NFL Playoff Prediction

Smells Like Team Spirit

```r
library(RCurl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(readr)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:RCurl':
##
##     complete
```

```r
library(knitr)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(purrr)
```

NFL teams that performed poorly in win column one year can rise to Super Bowl champions the next (like the Philadelphia Eagles). An NFL game has been sometimes referred to as a "game of inches" in which wins and losses can be determined by chance, hiding the true potential of a team. This could lead to the seemingly surprising rise of a team like the Eagles. We can use machine learning to look beyond just team record to determine which teams that performed poorly last year could compete for a Super Bowl this year. Our goal is to create a machine learning model that groups NFL teams together, predicting a set of playoff teams.

We'd look at NFL season (2000-2013) that we used to predict "Wins" with a column indicating whether a team makes the playoff or not. to test our model on predicting former playoff teams, and then later we will predict next season's.

Since we want to visualize the groupings of NFL teams, we must reduce the dimensionality of all the variable data we collected. To reduce dimensionality, we can use Principal Component Analysis (PCA), which is a statistical procedure that converts a set of variables into a new smaller set of variables that still captures the essence of all the original variables.

I. PCA

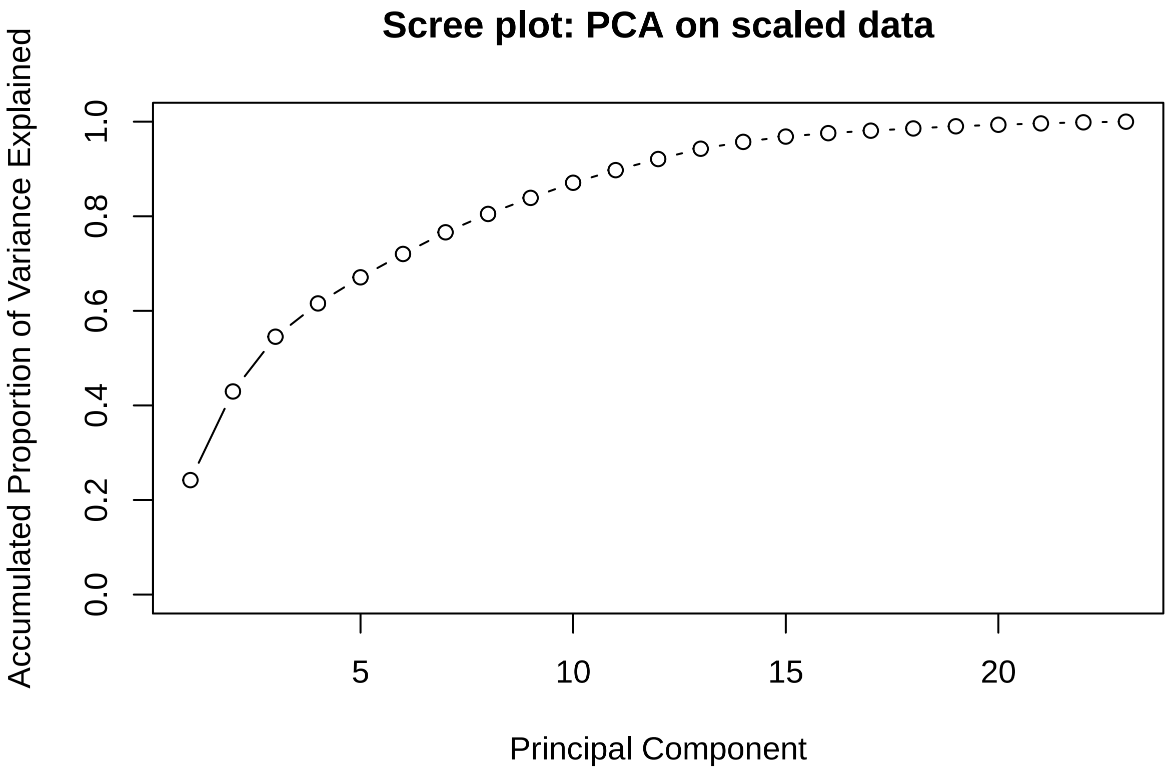
```
nfl_playoff_all_data <- read.csv("nfl_playoff_all_data.csv")
head(nfl_playoff_all_data)
```

```
##                TeamYear          TeamName Year YearF Playoff Wins ScoreOff
## 1 Arizona Cardinals 00' Arizona Cardinals  00'  2000       0    9      178
## 2   Atlanta Falcons 00'   Atlanta Falcons  00'  2000       0    7      238
## 3  Baltimore Ravens 00'  Baltimore Ravens  00'  2000       1   12      355
## 4     Buffalo Bills 00'     Buffalo Bills  00'  2000       0    8      288
## 5 Carolina Panthers 00' Carolina Panthers  00'  2000       0    7      272
## 6     Chicago Bears 00'     Chicago Bears  00'  2000       0    9      216
##   FirstDown RushAttOff RushYdsOff PassAttOff PassCompOff PassYdsOff PassIntOff
## 1       253        342       1284        554         316       3478         24
## 2       256        350       1214        515         285       3166         20
## 3       319        619       2480        553         309       3539         20
## 4       309        476       1921        546         312       3936         10
## 5       304        363       1186        566         340       3850         19
## 6       238        416       1736        542         304       3005         16
##   FumblesOff SackYdsOff PenYdsOff PuntAvgOff ScoreDef FirstDownDef RushAttDef
## 1         20        239       756        710      443          344        580
## 2         14        386       720        654      413          308        453
## 3          8        349       905        741      181          260        430
## 4         12        359       913        610      350          252        444
## 5         16        382       683        607      310          304        425
## 6         13        206       696        593      355          297        469
##   RushYdsDef PassAttDef PassCompDef PassYdsDef PassIntDef FumblesDef SackYdsDef
## 1       2609        458         295       3263         10         10        126
## 2       1983        515         306       3766         15         10        142
## 3       1162        650         357       3735         29         27        245
## 4       1559        480         283       3175         16         13        308
## 5       1949        552         352       3938         17         21        231
## 6       1828        530         332       3635         11          9        231
##   PenYdsDef
## 1        32
## 2        14
## 3        39
## 4        27
```

```
## 5          38
## 6           0
```

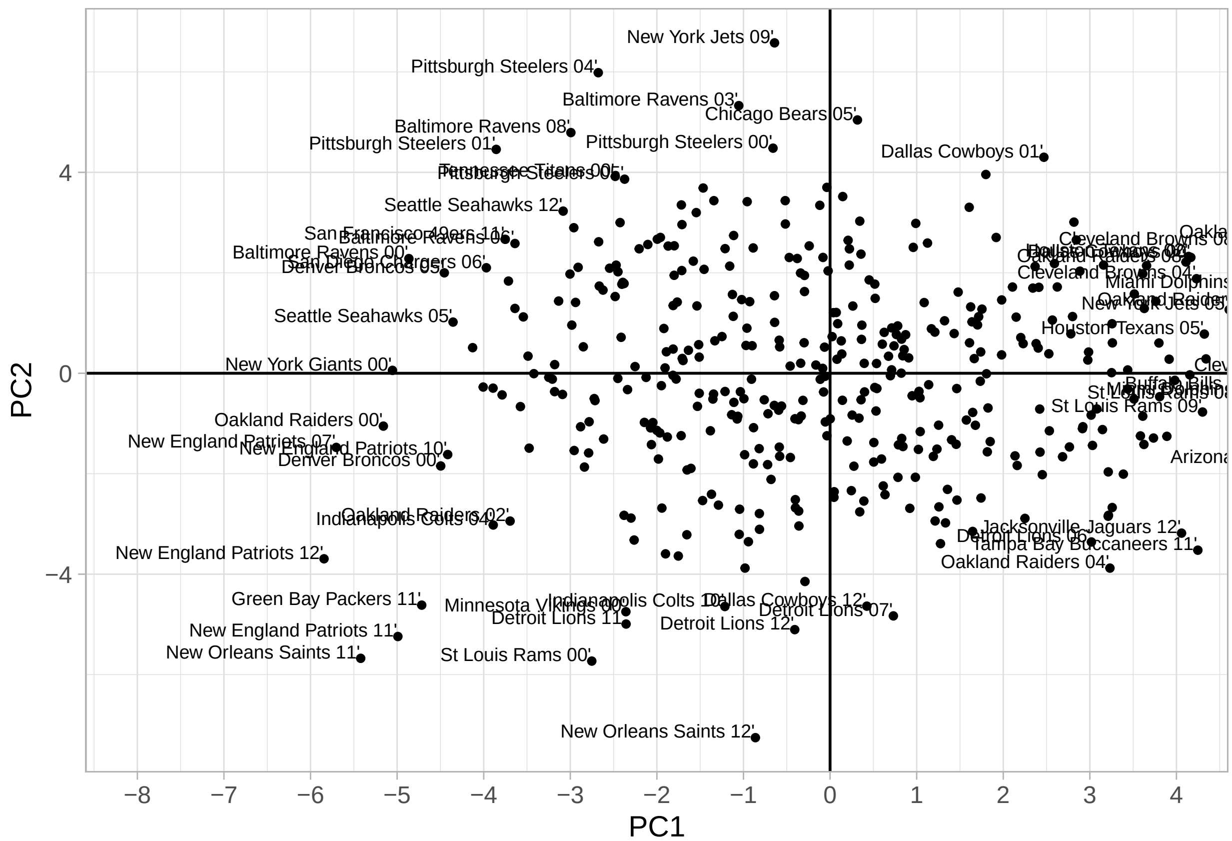
```
nfl_data_TY <- subset(nfl_playoff_all_data, YearF != 2013)[c(1,5:29)]
nfl_data <- subset(nfl_playoff_all_data, YearF != 2013)[c(5:29)]
nfl_2013 <- subset(nfl_playoff_all_data, YearF == 2013)[c(1,5:29)]
```

```
nfl_pca <- prcomp(nfl_data[2:24] , scale = TRUE)

pr.var <- nfl_pca$sdev^2
pve = pr.var / sum(pr.var)
plot(cumsum(pve), xlab = "Principal Component", ylab = "Accumulated Proportion of Variance Explained", 
     main = "Scree plot: PCA on scaled data")
```




```
nfl_pca_scores <- nfl_pca$x
low_dim_rep <- nfl_pca_scores %>%
data.frame() %>%
mutate(TeamYear = nfl_data_TY$TeamYear) %>%
select(TeamYear, everything())

ggplot(low_dim_rep, aes(x = PC1, y = PC2)) +
geom_vline(xintercept = 0) +
geom_hline(yintercept = 0) +
geom_point(size = 1) + geom_text(aes(label=ifelse(PC1^2+PC2^2 > 19  ,as.character(TeamYear),'')),hjust=
scale_x_continuous(breaks = -10:10) +
```

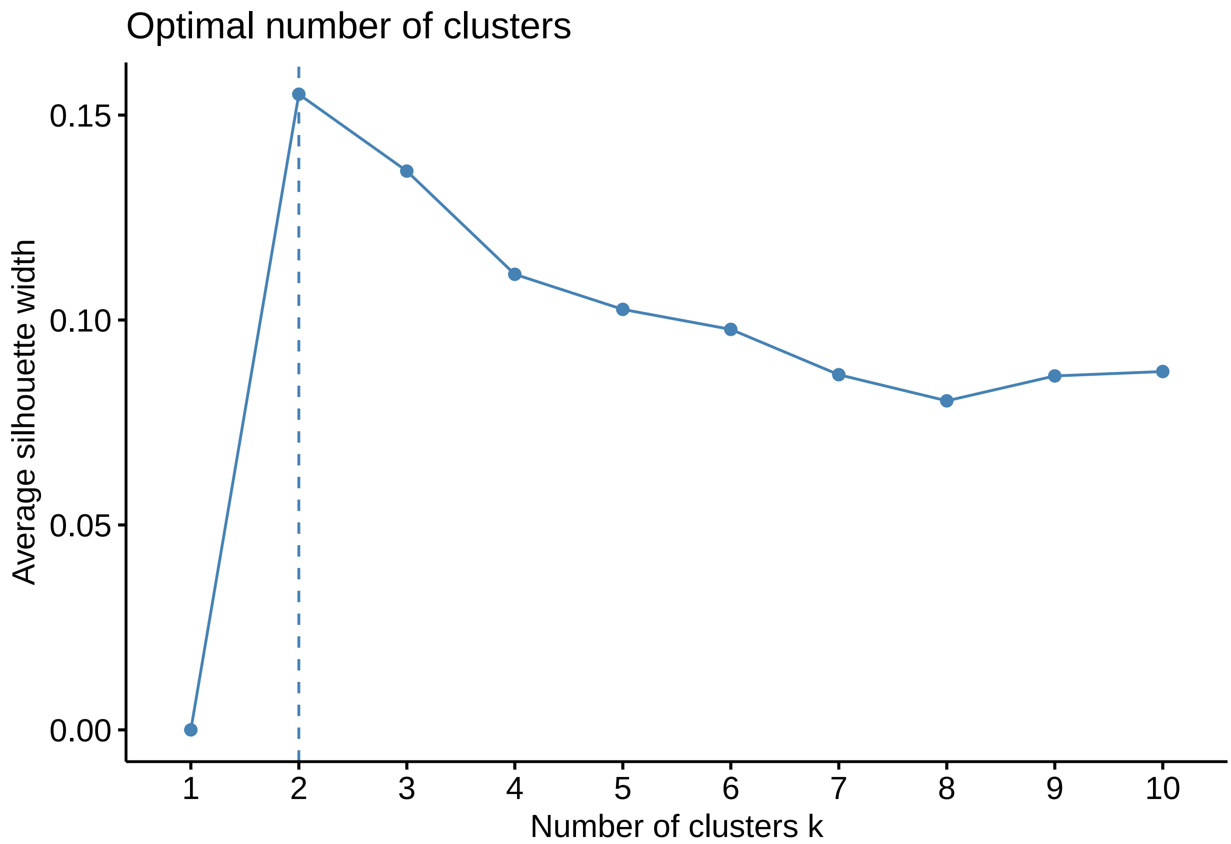
```r
coord_cartesian(xlim = c(-8, 4)) +
theme_light()
```



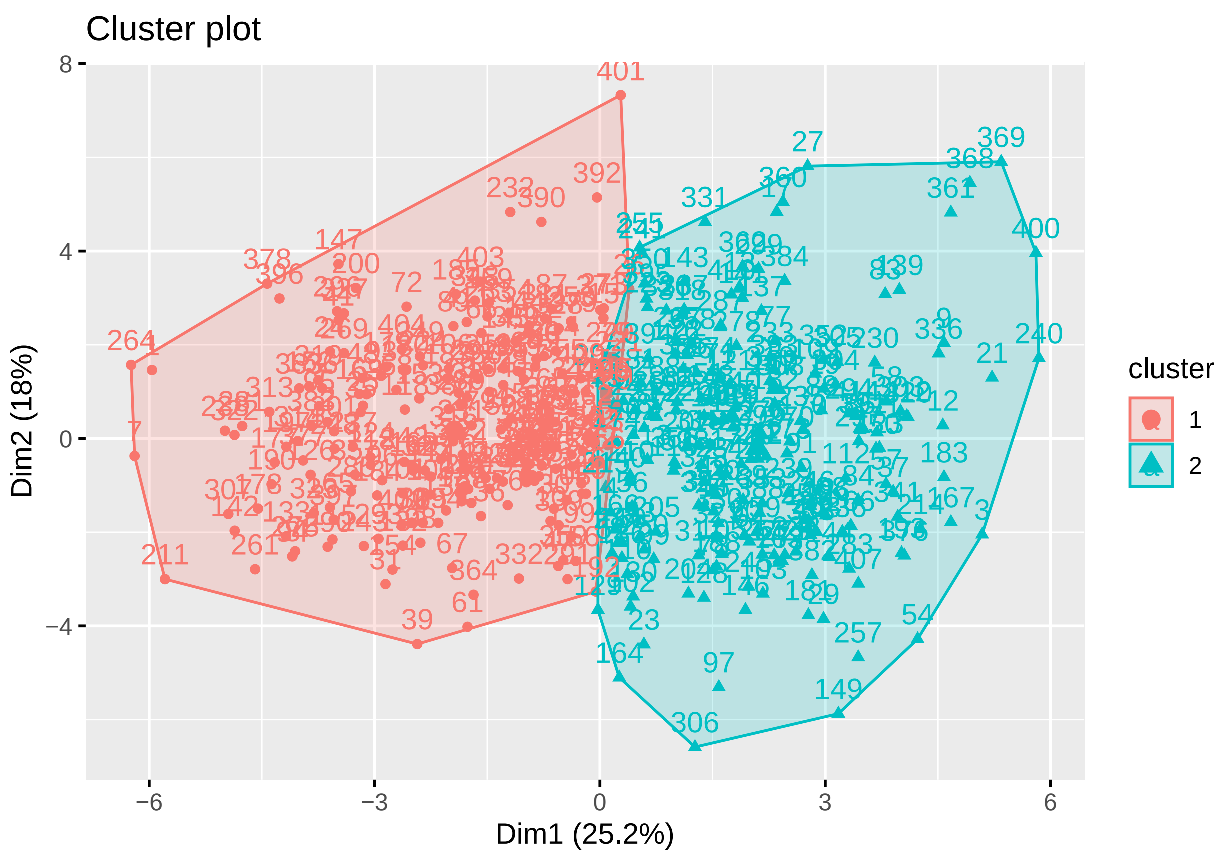
The axes labeled 'PC1' and 'PC2' represent all the variables we have reduced through PCA. For visual clarity, only some of the teams (plus year) have been labeled.

II. k-means Clustering Using the K-Means Elbow Method, we found the ideal number of clusters to be 2. Finally, we can use the K-Means Algorithm to determine and plot the clusters of different types of NFL teams, shown below:

```r
#fviz_nbclust(nfl_data_TY, kmeans, method = "wss")
df <- scale(nfl_data_TY[2:25])

fviz_nbclust(df, kmeans, method = "silhouette")
```




```
final2 <- kmeans(df, 2, nstart = 25)
fviz_cluster(final2, data = df)
```



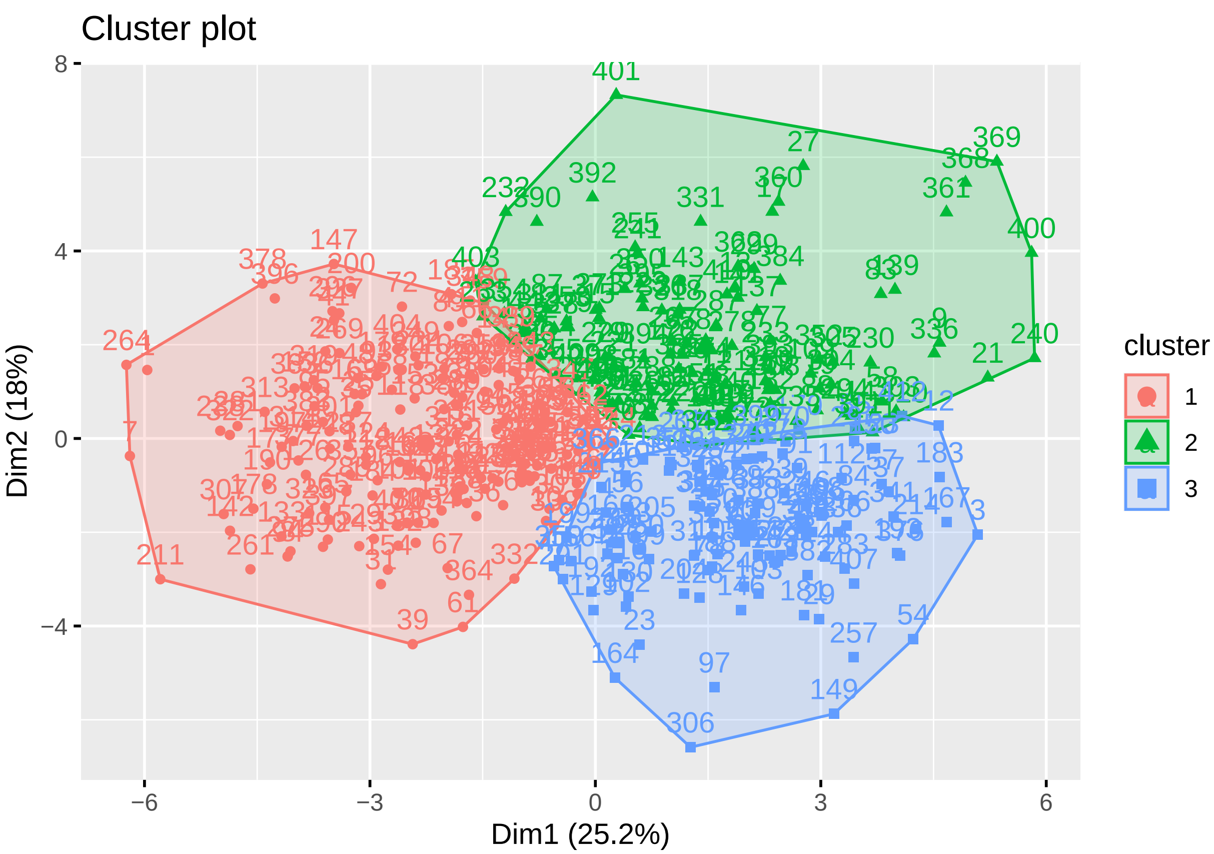

```r
table(final2$cluster, nfl_data_TY$Playoff)
```

```
##
##       0   1
##   1 210   5
##   2  49 150
```

Clusters represent the quality of teams based on collected input variables. According to the table (0 and 1 represent whether the team makes the playoff, 1 = yes), cluster 1 represents the non-playoff teams and cluster 2 are playoff caliber teams. Cluster 1 contained 5/215 = 2.32% of the playoff teams, while cluster 2 contained 75.4% of the playoff teams. This indicates that teams in cluster 2 were more than 30 times more likely to make the playoffs than cluster 1 teams.

We also tried something new as we manually changed the number of clusters from 2 to 3, which gives us the following result:

```r
final3 <- kmeans(df, 3, nstart = 25)
fviz_cluster(final3, data = df)
```



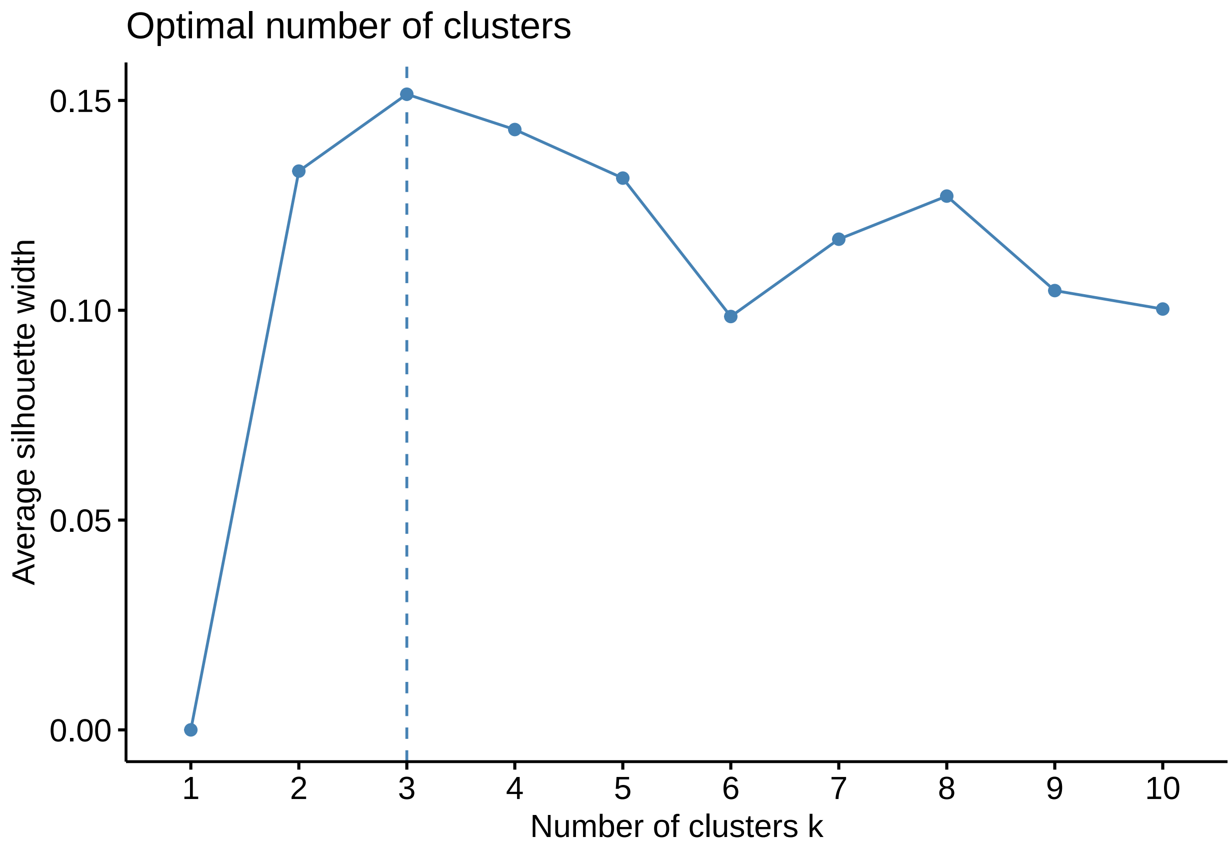

```r
table(final3$cluster, nfl_data_TY$Playoff)
```

```
##
##       0   1
##   1 175   3
##   2  55  68
##   3  29  84
```

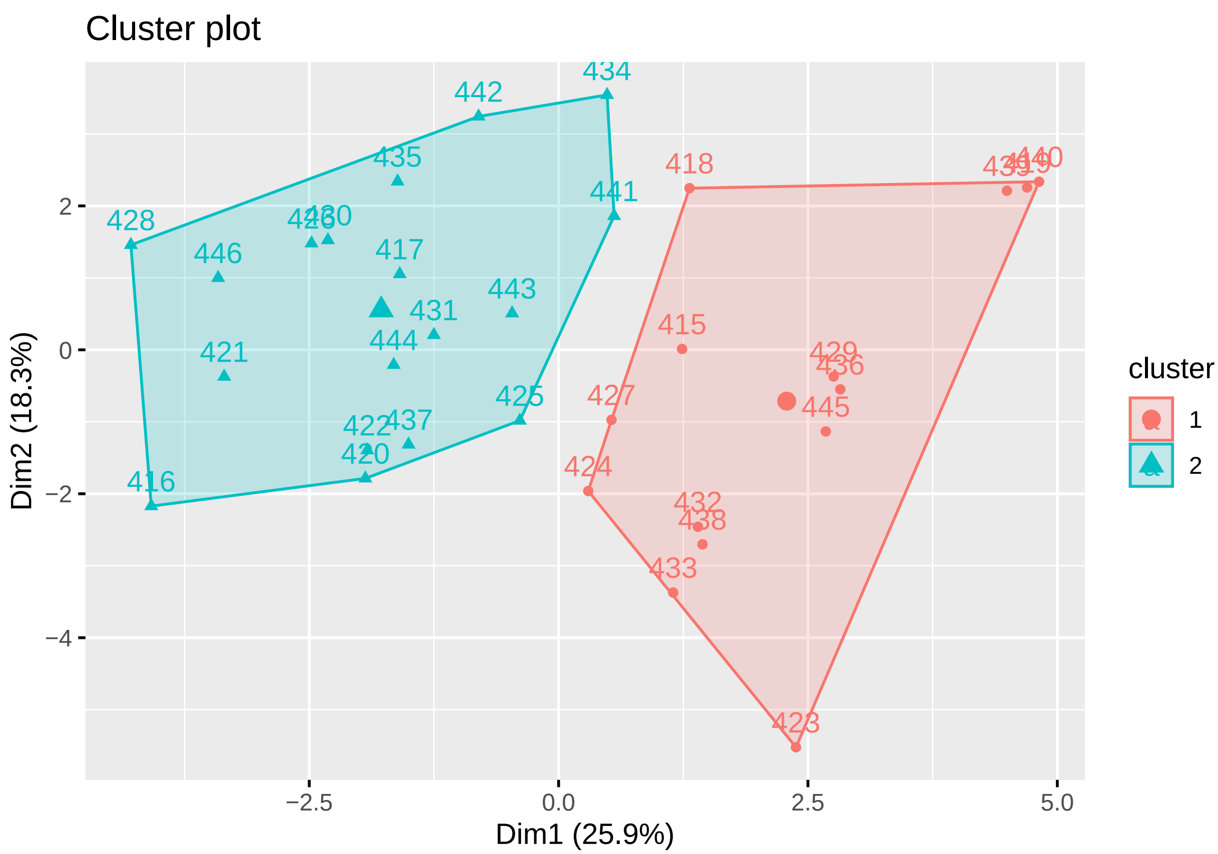
Cluster 1 represents the non-playoff teams, cluster 2 are borderline playoff teams, and cluster 3 are playoff caliber teams. Cluster 1 contained only 1.7% of the playoff teams, cluster 2 contained 44.7% of the playoff teams, while cluster 3 contained only 74.3% of the playoff teams.
The clustering method with 3 clusters provide more detailed description for mid-table teams, which could be also useful to predict whether a team can make the playoff based on its performance.

Lastly, we apply the K-Means algorithm to the 2013 season to predict its playoff teams, shown below:

```r
dfnew <- scale(nfl_2013[2:25])
fviz_nbclust(dfnew, kmeans, method = "silhouette")
```




```
testCluster <- kmeans(dfnew, 2, nstart = 10)
fviz_cluster(testCluster, data = dfnew)
```




```
testCluster$cluster
```

```
## 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434
##   1   2   2   1   1   2   2   2   1   1   2   2   1   2   1   2   2   1   1   2
## 435 436 437 438 439 440 441 442 443 444 445 446
##   2   1   2   1   1   1   2   2   2   2   1   2
```

```
nfl_2013$Playoff
```

```
##  [1] 0 0 0 0 1 0 0 0 1 0 1 0 1 0 1 0 0 1 1 0 0 1 0 1 1 1 0 0 0 0 1 0
```

```
table(testCluster$cluster, nfl_2013$Playoff)
```

```
##
##      0  1
##   1  3 11
##   2 17  1
```

The result demonstrates that Cluster 1 of our model predicts 79% of the teams correctly and does represent the playoff caliber teams. Cluster 2 yields a 5.5% of playoff team, showing that clustering really gives an accurate prediction on whether a team makes the NFL playoff based on its performance.