Project4

Audrey Holloman and Nicole Kadosh

10/18/2018

```
if (!require('Lahman'))
{
  install.packages('Lahman');
  library(Lahman);
}
```

```
## Loading required package: Lahman
```

```
# install.packages(Lahman)
# library(Lahman)
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
```

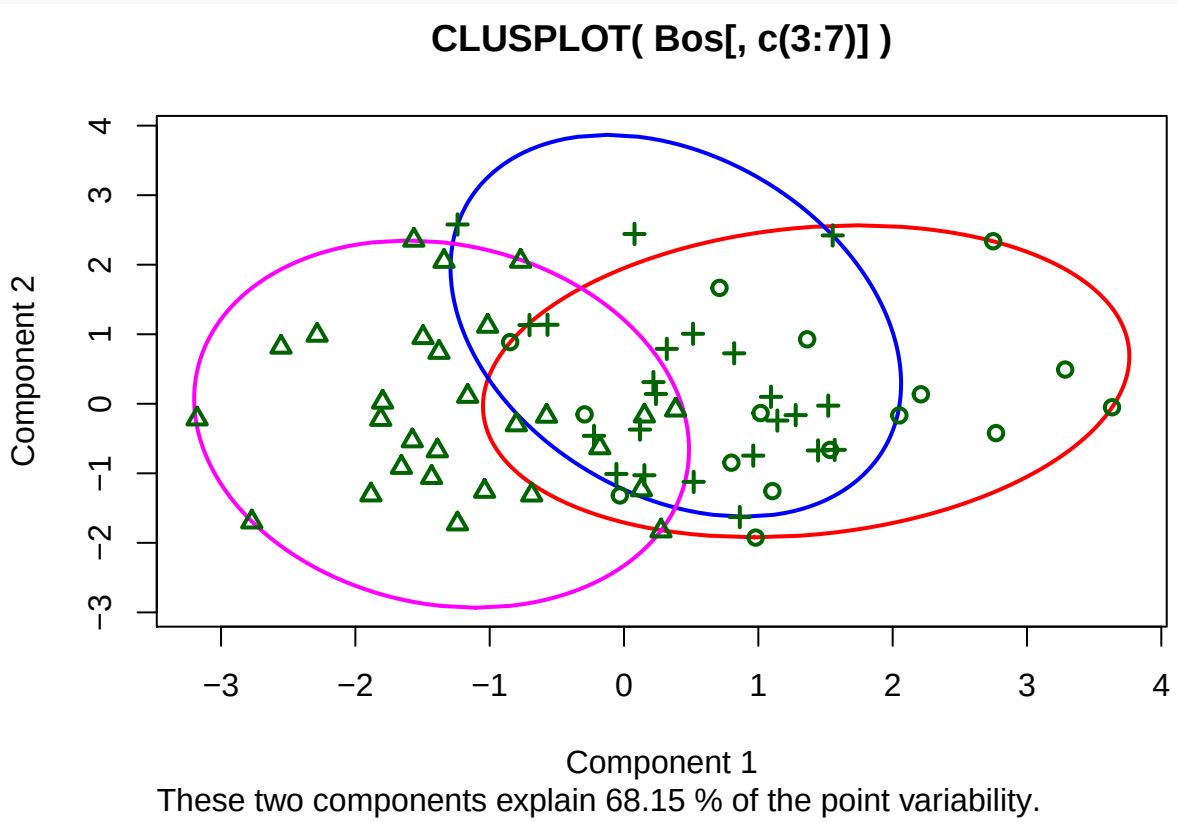
```
library(gapminder)
library(ggplot2)
library(cluster)
data <- Teams %>% filter (yearID > 1949, franchID %in% c("BOS", "CLE", "ATL", "NYY", "CHC"))
attach(data)
data$HG <- (H / G)
data$HRG <- (HR / G)
data$BBG <- (BB / G)
data$WG <- (W / G)
data$RAG <- (RA / G)
data$ABG <- (AB / G)
data$SOG <- (SO / G)
NewData <- data %>% select("franchID", "yearID", "HG", "HRG", "BBG", "WG", "RAG", "ABG", "SOG")
Bos <- NewData %>% filter (franchID %in% c("BOS"))
Cle <- NewData %>% filter (franchID %in% c("CLE"))
Atl <- NewData %>% filter (franchID %in% c("ATL"))
Nyy <- NewData %>% filter (franchID %in% c("NYY"))
Chc <- NewData %>% filter (franchID %in% c("CHC"))

#BOS
TeamClusterBOS <- kmeans(Bos[,3:7],3, nstart = 20)
TeamClusterBOS
```

```
## K-means clustering with 3 clusters of sizes 16, 28, 23
##
## Cluster means:
##         HG      HRG      BBG        WG      RAG
```

```
## 1 9.671534 1.084090 4.187244 0.5559273 4.782201
## 2 8.729978 0.875294 3.546723 0.5052396 4.340154
## 3 9.589751 1.071854 3.366767 0.5361103 4.572536
##
## Clustering vector:
##  [1] 1 1 2 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 2 3 3 3 3 3 3
## [36] 3 2 1 3 1 2 2 2 2 3 1 1 3 3 3 3 3 3 1 1 1 1 1 1 1 3 3 3 3 2 3 3
##
## Within cluster sum of squares by cluster:
## [1]  8.161162 10.240309  7.091763
##  (between_SS / total_SS =  47.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
ClusplotBOS <- clusplot(Bos[,c(3:7)], TeamClusterBOS$cluster, color=T, lwd=2)
```

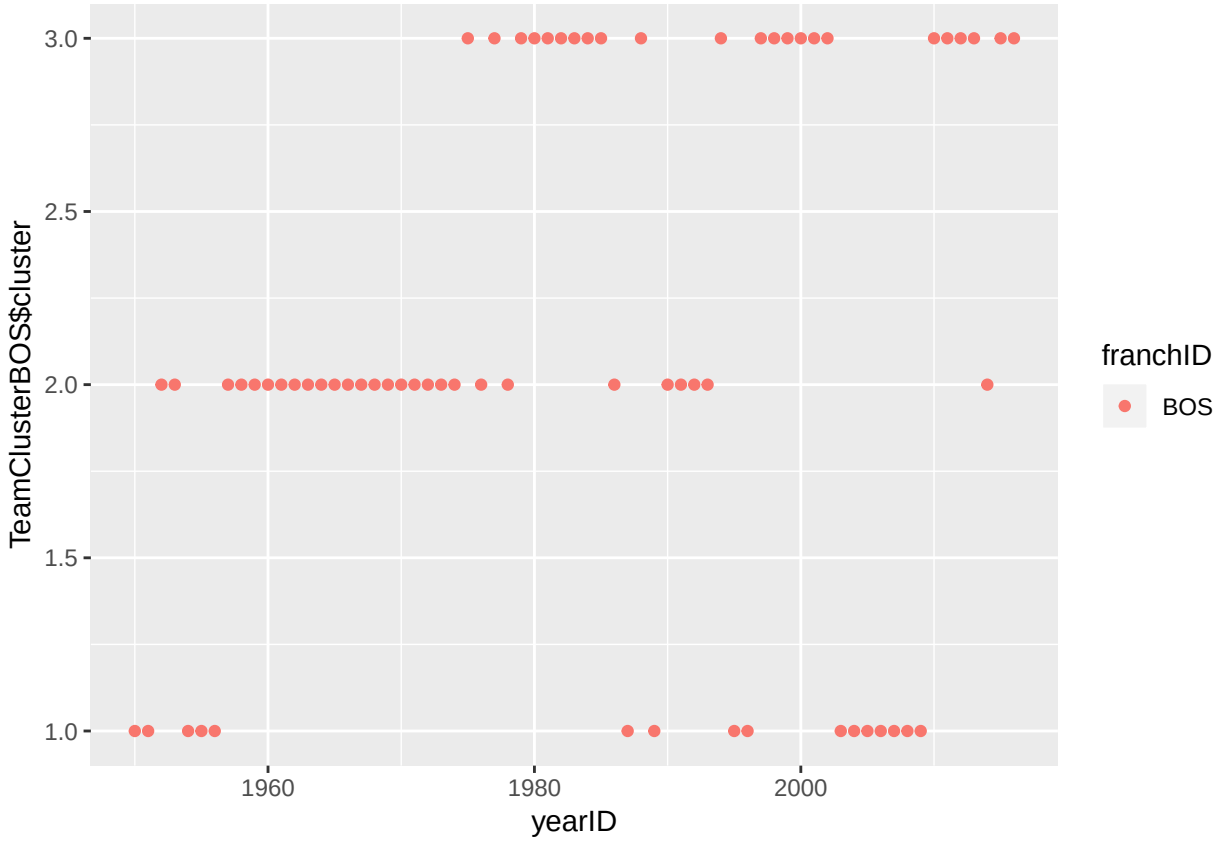



```
ClusplotBOS
```

```
## $Distances
##      [,1] [,2] [,3]
## [1,]    0   NA   NA
## [2,]   NA    0   NA
## [3,]   NA   NA    0
##
## $Shading
```

```
## [1] 10.90519 18.29942 16.79539
```

```
ggplot(Bos, aes(x = yearID, y = TeamClusterBOS$cluster, color = franchID)) + geom_point()
```

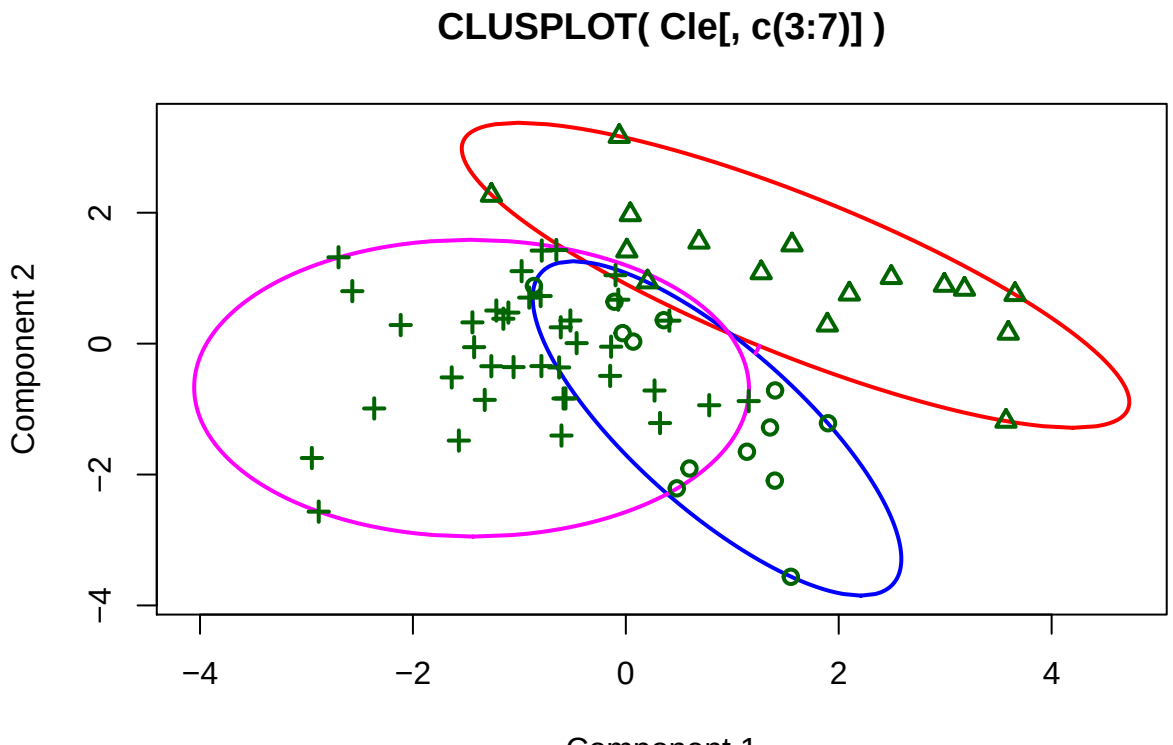



```
#CLE
TeamClusterCLE <- kmeans(Cle[,3:7],3, nstart = 20)
TeamClusterCLE
```

```
## K-means clustering with 3 clusters of sizes 13, 16, 38
##
## Cluster means:
##         HG       HRG      BBG        WG      RAG
## 1 8.854068 0.9019822 4.041508 0.5566466 4.239630
## 2 9.717607 1.1512966 3.618561 0.5184185 5.071891
## 3 8.614435 0.8658854 3.040056 0.4799239 4.323823
##
## Clustering vector:
##  [1] 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 2 3 1 1 1
## [36] 2 2 2 3 3 3 3 3 2 2 2 2 2 2 2 2 3 3 2 3 2 1 3 2 3 3 3 3 3 3 3
##
## Within cluster sum of squares by cluster:
## [1]  6.187516  9.427750 18.855428
##  (between_SS / total_SS =  49.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
```

```
## [9] "ifault"
```

```
ClusplotCLE <- clusplot(Cle[,c(3:7)], TeamClusterCLE$cluster, color=T, lwd=2)
```




```
ClusplotCLE
```

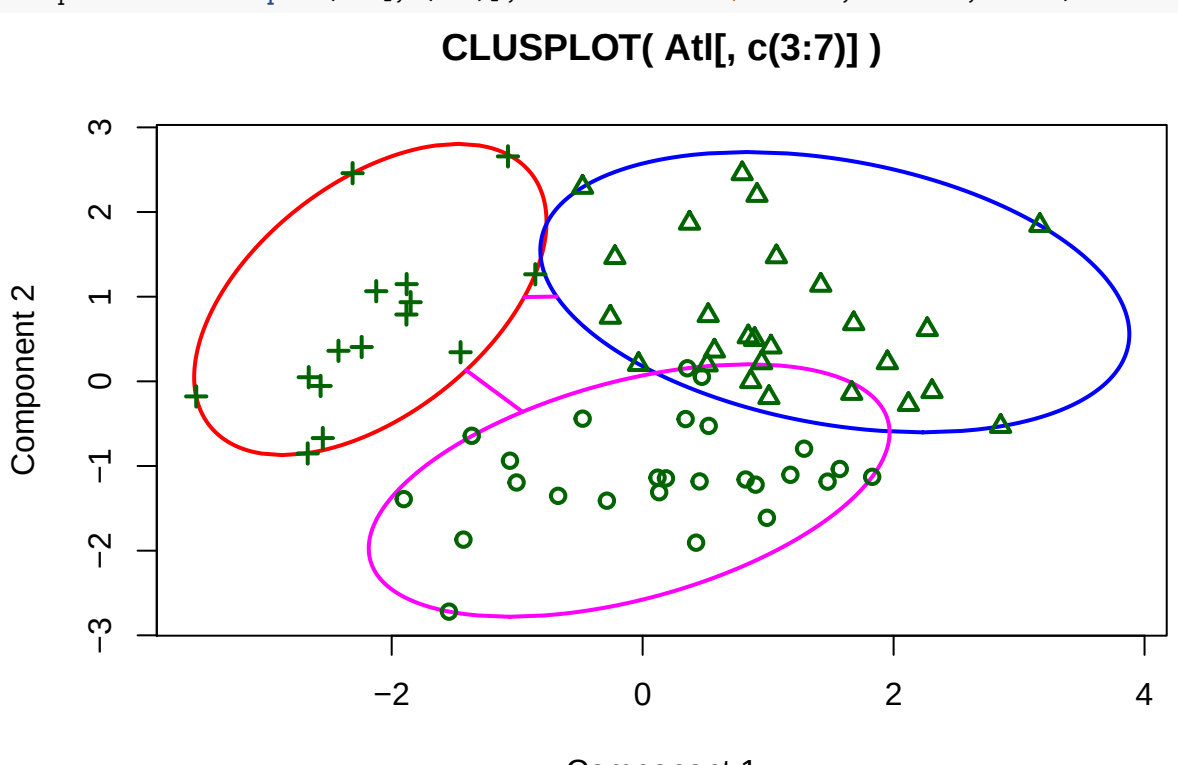
```
## $Distances
##        [,1]   [,2] [,3]
## [1,] 0.0000 0.1185   NA
## [2,] 0.1185 0.0000   NA
## [3,]     NA     NA    0
##
## $Shading
## [1] 14.53777 12.66195 18.80028
```

```
ggplot(Cle, aes(x = yearID, y = TeamClusterCLE$cluster, color = franchID)) + geom_point()
```


```
#ATL
TeamClusterATL <- kmeans(Atl[,3:7],3, nstart = 20)
TeamClusterATL
```

```
## K-means clustering with 3 clusters of sizes 26, 26, 15
##
## Cluster means:
##         HG       HRG      BBG        WG      RAG
## 1 8.574666 0.9324759 3.165272 0.5531853 3.762550
## 2 9.183088 1.1121504 3.421291 0.5530625 4.271529
## 3 8.377037 0.7708908 3.248226 0.4117931 4.708821
##
## Clustering vector:
##  [1] 2 2 3 1 1 2 1 2 1 2 2 2 1 1 2 1 2 1 1 1 2 1 3 2 1 3 3 3 3 3 1 1 2 2 1
## [36] 3 3 3 3 3 3 1 1 1 2 1 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 1 1 1 1 1 3 3
##
## Within cluster sum of squares by cluster:
## [1] 7.678830 7.695103 4.739724
##  (between_SS / total_SS =  48.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
ClusplotATL <- clusplot(Atl[,c(3:7)], TeamClusterATL$cluster, color=T, lwd=2)
```




```
ClusplotATL
```

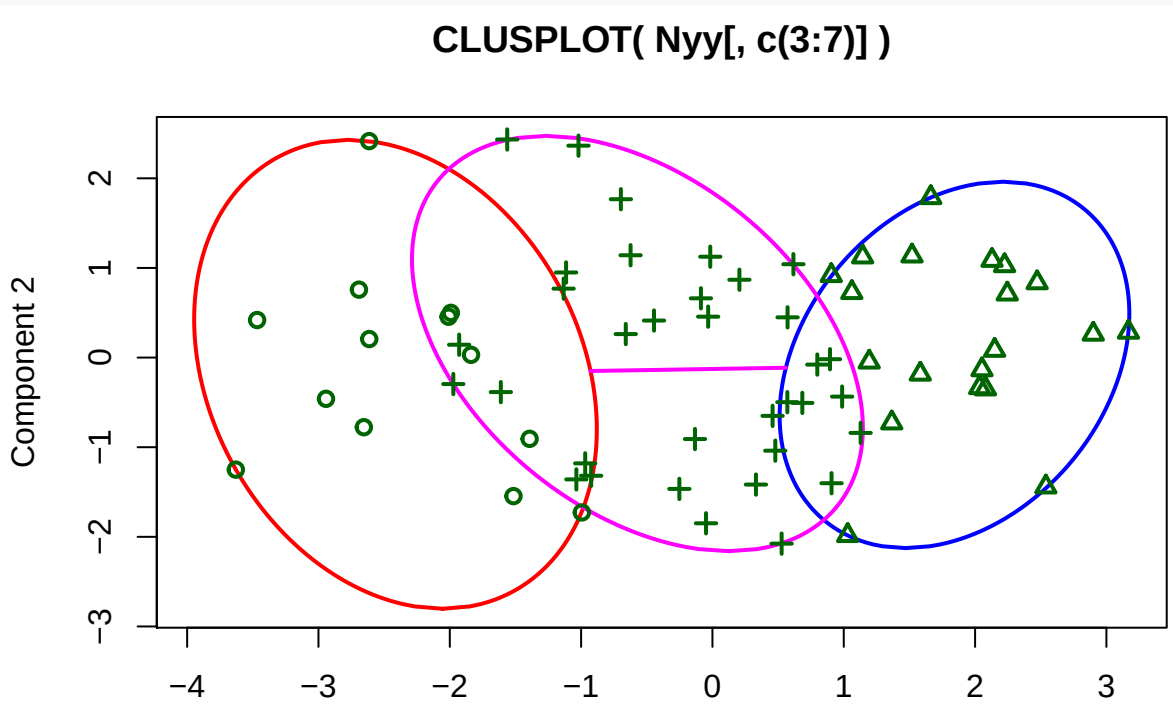
```
## $Distances
##           [,1]      [,2]      [,3]
## [1,] 0.0000000        NA 0.6661721
## [2,]        NA 0.0000000 0.2586315
## [3,] 0.6661721 0.2586315 0.0000000
##
## $Shading
## [1] 18.08993 14.19249 13.71758
```

```
ggplot(Atl, aes(x = yearID, y = TeamClusterATL$cluster, color = franchID)) + geom_point()
```


```
#NYY
TeamClusterNYY <- kmeans(Nyy[,3:7],3, nstart = 20)
TeamClusterNYY
```

```
## K-means clustering with 3 clusters of sizes 13, 20, 34
##
## Cluster means:
##         HG       HRG      BBG        WG      RAG
## 1 8.059148 0.8108369 3.174331 0.5073083 3.754091
## 2 9.665995 1.2117289 4.122395 0.6030411 4.533039
## 3 9.059074 1.0270545 3.416401 0.5699945 4.090204
##
## Clustering vector:
##  [1] 2 3 3 2 3 3 2 3 3 3 3 3 3 1 3 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 1 3 3 3
## [36] 3 2 3 3 3 1 3 3 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 3 2 2 3 3 1 1 3 3
##
## Within cluster sum of squares by cluster:
## [1]  6.592702  6.220640 13.018360
##  (between_SS / total_SS =  58.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

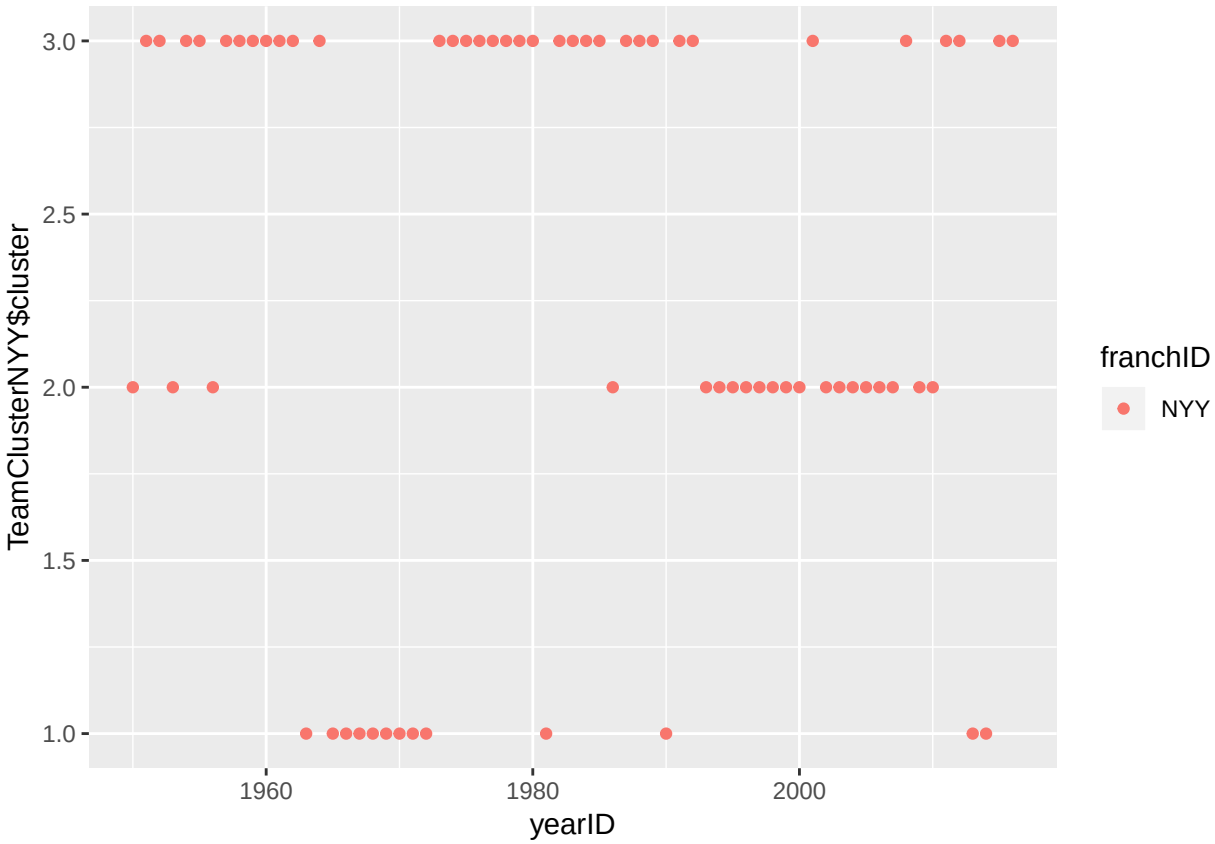
```
ClusplotNYY <- clusplot(Nyy[,c(3:7)], TeamClusterNYY$cluster, color=T, lwd=2)
```




```
ClusplotNYY
```

```
## $Distances
##          [,1]     [,2] [,3]
## [1,] 0.000000 1.487989   NA
## [2,] 1.487989 0.000000   NA
## [3,]       NA       NA    0
##
## $Shading
## [1]  9.064862 16.920234 20.014904
```

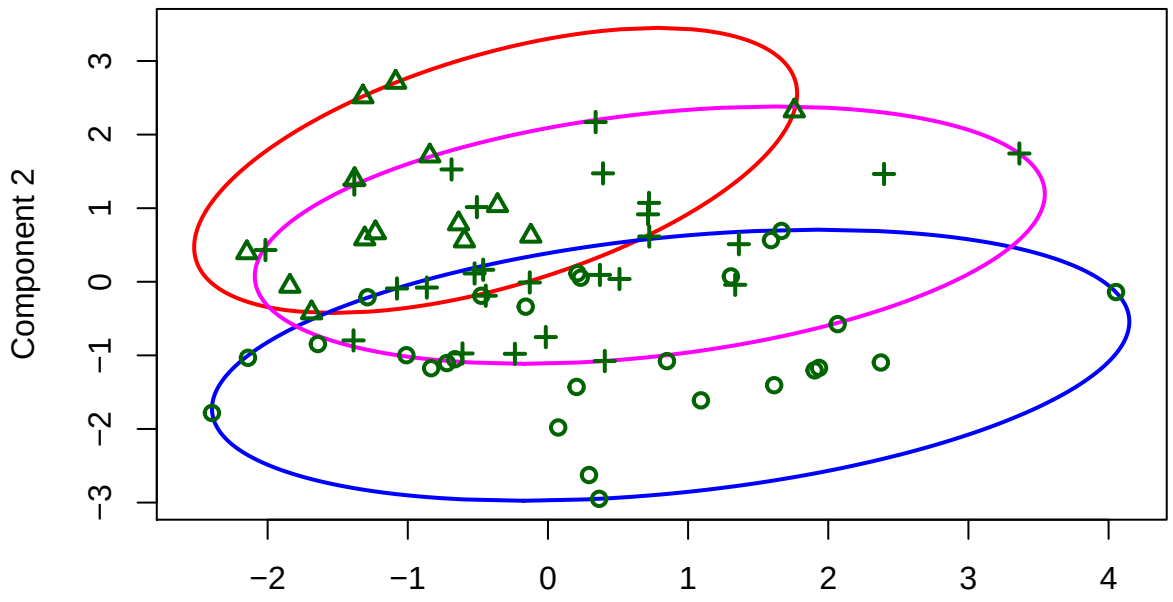
```
ggplot(Nyy, aes(x = yearID, y = TeamClusterNYY$cluster, color = franchID)) + geom_point()
```




```
#CHC
TeamClusterCHC <- kmeans(Chc[,3:7],3, nstart = 20)
TeamClusterCHC
```

```
## K-means clustering with 3 clusters of sizes 27, 14, 26
##
## Cluster means:
##         HG       HRG      BBG        WG      RAG
## 1 8.425746 0.8931308 3.179555 0.4920792 4.156246
## 2 8.662680 0.9577537 3.460518 0.4251112 5.078132
## 3 9.064363 0.9302729 2.943833 0.4731634 4.557316
##
## Clustering vector:
##  [1] 2 2 3 2 3 1 1 1 3 1 2 2 2 1 1 1 3 1 1 1 1 1 1 1 2 2 1 3 3 3 3 1 3 3 1
## [36] 1 2 3 3 1 3 3 1 3 3 3 2 3 2 2 2 1 2 3 3 3 3 3 3 1 3 3 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 10.937534  3.786966  7.658825
##  (between_SS / total_SS =  41.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

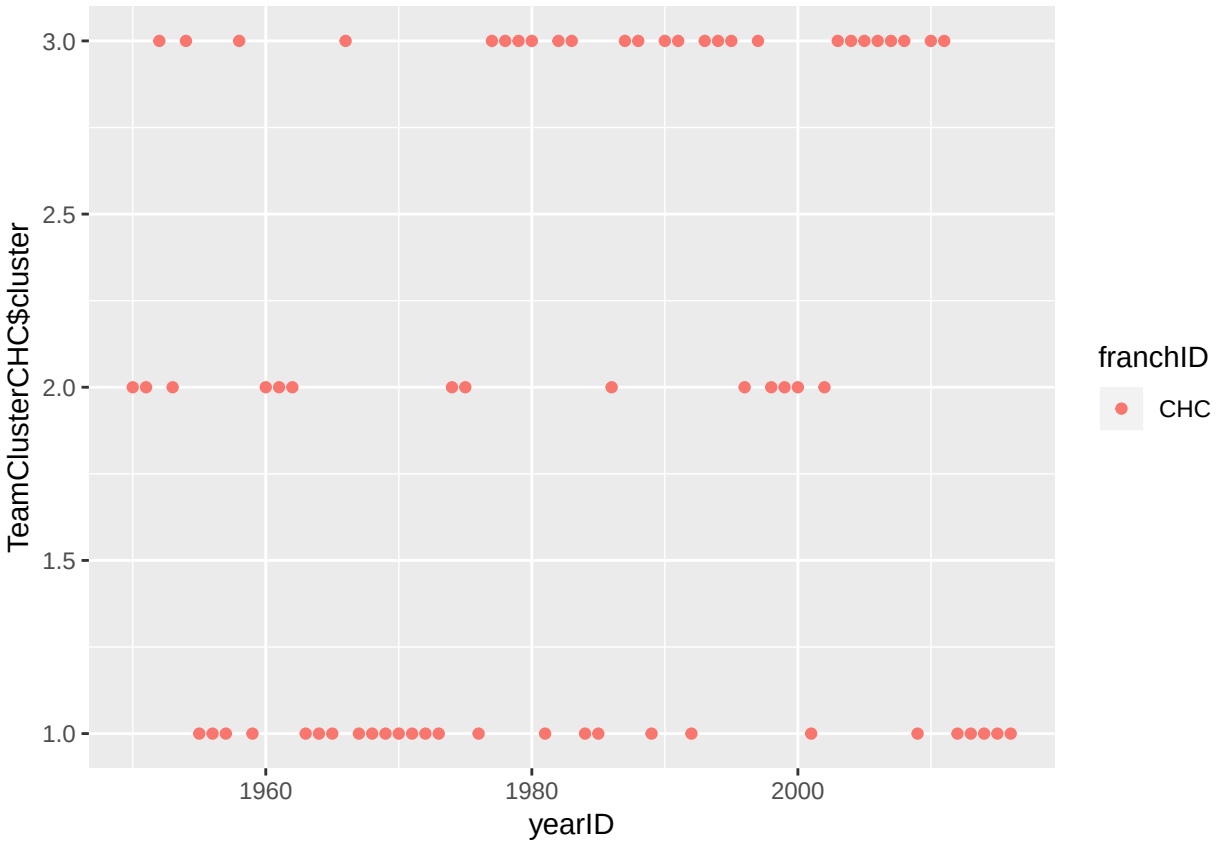
```
ClusplotCHC <- clusplot(Chc[,c(3:7)], TeamClusterCHC$cluster, color=T, lwd=2)
```




```
ClusplotCHC
```

```
## $Distances
##      [,1] [,2] [,3]
## [1,]    0   NA   NA
## [2,]   NA    0   NA
## [3,]   NA   NA    0
##
## $Shading
## [1] 15.25116 13.32623 17.42262
```

```
ggplot(Chc, aes(x = yearID, y = TeamClusterCHC$cluster, color = franchID)) + geom_point()
```



After analyzing the cluster plots and time series graphs we chose to only talk about two teams because they had the best cluster results. The first team is Boston, it contains only one main cluster from the late 1950s to the mid 1970s. We first determined that this is a cluster because in the time series there's an obvious pattern. This corresponded to cluster number two, which led us to the cluster means. We decided that this cluster is relatively good because the cluster means resulted in good values compared to the other team's clusters. The next team we chose was the New York Yankees, which contained two major clusters and overall had the best cluster. It also had the best looking cluster plot with the highest percent of point variability, 78.09%. The first cluster ranged from the mid 1970s to the early 1990s, this is the better cluster of the two as it had better cluster mean values. The second cluster ranged from the late 1990s to around 2010, which also had good cluster mean values but were not as good as the previous. These clusters are showing that the Yankees had a good streak from the mid 1970s to around 2010, considering how close in years the two clusters are.