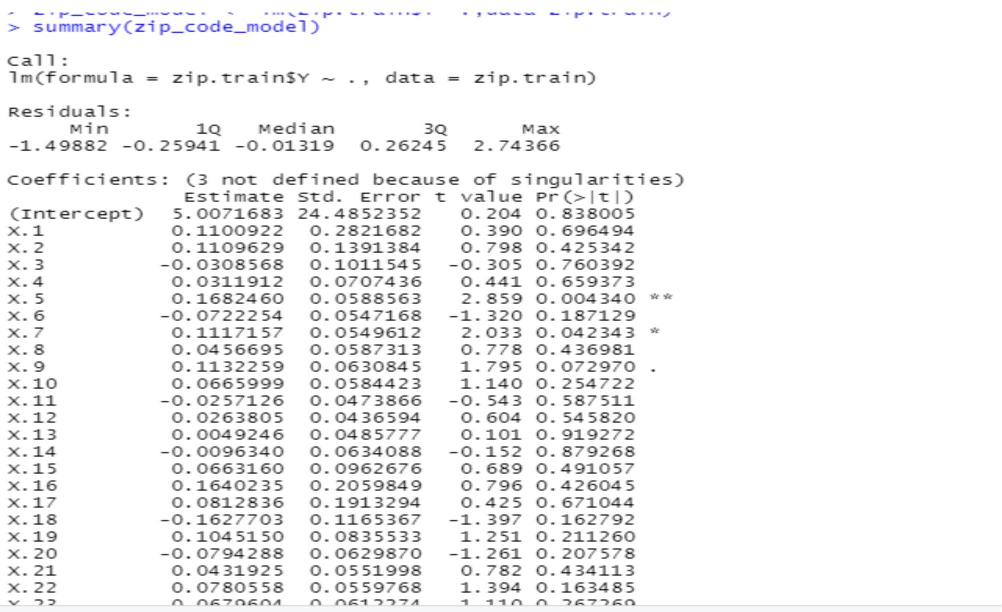


Residual error plot in **test data** of Zipcode after trained with linear regression model



Outcome: There is no such pattern in the residuals and all the data points are scattered in the both the above plot which is again a good indication that the model is fitting well.

Summary of linear regression model of zip code data

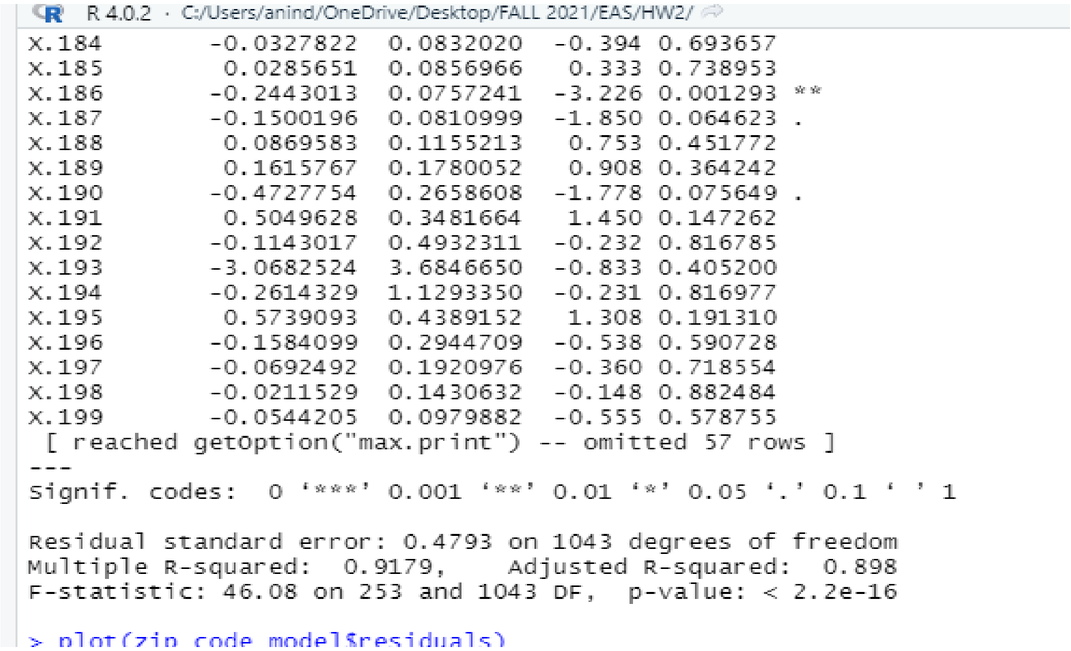


```
> zip_code_model <- lm(zip.train$Y ~.,data=zip.train)
> summary(zip_code_model)

Call:
lm(formula = zip.train$Y ~ ., data = zip.train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.49882 -0.25941 -0.01319  0.26245  2.74366 

Coefficients: (3 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  5.0071683 24.4852352   0.204 0.838005   
X.1          0.1100922  0.2821682   0.390 0.696494   
X.2          0.1109629  0.1391384   0.798 0.425342   
X.3         -0.0308568  0.1011545  -0.305 0.760392   
X.4          0.0311912  0.0707436   0.441 0.659373   
X.5          0.1682460  0.0588563   2.859 0.004340 **
X.6         -0.0722254  0.0547168  -1.320 0.187129   
X.7          0.1117157  0.0549612   2.033 0.042343 * 
X.8          0.0456695  0.0587313   0.778 0.436981   
X.9          0.1132259  0.0630845   1.795 0.072970 . 
X.10         0.0665999  0.0584423   1.140 0.254722   
X.11        -0.0257126  0.0473866  -0.543 0.587511   
X.12         0.0263805  0.0436594   0.604 0.545820   
X.13         0.0049246  0.0485777   0.101 0.919272   
X.14        -0.0096340  0.0634088  -0.152 0.879268   
X.15         0.0663160  0.0962676   0.689 0.491057   
X.16         0.1640235  0.2059849   0.796 0.426045   
X.17         0.0812836  0.1913294   0.425 0.671044   
X.18        -0.1627703  0.1165367  -1.397 0.162792   
X.19         0.1045150  0.0835533   1.251 0.211260   
X.20        -0.0794288  0.0629870  -1.261 0.207578   
X.21         0.0431925  0.0551998   0.782 0.434113   
X.22         0.0780558  0.0559768   1.394 0.163485   
X.23         0.0679604  0.0612274   1.110 0.267269   
```



R 4.0.2 · C:/Users/anind/OneDrive/Desktop/FALL 2021/EAS/HW2/

```
X.184       -0.0327822  0.0832020  -0.394 0.693657   
X.185        0.0285651  0.0856966   0.333 0.738953   
X.186       -0.2443013  0.0757241  -3.226 0.001293 **
X.187       -0.1500196  0.0810999  -1.850 0.064623 . 
X.188        0.0869583  0.1155213   0.753 0.451772   
X.189        0.1615767  0.1780052   0.908 0.364242   
X.190       -0.4727754  0.2658608  -1.778 0.075649 . 
X.191        0.5049628  0.3481664   1.450 0.147262   
X.192       -0.1143017  0.4932311  -0.232 0.816785   
X.193       -3.0682524  3.6846650  -0.833 0.405200   
X.194       -0.2614329  1.1293350  -0.231 0.816977   
X.195        0.5739093  0.4389152   1.308 0.191310   
X.196       -0.1584099  0.2944709  -0.538 0.590728   
X.197       -0.0692492  0.1920976  -0.360 0.718554   
X.198       -0.0211529  0.1430632  -0.148 0.882484   
X.199       -0.0544205  0.0979882  -0.555 0.578755   
 [ reached getOption("max.print") -- omitted 57 rows ]
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4793 on 1043 degrees of freedom
Multiple R-squared:  0.9179,	Adjusted R-squared:  0.898 
F-statistic: 46.08 on 253 and 1043 DF,  p-value: < 2.2e-16

> plot(zip_code_model$residuals)
```

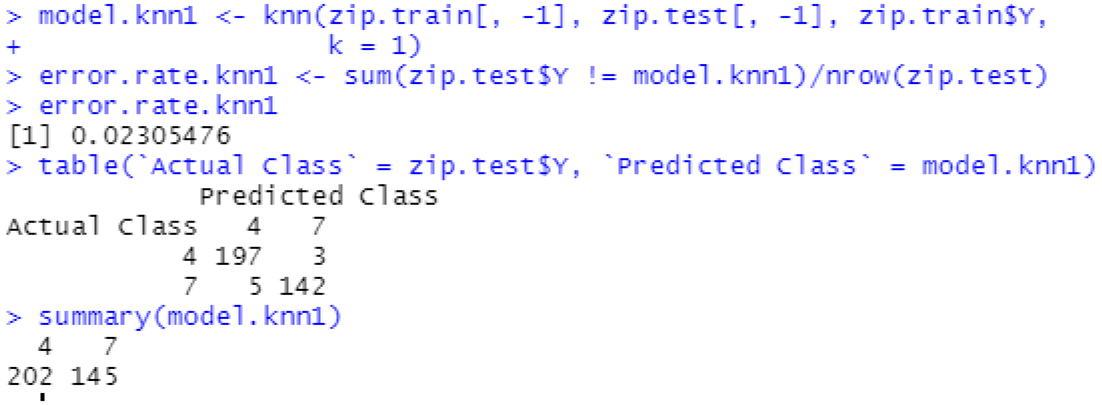
Analysis from the summary of multilinear regression model

1)The median, 1st quartile,3rd quartile is quite close to 0, Max and Min are close to 0 overall the but still the model is fitting well

2)Adjusted R square is 0.89 which signifies that 89 percent of the variance in Rating Data is explained by the predictors.

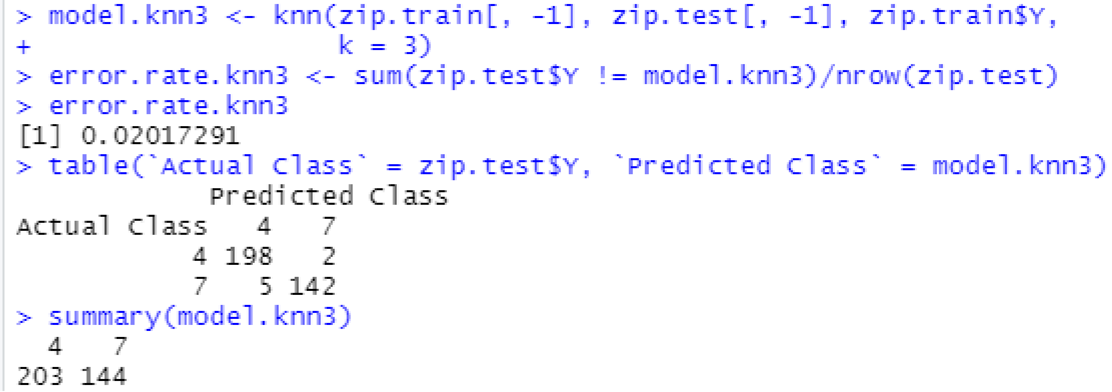
Fitting and plotting using K nearest neighbors using different values of k.

K=1



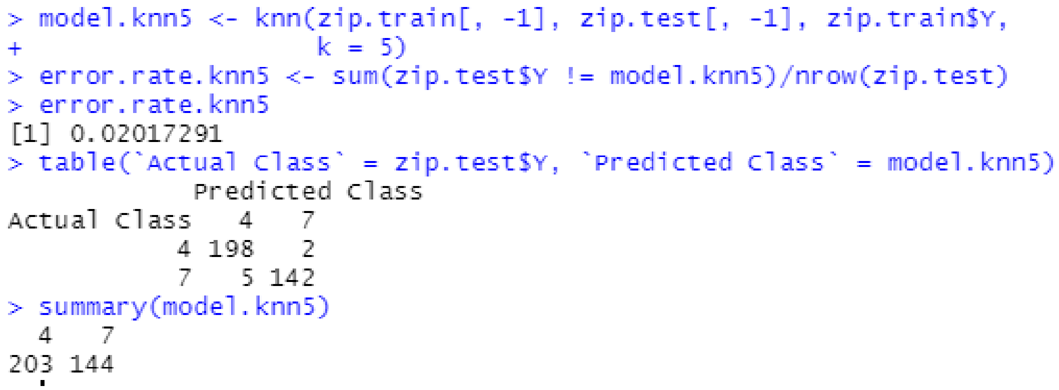
```
> model.knn1 <- knn(zip.train[, -1], zip.test[, -1], zip.train$Y,
+                  k = 1)
> error.rate.knn1 <- sum(zip.test$Y != model.knn1)/nrow(zip.test)
> error.rate.knn1
[1] 0.02305476
> table(`Actual Class` = zip.test$Y, `Predicted Class` = model.knn1)
            Predicted Class
Actual Class   4   7
           4 197   3
           7   5 142
> summary(model.knn1)
  4   7 
202 145 
```

K=3



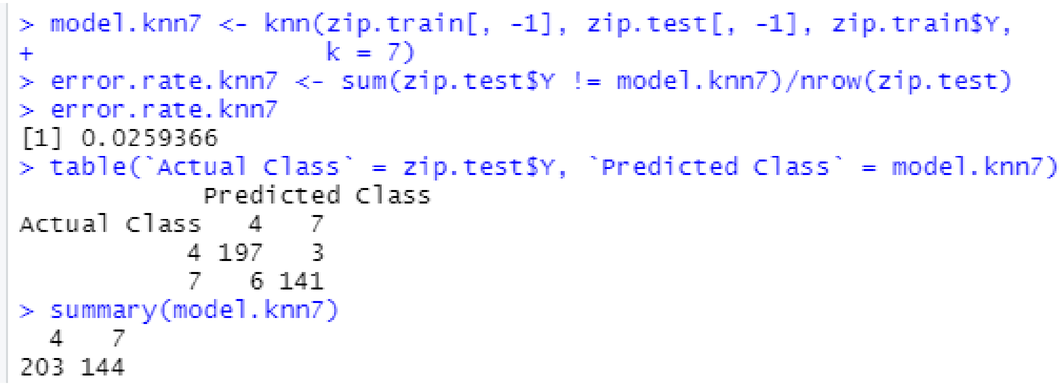
```
> model.knn3 <- knn(zip.train[, -1], zip.test[, -1], zip.train$Y,
+                  k = 3)
> error.rate.knn3 <- sum(zip.test$Y != model.knn3)/nrow(zip.test)
> error.rate.knn3
[1] 0.02017291
> table(`Actual Class` = zip.test$Y, `Predicted Class` = model.knn3)
            Predicted Class
Actual Class   4   7
           4 198   2
           7   5 142
> summary(model.knn3)
  4   7 
203 144 
```

K=5



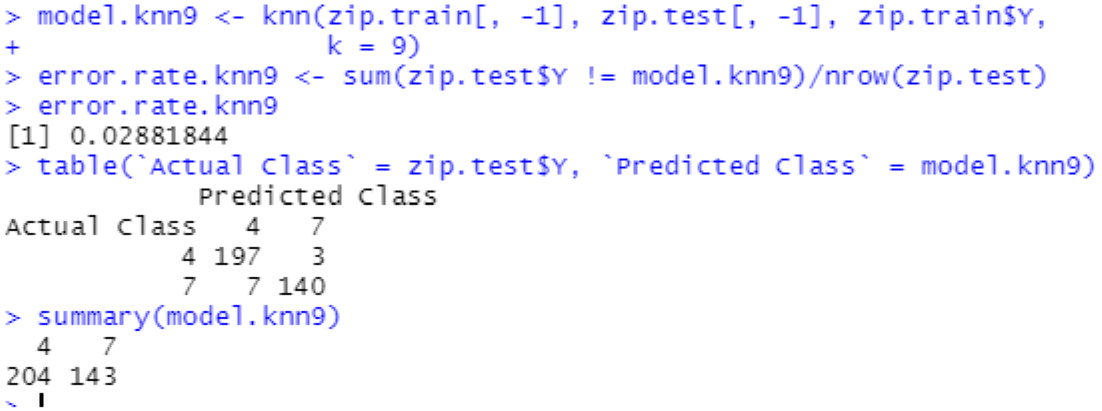
```
> model.knn5 <- knn(zip.train[, -1], zip.test[, -1], zip.train$Y,
+                  k = 5)
> error.rate.knn5 <- sum(zip.test$Y != model.knn5)/nrow(zip.test)
> error.rate.knn5
[1] 0.02017291
> table(`Actual Class` = zip.test$Y, `Predicted Class` = model.knn5)
            Predicted Class
Actual Class   4   7
           4 198   2
           7   5 142
> summary(model.knn5)
  4   7 
203 144 
```

K=7



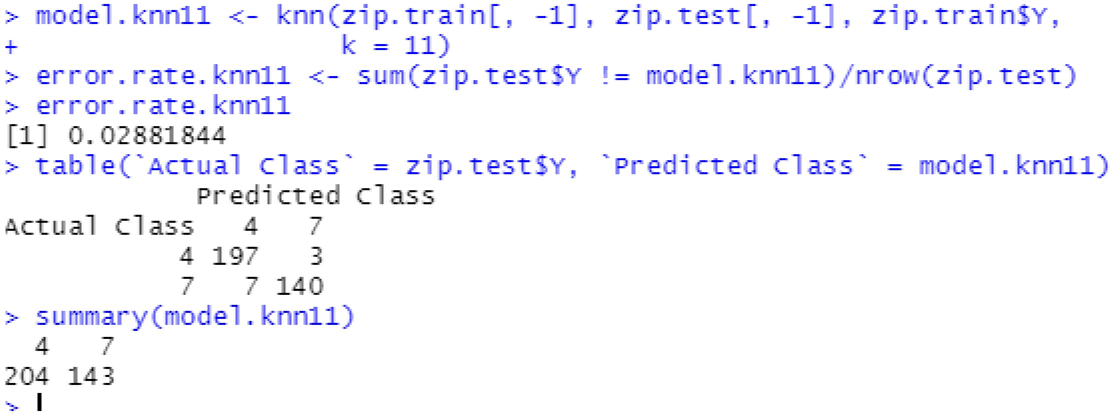
```
> model.knn7 <- knn(zip.train[, -1], zip.test[, -1], zip.train$Y,
+                    k = 7)
> error.rate.knn7 <- sum(zip.test$Y != model.knn7)/nrow(zip.test)
> error.rate.knn7
[1] 0.0259366
> table(`Actual Class` = zip.test$Y, `Predicted Class` = model.knn7)
            Predicted Class
Actual Class   4   7
           4 197   3
           7   6 141
> summary(model.knn7)
  4   7 
203 144 
```

K=9



```
> model.knn9 <- knn(zip.train[, -1], zip.test[, -1], zip.train$Y,
+                    k = 9)
> error.rate.knn9 <- sum(zip.test$Y != model.knn9)/nrow(zip.test)
> error.rate.knn9
[1] 0.02881844
> table(`Actual Class` = zip.test$Y, `Predicted Class` = model.knn9)
            Predicted Class
Actual Class   4   7
           4 197   3
           7   7 140
> summary(model.knn9)
  4   7 
204 143 
> 
```

K=11



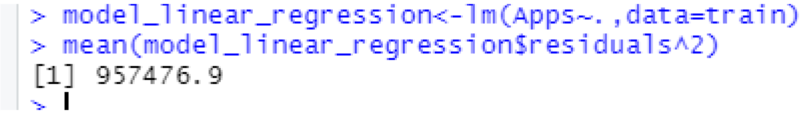
```
> model.knn11 <- knn(zip.train[, -1], zip.test[, -1], zip.train$Y,
+                     k = 11)
> error.rate.knn11 <- sum(zip.test$Y != model.knn11)/nrow(zip.test)
> error.rate.knn11
[1] 0.02881844
> table(`Actual Class` = zip.test$Y, `Predicted Class` = model.knn11)
            Predicted Class
Actual Class   4   7
           4 197   3
           7   7 140
> summary(model.knn11)
  4   7 
204 143 
> 
```

Going through each of this tables it's obvious that the testing error is minimum for K=3 and K=5

Hence its neither underfitting with nor overfitting with K=3 or K=5 compared to the other lower values and higher values of K thus giving us an optimum solution. Unlike training data where generally with increase of K the error increases because its underfitting the data and with smaller values of K generally its overfitting. However, the same doesn't apply for test data as we can see for very small values as well as for very large values of K the error is comparatively higher in this case

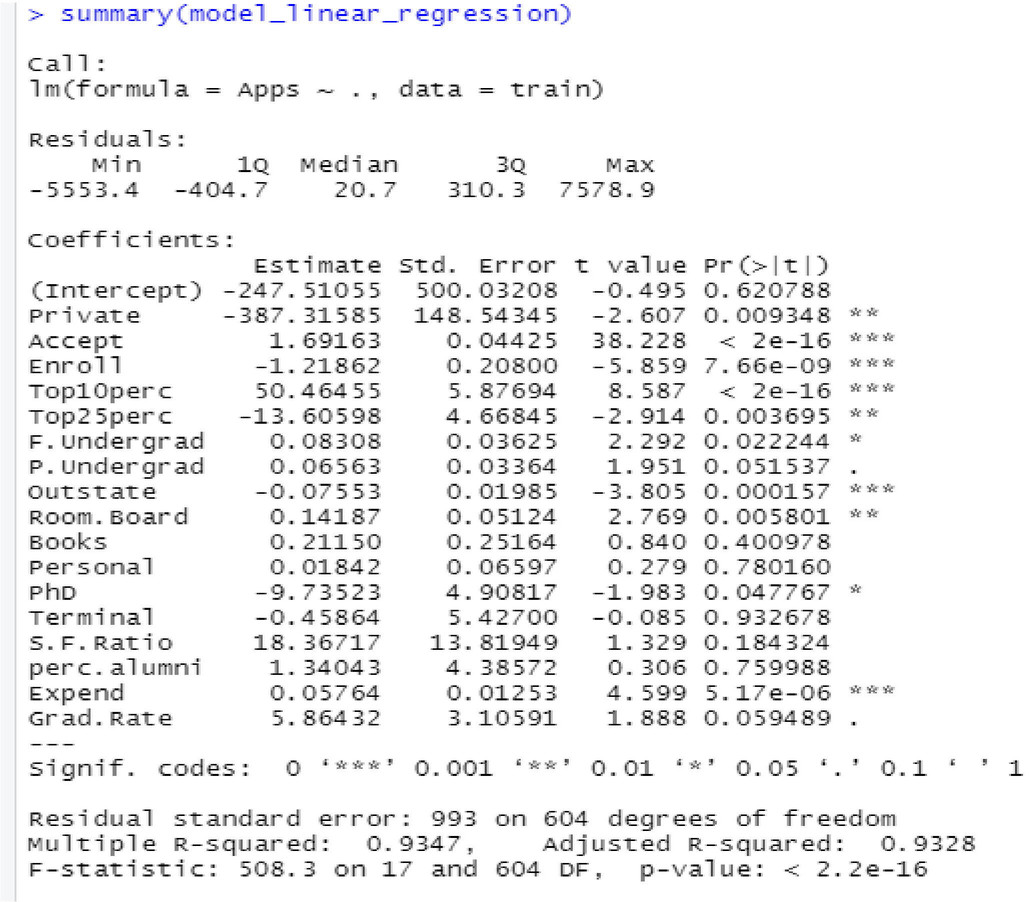
Task3:

After fitting the college data with multiple linear regression model the mean square error is:



```
> model_linear_regression<-lm(Apps~.,data=train)
> mean(model_linear_regression$residuals^2)
[1] 957476.9
> 
```

Following is the summary of linear model:



```
> summary(model_linear_regression)

Call:
lm(formula = Apps ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-5553.4  -404.7    20.7   310.3  7578.9 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -247.51055  500.03208  -0.495 0.620788    
Private     -387.31585  148.54345  -2.607 0.009348 ** 
Accept         1.69163    0.04425  38.228  < 2e-16 ***
Enroll        -1.21862    0.20800  -5.859 7.66e-09 ***
Top10perc     50.46455    5.87694   8.587  < 2e-16 ***
Top25perc    -13.60598    4.66845  -2.914 0.003695 ** 
F.Undergrad    0.08308    0.03625   2.292 0.022244 *  
P.Undergrad    0.06563    0.03364   1.951 0.051537 .  
Outstate      -0.07553    0.01985  -3.805 0.000157 ***
Room.Board     0.14187    0.05124   2.769 0.005801 ** 
Books          0.21150    0.25164   0.840 0.400978    
Personal       0.01842    0.06597   0.279 0.780160    
PhD           -9.73523    4.90817  -1.983 0.047767 *  
Terminal      -0.45864    5.42700  -0.085 0.932678    
S.F.Ratio     18.36717   13.81949   1.329 0.184324    
perc.alumni    1.34043    4.38572   0.306 0.759988    
Expend         0.05764    0.01253   4.599 5.17e-06 ***
Grad.Rate      5.86432    3.10591   1.888 0.059489 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 993 on 604 degrees of freedom
Multiple R-squared:  0.9347,	Adjusted R-squared:  0.9328 
F-statistic: 508.3 on 17 and 604 DF,  p-value: < 2.2e-16
```

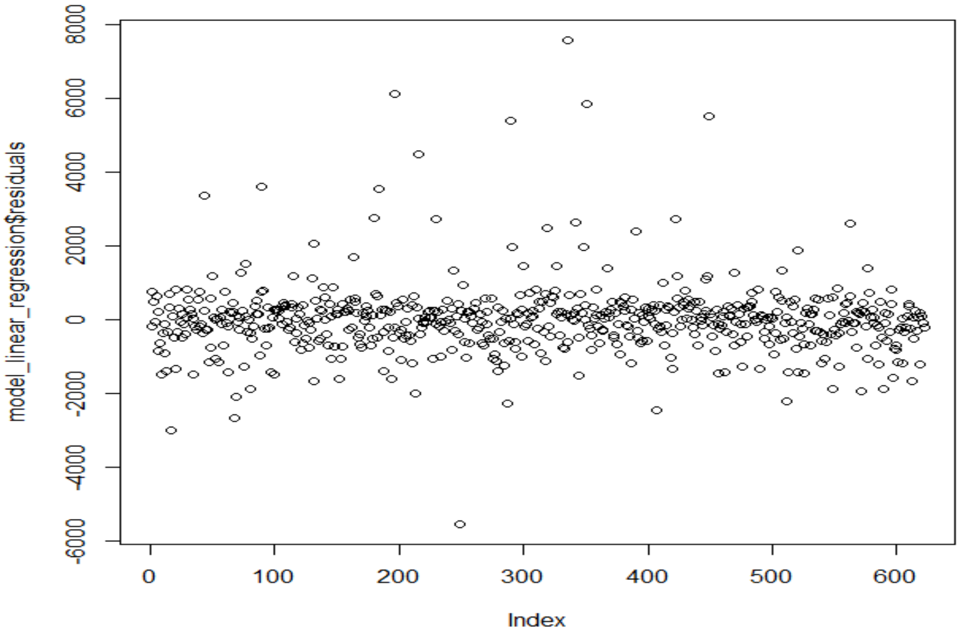
Analysis from the summary of multilinear regression model

1)The median, 1st quartile,3rd quartile is quite close to 0, however Max and Min are not so close to 0 hence overall the parameters seem to be ok but still there is chance of improvement in the model

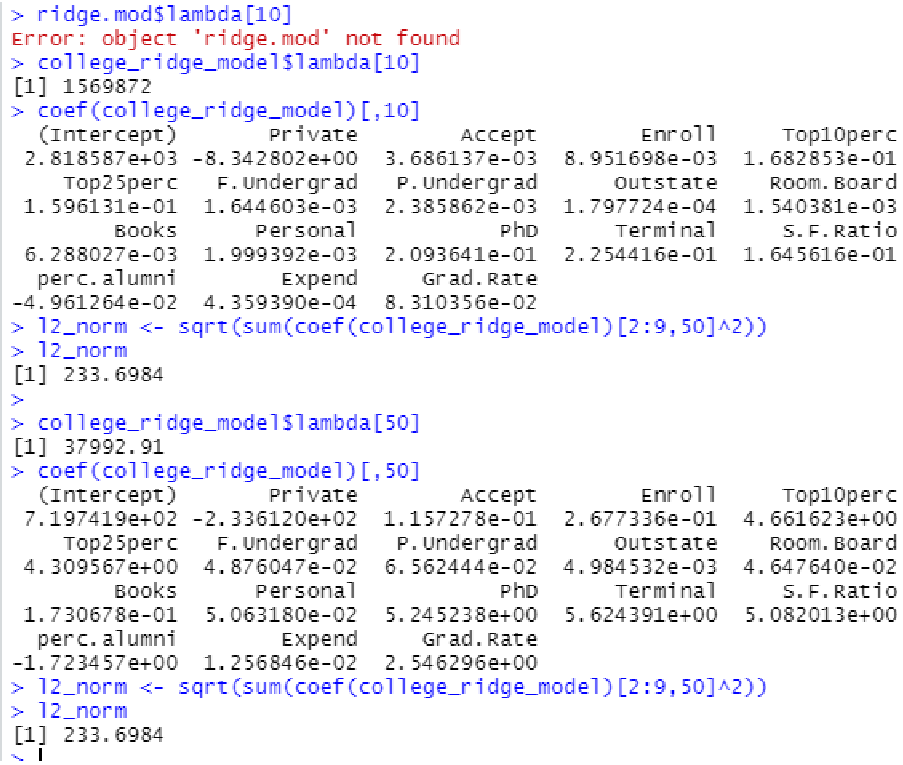
2)Adjusted R square is 0.93 which signifies that 93 percent of the variance in Applications response variable is explained by the predictors.

3)The p value for predictors Private, Accept, Enroll, Top10Perc,Top25Perc,Outstate,Expenditure,Room.board and sodium are <=0.05 indicating that these variables are significant to prediction of response variable that is Applications received per year.

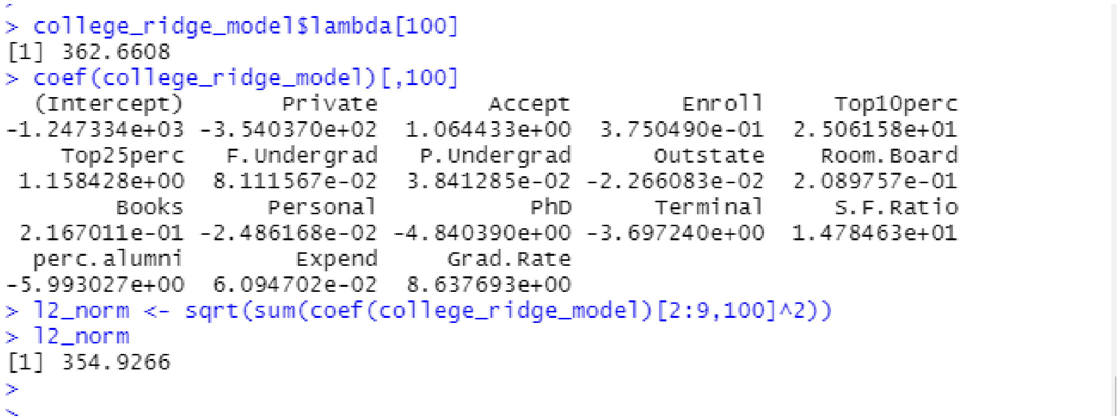
Plot of linear model residuals:



The data point appears to be very clumsy in the residuals which is not so indicative of a good fit ,had it been a good model then the residual plot would have been scattered.

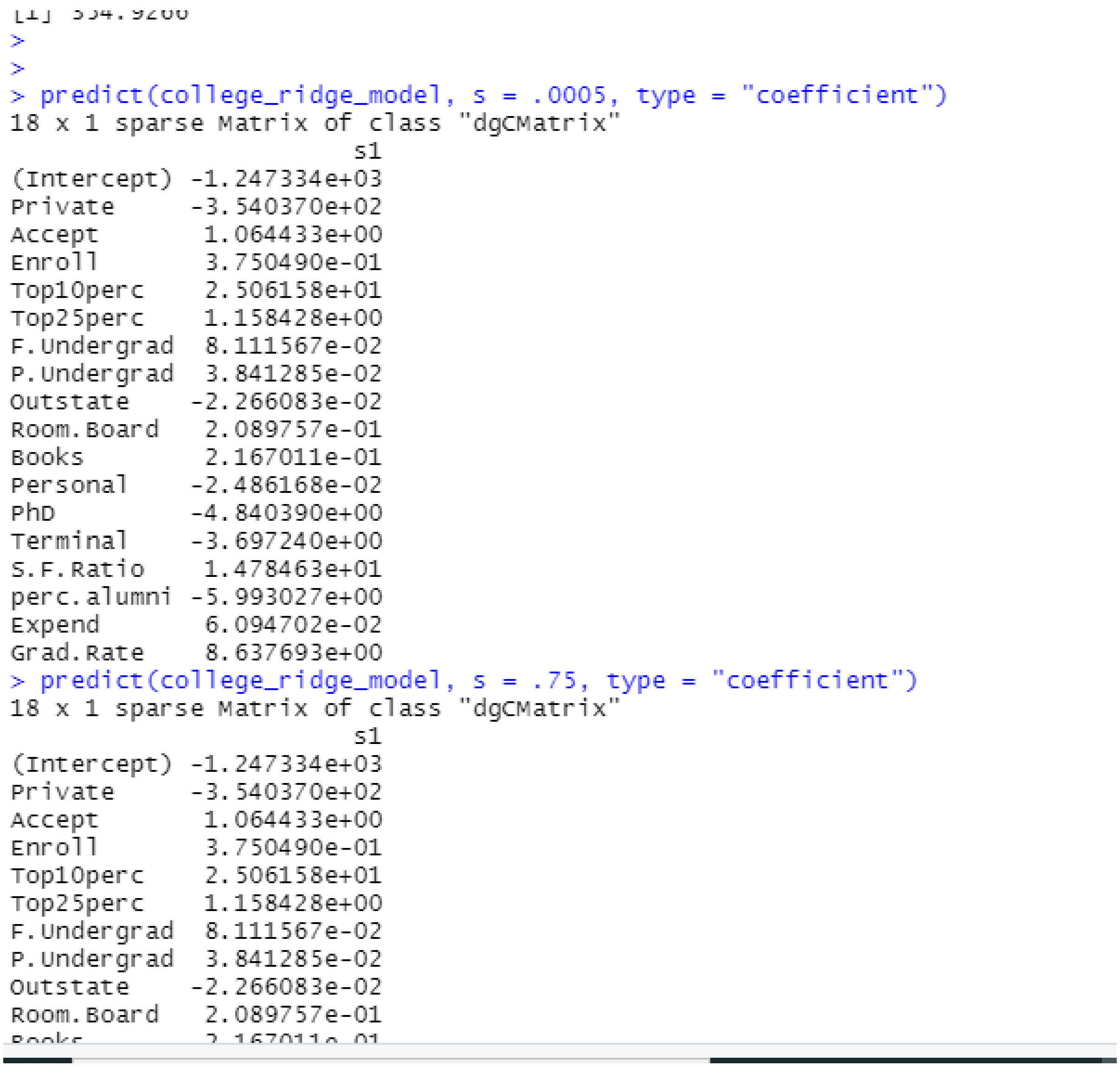


```
> ridge.mod$lambda[10]
Error: object 'ridge.mod' not found
> college_ridge_model$lambda[10]
[1] 1569872
> coef(college_ridge_model)[,10]
  (Intercept)       Private        Accept        Enroll     Top10perc 
 2.818587e+03 -8.342802e+00  3.686137e-03  8.951698e-03  1.682853e-01 
    Top25perc   F.Undergrad   P.Undergrad      Outstate    Room.Board 
 1.596131e-01  1.644603e-03  2.385862e-03  1.797724e-04  1.540381e-03 
        Books      Personal           PhD      Terminal     S.F.Ratio 
 6.288027e-03  1.999392e-03  2.093641e-01  2.254416e-01  1.645616e-01 
  perc.alumni        Expend     Grad.Rate 
-4.961264e-02  4.359390e-04  8.310356e-02 
> l2_norm <- sqrt(sum(coef(college_ridge_model)[2:9,50]^2))
> l2_norm
[1] 233.6984
> 
> college_ridge_model$lambda[50]
[1] 37992.91
> coef(college_ridge_model)[,50]
  (Intercept)       Private        Accept        Enroll     Top10perc 
 7.197419e+02 -2.336120e+02  1.157278e-01  2.677336e-01  4.661623e+00 
    Top25perc   F.Undergrad   P.Undergrad      Outstate    Room.Board 
 4.309567e+00  4.876047e-02  6.562444e-02  4.984532e-03  4.647640e-02 
        Books      Personal           PhD      Terminal     S.F.Ratio 
 1.730678e-01  5.063180e-02  5.245238e+00  5.624391e+00  5.082013e+00 
  perc.alumni        Expend     Grad.Rate 
-1.723457e+00  1.256846e-02  2.546296e+00 
> l2_norm <- sqrt(sum(coef(college_ridge_model)[2:9,50]^2))
> l2_norm
[1] 233.6984
>
```

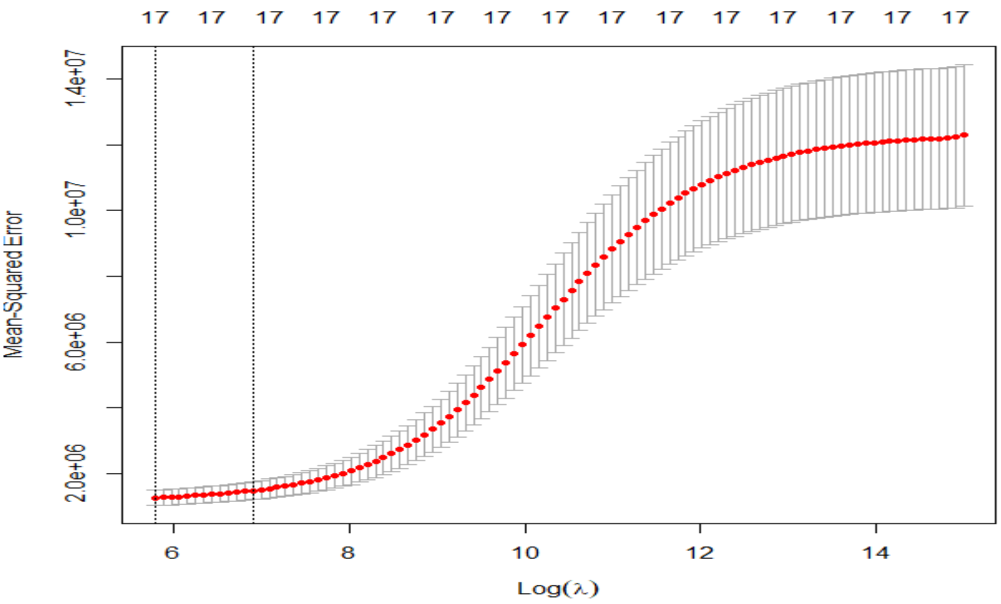


```
> college_ridge_model$lambda[100]
[1] 362.6608
> coef(college_ridge_model)[,100]
  (Intercept)       Private        Accept        Enroll     Top10perc 
-1.247334e+03 -3.540370e+02  1.064433e+00  3.750490e-01  2.506158e+01 
    Top25perc   F.Undergrad   P.Undergrad      Outstate    Room.Board 
 1.158428e+00  8.111567e-02  3.841285e-02 -2.266083e-02  2.089757e-01 
        Books      Personal           PhD      Terminal     S.F.Ratio 
 2.167011e-01 -2.486168e-02 -4.840390e+00 -3.697240e+00  1.478463e+01 
  perc.alumni        Expend     Grad.Rate 
-5.993027e+00  6.094702e-02  8.637693e+00 
> l2_norm <- sqrt(sum(coef(college_ridge_model)[2:9,100]^2))
> l2_norm
[1] 354.9266
> 
```

Predicting the ridge regression for a new value of lambda

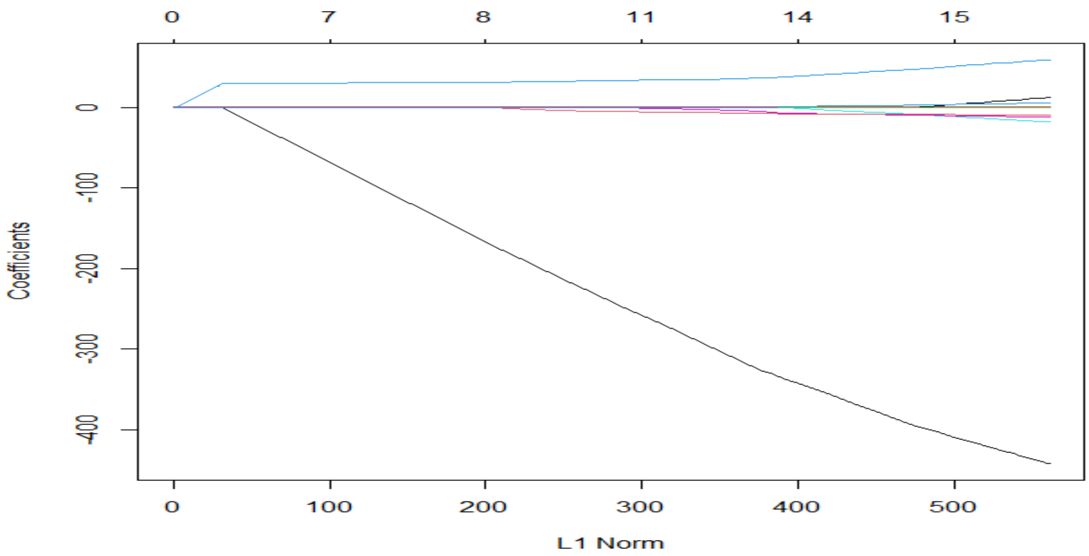


```
[1] 354.9266
> 
> 
> predict(college_ridge_model, s = .0005, type = "coefficient")
18 x 1 sparse Matrix of class "dgCMatrix"
                       s1
(Intercept) -1.247334e+03
Private     -3.540370e+02
Accept       1.064433e+00
Enroll       3.750490e-01
Top10perc    2.506158e+01
Top25perc    1.158428e+00
F.Undergrad  8.111567e-02
P.Undergrad  3.841285e-02
Outstate    -2.266083e-02
Room.Board   2.089757e-01
Books        2.167011e-01
Personal    -2.486168e-02
PhD         -4.840390e+00
Terminal    -3.697240e+00
S.F.Ratio    1.478463e+01
perc.alumni -5.993027e+00
Expend       6.094702e-02
Grad.Rate    8.637693e+00
> predict(college_ridge_model, s = .75, type = "coefficient")
18 x 1 sparse Matrix of class "dgCMatrix"
                       s1
(Intercept) -1.247334e+03
Private     -3.540370e+02
Accept       1.064433e+00
Enroll       3.750490e-01
Top10perc    2.506158e+01
Top25perc    1.158428e+00
F.Undergrad  8.111567e-02
P.Undergrad  3.841285e-02
Outstate    -2.266083e-02
Room.Board   2.089757e-01
Books        2.167011e-01
```



Plotting of Ridge Regression model with different values of lambda .

Test error is :47.3



Po

Test set error for lasso regression:

41.31

No. of non-zero coefficient estimates =5