Pitch Clustering Example

Ronald Michaels

11/6/2020

Import Libraries

```
# library needed to unscale clustering centers
library(DMwR)
```

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

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

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
# library needed for clustering plots
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 3.6.2
```

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

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

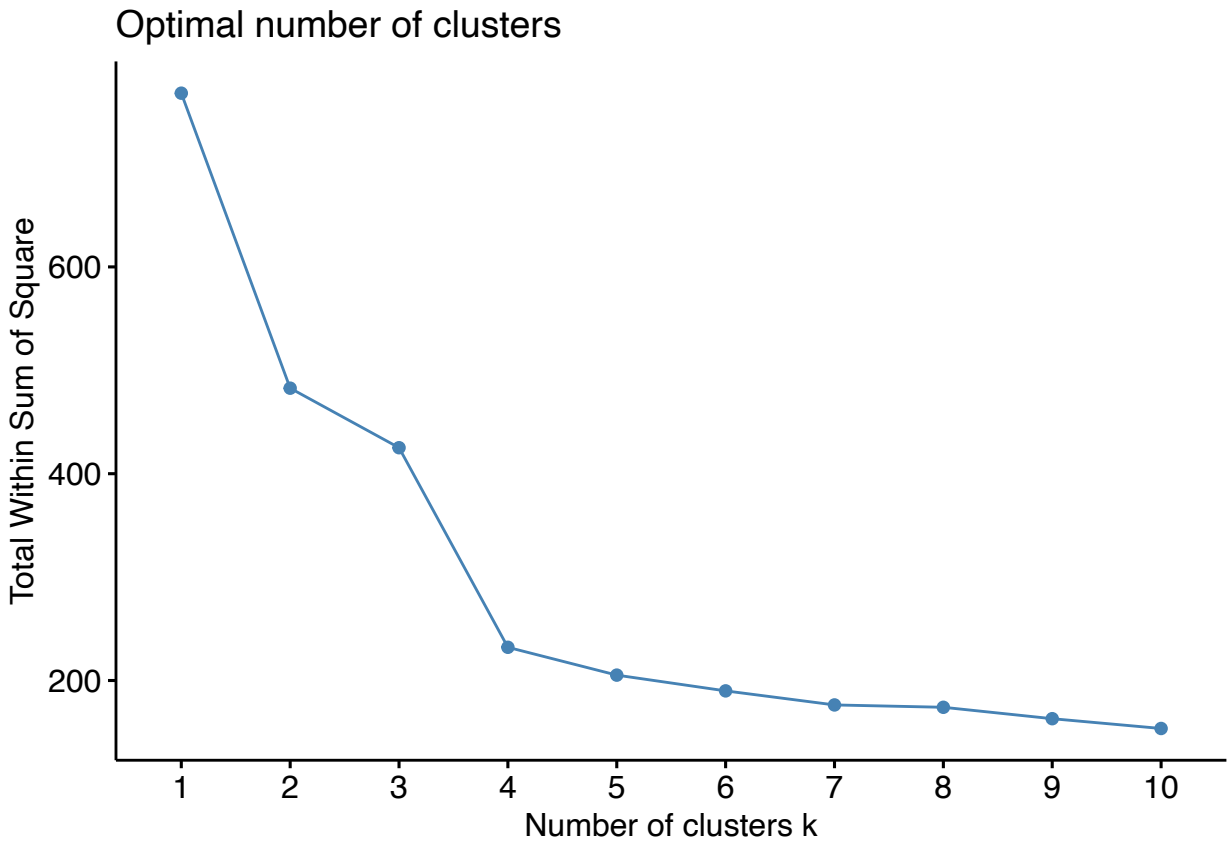
Load Dataset

```
# read in csv file
pitch_data =
  read.csv("Pitch_Clustering_Practice.csv")
# scale values for clustering
pitch_data_scale = scale(pitch_data[,2:9])
# create dataframe
pitch_data_df = as.data.frame(pitch_data_scale)
# alter row title for clustering
row.names(pitch_data_df) = pitch_data[,1]
# display head of scaled dataframe
head(pitch_data_df)
```

```
##    Velocity Total_Spin  True_Spin Spin_Efficiency Horizontal_Break
## 1 -2.810594 -0.1071962 -0.1350068      -0.1169504        2.9316266
## 2 -2.586367 -2.1811084 -2.6821448      -1.4693711        1.7934810
## 3 -1.689460  0.1693255 -1.2513732      -3.4677242        1.3719456
## 4 -1.497266  0.5231105 -0.2717047      -1.9134496        2.2571699
## 5  1.577842  0.8524966  0.6669543      -0.5408434        0.3181071
## 6  1.449713  0.9704249  0.7945391      -0.5206581        0.1073394
##   Vertical_Break Release_Height Release_Horizontal_Extension
## 1    -1.84853921     -2.9329945                    3.4826437
## 2    -3.26399804     -2.9329945                    3.7682028
## 3    -2.20240392     -4.1672698                    4.3393211
## 4    -2.40461232     -1.6987191                    3.4826437
## 5     0.02188854      0.1526939                   -0.2296249
## 6     0.32520115      0.1526939                   -0.2296249
```

Determine Optimal Number of Clusters

```r
# Use Elbow Method
fviz_nbclust(pitch_data_df, kmeans, method = "wss")
```

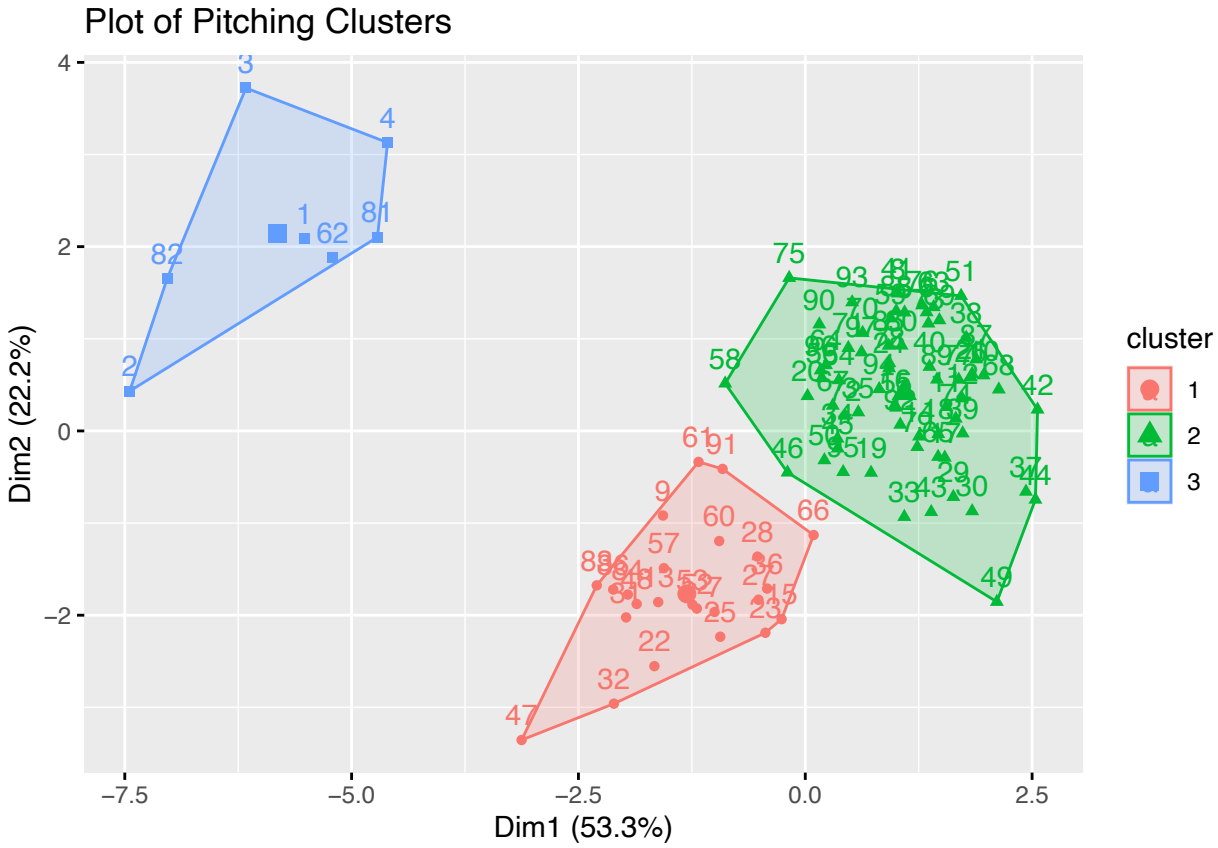


Perform Cluster Analysis With 3 Centers

```r
clustering = kmeans(pitch_data_df, centers = 3, nstart = 25)
# unscale centers to view pitch information
result = unscale(as.matrix(clustering$centers), as.matrix(pitch_data_scale))
# display centers
result
```

```
##   Velocity Total_Spin True_Spin Spin_Efficiency Horizontal_Break
## 1 79.37083   1464.333  1416.167        96.82500         6.970833
## 2 83.50152   1946.561  1802.303        92.60455         6.381818
## 3 75.11429   1712.571  1495.571        87.48571        13.557143
##   Vertical_Break Release_Height Release_Horizontal_Extension
## 1       14.62917       5.329167                    0.5375000
## 2       16.18030       5.436364                    0.4742424
## 3       10.08571       4.957143                    1.7285714
```

```r
# graph of clusters
fviz_cluster(clustering, data = pitch_data_df, main = "Plot of Pitching Clusters")
```



Assign Cluster Values as Final Column

```
# create final column
pitch_data$Pitch_Name = clustering$cluster
# change to character values
pitch_data$Pitch_Name = as.character(pitch_data$Pitch_Name)
# assign pitch name to cluster value
for (i in 1:nrow(pitch_data)) {
  if (pitch_data$Pitch_Name[i] == "1"){
    pitch_data$Pitch_Name[i][pitch_data$Pitch_Name[i] == "1"] = "Changeup"
  }
  else if (pitch_data$Pitch_Name[i] == "2"){
   pitch_data$Pitch_Name[i][pitch_data$Pitch_Name[i] == "2"] = "Fastball"
  }
  else if (pitch_data$Pitch_Name[i] == "3"){
    pitch_data$Pitch_Name[i][pitch_data$Pitch_Name[i] == "3"] = "Curveball"
  }
}
# view first 10 rows of dataset
head(pitch_data,10)
```

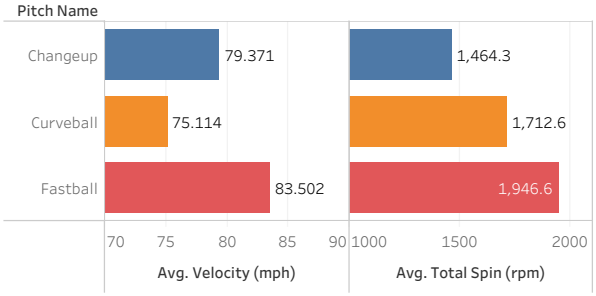
```
##    Pitch_ID Velocity Total_Spin True_Spin Spin_Efficiency Horizontal_Break
## 1         1     73.1       1784      1655            92.7             14.0
## 2         2     73.8       1274      1096            86.0             11.3
## 3         3     76.6       1852      1410            76.1             10.3
## 4         4     77.2       1939      1625            83.8             12.4
## 5         5     86.8       2020      1831            90.6              7.8
## 6         6     86.4       2049      1859            90.7              7.3
```

```
## 7         7     80.6       1441      1418            98.4             6.1
## 8         8     83.2       2114      1902            90.0             7.2
## 9         9     78.9       1547      1383            89.4             6.0
## 10       10     84.9       2032      1983            97.6             6.7
##    Vertical_Break Release_Height Release_Horizontal_Extension Pitch_Name
## 1            11.7            4.9                          1.8  Curveball
## 2             8.9            4.9                          1.9  Curveball
## 3            11.0            4.7                          2.1  Curveball
## 4            10.6            5.1                          1.8  Curveball
## 5            15.4            5.4                          0.5   Fastball
## 6            16.0            5.4                          0.5   Fastball
## 7            15.3            5.3                          0.6   Changeup
## 8            16.2            5.4                          0.6   Fastball
## 9            14.2            5.3                          0.6   Changeup
## 10           17.7            5.5                          0.7   Fastball
```

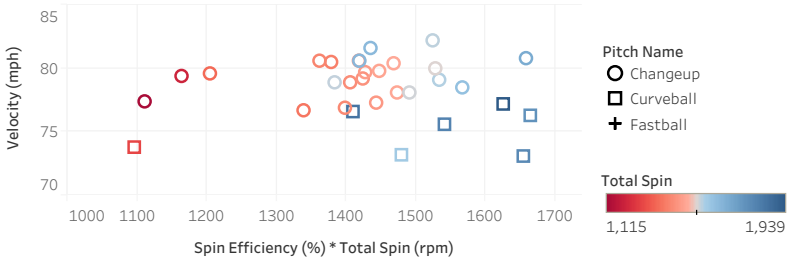
Export Dataset For Tableau Visualizations

```
# export as csv
write.csv(pitch_data,
          "Pitch_Clustering_Final_Data.csv",
          row.names = FALSE)
```

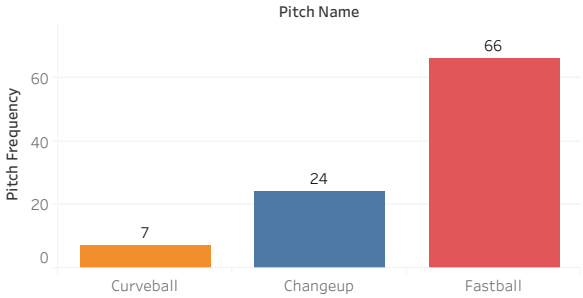
Velocity and Avg Total Spin



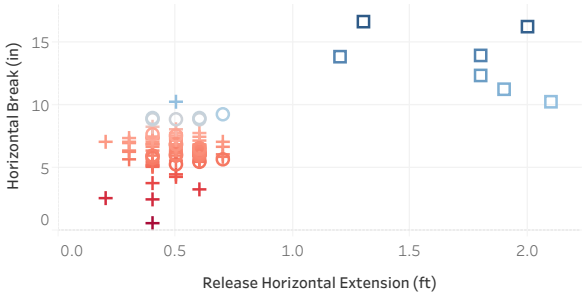
Spin Analysis: What Separates Changeup Velocity vs Curveball Velocity?



Pitch Frequency



Horizontal Extension vs Break



Ronald Michaels
Pitch Clustering Analysis

This project takes a look at pitch data collected through rapsodo from a starting pitcher at the University of Rochester with the goal of clustering the unclassified pitches into their pitch types. The initial data contained 97 unique pitches with the following attributes:

- Velocity (mph)
- Total_Spin (rpm)
- True_Spin (rpm)
- Spin_Efficiency (0-100 scale)
- Horizontal_Break (in)
- Vertical_Break (in)
- Release_Height (ft)
- Release_Horizontal_Extension (ft)

I used the clustering method of KMeans clustering and utilized the Elbow Method to determine the optimal number of clusters. After running the Elbow Method visualization, it was clear that the optimal number of centers for the KMeans algorithm was three. Once the value of three was used for analysis, I unscaled the clustering centers to view the center values for each cluster's attributes.

Afterwards, the next step was to determine which pitches can be used to classify the three clusters. With the highest cluster velocity being 83.50152 miles per hour, it makes logical sense that cluster two can be classified as a fastball. This assumption is further backed up by the value of total spin coupled with true spin. A fastball travels to the plate with the most direct backspin out of the pitch types, resulting in the highest total spin value out of possible pitches. In addition, due to the ball being released with top backspin, almost all of the spin of the ball plays a role in its movement as it travels towards the plate, leading to a high true spin value as well.

The classification of the remaining two clusters is a little more difficult. The key to identifying a changeup is a large reduction in total spin combined with a slight drop in velocity. This key lends itself for the identification of cluster one as a changeup. Lastly, to identify cluster three, a drop as drastic of an average of over eight miles per hour eliminates any fastball types. Furthermore, we have already identified cluster one as a changeup, meaning that the remaining logical options are either a slider or a curveball. The average velocity of a slider, according to the rapsodo website, is higher than the velocity of a changeup. Thus, the slider can be eliminated from consideration, leading to the classification of cluster three as a curveball.

The plot titled “Plot of Pitching Clusters” displays the dispersion of values grouped by cluster. Seeing as there are no overlapping values and there is a clear difference in identifying clusters, the use of KMeans with a center of three produced a useful model.

Finally, I want to talk about the dashboard of visuals which display not only the frequency of pitch usage by the pitcher, but also an explanation as to why certain pitches move, and travel at speeds, the way they do. As per the top left visual, total spin does not directly correlate with an increase in velocity, leading to the realization that displacement of the ball relative to the body upon release, in addition to the pitcher's grip on the ball, impacts the velocity of the baseball as well. The bottom right chart shows horizontal extension relative to the middle of the pitching mound (ft) versus the pitches horizontal break (in). It becomes clear that as the baseball moves further away from the body, it is more difficult to get power behind the pitch, leading to slower velocities. However, the advantage of an increase in displacement is to create larger horizontal breaks. When the arm becomes outstretched, the rotation of the wrist, in order to hit the strike zone, must be externally tilted from the vertical axis to maintain a straight line from the elbom to the pitcher's finger tips. This external tilt causes an increase in horizontal break.

To sum up this analysis, the top right visual filters for only changeups and curveballs and measures the pitch's total spim (rpm) multiplied by its spin efficiency versus pitch velocity (mph). This visual reiterates the message that total spin does not directly determine velocity. When accompanied with a slow spin efficiency (the rate at which the spin put on the ball by the pitcher impacts the effectiveness of the pitches movement), we are able to better understand why a changeup, which has a lower spin value than a curveball, still moves at a greater velocity. Put all these factors together and we are able to see how the types of pitches relate to one another, and why they move the way they do.