Midterm Project

Purpose:

This assignment studied the application of K-Means, Hierarchical Clustering and Cross Validation techniques. We applied the above techniques on two files: RegularSeasonDetailedResults.csv, which contains the results of all NCAA regular season basketball games from 2003-2017; and TourneyCompactResults.csv, which contains tournament results from 2003-2016.

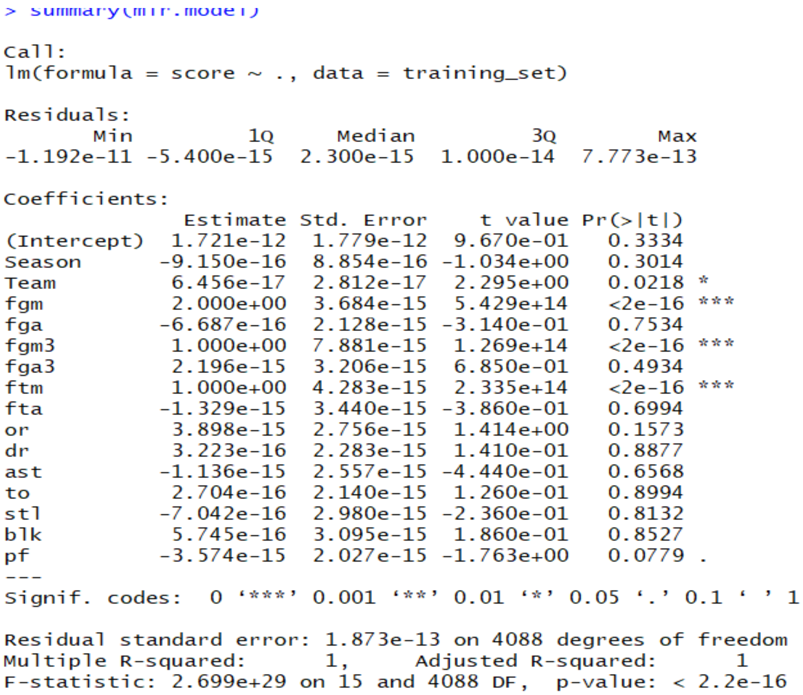
Dataset(s):

The datasets used in this assignment are RegularSeasonDetailedResults, Teams and TourneyCompactResults.

Approach:

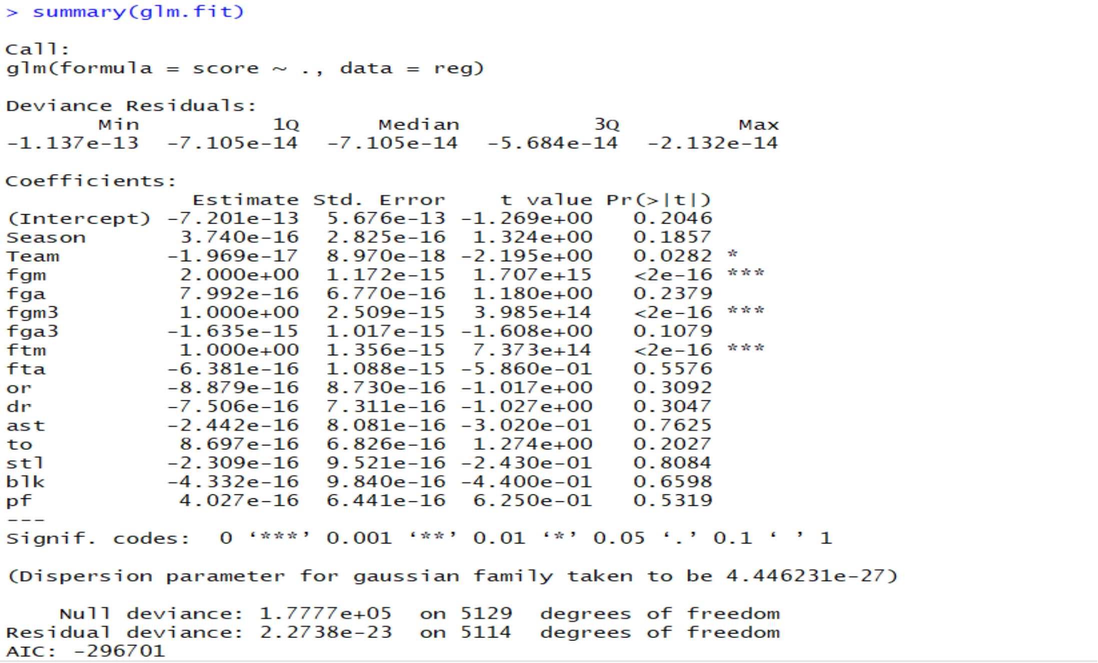
- First, we performed data wrangling by using RegularSeasonDetailedResults.csv to create a dataset that contains for each team and each season a list of team's overall statistics for that season. So, each row in this dataset should contain a column for the season, a column for the team number, and then columns for the team's statistics for that season.
- Next, we applied K-mean clustering for various values of K ranging from 2 to 15. On applying the Elbow method we found that the optimum number of clusters would be 3. As can be viewed in graph 1.
  We found the values as follows:
  betweenSS : 101819.8
  totSS : 295316.5
  withinSS : 60892.15 65338.52 67266.06
  tot.withinSS : 193496.7
- Next, we applied hierarchical clustering on the same dataset, using different methods and plot the same(refer Graphs 2-4).
- Next, we created a dataset that includes the season, the team number, their statistics for that season and the points they scored in the game of the tournament.

  Then, we split the dataset into training and testing datasets(in 80:20 ratio) and used the training dataset to build a model using regression techniques.



```
> summary(mlr.model)

Call:
lm(formula = score ~ ., data = training_set)

Residuals:
       Min         1Q     Median         3Q        Max 
-1.192e-11 -5.400e-15  2.300e-15  1.000e-14  7.773e-13 

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)    
(Intercept)  1.721e-12  1.779e-12  9.670e-01   0.3334    
Season      -9.150e-16  8.854e-16 -1.034e+00   0.3014    
Team         6.456e-17  2.812e-17  2.295e+00   0.0218 *  
fgm          2.000e+00  3.684e-15  5.429e+14   <2e-16 ***
fga         -6.687e-16  2.128e-15 -3.140e-01   0.7534    
fgm3         1.000e+00  7.881e-15  1.269e+14   <2e-16 ***
fga3         2.196e-15  3.206e-15  6.850e-01   0.4934    
ftm          1.000e+00  4.283e-15  2.335e+14   <2e-16 ***
fta         -1.329e-15  3.440e-15 -3.860e-01   0.6994    
or           3.898e-15  2.756e-15  1.414e+00   0.1573    
dr           3.223e-16  2.283e-15  1.410e-01   0.8877    
ast         -1.136e-15  2.557e-15 -4.440e-01   0.6568    
to           2.704e-16  2.140e-15  1.260e-01   0.8994    
stl         -7.042e-16  2.980e-15 -2.360e-01   0.8132    
blk          5.745e-16  3.095e-15  1.860e-01   0.8527    
pf          -3.574e-15  2.027e-15 -1.763e+00   0.0779 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.873e-13 on 4088 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 2.699e+29 on 15 and 4088 DF,  p-value: < 2.2e-16
```

- Then we apply the model to the testing data and make a prediction. We then output the values to an output file output.csv.
- After cross validation, we get the following results(also refer Graph5)

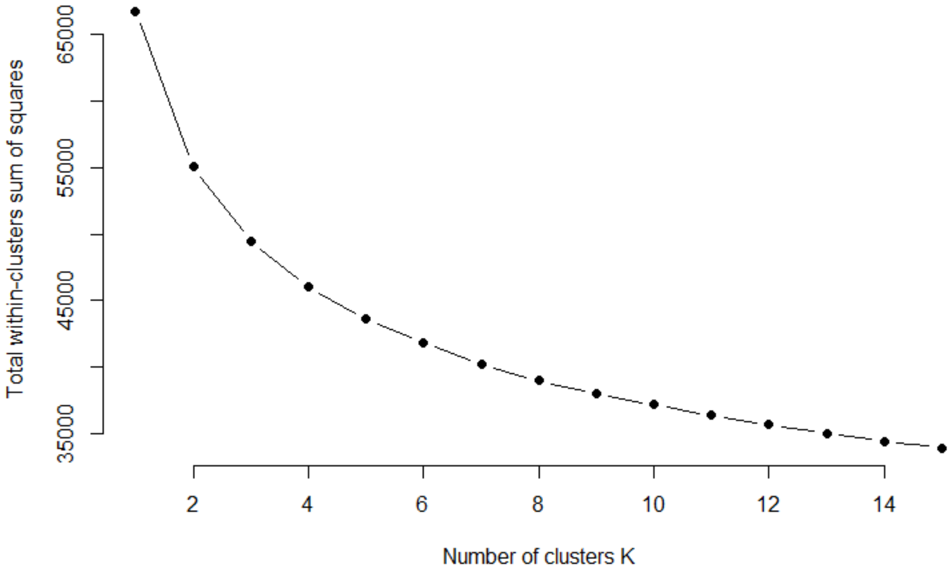


```
> summary(glm.fit)

Call:
glm(formula = score ~ ., data = reg)

Deviance Residuals: 
       Min          1Q      Median          3Q         Max  
-1.137e-13  -7.105e-14  -7.105e-14  -5.684e-14  -2.132e-14  

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)    
(Intercept) -7.201e-13  5.676e-13 -1.269e+00   0.2046    
Season       3.740e-16  2.825e-16  1.324e+00   0.1857    
Team        -1.969e-17  8.970e-18 -2.195e+00   0.0282 *  
fgm          2.000e+00  1.172e-15  1.707e+15   <2e-16 ***
fga          7.992e-16  6.770e-16  1.180e+00   0.2379    
fgm3         1.000e+00  2.509e-15  3.985e+14   <2e-16 ***
fga3        -1.635e-15  1.017e-15 -1.608e+00   0.1079    
ftm          1.000e+00  1.356e-15  7.373e+14   <2e-16 ***
fta         -6.381e-16  1.088e-15 -5.860e-01   0.5576    
or          -8.879e-16  8.730e-16 -1.017e+00   0.3092    
dr          -7.506e-16  7.311e-16 -1.027e+00   0.3047    
ast         -2.442e-16  8.081e-16 -3.020e-01   0.7625    
to           8.697e-16  6.826e-16  1.274e+00   0.2027    
stl         -2.309e-16  9.521e-16 -2.430e-01   0.8084    
blk         -4.332e-16  9.840e-16 -4.400e-01   0.6598    
pf           4.027e-16  6.441e-16  6.250e-01   0.5319    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 4.446231e-27)

    Null deviance: 1.7777e+05  on 5129  degrees of freedom
Residual deviance: 2.2738e-23  on 5114  degrees of freedom
AIC: -296701
```

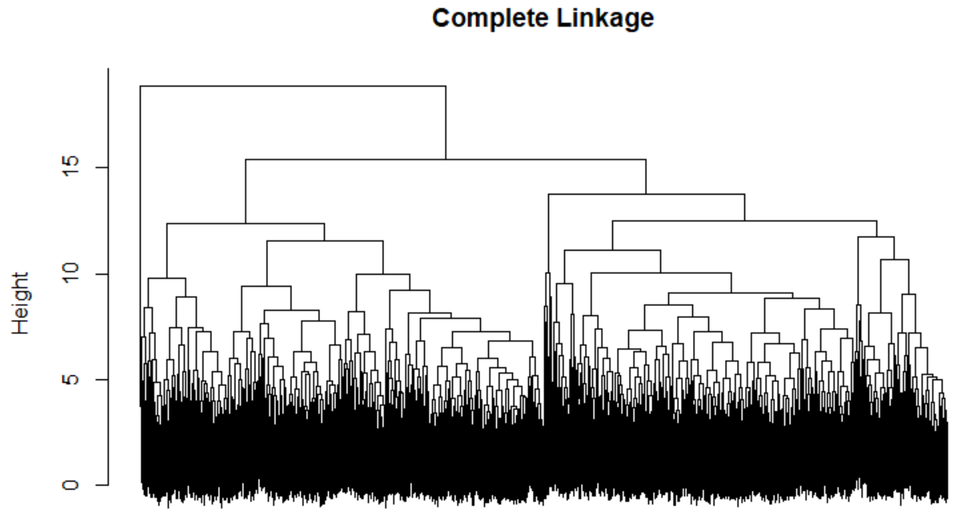
- We get a RMSE value of 1.98238e-13 and the model built using cross validation gives raw cross-validation estimate of prediction error as 8.569462e-27 and adjusted crossvalidation estimate of prediction error as 4.452348e-27 which is good.

Graphs:

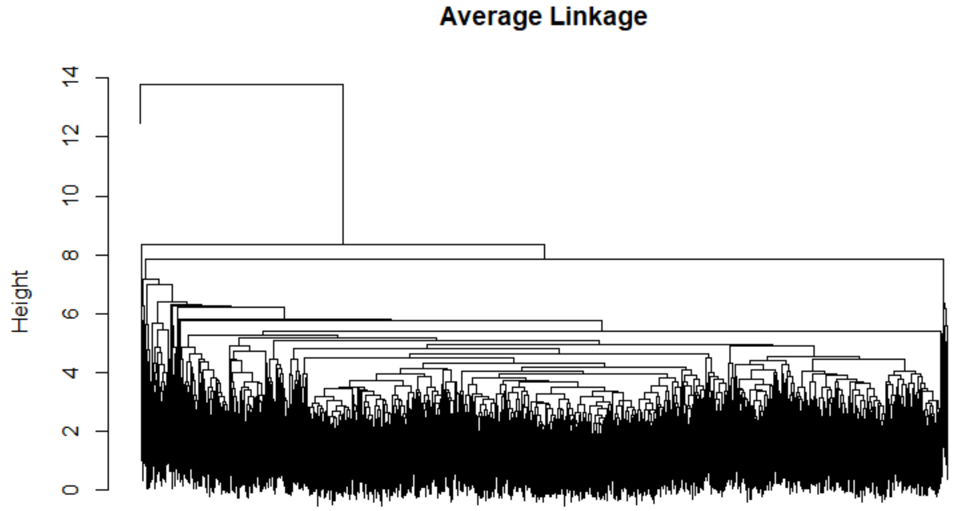
Graph 1: Compute and plot wss for k = 2 to k = 15



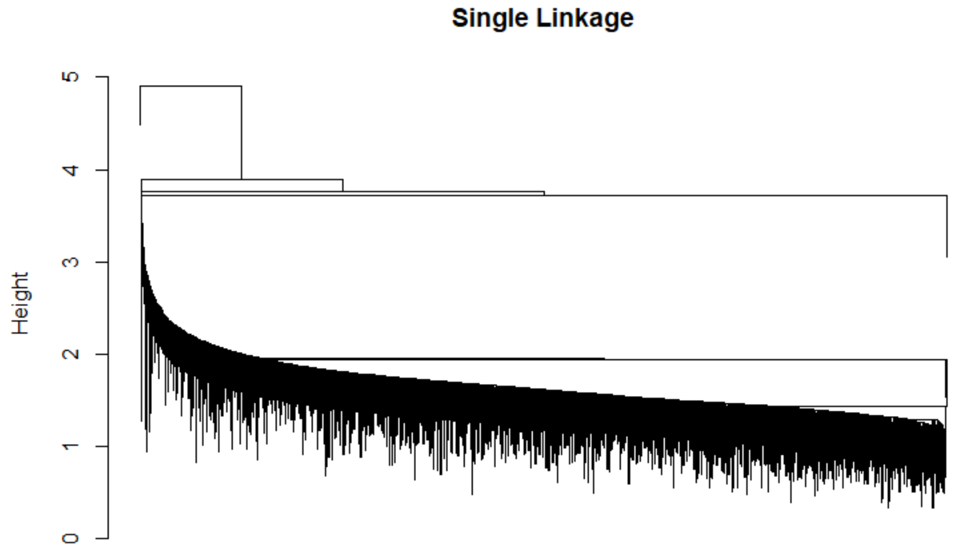
Graph 2:Hierarchical clustering



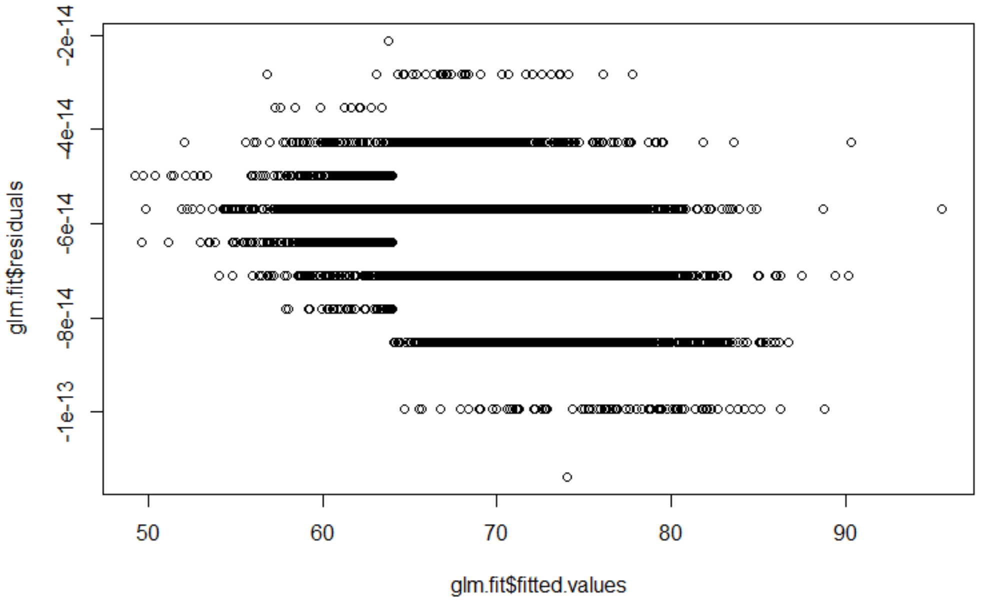
Graph 3: Hierarchical clustering



Graph 4: Hierarchical clustering



Graph 5:Model Building with cross validation



Summary:

- We observed that Complete Linkage method of H-Clustering provides a compact plot compared to the other methods.

- It gives a good observation of the optimal number of clusters for our dataset along with the elbow method.
- The model built and tested using cross validation performs well with low prediction error.