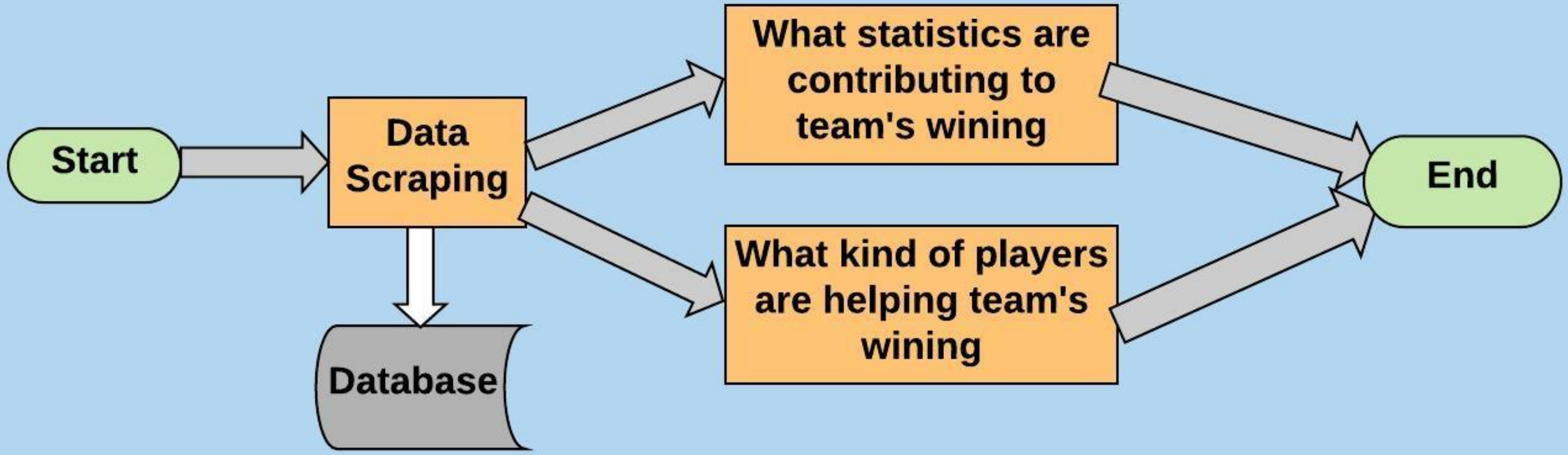
NBA Data Analysis

Chenjie Li, Qiao Qiao, Dilruba Palabiyik

Project Outline



Start

Data Scraping

Database

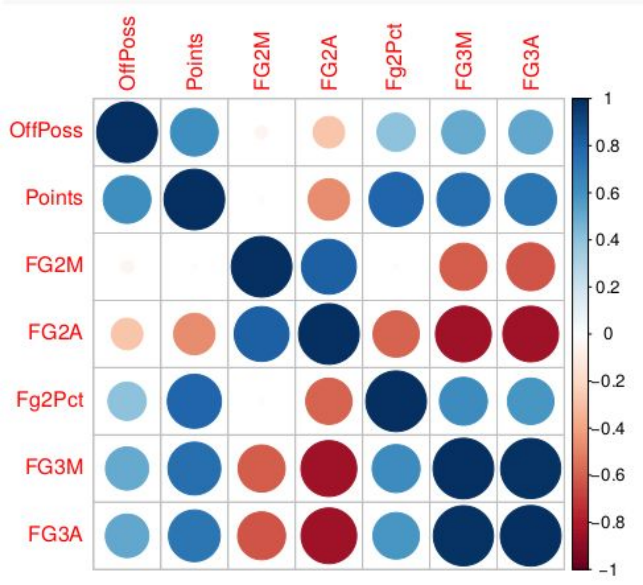
What statistics are contributing to team's wining

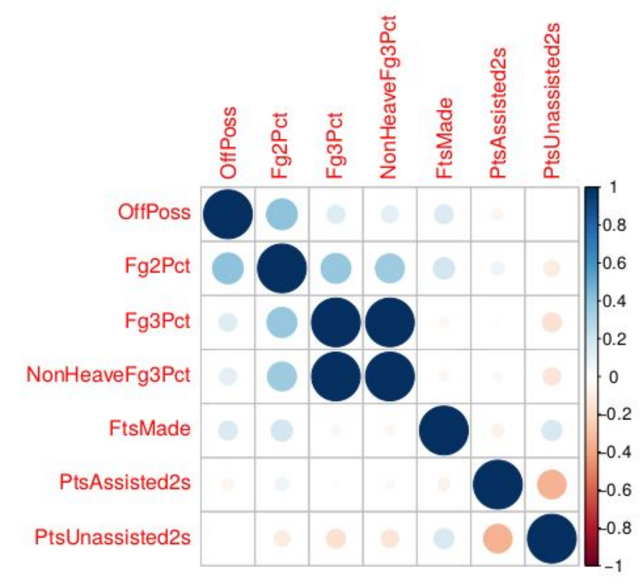
What kind of players are helping team's wining

End

Data Cleaning

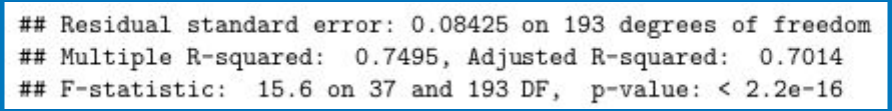
Collinearity and Null Value



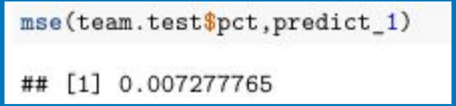


Model Selection

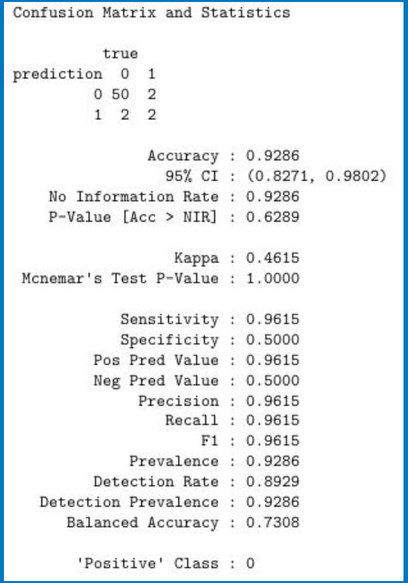
Linear Regression



```
## Residual standard error: 0.08425 on 193 degrees of freedom
## Multiple R-squared:  0.7495, Adjusted R-squared:  0.7014
## F-statistic:  15.6 on 37 and 193 DF,  p-value: < 2.2e-16
```



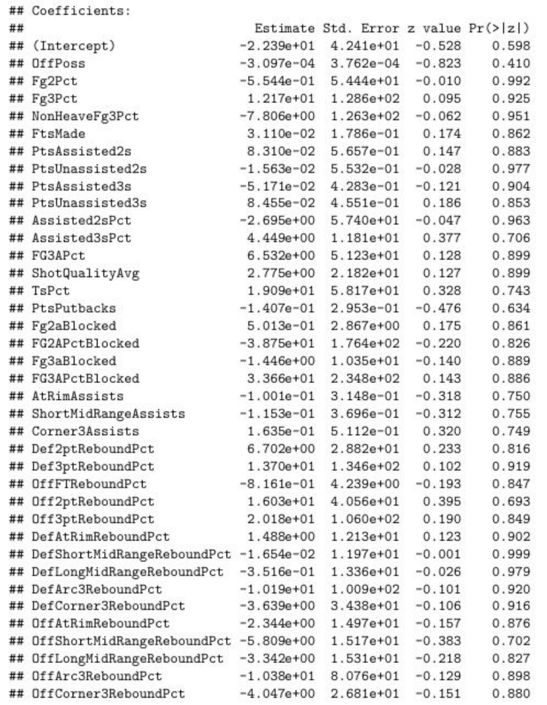
```
mse(team.test$pct,predict_1)

## [1] 0.007277765
```



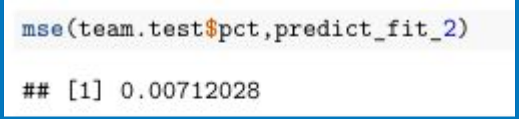
```
Confusion Matrix and Statistics

          true
prediction  0  1
         0 50  2
         1  2  2

               Accuracy : 0.9286
                 95% CI : (0.8271, 0.9802)
    No Information Rate : 0.9286
    P-Value [Acc > NIR] : 0.6289

                  Kappa : 0.4615
 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.9615
            Specificity : 0.5000
         Pos Pred Value : 0.9615
         Neg Pred Value : 0.5000
              Precision : 0.9615
                 Recall : 0.9615
                     F1 : 0.9615
             Prevalence : 0.9286
         Detection Rate : 0.8929
   Detection Prevalence : 0.9286
      Balanced Accuracy : 0.7308

       'Positive' Class : 0
```

Model Selection

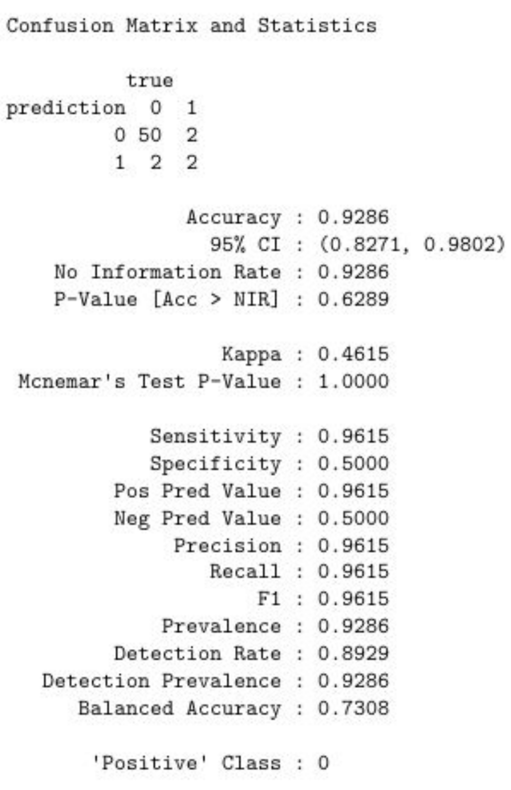
Logistic Regression



```
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.239e+01  4.241e+01  -0.528    0.598
## OffPoss                    -3.097e-04  3.762e-04  -0.823    0.410
## Fg2Pct                     -5.544e-01  5.444e+01  -0.010    0.992
## Fg3Pct                      1.217e+01  1.286e+02   0.095    0.925
## NonHeaveFg3Pct             -7.806e+00  1.263e+02  -0.062    0.951
## FtsMade                     3.110e-02  1.786e-01   0.174    0.862
## PtsAssisted2s               8.310e-02  5.657e-01   0.147    0.883
## PtsUnassisted2s            -1.563e-02  5.532e-01  -0.028    0.977
## PtsAssisted3s              -5.171e-02  4.283e-01  -0.121    0.904
## PtsUnassisted3s             8.455e-02  4.551e-01   0.186    0.853
## Assisted2sPct              -2.695e+00  5.740e+01  -0.047    0.963
## Assisted3sPct               4.449e+00  1.181e+01   0.377    0.706
## FG3APct                     6.532e+00  5.123e+01   0.128    0.899
## ShotQualityAvg              2.775e+00  2.182e+01   0.127    0.899
## TsPct                       1.909e+01  5.817e+01   0.328    0.743
## PtsPutbacks                -1.407e-01  2.953e-01  -0.476    0.634
## Fg2aBlocked                 5.013e-01  2.867e+00   0.175    0.861
## FG2APctBlocked             -3.875e+01  1.764e+02  -0.220    0.826
## Fg3aBlocked                -1.446e+00  1.035e+01  -0.140    0.889
## FG3APctBlocked              3.366e+01  2.348e+02   0.143    0.886
## AtRimAssists               -1.001e-01  3.148e-01  -0.318    0.750
## ShortMidRangeAssists       -1.153e-01  3.696e-01  -0.312    0.755
## Corner3Assists              1.635e-01  5.112e-01   0.320    0.749
## Def2ptReboundPct            6.702e+00  2.882e+01   0.233    0.816
## Def3ptReboundPct            1.370e+01  1.346e+02   0.102    0.919
## OffFTReboundPct            -8.161e-01  4.239e+00  -0.193    0.847
## Off2ptReboundPct            1.603e+01  4.056e+01   0.395    0.693
## Off3ptReboundPct            2.018e+01  1.060e+02   0.190    0.849
## DefAtRimReboundPct          1.488e+00  1.213e+01   0.123    0.902
## DefShortMidRangeReboundPct -1.654e-02  1.197e+01  -0.001    0.999
## DefLongMidRangeReboundPct  -3.516e-01  1.336e+01  -0.026    0.979
## DefArc3ReboundPct          -1.019e+01  1.009e+02  -0.101    0.920
## DefCorner3ReboundPct       -3.639e+00  3.438e+01  -0.106    0.916
## OffAtRimReboundPct         -2.344e+00  1.497e+01  -0.157    0.876
## OffShortMidRangeReboundPct -5.809e+00  1.517e+01  -0.383    0.702
## OffLongMidRangeReboundPct  -3.342e+00  1.531e+01  -0.218    0.827
## OffArc3ReboundPct          -1.038e+01  8.076e+01  -0.129    0.898
## OffCorner3ReboundPct       -4.047e+00  2.681e+01  -0.151    0.880
```

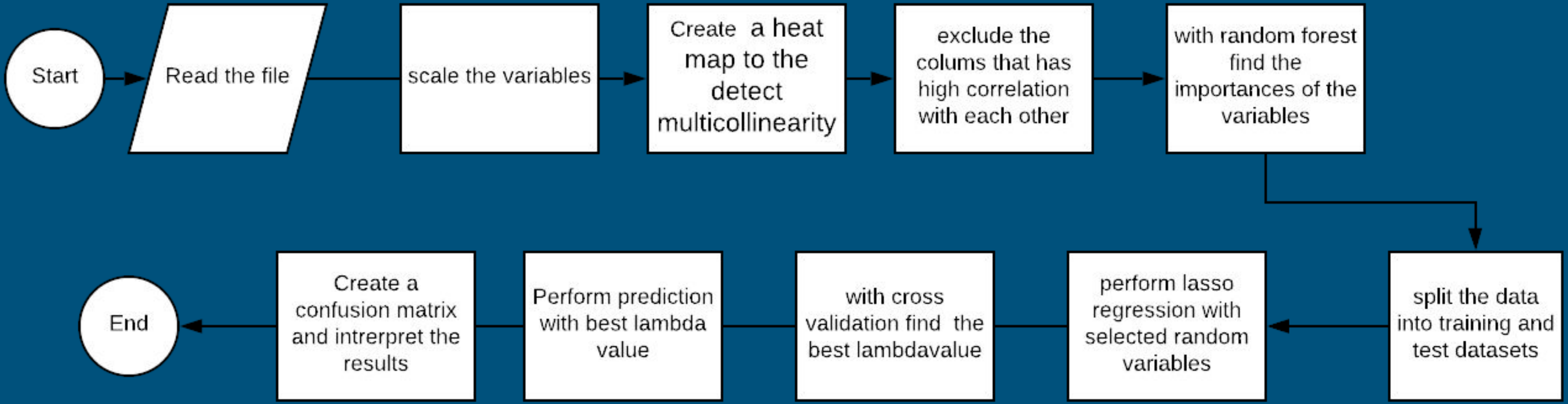


```
mse(team.test$pct,predict_fit_2)

## [1] 0.00712028
```



```
Confusion Matrix and Statistics

          true
prediction  0  1
         0 50  2
         1  2  2

               Accuracy : 0.9286
                 95% CI : (0.8271, 0.9802)
    No Information Rate : 0.9286
    P-Value [Acc > NIR] : 0.6289

                  Kappa : 0.4615
 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.9615
            Specificity : 0.5000
         Pos Pred Value : 0.9615
         Neg Pred Value : 0.5000
              Precision : 0.9615
                 Recall : 0.9615
                     F1 : 0.9615
             Prevalence : 0.9286
         Detection Rate : 0.8929
   Detection Prevalence : 0.9286
      Balanced Accuracy : 0.7308

       'Positive' Class : 0
```

Overview of the Algorithm



Start

Read the file

scale the variables

Create a heat map to the detect multicollinearity

exclude the colums that has high correlation with each other

with random forest find the importances of the variables

split the data into training and test datasets

perform lasso regression with selected random variables

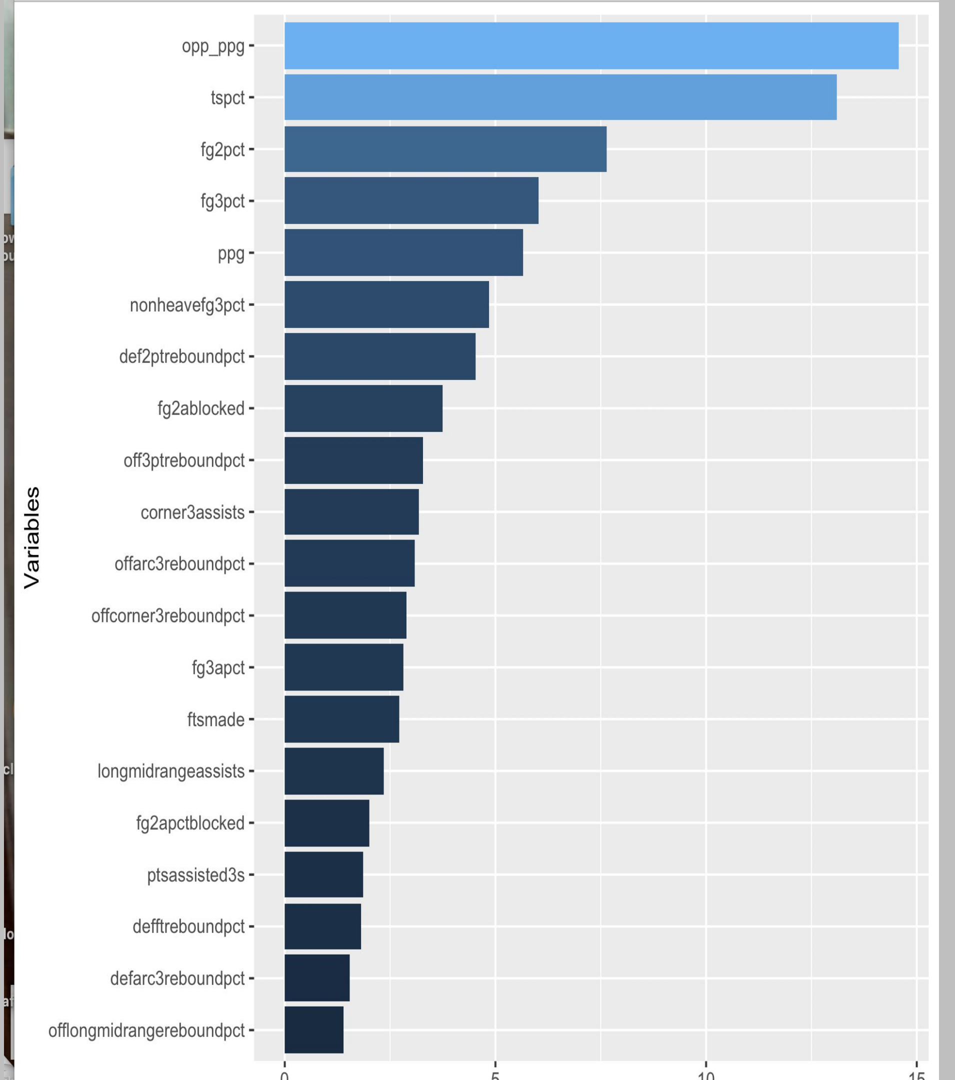
with cross validation find the best lambdavalue

Perform prediction with best lambda value

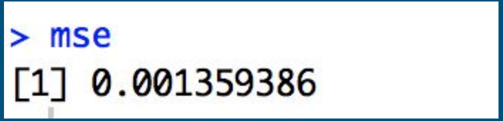
Create a confusion matrix and intrerpret the results

End

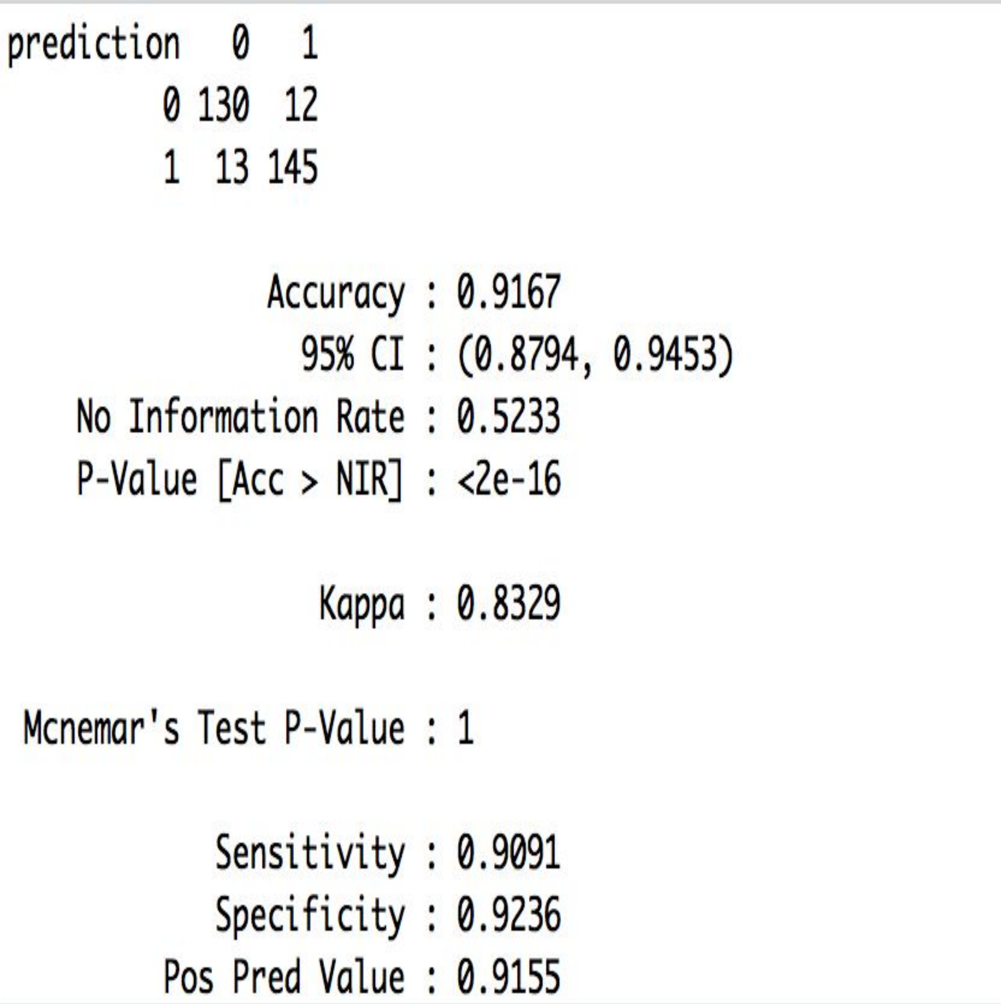
Variable selection based on variable importances



Selecting the Tuning Parameter and interpreting the results



```
> mse
[1] 0.001359386
```



```
prediction   0   1
         0 130  12
         1  13 145

               Accuracy : 0.9167
                 95% CI : (0.8794, 0.9453)
    No Information Rate : 0.5233
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8329

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9091
            Specificity : 0.9236
         Pos Pred Value : 0.9155
```

Players Clustering

❑ How are the “good teams” constructed?

❑ What trend changes could we observe during last 10 seasons?

❑ How the role of a certain player changes as the time goes by?

Players Clustering

Process

- ❑ **Query to get desired data**
- ❑ **Outlier removal (based on GP, and MPG).**
- ❑ **Detect and resolve collinearity (Domain and Also Cor Matrix)**
- ❑ **LDA on filtered attributes, training and testing get accuracy**
- ❑ **PCA on filtered attributes, extract attributes with which we can exceed the prediction accuracy we got from LDA**
- ❑ **K-Means clustering (using Gap Statistic * to decide appropriate K)**

* Robert Tibshirani , Guenther Walther , Trevor Hastie, Estimating the number of clusters in a dataset via the Gap statistic

Players Clustering

Elbow Method:

Elbow method maps the within-cluster sum of squares onto the number of possible clusters.

Silhouette Method:

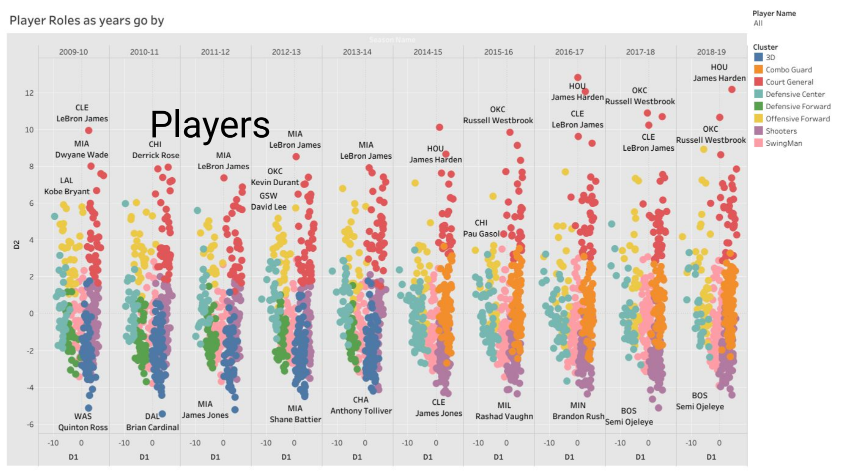
Silhouette plots display a measure of how close each point in one cluster is to points in the neighboring clusters.

Gap Statistic:

Gap statistic is a goodness of clustering measure, where for each hypothetical number of clusters k, it compares two functions: log of within-cluster sum of squares (wss) with its expectation under the null reference distribution of the data. In essence, it standardizes wss. It chooses the value where the log(wss) is the farthest below the reference curve

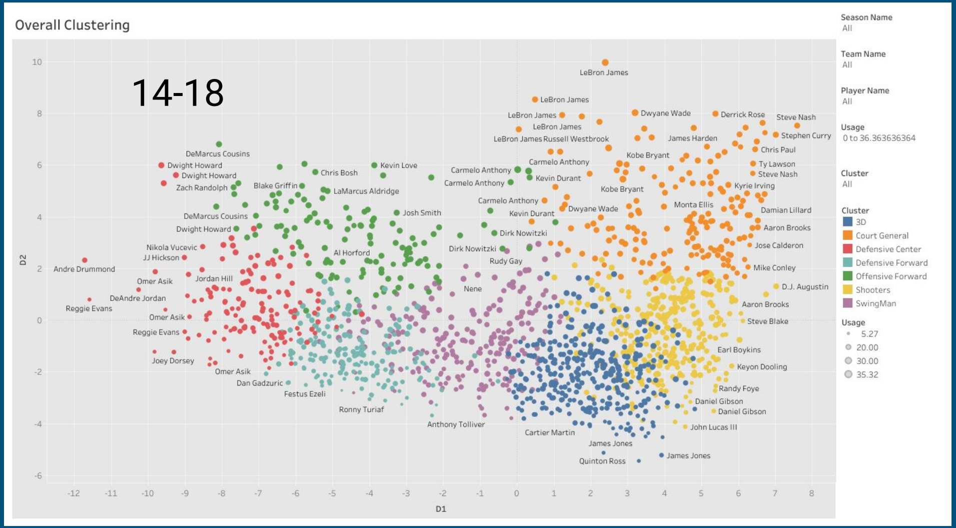
Players Clustering

Results:


Player Roles as years go by

Player Name
All

Cluster
🟦 3D
🟧 Combo Guard
🟥 Court General
🟩 Defensive Center
🟩 Defensive Forward
🟨 Offensive Forward
🟪 Shooters
🟥 SwingMan


Overall Clustering

Season Name
All

Team Name
All

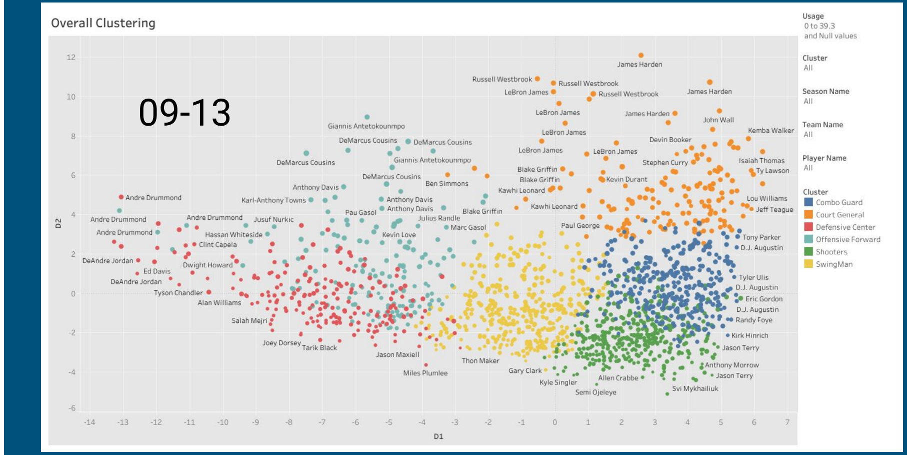
Player Name
All

Usage
0 to 36.363636364

Cluster
All

Cluster
🟦 3D
🟧 Court General
🟥 Defensive Center
🟩 Defensive Forward
🟩 Offensive Forward
🟨 Shooters
🟪 SwingMan

Usage
● 5.27
● 20.00
● 30.00
● 35.32


Overall Clustering

Usage
0 to 39.3
and Null values

Cluster
All

Season Name
All

Team Name
All

Player Name
All

Cluster
🟦 Combo Guard
🟧 Court General
🟥 Defensive Center
🟩 Offensive Forward
🟩 Shooters
🟨 SwingMan

Conclusion

- Among linear, logistic and lasso regression, we obtained the best results with lasso regression.
- With K Means Clustering result, we found the clustering distribution difference between 2 five-year periods and also we found some interesting evolvement of players as the year goes by