Stephen Imhoff

Data Analysis

11/30/2020

Four Big Baseball Questions and Answers

This report is an exploration of baseball data recorded on every single Major League game played from 1871 to 2016. This data was found on data.world and was compiled by Retrosheet. It consists of 171908 rows and 161 columns, each recording various stats about each game played. Of this massive data set, I focused on offensive statistics and pitching statistics to answer four major questions. Data manipulation to clean and prepare the data involved extracting the year from the date attribute and creating dozens of aggregations around attributes to study. There are numerous structural issues with the data. First, the players who played in 1871 and those who played in 2016 are playing different games entirely (see Fig 1 and Fig 2 in appendix). The way data was recorded changed fundamentally, strategies and rules have changed, and players have gone from day laborers playing on the weekends to professional athletes earning hundreds of millions of dollars today. Other issues were prevalent, such as the data not being extremely well suited to some types of analysis. The data set does not include individual player statistics and so any analysis of individuals is flawed. Note that all supplemental data was found on Baseball-Reference.com.

This report examines four questions:

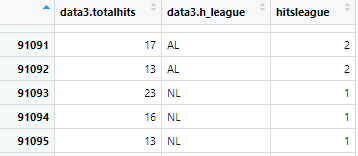
1. Is there a statistically significant difference in hits per game between the National League (NL) and the American League (AL)? This question is particularly pertinent as the NL this year considers adopting a Designated Hitter like the AL which should theoretically increase offense output.
2. Does more pitchers used in a game by a team correlate with better performance by that team? Modern baseball strategy is increasingly requiring more and more pitchers to be used from the bullpen to get through the opposing lineup. Is this actually a good idea?
3. How can runs scored by a team be best predicted? (Or, how many runs are certain occurrences worth?) In doing this analysis I can identify the most important traits for a hitter to have and what statistics teams should focus on.
4. Can I identify and cluster quality starting pitchers from the dataset based on the pitching statistics of the game? It is difficult to differentiate between the quality of pitchers in watching a baseball game. It’s difficult to tell if a strikeout is the pitcher performing well or the hitter poorly, so I will quantify pitching performance.

The methodologies are:

1. T-test-involved isolating the NL and AL games and separating them into two bins, then doing a t-test to see if there was a statistically significant difference in the total hits per game between the two leagues. The ggplot library was used to test for normality in the dataset. All code is in the appendix. Review lines 18-66 for this first question.
2. Correlation test/matrix-involved removing all rows with null values, measuring the correlation and creating a correlation plot to measure individual elements of pitching performance, such as walks, strikeouts, and hits. The correlation plot required the corrplot library. See lines 66-107.
3. Linear Regression and Model Selection- involved creating a list of statistics to measure and ran a linear regression to see how many runs each statistic created. Regression and model selection required the leaps library. Ran this model with home pitchers as well as away pitchers to validate the results. See lines 109-135.
4. Clustering-involved aggregating various pitching statistics around the home pitcher’s name and clustering the name of the pitchers based on their performance. This was initially done with all pitchers from 2007 onward which was far too many pitchers to make anything of. Then, 12 pitchers were selected with various levels of fame and success and were clustered based on their performance. Clustering required the NbClust library. See lines 136-191.

For the first question, I am examined the offensive difference between the NL and AL by assessing total hits by league. A new dataframe was constructed with calculations for total runs, home runs (hrs), and hits in a game added to the original dataset. Hrs and runs were used to get a “feel” for the data while I will be using hits to compare the AL and NL. Then, using aggregate(), the average hits by year was calculated. I added a new column with a value of 1 for NL and 2 for AL and added it to the hits data so the resulting table looked like this:

**Table 1**



A t-test was done to see if there is a significant difference between the mean total hits and the two leagues. The mean AL hits is 17.97 and mean NL hits is 17.82 with a p value of 1.043e-07, indicating high levels of significance even though the difference is slight. A test for normality looked good:

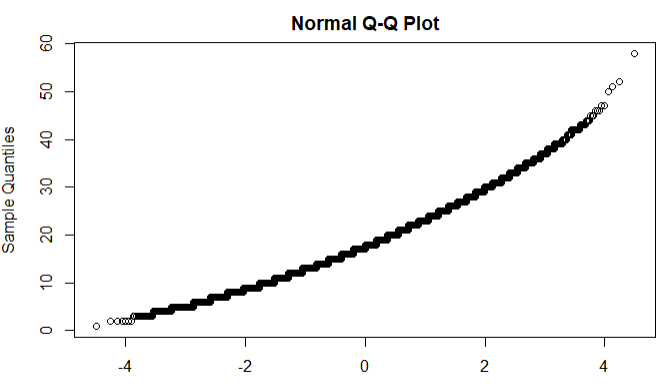


Fig. 5

The implied reason for the difference is the AL’s use of a designated hitter instead of having pitchers hit who are notoriously bad hitters. A potential reason why the averages are close is the AL didn’t adopt the DH until 1973 even though it was founded in 1901, meaning there is lots of data where the games were played essentially the same way.

For the second problem, I am challenging the modern assertion that using more pitchers in a game is better at limiting the opposing team’s offensive performance. First, all NA values were omitted from the data set. The variables I am assessing to measure pitching performance are hits, strikeouts, hit by pitch, and walks. (A pitcher hitting a batter is generally a mistake and should be avoided, as with walks and hits. Strikeouts are good, meaning more strikeouts is better). Using the princomp() function, the correlation between number of pitchers used and hits of the opposing team is only .33 and the correlation between pitchers used and strikeouts is .345 (averaging the away and home pitching performance). This weak correlation indicates that using more pitchers is perhaps less effective than modern baseball thinking suggests. A correlation plot was made using all of those variables to see if there is any significant correlation between number of pitchers used and performance.

**Pitcher Correlation Plot**

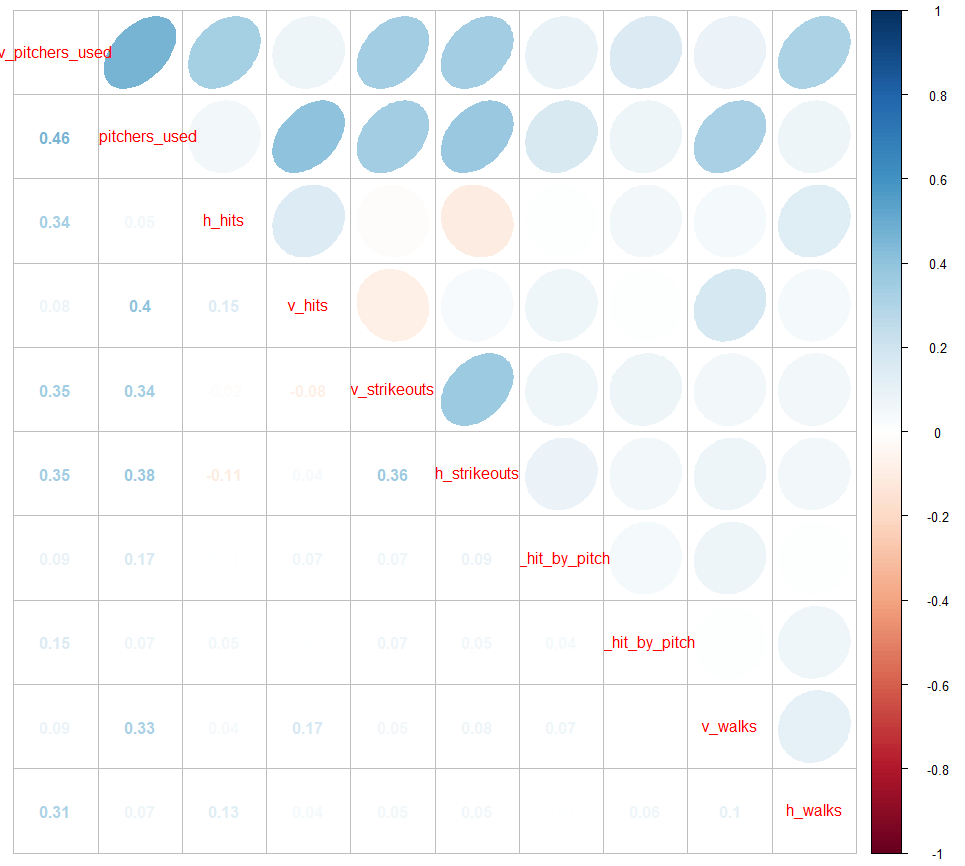


Fig. 4

While this chart is hard to read, the first two rows are v\_pitchers\_used and h\_pitchers\_used, meaning the visiting team and the home team. The strongest correlation was between these two variables rather than any performance metric. There didn’t seem to be any significant correlation between using more pitchers and better performance.

The third problem is assessing which model is best to predict number of runs scored in a game by a particular team. The assessed variables using regsubsets() were visiting pitchers used, visiting errors, visiting balks, visiting passed balls, home stolen bases, home hits, home walks, home intentional walks, home hrs, and home strikeouts. The most significant factors (using 6) were pitchers used, errors, hits, walks, hrs, and strikeouts. The model created by lm() was runs scored = -2.4 + .2\*pitchers used+ .5\* errors + .5\*hits + .3\*walks + .85\*homeruns - .08\*Strikeouts. I recreated this model from the perspective of the visiting team and flipped the team of all of the variables and the results were nearly identical. You can see both models on lines 129-130. This model is confusing because a home run is only worth .85 runs, but a home run is also a hit. This means a home run is worth roughly 1.35 runs, which is feasible.

The fourth problem is meaningfully clustering pitchers based on their performance. I will be checking the results of this clustering against their wins above replacement (WAR), considered a holistic measure of a player’s value. I began by attempting to just cluster pitchers from 2007 on using aggregate(), dist(), hclust(), and scale(). The clustering is based on the opposing team’s score, hits, doubles, triples, hrs, hit by pitch, walks, and strikeouts. The resulting cluster was far too messy to read with a monitor of my size because it encompasses hundreds of starting pitchers (cluster viewable in appendix as Fig. 6). As an alternative, I used all of the data but only selected 13 pitchers to compare. This clustering was much more successful.

**Pitcher Clustering**

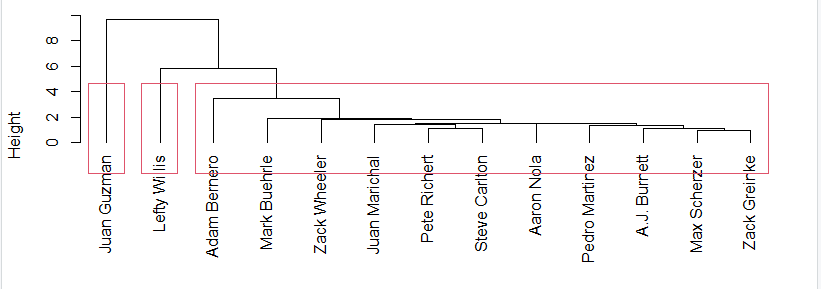


Fig. 5

I intentionally chose pitchers who were known for being incredibly good as well as pitchers I had never heard of. Here is their ranking in the order of their clustering and by Average WAR

**Table 2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Ranking by** | **Clustering** | **Ranking by** | **WAR** |
| Pitcher Name | Average WAR | Pitcher Name | Average WAR |
| Juan Guzman | 2.43 | Adam Bernero | -0.2 |
| Lefty Willis | 0.266667 | Lefty Willis | 0.266667 |
| Adam Bernero | -0.2 | Pete Richert | 1.038462 |
| Mark Buehrle | 3.69375 | AJ Burnett | 1.694118 |
| Zack Wheeler | 2.233333 | Zack Wheeler | 2.233333 |
| Juan Marichal | 3.93125 | Juan Guzman | 2.43 |
| Pete Richert | 1.038462 | Aaron Nola | 3.616667 |
| Steve Carlton | 3.758333 | Mark Buehrle | 3.69375 |
| Aaron Nola | 3.616667 | Steve Carlton | 3.758333 |
| Pedro Martinez | 4.661111 | Juan Marichal | 3.93125 |
| AJ Burnett | 1.694118 | Zack Greinke | 4.247059 |
| Max Scherzer | 4.792308 | Pedro Martinez | 4.661111 |
| Zack Greinke | 4.247059 | Max Scherzer | 4.792308 |

The clustering did a decent job of arranging the pitchers by their value. The one major correction was AJ Burnett who’s ranking was off by 7 positions. Other than him, the greatest variation was by 4 places. There are also only slight differences in WAR between the 7 best pitchers so any clustering variation is understandable.

I learned that R can easily create meaningful and interesting answers to questions I have. The AL, I found, has slightly more hits per game than the NL. The NL adopting the designated hitter will likely even that difference going forward. I will miss seeing pitchers with no business holding a bat strike out pitifully (or hit bombs, see appendix for an example). The second problem had disappointing results and wasn’t able to show a correlation between more pitchers and better performance. If a pitcher is left in for a while, it’s because they are doing well. If your pitchers are always being substituted, generally they aren’t performing well. This phenomenon may explain the low correlation. Also, the pitcher rotation phenomena developed in the last 5-10 years (see fig. 7 in appendix) meaning the data needs to be studied with greater focus on the last few years. I was able to validate the accuracy of the importance of hrs in modern baseball thinking as they are worth 1.35 runs versus the average .5 runs for the average hit. One conclusion is that a line-drive hitter needs to get at least three hits for every home run the power hitter can club to be more valuable. I was unable to satisfyingly cluster pitchers based on performance. The best pitchers were all grouped together but there were a few outliers. A huge issue with this clustering is the performance data was tracked from the entire game, not just when that pitcher was pitching. For future analysis, this clustering should only use statistics from when that pitcher was on the mound because the data is being skewed by the bullpen’s pitching as well. To better improve upon all of these results, more data should be analyzed more specifically. Instead of looking at all time periods, examining specific time periods and comparing players in their day instead of across time periods will be more precise.

Appendix

**Average Runs Per Game**

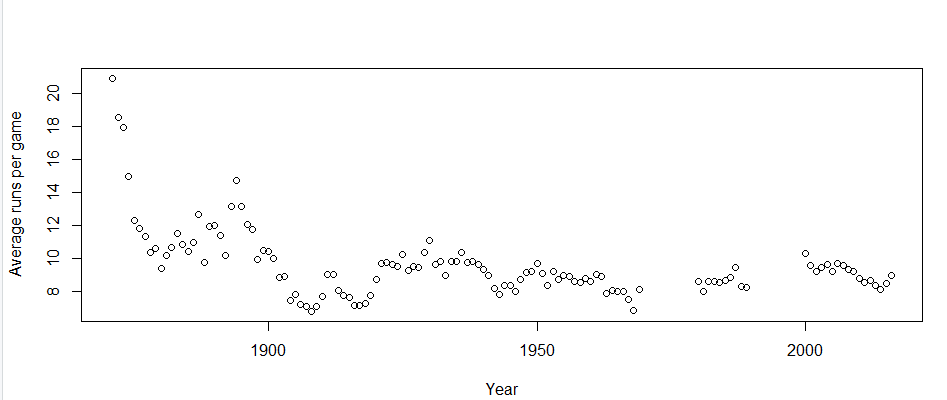


Fig. 1

Fig. 1 is a chart of the average runs scored (sum of both the home and away team) over the years. The game played before 1900 is fundamentally different from the game played in 1950. Games in 1871 averaged over 20 runs per game. Compare that slugfest with the roughly 9 runs per game over the last 20 years and you’ll see the game has changed drastically over time.

**Average Home Runs Per Game**

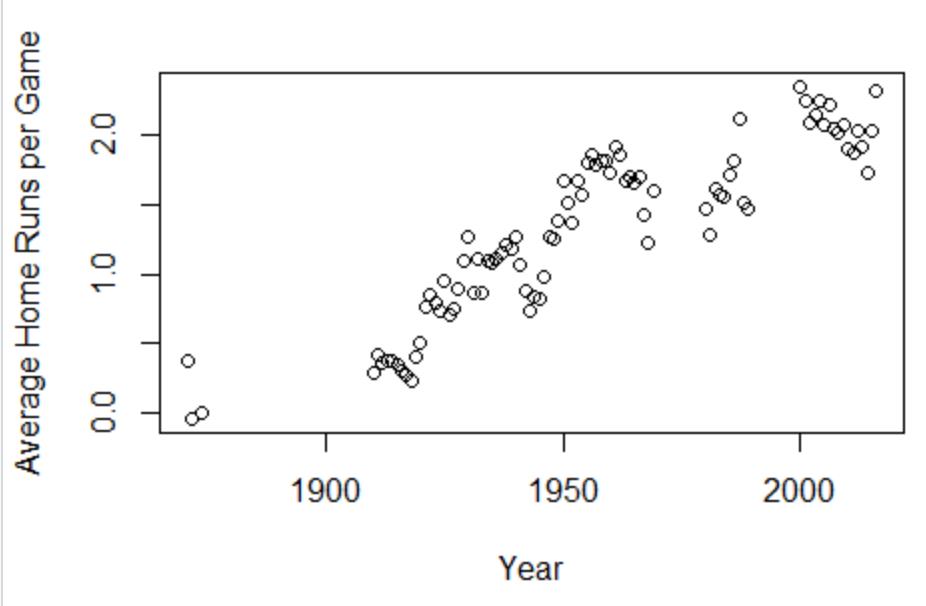


Fig. 2

Although the average number of runs per game has been relatively stagnant over the last 80+ years, Fig. 2 shows the number of home runs per game has skyrocketed and is currently at an all-time high.

**Initial Cluster**

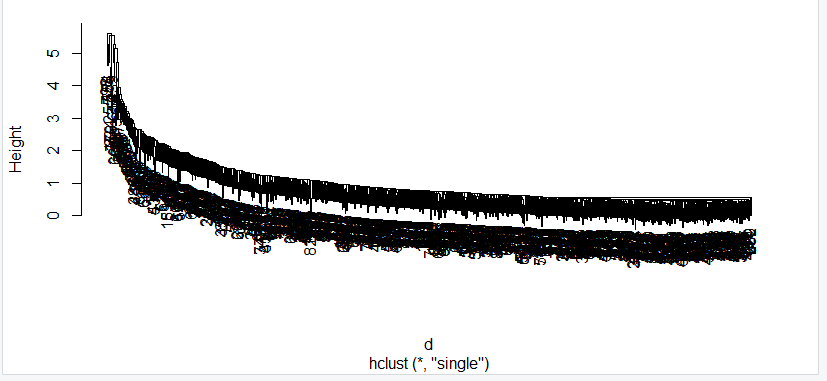


Fig. 6

**Mean Pitchers Used Annually**

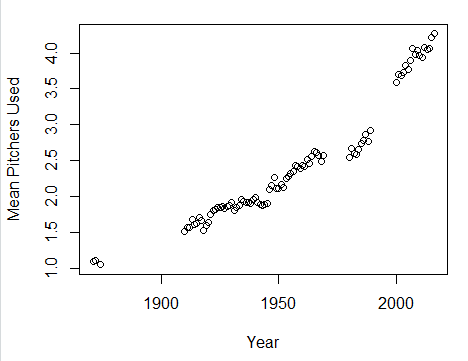


Fig. 7

Pitcher home run: <https://www.youtube.com/watch?v=OVFsq9FQBlc>

Data found on data.world: <https://data.world/dataquest/mlb-game-logs>

Baseball supplemental data: Baseball-Reference.com

---

title: "SI-Final Project"

output: html\_notebook

---

```{r}

setwd("~/Data Analytics-R") # My directory where I store the data

data = read.csv("game\_logs.csv", header = T)

```

```{r}

names(data)

```

```{r}

summary(data)

```

```{r}

#does scoring change over time?

totalruns = data$v\_score + data$h\_score

totalhrs = data$v\_homeruns + data$h\_homeruns

totalhits = data$v\_hits + data$h\_hits

data\_2 = data.frame(data, totalruns, totalhrs, totalhits)

year = substr(data\_2$date, start = 1, stop = 4)

data3 = data.frame(data\_2, year)

#plot average runs per game over time

#do by year

library(ggplot2)

#create average annual runs and fill it with the average runs by year, plot

data3$avgannruns <- NA

avgannnualruns = aggregate(totalruns~year, data3,mean)

plot(avgannnualruns, xlab = "year", ylab = "Average runs per game")

#same with average Home runs and hits

avghrs = aggregate(totalhrs~year, data3,mean)

plot(avghrs, xlab = "Year", ylab = "Average Home Runs per Game")

avghits = aggregate(totalhits~year, data3, mean)

plot(avghits)

#is there a statistically significant difference in hits/walks in the NL vs AL? (based on when the DH was instated)

#hits analysis

leaguehits = aggregate(totalhits~h\_league, data3, mean)

#this analysis shows that there are on average .2 more hits in the AL vs the NL. Let's test if that is statistically significant

hitsdata = data.frame(data3$totalhits, data3$h\_league)

hitsleague = rep(0, nrow(hitsdata))

#assign 1 to NL and AL values

hitsleague[grepl("NL",hitsdata$data3.h\_league, fixed = TRUE)] = 1

hitsleague[grepl("AL",hitsdata$data3.h\_league, fixed = TRUE)] = 2

hitsdata = data.frame(hitsdata, hitsleague)

hitssub = apply(hitsdata, 1, function(row) all(row != 0))

hitsdata = hitsdata[hitssub,]

#run two sided t test on the data to see if there is a significant difference between the leagues

t.test(hitsdata$data3.totalhits ~ hitsdata$data3.h\_league, alternative = "two.sided", conf.level = .95)

qqnorm(hitsdata$data3.totalhits)

#test for normality looks good

#there is a minor but statistically very significant difference in the means of hits between the AL and NL

#hypothesis testing and plots

```

```{r}

#does pitchers used correlate with better performance?

#is there a correlation between number of pitchers used and the runs scored by the opposite team

#add in strikeouts, hit by pitch, walks

cleand2 = data.frame(na.omit(data3$v\_pitchers\_used), na.omit(data3$h\_pitchers\_used), na.omit(data3$h\_hits), na.omit(data3$v\_hits), na.omit(data3$v\_strikeouts), na.omit(data3$h\_strikeouts), na.omit(data3$v\_hit\_by\_pitch), na.omit(data3$h\_hit\_by\_pitch), na.omit(data3$v\_walks), na.omit(data3$h\_walks))

#rename

names(cleand2)[1] = "v\_pitchers\_used"

names(cleand2)[2] = "h\_pitchers\_used"

names(cleand2)[3] = "h\_hits"

names(cleand2)[4] = "v\_hits"

names(cleand2)[5] = "v\_strikeouts"

names(cleand2)[6] = "h\_strikeouts"

names(cleand2)[7] = "v\_hit\_by\_pitch"

names(cleand2)[8] = "h\_hit\_by\_pitch"

names(cleand2)[9] = "v\_walks"

names(cleand2)[10] = "h\_walks"

pc1 = princomp(cleand2[,c("v\_pitchers\_used", "h\_hits")])

pc2 = princomp(cleand2[,c("h\_pitchers\_used", "v\_hits")])

summary(pc1)

summary(pc2)

cor(cleand2$v\_pitchers\_used, cleand2$h\_hits, method = "pearson")

#correlation only .33

cor(cleand2$h\_pitchers\_used, cleand2$v\_hits, method = "pearson")

library(corrplot)

cor(cleand2$v\_pitchers\_used, cleand2$h\_strikeouts, method = "pearson")

#.35

cor(cleand2$h\_pitchers\_used, cleand2$v\_strikeouts, method = "pearson")

#.34

corrplot.mixed(cor(cleand2),lower = "number", upper = "ellipse")

#results are disappointing, I'm not noticing significant changes in any result given a different number of pitchers used

# more pitchers used might be correlated with more hits because the pitcher has to get pulled if they give up hits (a sign of traditions being followed) as opposed to having more pitchers actually improving performance

#correlation higher at .4, home pitchers have an advantage

```

```{r}

#linear

library(leaps)

runmodel = regsubsets(h\_score ~ v\_pitchers\_used + v\_errors + v\_balks + v\_passed\_balls + h\_stolen\_bases + h\_hits + h\_walks + h\_intentional.walks + h\_homeruns + h\_strikeouts, data = data3)

review = summary(runmodel)

review

#use pitchersused, errors, hits, walks, home runs, strikeouts

bestmodel=lm(h\_score ~ v\_pitchers\_used + v\_errors + h\_hits + h\_walks + h\_homeruns + h\_strikeouts, data = data3)

summary(bestmodel)

#model: runs scored = -2.4 + .2\*pitchers used by opposing team + .5\*opposing errors + .5\*hits + .3\*walks + .85\*homeruns - .08\*Strikeouts

#will test the inverse of this data to see if it works for opposing teams as well, also will validate results

awaymodel = lm(v\_score ~ h\_pitchers\_used + + h\_errors + v\_hits + v\_walks + v\_homeruns + v\_strikeouts, data = data3)

summary(awaymodel)

#model: runs scored = -2.6 + .2\*pitchers used by opposing team + .5\*opposing errors + .5\*errors + .5\*hits + .3\*walks + .79\*home runs -.05\*strikeouts

#results are incredibly similar to previous model and indicate the validity of the model

#predict runs scored by a team based on: errors, passed balls, stolen bases, balks

#do model selection to create the optimal model measuring baseball success

```

```{r}

data7 = read.csv("GL2007.csv", header = T)

#data using: v\_score, v\_hits, v\_doubles, v\_triples, v\_homeruns, v\_Hit\_by\_pitch, v\_walks, v\_strikeouts, h\_starting\_pitcher\_name

data8 = data.frame(na.omit(data7[c(10,23,24,25,26,30,31,33,105)]))

scoreagg = aggregate(v\_score ~ h\_starting\_pitcher\_name, data8, mean)

hitsagg = aggregate(v\_hits ~ h\_starting\_pitcher\_name, data8, mean)

doubagg = aggregate(v\_doubles~ h\_starting\_pitcher\_name, data8, mean)

tripagg = aggregate(v\_triples~ h\_starting\_pitcher\_name, data8, mean)

hragg = aggregate(v\_homeruns ~ h\_starting\_pitcher\_name, data8, mean)

hbpagg = aggregate(v\_hit\_by\_pitch ~ h\_starting\_pitcher\_name, data8, mean)

walksagg = aggregate(v\_walks~ h\_starting\_pitcher\_name, data8, mean)

SOagg = aggregate(v\_strikeouts~ h\_starting\_pitcher\_name, data8, mean)

clustdata = data.frame(scoreagg$h\_starting\_pitcher\_name, scoreagg$v\_score,hitsagg$v\_hits,doubagg$v\_doubles,tripagg$v\_triples,hragg$v\_homeruns,hbpagg$v\_hit\_by\_pitch,walksagg$v\_walks,SOagg$v\_strikeouts)

library(NbClust)

d = dist(as.matrix(clustdata[,2:9]), method = "euclidean")

hc\_1 = hclust(d, method = "single", members = as.factor(clustdata[,1]))

plot(hc\_1)

#unfortunately, the dataset is far too large to make anything of this mess

#I'm going to massively limit my sample size in order to create some meaningful data

data9 = read.csv("clust.csv", header = T)

#data using: v\_score, v\_hits, v\_doubles, v\_triples, v\_homeruns, v\_Hit\_by\_pitch, v\_walks, v\_strikeouts, h\_starting\_pitcher\_name

data10 = data.frame(na.omit(data9[c(10,23,24,25,26,30,31,33,105)]))

scoreagg = aggregate(v\_score ~ h\_starting\_pitcher\_name, data10, mean)

hitsagg = aggregate(v\_hits ~ h\_starting\_pitcher\_name, data10, mean)

doubagg = aggregate(v\_doubles~ h\_starting\_pitcher\_name, data10, mean)

tripagg = aggregate(v\_triples~ h\_starting\_pitcher\_name, data10, mean)

hragg = aggregate(v\_homeruns ~ h\_starting\_pitcher\_name, data10, mean)

hbpagg = aggregate(v\_hit\_by\_pitch ~ h\_starting\_pitcher\_name, data10, mean)

walksagg = aggregate(v\_walks~ h\_starting\_pitcher\_name, data10, mean)

SOagg = aggregate(v\_strikeouts~ h\_starting\_pitcher\_name, data10, mean)

clustdata2 = data.frame(scoreagg$h\_starting\_pitcher\_name, scoreagg$v\_score,hitsagg$v\_hits,doubagg$v\_doubles,tripagg$v\_triples,hragg$v\_homeruns,hbpagg$v\_hit\_by\_pitch,walksagg$v\_walks,SOagg$v\_strikeouts)

d2 = dist(as.matrix(clustdata2[,2:9]), method = "euclidean")

hc\_2 = hclust(d2, method = "single", members = as.factor(clustdata2[,1]))

plot(hc\_2, labels = clustdata2[,1], hang = -1)

cluststd = scale(clustdata2[,2:9])

nclust = NbClust(cluststd, distance = "euclidean", min.nc = 2, max.nc = 5, method = "single")

#3 clusters seems ideal

plot(hc\_2, labels = clustdata2[,1], hang = -1)

rect.hclust(hc\_2, k = 3)

#this is a plot of the mean pitchers used per game

avgpitchers = aggregate(h\_pitchers\_used~year, data3, mean)

plot(avgpitchers, ylab = "Mean Pitchers Used", xlab = "Year")

```