# Cluster Analysis HW

Kyle Barisone

4/21/2020

## Introduction

This set involves data based on MLB teams and has information including the amount of certain awards that team won in the previous year as well as their totals for certain statistics like hits, homeruns, etc.The data set includes both statistics on the teams hitting as well as pitching ability for that year. Using cluster analysis for our project, i want to see how teams are clustered based on their performance and statistics from each individual year.

## Data Preparation

```
team.Data <- readRDS("C:/Users/KBari/OneDrive/Desktop/Math 456/project/team.Data.rds")
team.Data$salary.x = NULL
team.Data$salary.y = NULL
```

I decided to start with the year 2016 and analyze the groupings of the american league, national league, then as the whole MLB. I filtered that data from our original data set to show 2016 data only then filtered the leagues into two separate data sets. I then only used columns 37-55 which contain numerical data representing certain stats for each of the teams. The columns that i cut out are all columns based on awards and i wanted to focus more on the stats of each team.Wins and losses were not included in the data because i want to see if clustering the teams based on their statistics (Hits, Runs, Runs allowed, etc.) for that year will separate them based on which teams made the playoffs.

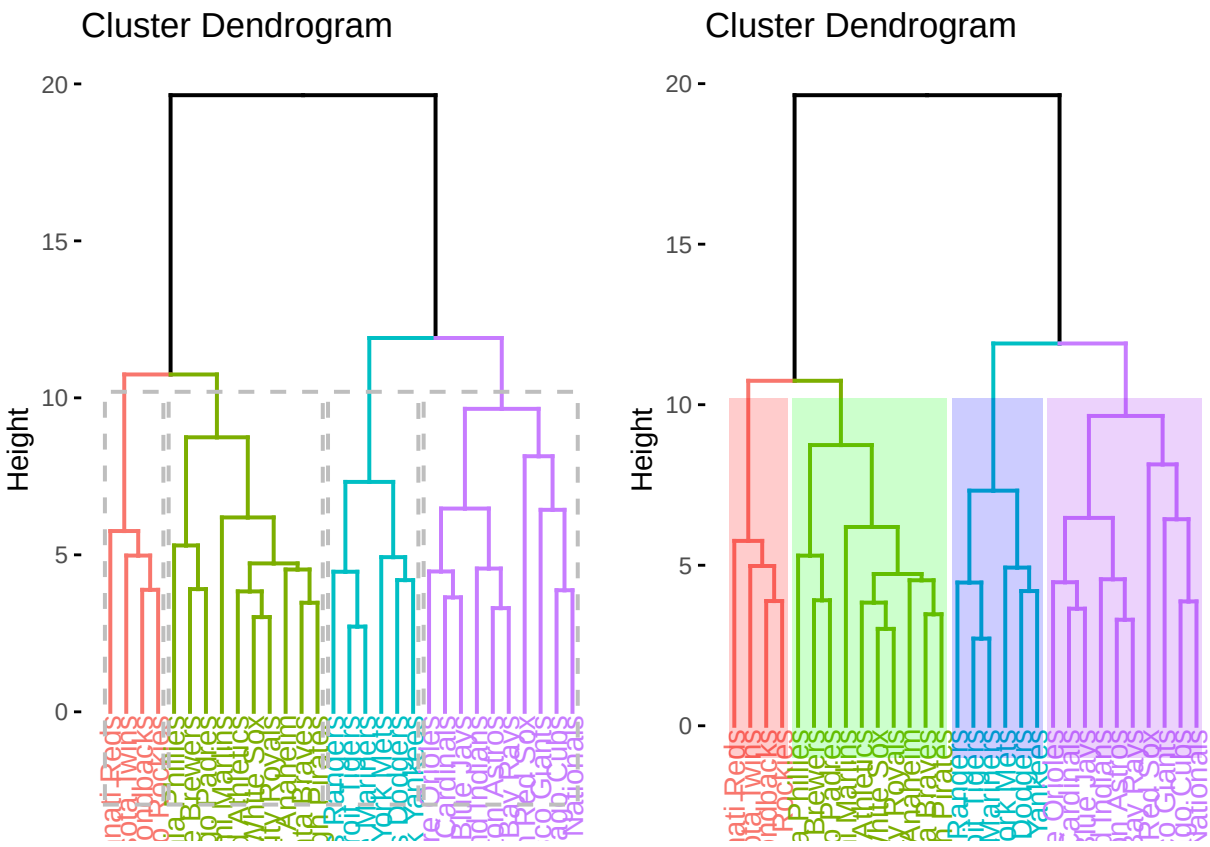
## Method 1 Hierarchical

For this method, I used Hierarchical clustering with euclidean distance and ward.D linkage to produce the following dendogram below.

```
team.Data1 <- team.Data %>% filter(yearID == 2016)
team.DataNL <- team.Data1 %>% filter(lgID == "NL")
team.DataAL <- team.Data1 %>% filter(lgID == "AL")

team.all <- team.Data[,c(3,4,37:55)]
team.Data1 <- team.Data1[,c(3,4,37:55)]
team.DataNL <- team.DataNL[,c(3,4,37:55)]
team.DataAL <- team.DataAL[,c(3,4,37:55)]
```

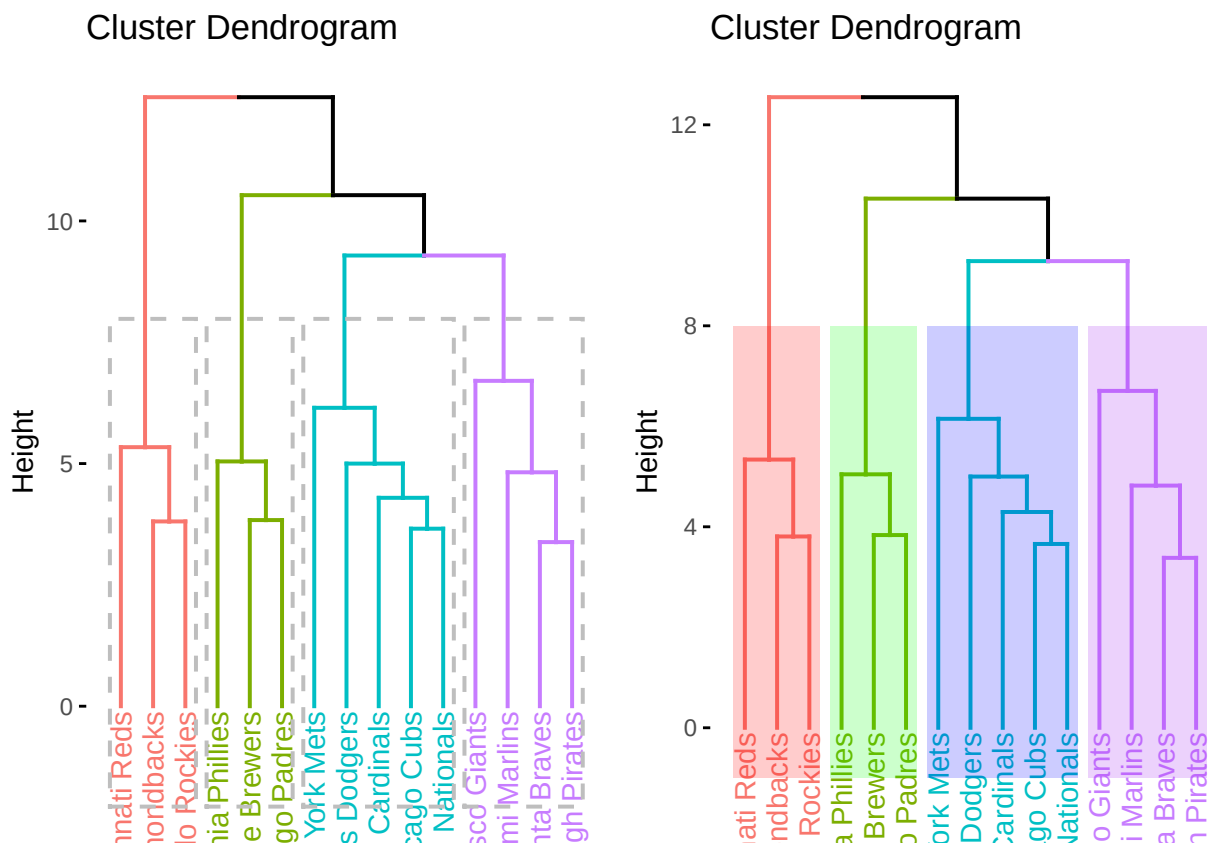
```
cluster.NL <- cbind(team.DataNL[,1:2],scale(team.DataNL[,3:21]))
cluster.AL <- cbind(team.DataAL[,1:2],scale(team.DataAL[,3:21]))
cluster.dta <- cbind(team.Data1[,1:2],scale(team.Data1[,3:21]))
cluster.all <- cbind(team.all[,1:2],scale(team.all[,3:21]))

cluster.NL <- cluster.NL %>% select(3:21)
cluster.AL <- cluster.AL %>% select(3:21)
cluster.dta <- cluster.dta %>% select(3:21)
cluster.all <- cluster.all %>% select(3:21)

rownames(cluster.NL) <- team.DataNL$name
rownames(cluster.AL) <- team.DataAL$name
rownames(cluster.dta) <- team.Data1$name

d.NL <- dist(cluster.NL, method="euclidean")
d.AL <- dist(cluster.AL, method="euclidean")
d <- dist(cluster.dta, method="euclidean")

clust.wardNL <- hclust(d.NL, method="ward.D")
clust.wardAL <- hclust(d.AL, method="ward.D")
clust.ward <- hclust(d, method="ward.D")
```

```
set.seed(4567)
a <- fviz_dend(clust.ward, rect=TRUE, k=4)
b <- fviz_dend(clust.ward,  k =4, rect = TRUE, color_labels_by_k = TRUE,
               rect_border = c("red", "green", "blue", "purple"),
               rect_fill = TRUE, lower_rect = -1)
grid.arrange(a, b, ncol=2)
```



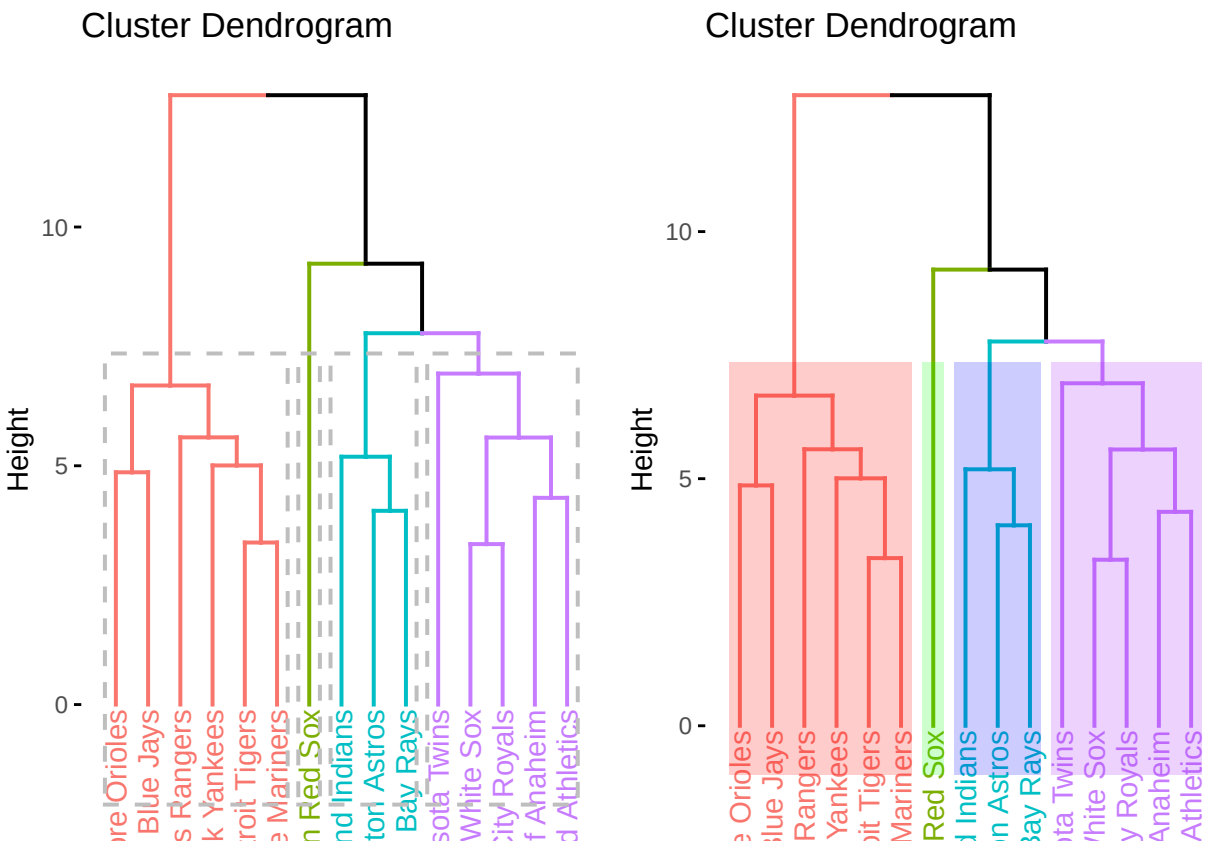

The clustering for all MLB teams in the year 2016 based on their stats is interesting because in both graphs, 2 clusters contain all 10 teams that made the playoffs for the year 2016. So it seems to be the case that teams that did not do as well(The White sox, Marlins, Reds) are also clustered together based on the provided statistics. However the data is very cluttered and hard to read when i knit so i decided to split the MLB into the national league and american league.

```
set.seed(4567)
a.NL <- fviz_dend(clust.wardNL, rect=TRUE, k=4)
b.NL <- fviz_dend(clust.wardNL,  k =4, rect = TRUE, color_labels_by_k = TRUE,
               rect_border = c("red", "green", "blue", "purple"),
               rect_fill = TRUE, lower_rect = -1)
grid.arrange(a.NL, b.NL, ncol=2)
```




Similar to the tree above, when we just look at the national league 2 clusters have all 5 of the national league teams that made the playoffs.(Cubs,Giants,Dodgers,Nationals,Mets).

```
set.seed(4567)
a.AL <- fviz_dend(clust.wardAL, rect=TRUE, k=4)
b.AL <- fviz_dend(clust.wardAL,  k =4, rect = TRUE, color_labels_by_k = TRUE,
               rect_border = c("red", "green", "blue", "purple"),
               rect_fill = TRUE, lower_rect = -1)
grid.arrange(a.AL, b.AL, ncol=2)
```

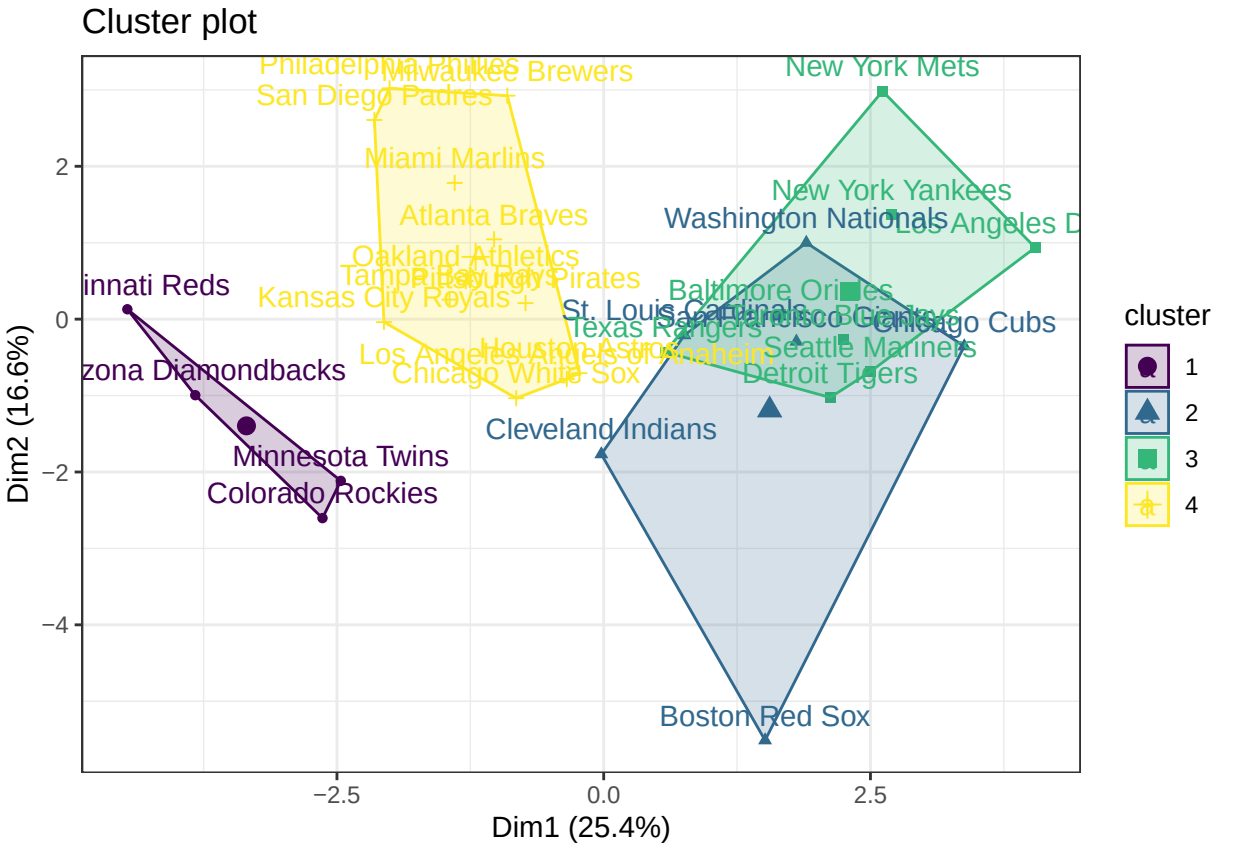



Here the left two clusters(red and green) contain 4 out of the 5 teams that made it to the playoffs and the purple represents teams that did not perform well. Interestingly the Red Sox are given their own cluster when the league is separated into AL and NL.

## Method 2

For my second method, I used k-means clustering i tested with 2,3, and 4 different clusters.

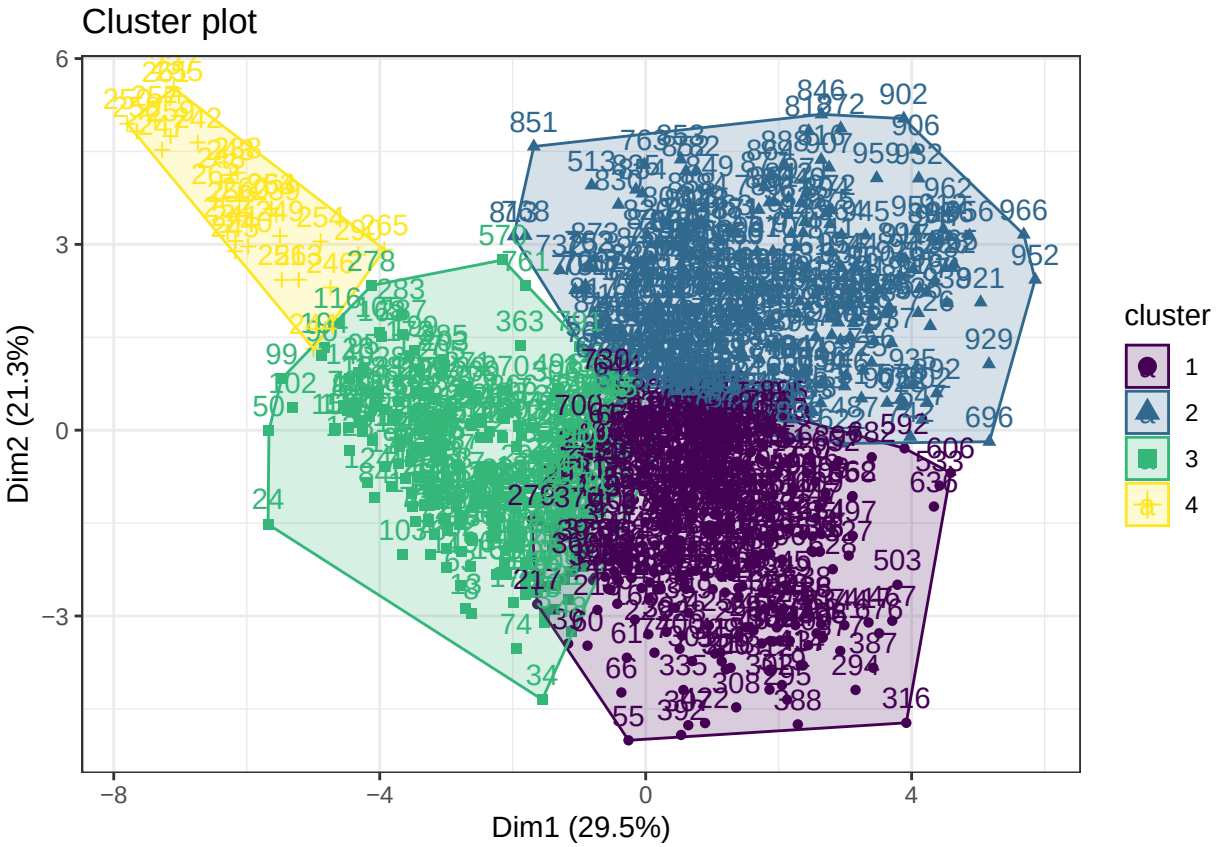
```
set.seed(4567)
k <- kmeans(cluster.dta, centers=4,nstart=10)


fviz_cluster(object=k, data=cluster.dta) +
theme_bw() + scale_colour_viridis_d() + scale_fill_viridis_d()
```




Again, all 10 teams that made the playoffs are in the clusters which score high on PC1. In addition teams with worse hitting and pitching stats are grouped together in purple. The Boston Red Sox look like outliers and score extremely high on PC2.

### Including data from all years

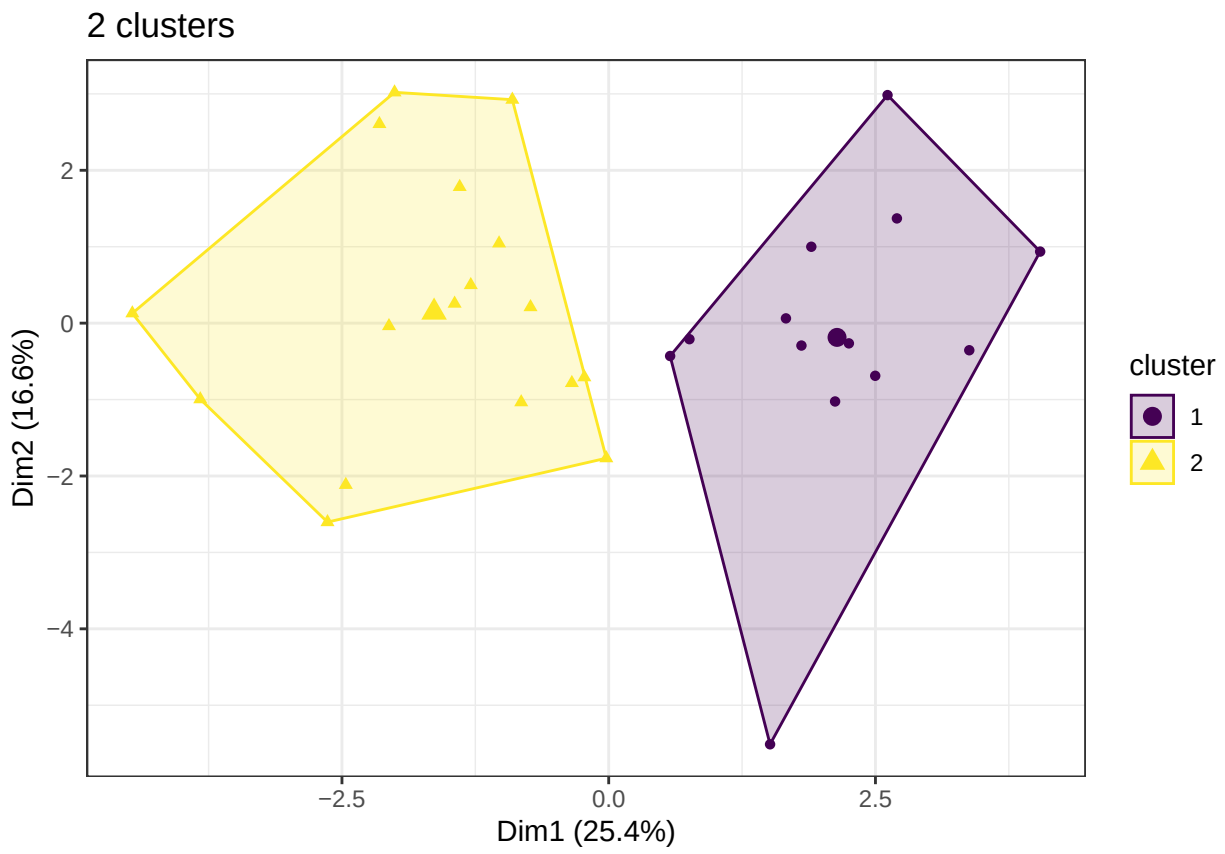
```
set.seed(4567)
k.all <- kmeans(cluster.all, centers=4,nstart=10)
fviz_cluster(object=k.all, data=cluster.all) +
theme_bw() + scale_colour_viridis_d() + scale_fill_viridis_d()
```



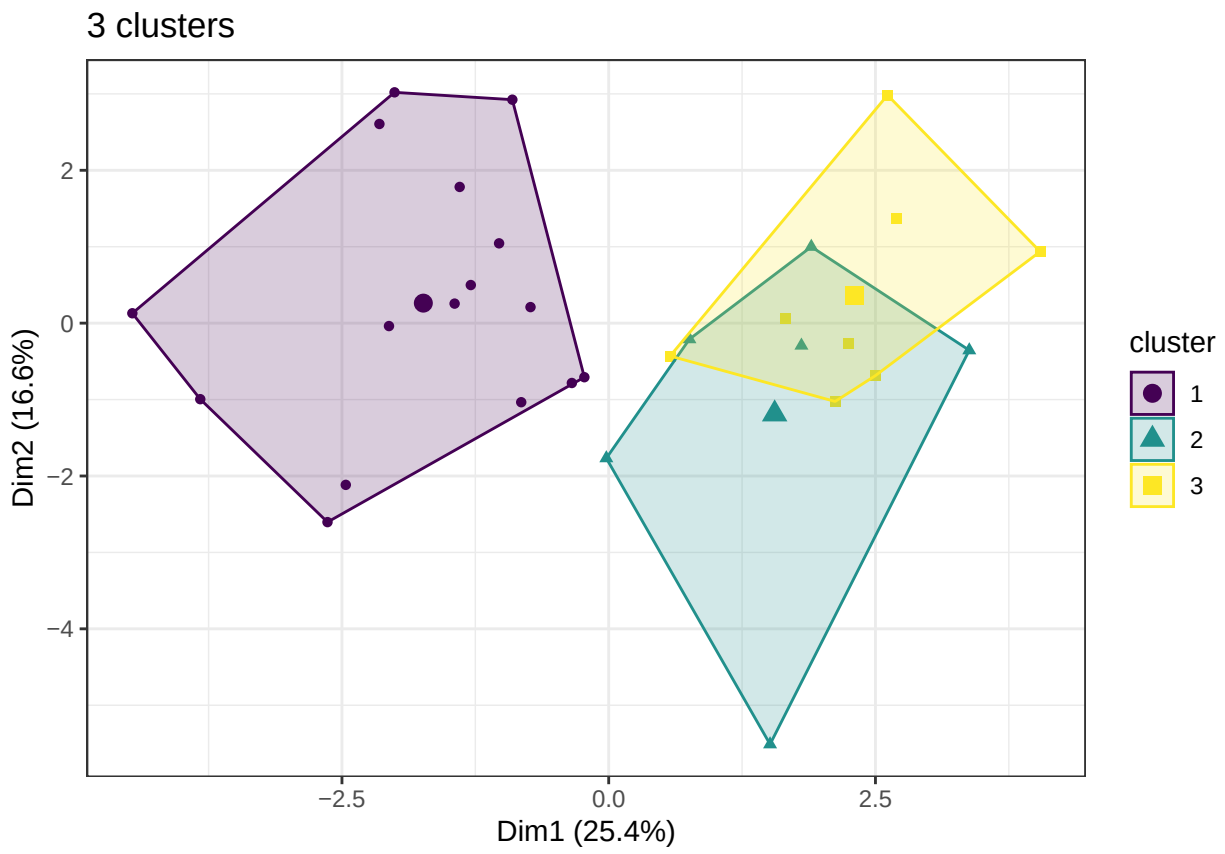

Below is what the clusters look like when k = 4.The blue cluster scores high on both pc's, while the purple cluster is high positive on pc1 but high negative on pc2. Finally, the yellow cluster is high on pc2 but high negative on pc1 There is a small amount of overlap but the clusters are separated and stand out on their own for the most part.

### Data from only 2016

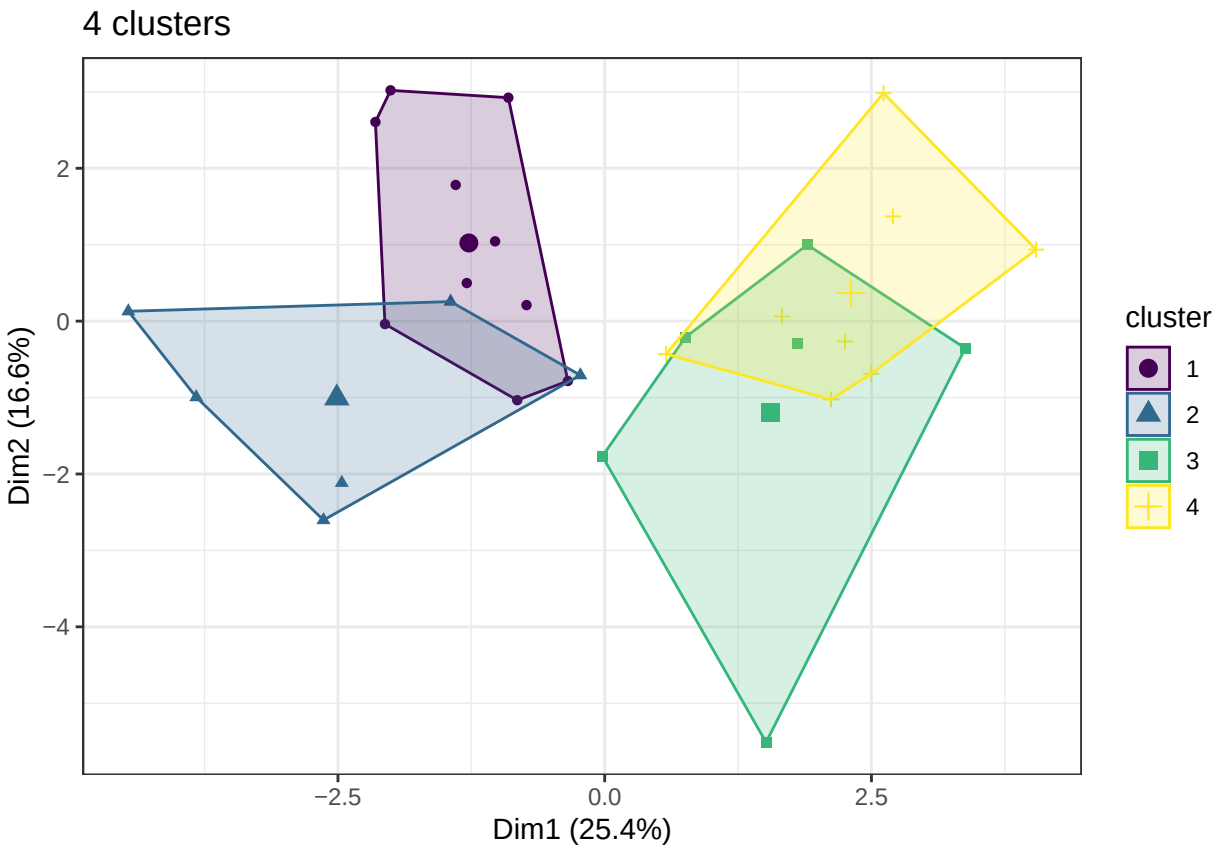
```
set.seed(4567)
nclust.2 <- kmeans(cluster.dta, centers=2, nstart=10) %>%
              fviz_cluster(data=cluster.dta, geom="point") + theme_bw() +
              ggtitle("2 clusters")+
              scale_colour_viridis_d() + scale_fill_viridis_d()
nclust.2
```



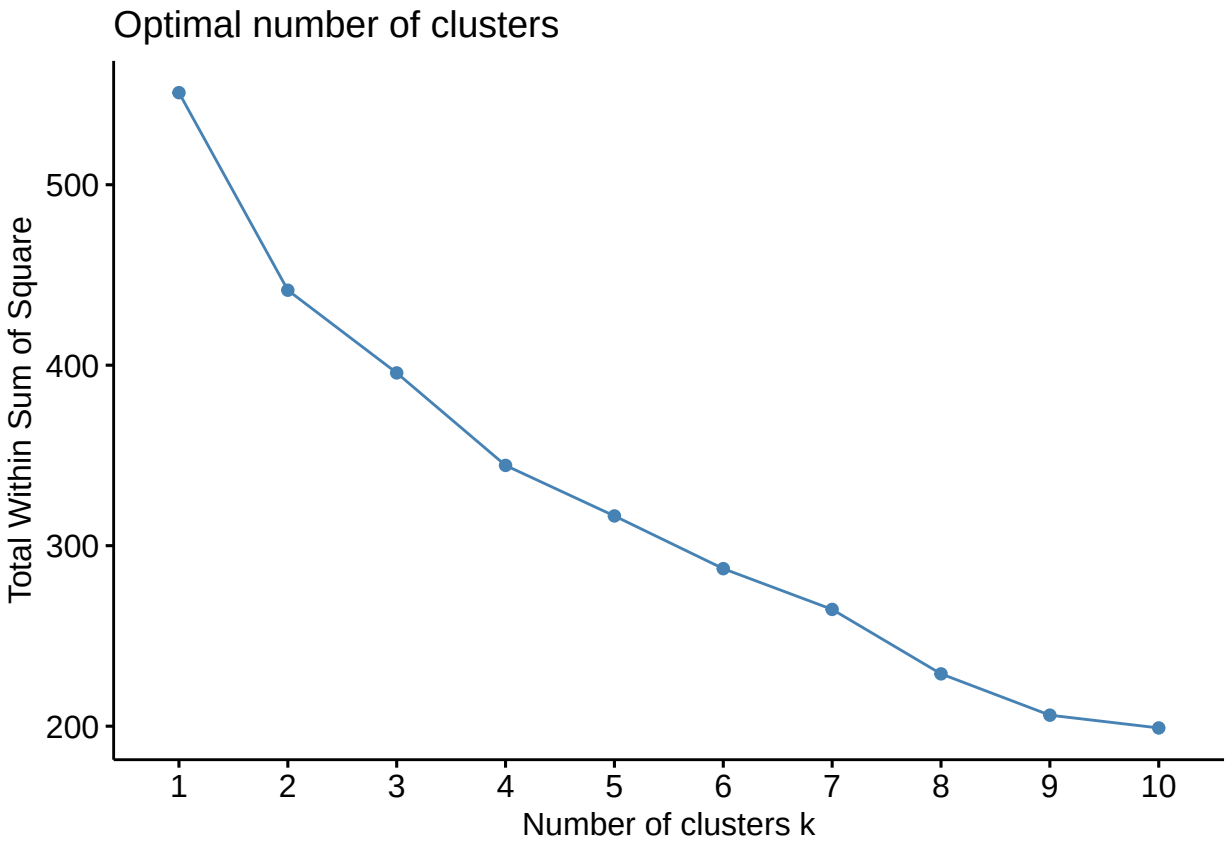

```
nclust.3 <- kmeans(cluster.dta, centers=3, nstart=10) %>%
              fviz_cluster(data=cluster.dta, geom="point") + theme_bw() +
              ggtitle("3 clusters")+
              scale_colour_viridis_d() + scale_fill_viridis_d()
nclust.3
```




```
nclust.4 <- kmeans(cluster.dta, centers=4, nstart=10) %>%
              fviz_cluster(data=cluster.dta, geom="point") + theme_bw() +
              ggtitle("4 clusters")+
              scale_colour_viridis_d() + scale_fill_viridis_d()
nclust.4
```




```
fviz_nbclust(cluster.dta, kmeans, method="wss")
```




## Conclusion:

Four clusters seems to be the best option when considering the value we should choose for k. There is a small amount of overlap between the clusters. The left 2 clusters are well separated from the others (high on PC1) and stand out on their own. The elbow method does not really flatten out at any point on the curve so relying on this plot does not seem helpful in this situation. Clustering the teams by the statistics for a particular year is very accurate when predicting which teams made the playoffs. When including data from all years, the clusters become harder to interpret because we dont know which data points represent which teams.