Final Project

Rohith Desamseety

2022-12-17

```
library(readr)
Universal_RD <- read.csv("~/Downloads/mlb-umpire-scorecard.csv", header=TRUE, stringsAsFactors=FALSE)
View(Universal_RD)
#Downloaded from https://www.kaggle.com/datasets/mattop/mlb-baseball-umpire-scorecards-2015-2022

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(factoextra)
```

```
## Loading required package: ggplot2
```

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

```
library(flexclust)
```

```
## Loading required package: grid
```

```
## Loading required package: lattice
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
library(ggsignif)
library(backports)
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(tinytex)
```

```
summary(Universal_RD)
```

```
##        id            date              umpire              home
##  Min.   :    1   Length:18213       Length:18213       Length:18213
##  1st Qu.: 4554   Class :character   Class :character   Class :character
##  Median : 9107   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 9107
##  3rd Qu.:13660
##  Max.   :18213
##      away           home_team_runs   away_team_runs   pitches_called
##  Length:18213       Min.   : 0.000   Min.   : 0.000   Length:18213
##  Class :character   1st Qu.: 2.000   1st Qu.: 2.000   Class :character
##  Mode  :character   Median : 4.000   Median : 4.000   Mode  :character
##                     Mean   : 4.559   Mean   : 4.433
##                     3rd Qu.: 6.000   3rd Qu.: 6.000
##                     Max.   :29.000   Max.   :28.000
##  incorrect_calls    expected_incorrect_calls correct_calls
##  Length:18213       Length:18213             Length:18213
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##  expected_correct_calls correct_calls_above_expected   accuracy
##  Length:18213           Length:18213                 Length:18213
##  Class :character       Class :character             Class :character
##  Mode  :character       Mode  :character             Mode  :character
##
##
##
##  expected_accuracy  accuracy_above_expected consistency
##  Length:18213       Length:18213            Length:18213
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##   favor_home        total_run_impact
##  Length:18213       Length:18213
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
Data <-  subset(Universal_RD, select = -c(1,2,4,5,6,7))
#Remove irrelevant columns and convert all data to numbers.
Data$pitches_called <- as.numeric(Data$pitches_called)
```

```
## Warning: NAs introduced by coercion
```

```
Data$incorrect_calls <- as.numeric(Data$incorrect_calls)
```

```
## Warning: NAs introduced by coercion
```

```
Data$expected_incorrect_calls <- as.numeric(Data$expected_incorrect_calls)
```

```
## Warning: NAs introduced by coercion
```

```
Data$correct_calls <- as.numeric(Data$correct_calls)
```

```
## Warning: NAs introduced by coercion
```

```
Data$expected_correct_calls <- as.numeric(Data$expected_correct_calls)
```

```
## Warning: NAs introduced by coercion
```

```
Data$correct_calls_above_expected <- as.numeric(Data$correct_calls_above_expected)
```

```
## Warning: NAs introduced by coercion
```

```
Data$accuracy <- as.numeric(Data$accuracy)
```

```
## Warning: NAs introduced by coercion
```

```
Data$expected_accuracy <- as.numeric(Data$expected_accuracy)
```

```
## Warning: NAs introduced by coercion
```

```
Data$accuracy_above_expected <- as.numeric(Data$accuracy_above_expected)
```

```
## Warning: NAs introduced by coercion
```

```
Data$consistency <- as.numeric(Data$consistency)
```

```
## Warning: NAs introduced by coercion
```

```
Data$favor_home <- as.numeric(Data$favor_home)
```

```
## Warning: NAs introduced by coercion
```

```
Data$total_run_impact <- as.numeric(Data$total_run_impact)
```

```
## Warning: NAs introduced by coercion
```

```
summary(Data)
```

```
##     umpire          pitches_called  incorrect_calls expected_incorrect_calls
##  Length:18213       Min.   : 68.0   Min.   : 0.0    Min.   : 3.10
##  Class :character   1st Qu.:138.0   1st Qu.: 8.0    1st Qu.: 9.60
##  Mode  :character   Median :153.0   Median :11.0    Median :11.60
##                     Mean   :154.6   Mean   :11.7    Mean   :11.92
##                     3rd Qu.:169.0   3rd Qu.:14.0    3rd Qu.:13.90
##                     Max.   :375.0   Max.   :45.0    Max.   :43.90
##                     NA's   :120     NA's   :120     NA's   :120
##  correct_calls   expected_correct_calls correct_calls_above_expected
##  Min.   : 63.0   Min.   : 63.0          Min.   :-24.5000
##  1st Qu.:127.0   1st Qu.:127.5          1st Qu.: -1.9000
##  Median :141.0   Median :141.1          Median :  0.4000
##  Mean   :142.9   Mean   :142.6          Mean   :  0.2114
##  3rd Qu.:156.0   3rd Qu.:155.9          3rd Qu.:  2.5000
##  Max.   :331.0   Max.   :331.1          Max.   : 16.1000
##  NA's   :120     NA's   :120            NA's   :120
##     accuracy      expected_accuracy accuracy_above_expected  consistency
##  Min.   : 78.40   Min.   :85.00     Min.   :-11.7000        Min.   : 81.40
##  1st Qu.: 90.70   1st Qu.:91.20     1st Qu.: -1.3000        1st Qu.: 91.70
##  Median : 92.70   Median :92.50     Median :  0.2000        Median : 93.30
##  Mean   : 92.42   Mean   :92.28     Mean   :  0.1356        Mean   : 93.17
##  3rd Qu.: 94.40   3rd Qu.:93.50     3rd Qu.:  1.7000        3rd Qu.: 94.70
```

```
##  Max.   :100.00   Max.   :97.40     Max.   :  9.4000        Max.   :100.00
##  NA's   :120      NA's   :120       NA's   :120             NA's   :120
##    favor_home       total_run_impact
##  Min.   :-3.45000   Min.   :0.000
##  1st Qu.:-0.33000   1st Qu.:0.970
##  Median : 0.03000   Median :1.410
##  Mean   : 0.03454   Mean   :1.532
##  3rd Qu.: 0.40000   3rd Qu.:1.950
##  Max.   : 3.40000   Max.   :7.140
##  NA's   :120        NA's   :120
```

```
nonaData <- na.omit(Data)
newdf <- nonaData %>% group_by(umpire) %>% summarise_each(funs(mean))
```

```
## Warning: `summarise_each_()` was deprecated in dplyr 0.7.0.
## Please use `across()` instead.
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```
head(newdf)
```

```
## # A tibble: 6 x 13
##   umpire         pitch~1 incor~2 expec~3 corre~4 expec~5 corre~6 accur~7 expec~8
##   <chr>            <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Adam Beck         153.    8.56   10.4     144.    142.   1.8      94.4    93.2
## 2 Adam Hamari       152.   10.5    11.7     141.    140.   1.20     93.0    92.3
## 3 Adrian Johnson    156.   13.0    12.2     143.    144.  -0.854    91.7    92.2
## 4 Alan Porter       156.   10.4    12.0     145.    143.   1.60     93.3    92.3
## 5 Alex MacKay       153.    8.83    9.02    144     144.   0.183    94.0    94.0
## 6 Alex Tosi         157.    8.63   11.3     148.    146.   2.63     94.5    92.8
## # ... with 4 more variables: accuracy_above_expected <dbl>, consistency <dbl>,
## #   favor_home <dbl>, total_run_impact <dbl>, and abbreviated variable names
## #   1: pitches_called, 2: incorrect_calls, 3: expected_incorrect_calls,
## #   4: correct_calls, 5: expected_correct_calls,
## #   6: correct_calls_above_expected, 7: accuracy, 8: expected_accuracy
```

```
metricsdf <- subset (newdf, select =-c(2,3,4,5,6,8,9))
#Because the primary focus is on umpire performance that is above or below the acceptable/expected leve
head(metricsdf)
```

```
## # A tibble: 6 x 6
##   umpire         correct_calls_above_expected accurac~1 consi~2 favor_~3 total~4
##   <chr>                                 <dbl>     <dbl>   <dbl>    <dbl>   <dbl>
## 1 Adam Beck                              1.8    1.20e+ 0    94.1  0.00177    1.18
## 2 Adam Hamari                            1.20   7.77e- 1    93.5 -0.0226     1.30
## 3 Adrian Johnson                        -0.854 -5.35e- 1    92.8  0.0710     1.65
## 4 Alan Porter                            1.60   1.04e+ 0    93.5 -0.0302     1.36
```

```
## 5 Alex MacKay                          0.183  7.40e-17    93.7  0.25       1.14
## 6 Alex Tosi                            2.63   1.67e+ 0    93.6  0.150      1.13
## # ... with abbreviated variable names 1: accuracy_above_expected,
## #   2: consistency, 3: favor_home, 4: total_run_impact
```

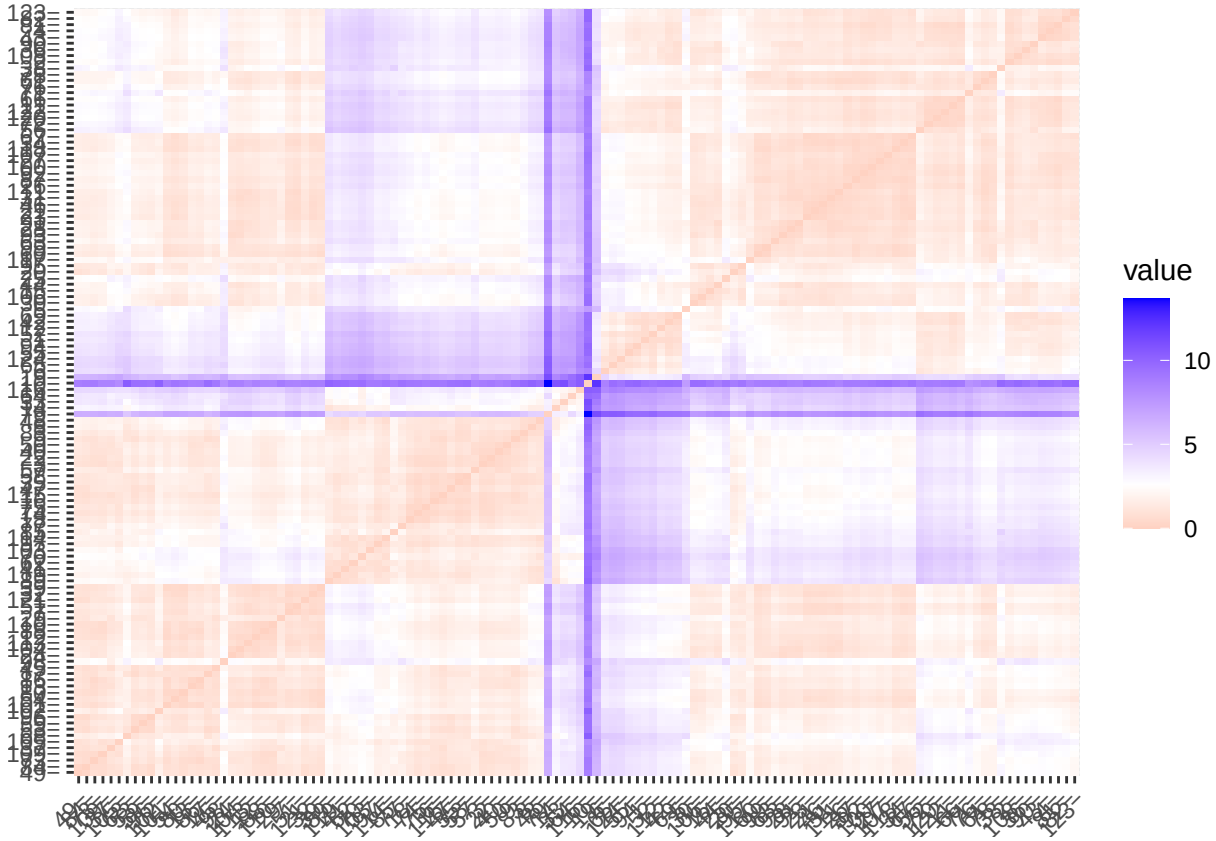
```
rownames(metricsdf) <- metricsdf$umpire
```

```
## Warning: Setting row names on a tibble is deprecated.
```

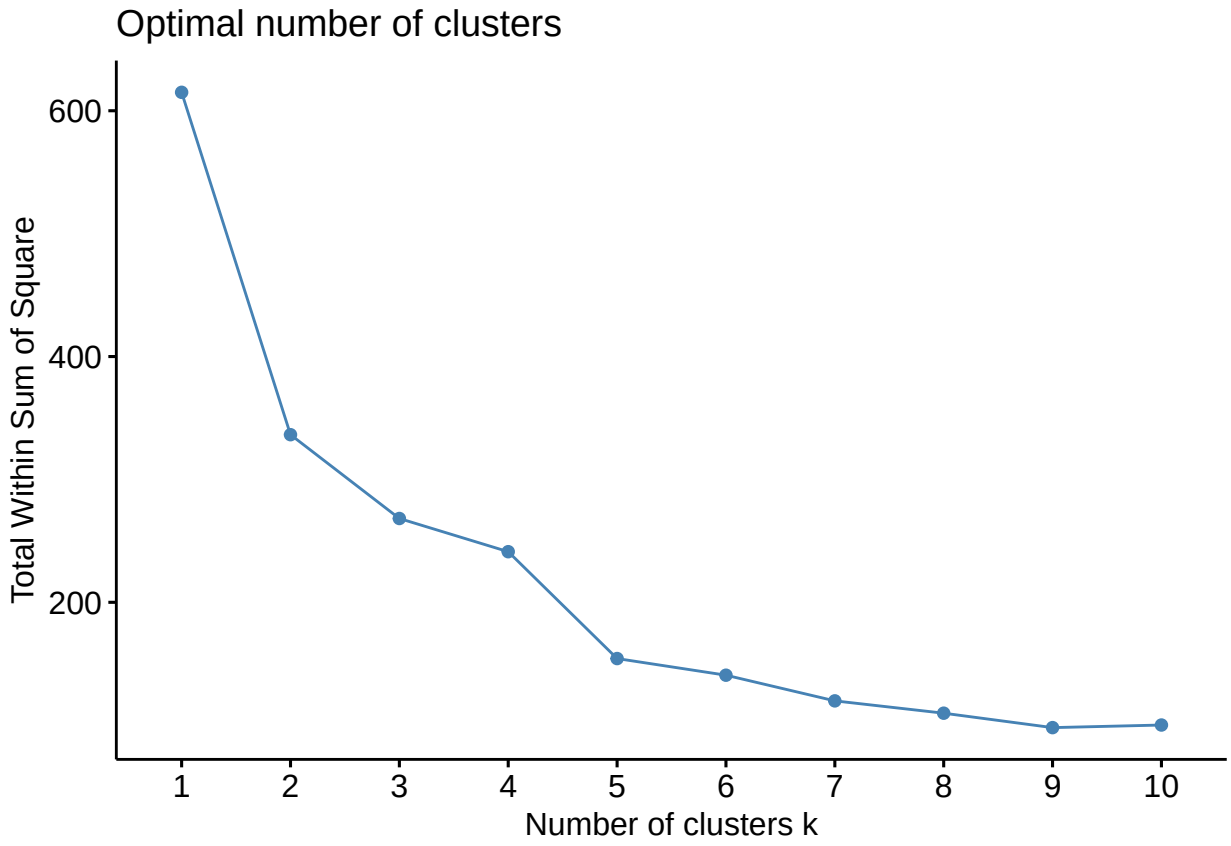
```
metricsdf$umpire <- NULL
normdf <- scale(metricsdf)
head(normdf)
```

```
##      correct_calls_above_expected accuracy_above_expected consistency
## [1,]                   1.28517120               1.3121123   1.6930681
## [2,]                   0.80117398               0.7932280   0.5287273
## [3,]                  -0.87271998              -0.8301581  -0.6791146
## [4,]                   1.12480251               1.1225877   0.6094562
## [5,]                  -0.02923043              -0.1684605   0.9439074
## [6,]                   1.95639523               1.9001465   0.8234086
##      favor_home total_run_impact
## [1,] -0.2910064       -1.4287472
## [2,] -0.4856616       -0.9388331
## [3,]  0.2607901        0.5402487
## [4,] -0.5460718       -0.6999696
## [5,]  1.6873307       -1.5986647
## [6,]  0.8866329       -1.6499882
```

```
distance <- get_dist(normdf)
fviz_dist(distance)
```

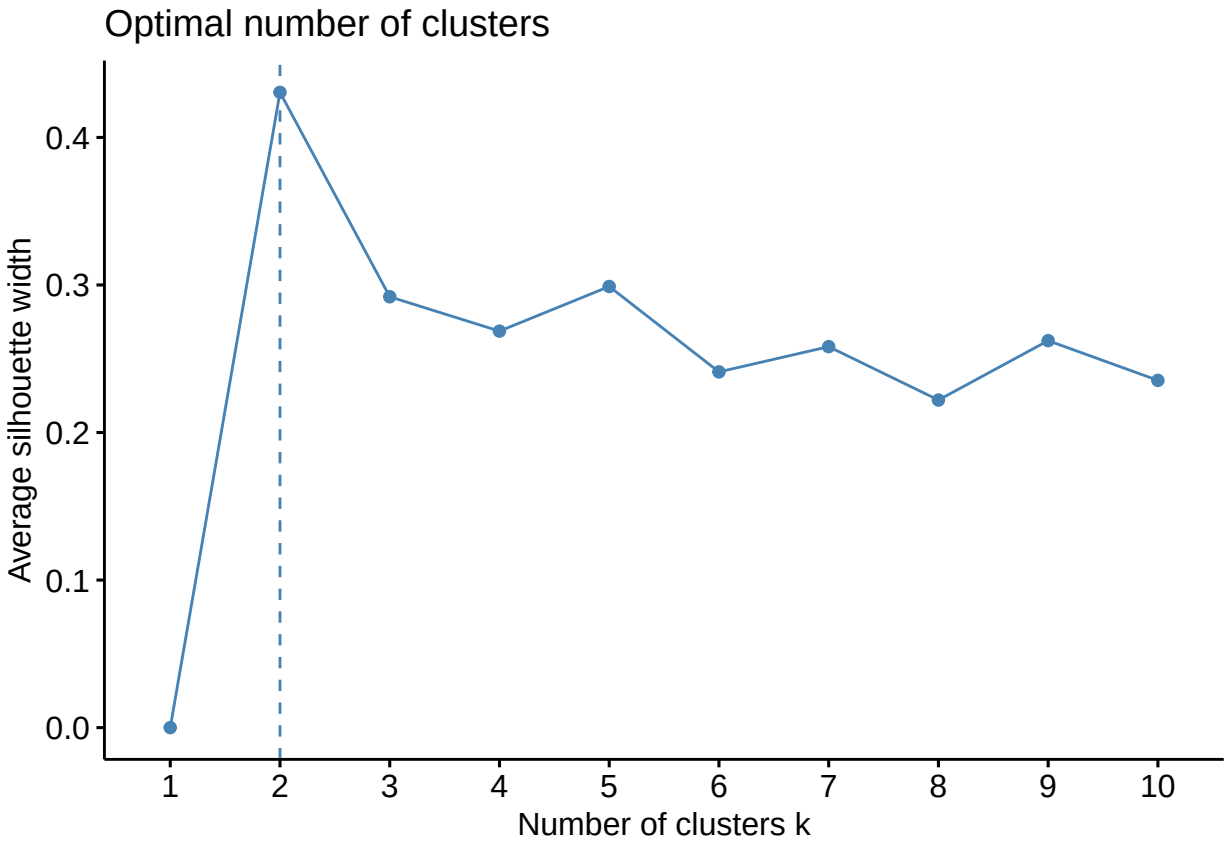



```
fviz_nbclust(normdf, kmeans, method="wss")
```



#There appear to be elbows between 2 and 5 as reasonable cluster numbers.

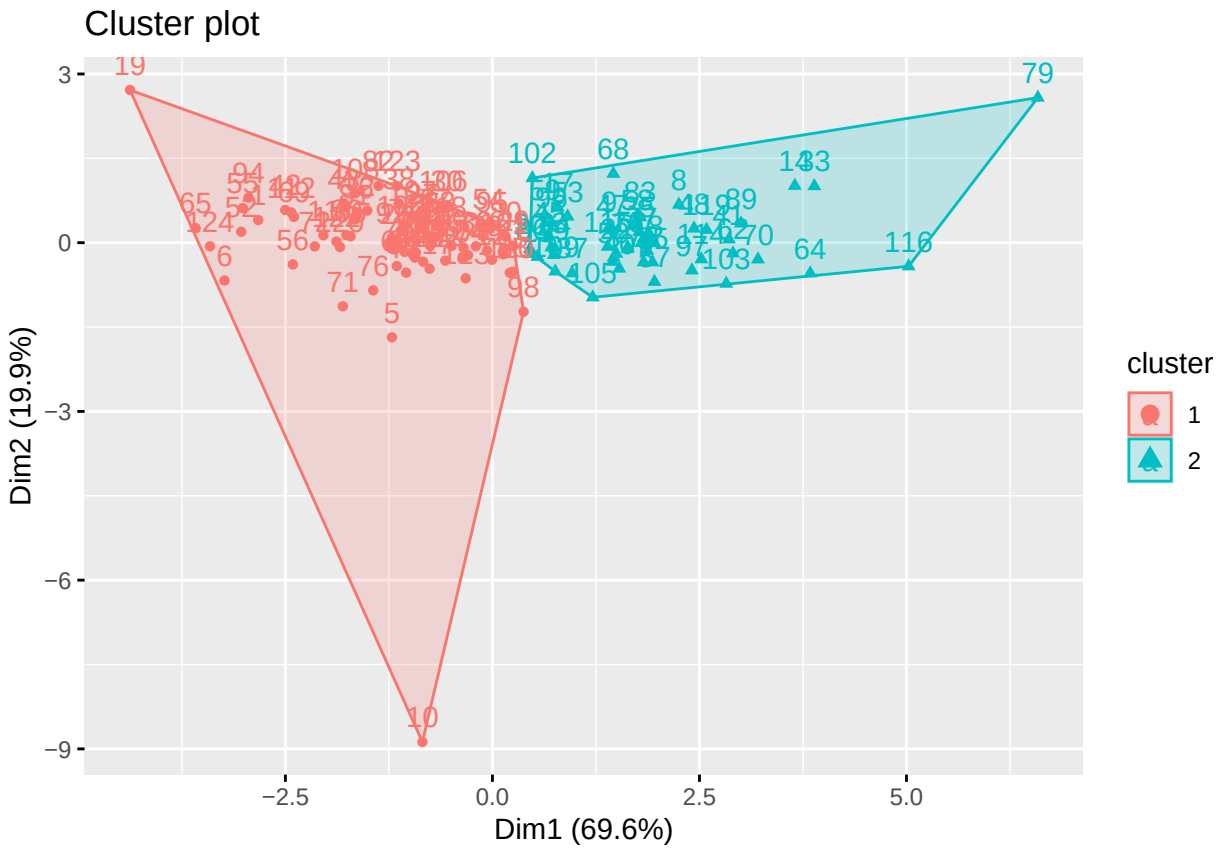
```
fviz_nbclust(normdf, kmeans, method="silhouette")
```



#The silhouette approach confirms that two clusters are the ideal number. This may be used to distinguish between umpires who are performing well and those who are underperforming. We could add more clusters if we wanted to include a few groups in the center.

```
k2 <- kmeans(normdf, centers = 2, nstart = 25)
```

```
fviz_cluster(k2, data = normdf)
```




```
k2$centers
```

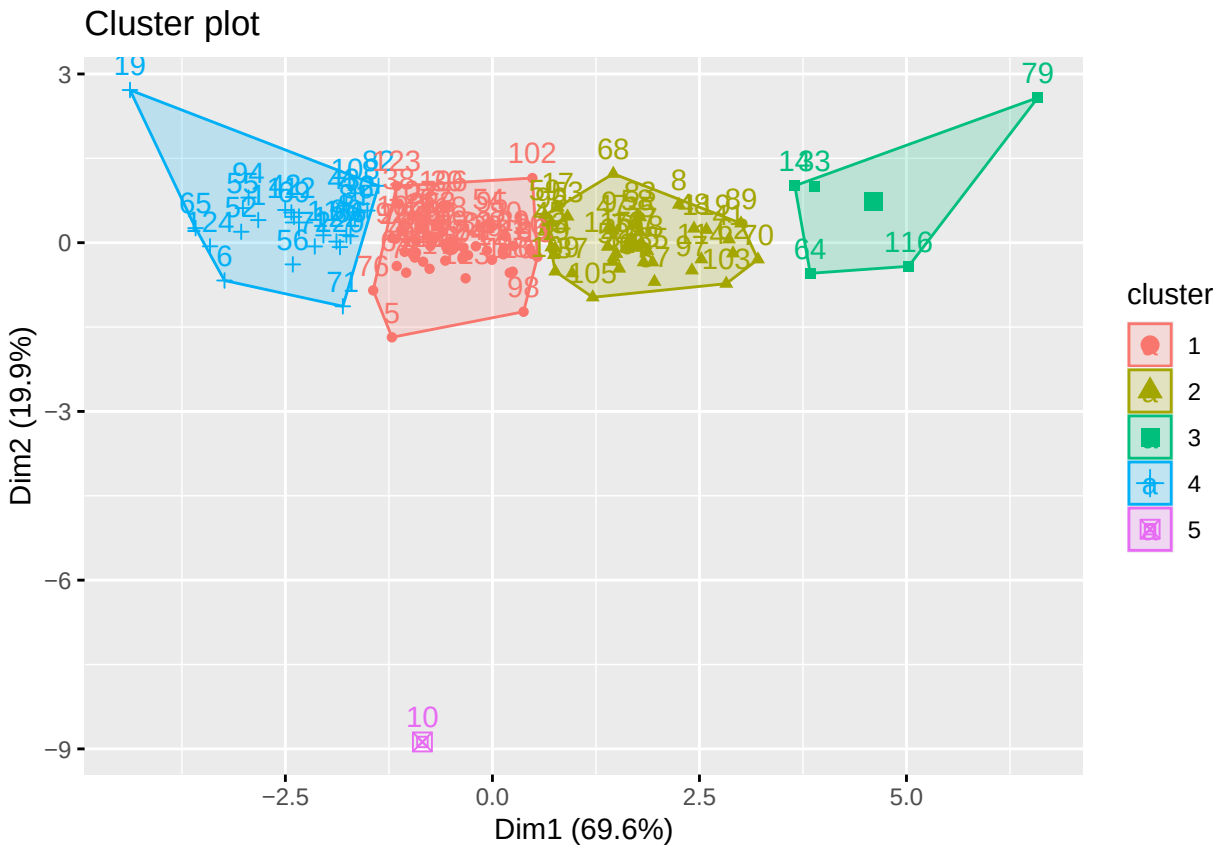
```
##   correct_calls_above_expected accuracy_above_expected consistency favor_home
## 1                    0.6309703               0.6262481   0.5328498  0.1013532
## 2                   -0.9990363              -0.9915595  -0.8436788 -0.1604759
##   total_run_impact
## 1       -0.5783485
## 2        0.9157184
```

#Cluster 1 includes umpires with stronger performance metrics, whereas Cluster 2 includes umpires with performance concerns.

#It appears that it is preferable to split the data down into further clusters in order to have a better notion of how to manage assignments for critical postseason games and summer training for the umpires.

```
k5 <- kmeans(normdf, centers = 5, nstart = 25)
```

```
fviz_cluster(k5, data = normdf)
```




```
k5$centers
```

```
##   correct_calls_above_expected accuracy_above_expected consistency  favor_home
## 1                   0.25541260              0.25167027   0.2095231  0.08145916
## 2                  -0.94583013             -0.93195415  -0.6525896 -0.06442308
## 3                  -1.93768962             -1.97355696  -2.7750502 -0.89991634
## 4                   1.27445917              1.26616228   1.0420217 -0.23713471
## 5                  -0.09698309             -0.04474683   1.1289590  8.86024885
##   total_run_impact
## 1       -0.2900646
## 2        0.8105035
## 3        2.4898540
## 4       -1.1074636
## 5        0.1085695
```

```
#There are four big clusters and one cluster with an outlier (Cluster 5 consisted of solely Anthony Joh
```