Project 2 Report

Dataset Choice

For this project, I've chosen to analyze a dataset outlining statistics for hitters playing in Major League Baseball (hereafter MLB). This information is widely available in the public domain and should provide an appropriate level of richness for the purposes of this assignment. Specifically, I pulled the data from the popular site www.fangraphs.com and downloaded a csv file which made for convenient import to the R analysis environment. The export from Fangraphs was favoured over Sean Lahman's public database as Fangraphs has a number of advanced, calculated measures pre-made that Lahman does not. Note that the number of records pulled (12335) satisfies the minimum requirement for the assignment (1000).

Data-Scrubbing Pre-R

The aformentioned CSV download came with a number of default column headers that are illegal in R. For example, K% is not a valid name and thus I manually changed the header to KPct. Likewise for BBPct and HRPerFBPct. Further, the values in these columns were downloaded in a percentage format. IE - 0.24 showed up as 24%. While converting these to the appropriate decimal values in R is not a difficult undertaking, doing a quick column operation in MS Excel was even quicker. The rest of the data manipulation and analysis was conducted within R. The data was loaded into R with the following code. Also note the packages imported at the beginning of the analysis that will be used in subsequent sections.

```
> library(ggplot2)
> library(plyr)
> library(data.table)
> library(dplyr)
> library(boot)
> library(reshape2)
> masterHitters = read.csv("Hitters.csv")
> hitterTable <- data.table(masterHitters)
>
```

Dataset Overview

The data has 27 variables (columns) per observation (row). Each row represents information about a player for a given season of games in the MLB. It includes summary information such as the player's name, team and age as well as a rather detailed look at their offensive and defensive performance. I've intentionally chosen a mix of traditional metrics (Batting Avg, RBIs, Runs) as well as more useful modern metrics (wOBA, wRC+, WAR etc.). An exact listing of the columns follows.

```
   Season      Name      Team       Age         G        PA        HR         R
"integer"  "factor"  "factor" "integer" "integer" "integer" "integer" "integer"
      RBI        SB     BBPct      KPct       ISO     BABIP       AVG       OBP
"integer" "integer" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
      SLG      wOBA   wRCPlus       BsR       Off       Def       WAR     HR.FB
"numeric" "numeric" "integer" "numeric" "numeric" "numeric" "numeric" "numeric"
      wRC       Pos  playerid
"integer" "numeric" "integer"
```

To use the terminology common to relational databases, the primary key of this table would be the integer "playerid" (unique for every individual player) combined with the integer "Season" column. A combination of these two items guarantees you to acquire at most one record from the table.

Analysis

Outliers: Discarding Small Sample Seasons

Traditional, basic outlier detection methods would flag a number of seasons in which a player has performed abnormally well or poorly in a very small number of plate appearances (hereafter PAs). To counter this, only seasons in which a player recorded a "qualifying" number of plate appearances are included in the analysis. A qualifying season is defined as one in which the player has recorded 3.1 PAs per game his team has played. Yes, that's a pretty arbitrary number but it's a legal MLB standard so we'll go with it as it roughly fits what we want.

Column Relationships and Correlation (R)

Most of the columns in the table are inherently correlated. This is because they are largely built off of one another. For example, SLG is a combination of AB, H, 2B, 3B, HR. OBP is a combination of BB, HBP, H and PA etc. Because of this, many of the column to column correlations are fairly uninteresting. They become more interesting (and useful) in the case in which the columns denote events that are completely independent of each other. For example, a hitter cannot take a Walk and strikeout in the same plate appearance.

The following code and subsequent resultant table show all correlations in a subset of columns. The subset of columns was chosen to highlight some useful independent relationships as well as how some of the new "advanced" metrics are constructed.

```
> noNAHitters <- filter(masterHitters, !is.na(KPct))
> d <- data.frame(KPct = noNAHitters$KPct, BBPct = noNAHitters$BBPct,
+                 HR = noNAHitters$HR, RBI = noNAHitters$RBI,
+                 R = noNAHitters$R, wRCPlus = noNAHitters$wRCPlus,
+                 Off = noNAHitters$Off, Def = noNAHitters$Def, WAR = noNAHitters$WAR)
> d_cor <- as.matrix(cor(d))
> d_cor_melt <- arrange(melt(d_cor), -abs(value))
> c <- ncol(d)
> r <- nrow(d_cor_melt)
> numTail <- r- c
> d_cor_melt<- tail(d_cor_melt, numTail)
> Nth.delete<-function(dataframe, n)dataframe[-(seq(n,to=nrow(dataframe),by=n)),]
> corFrame <- data.frame(d_cor_melt)
> row.names(corFrame) <- seq(nrow(corFrame))
> corFrame <- Nth.delete(corFrame, 2)
```

	Var1	Var2	value
1	Off	wRCPlus	0.979566741
3	WAR	Off	0.857376639
5	WAR	wRCPlus	0.826710966
7	RBI	HR	0.786685687
9	Off	RBI	0.655121220
11	Off	R	0.651296464
13	WAR	R	0.647958758
15	wRCPlus	RBI	0.641638185
17	Off	HR	0.630261534
19	wRCPlus	HR	0.623514848
21	wRCPlus	R	0.586841712
23	wRCPlus	BBPct	0.538598864
25	Off	BBPct	0.533026061
27	WAR	RBI	0.532142205
29	R	RBI	0.532131054
31	HR	KPct	0.517645187

```
33     WAR      HR  0.483877903
35       R      HR  0.464492032
37     WAR   BBPct  0.431413935
39       R   BBPct  0.339432392
41      HR   BBPct  0.331473268
43     WAR     Def  0.309956813
45     Def      HR -0.239336067
47     RBI   BBPct  0.231205826
49     Def wRCPlus -0.228090507
51     Def     RBI -0.226134263
53     Def     Off -0.205153706
55     RBI    KPct  0.202926222
57     Def    KPct -0.166877119
59   BBPct    KPct  0.162365494
61     Def   BBPct -0.146399732
63 wRCPlus    KPct  0.104662709
65     Off    KPct  0.094679546
67     Def       R -0.050919618
69     WAR    KPct -0.002876488
71       R    KPct -0.001504895
```

A couple of things to note here:

1. Offense and defense (the aggregated measures Off and Def) are negatively correlated. On average, it seems a challenge to be both an good hitter and good defender. (There are obvious exceptions - Evan Longoria, Mike Trout before he got big etc)

2. WAR and wRC+ are very closely correlated. This is not surprising given the offensive component that goes into WAR is essentially what wRC+ is (a park adjusted, linear weights driven measure of offensive output).

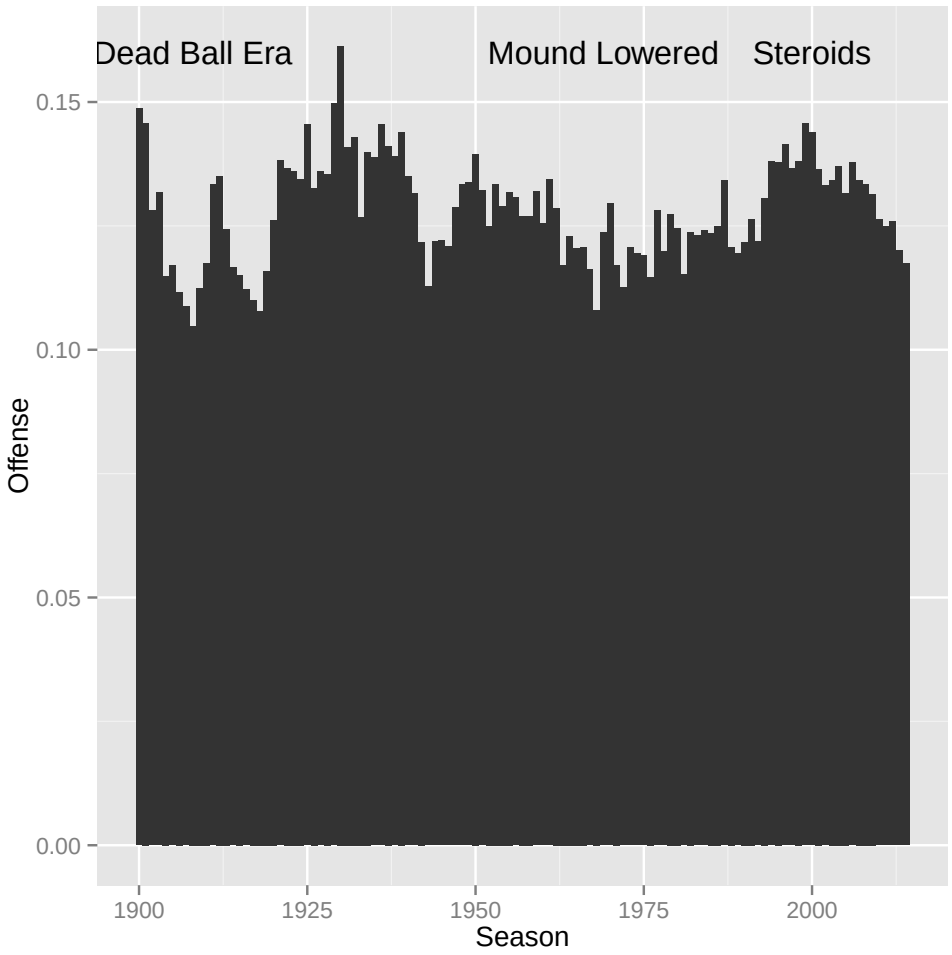
Raw Offense Through the Years

Much has been made of the recent decrease in offense throughout the came. When compared to the late 1990s and early 2000s, this is certainly true. However, when one zooms out and looks at offense since 1900 it becomes clear that there are natural ebbs and flows to the MLB run environment. The following code and chart illustrates this.

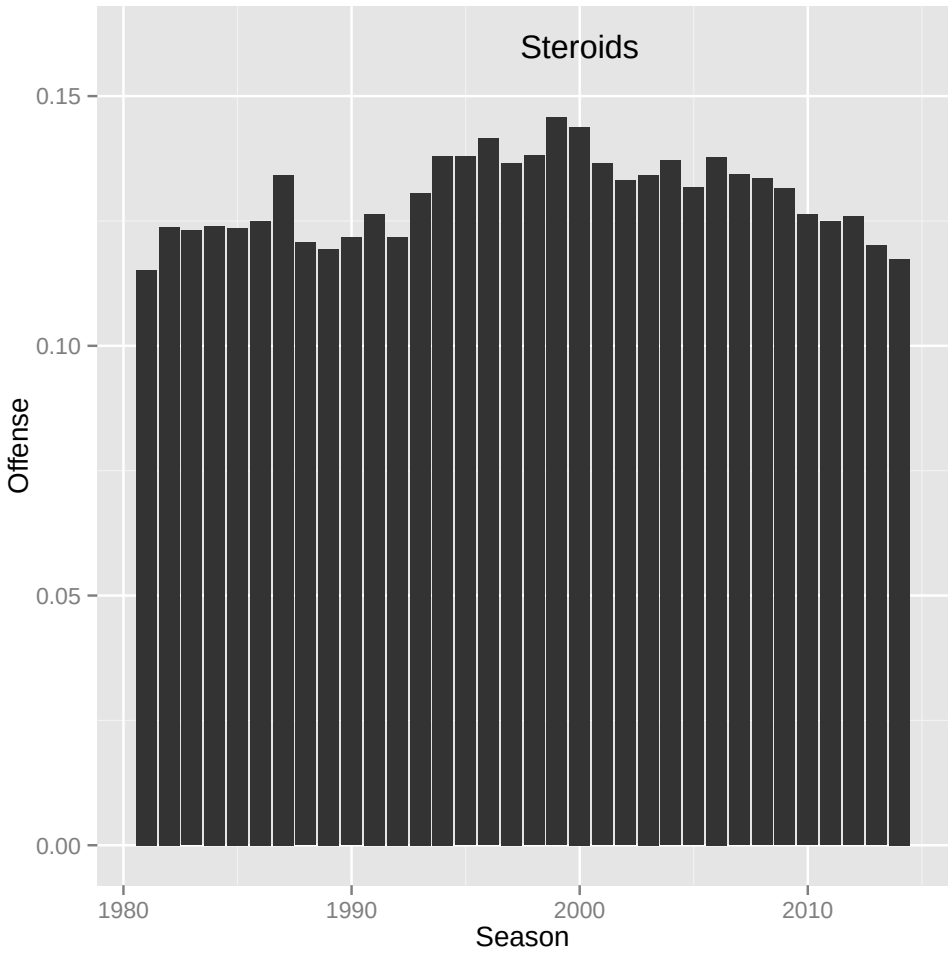
Note that the proxy used for offense in each season was league wide runs per plate appearance. Raw total runs won't work in this case as the number of teams (and thus the number of total runs scored league wide) has changed significantly throughout the years.

```
> yearlyOffenseProxyTable <- hitterTable[, sum(R) / sum(PA), by = Season]
> yearlyOffenseProxy <- data.frame(Season = yearlyOffenseProxyTable[[1]], Offense = yearlyOffenseProxyTa
> p1 <- ggplot(yearlyOffenseProxy, aes(x= Season, y = Offense)) + geom_bar(stat = "identity")
>
```

I've added labels at the peak of the steroid era, the era in which offense plummeted such that the powers at be decided they needed to lower the pitching mound to increase offense and the period at the start of the century colloquially known as the dead ball era (low offense).



Zoom the chart in a little further and the steroid era seems more isolated and more fitting of the narrative that juiced up pill-poppers had a giant effect on the game. If I were a journalist who really only cared about crafting a narrative, I might choose this chart.



Team Offense Over Years

Next, we'll take a look at team offense over the years using the same proxy for total offense. We bin the players into the teams that they played on in their given season and combine their runs scored per PA. The first chart shows the overall distribution. Most teams are relatively similar. At the tails of the distribution things start to spread out more.

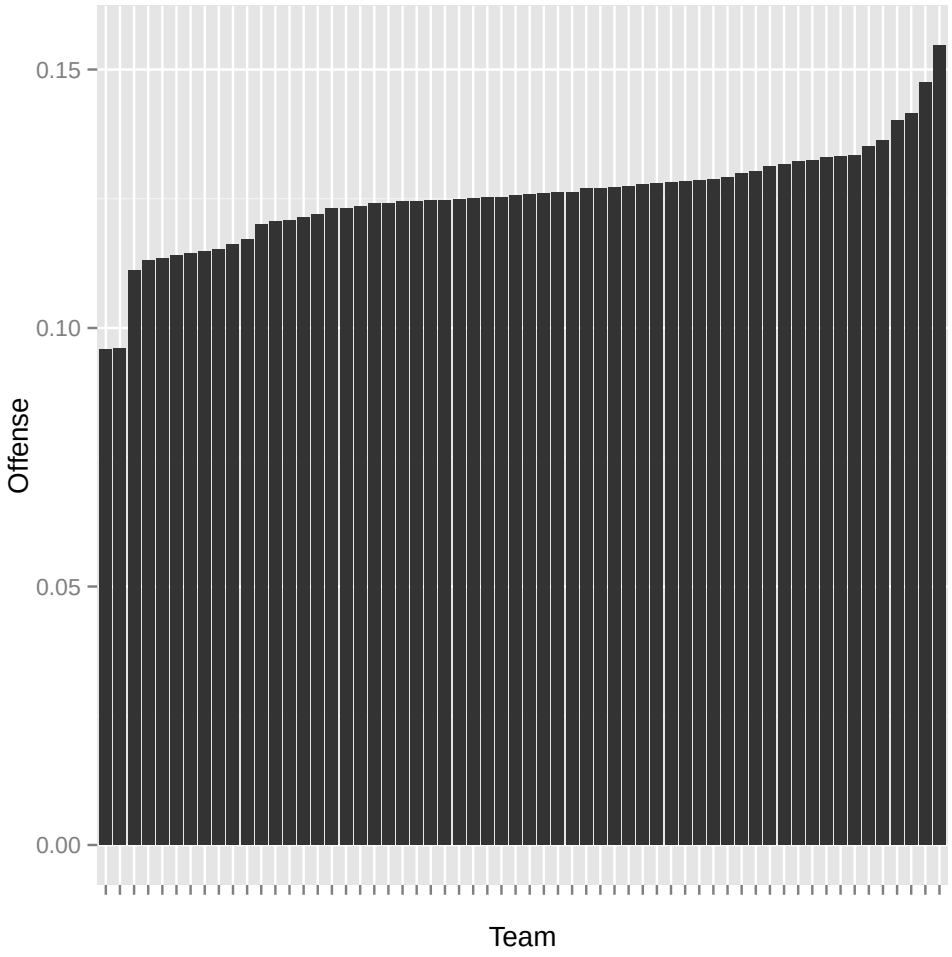
Note that this is really measuring combined team skill + their home park effect. This is not how you would go about creating park effects but it does glean some insight into which parks are hitter friendly and not.

```
>    teamOffenseProxyTable <- hitterTable[, sum(R) / sum(PA), by = Team]
>    setnames(teamOffenseProxyTable, c("Team","Offense") )
>    teamOffenseProxyTable$Team3 <- reorder(teamOffenseProxyTable$Team, teamOffenseProxyTable$Offense)
>    p2 <- ggplot(teamOffenseProxyTable, aes(y=Offense))
>    print(teamOffenseProxyTable[with(teamOffenseProxyTable, order(-Offense))])
```

```
           Team    Offense        Team3
 1:     Hoosiers 0.15456599     Hoosiers
 2:      Rockies 0.14744456      Rockies
 3:      Yankees 0.14153065      Yankees
 4:        Blues 0.14016412        Blues
 5:      Redlegs 0.13635393      Redlegs
```

```
 6:     Bronchos 0.13508613     Bronchos
 7: Diamondbacks 0.13343610 Diamondbacks
 8:       Tigers 0.13314146       Tigers
 9:       Giants 0.13291573       Giants
10:      Red Sox 0.13249361      Red Sox
11:      Indians 0.13219423      Indians
12:    Cardinals 0.13163231    Cardinals
13:      Rangers 0.13124406      Rangers
14:      Dodgers 0.13024321      Dodgers
15:    Blue Jays 0.12980603    Blue Jays
16:      Pirates 0.12903698      Pirates
17:    Athletics 0.12869289    Athletics
18:      Marlins 0.12852507      Marlins
19:      Brewers 0.12841977      Brewers
20:         Cubs 0.12807565         Cubs
21:         Rays 0.12785759         Rays
22:        Twins 0.12770497        Twins
23:       Whales 0.12739965       Whales
24:     Mariners 0.12717125     Mariners
25:       Pilots 0.12695652       Pilots
26:       Astros 0.12689144       Astros
27:       Braves 0.12628853       Braves
28:         Reds 0.12620435         Reds
29:    Nationals 0.12607571    Nationals
30:   Devil Rays 0.12585153   Devil Rays
31:    White Sox 0.12558786    White Sox
32:     Phillies 0.12525034     Phillies
33:       Royals 0.12515736       Royals
34:     Tip-Tops 0.12502709     Tip-Tops
35:       Angels 0.12477901       Angels
36:      Buffeds 0.12473647      Buffeds
37:         Naps 0.12469124         Naps
38:       Browns 0.12451258       Browns
39:       Robins 0.12440600       Robins
40:         Mets 0.12410652         Mets
41:      Orioles 0.12404526      Orioles
42:        Expos 0.12348342        Expos
43:     Senators 0.12318219     Senators
44:     Rustlers 0.12307692     Rustlers
45:    Americans 0.12186798    Americans
46:      Packers 0.12142857      Packers
47:       Padres 0.12088644       Padres
48:        - - - 0.12069577        - - -
49:       Rebels 0.12003817       Rebels
50:      Orphans 0.11708170      Orphans
51:  Highlanders 0.11624776  Highlanders
52:     Superbas 0.11527252     Superbas
53:         Bees 0.11471611         Bees
54:    Terrapins 0.11444292    Terrapins
55:     Chi-Feds 0.11403888     Chi-Feds
56:     Terriers 0.11349268     Terriers
57:   Beaneaters 0.11305857   Beaneaters
58:       Pepper 0.11108416       Pepper
59:   Colt .45's 0.09599075   Colt .45's
```

```
60:        Doves 0.09576516        Doves
            Team    Offense        Team3
```



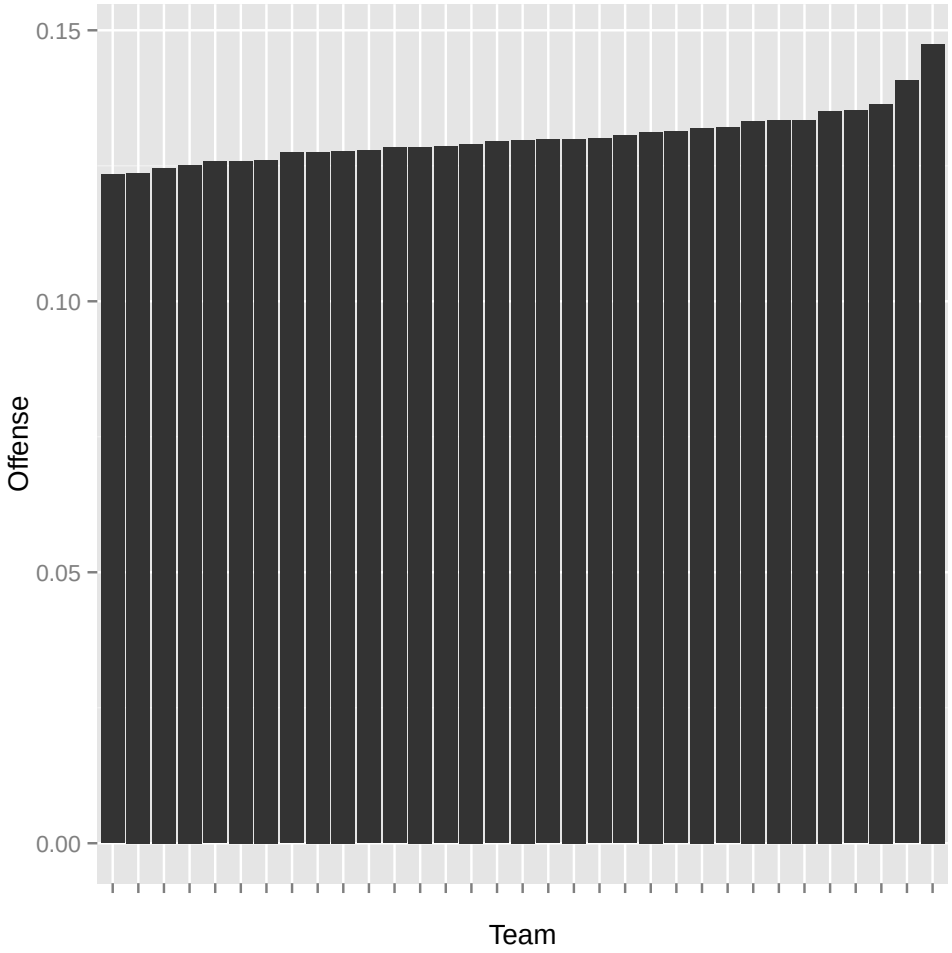
Not surprisingly at the top we see teams that are good (the Yankees) or that have known hitter friendly ballparks (the Rockies) or played in a weird, more unstable era (Indianapolis Hoosiers). The bottom of the distribution is almost completely made out of teams that played exclusively in the dead ball era.

It's probably worth taking a look at only modern teams as dead ball era teams and players don't really factor into any relevant analysis these days.

```
> modernHitterTable = subset(hitterTable, Season > 1980)
> teamOffenseProxyTable <- modernHitterTable[, sum(R) / sum(PA), by = Team]
> setnames(teamOffenseProxyTable, c("Team","Offense") )
> teamOffenseProxyTable$Team3 <- reorder(teamOffenseProxyTable$Team, teamOffenseProxyTable$Offense)
> p2 <- ggplot(teamOffenseProxyTable, aes(y=Offense))
> teamOffenseProxyTable <- arrange(teamOffenseProxyTable,desc(Offense))
> print(teamOffenseProxyTable[with(teamOffenseProxyTable, order(-Offense))])

           Team   Offense        Team3
 1:      Rockies 0.1474446      Rockies
 2:      Yankees 0.1407175      Yankees
 3:      Indians 0.1363888      Indians
 4:      Red Sox 0.1352022      Red Sox
 5:      Rangers 0.1350836      Rangers
```

```
 6:    Blue Jays 0.1334400    Blue Jays
 7: Diamondbacks 0.1334361 Diamondbacks
 8:    Cardinals 0.1333051    Cardinals
 9:       Tigers 0.1320228       Tigers
10:       Braves 0.1319747       Braves
11:     Phillies 0.1313592     Phillies
12:    White Sox 0.1312340    White Sox
13:       Astros 0.1305936       Astros
14:         Reds 0.1300678         Reds
15:      Brewers 0.1299216      Brewers
16:         Mets 0.1298876         Mets
17:    Athletics 0.1297793    Athletics
18:       Angels 0.1295011       Angels
19:       Giants 0.1290377       Giants
20:      Marlins 0.1285251      Marlins
21:     Mariners 0.1284541     Mariners
22:         Cubs 0.1283808         Cubs
23:         Rays 0.1278576         Rays
24:        Twins 0.1277621        Twins
25:        Expos 0.1275200        Expos
26:      Orioles 0.1274531      Orioles
27:    Nationals 0.1260757    Nationals
28:   Devil Rays 0.1258515   Devil Rays
29:      Dodgers 0.1258204      Dodgers
30:      Pirates 0.1251051      Pirates
31:       Royals 0.1246024       Royals
32:        - - - 0.1236922        - - -
33:       Padres 0.1234353       Padres
           Team   Offense        Team3
```



The low-run tail of the distribution has smoothed considerably without the inclusion of deadball era teams. Not the Yankees (astronomical payroll) and the Rockies (ballpark at the highest elevation) occupy the top two spots. Heuristic: If you want your team to score more runs - either spend more or relocate to your closest mountain range.

Year to Year Player Performance

In baseball, more interesting analysis can be peformed when looking at a player's performance through the years and how much predictive power past outcomes have on future outcomes. The most simplistic way to do this is to look at the year to year correlations for a number of offensive metrics. To do this, we need to run correlations on the data row to row (as opposed to column to column). The following code pairs up subsequent seasons of player performance and meshes them together into a single row thus making the correlation calculations easier (once again column to column).

```
> masterHitters2 <- masterHitters
> masterHitters2$Season <- masterHitters2$Season + 1
> year2year <- merge(masterHitters2, masterHitters, by=c("playerid", "Season"))
```

Additionally, we're going to perform a weighted correlation on the data so as to properly balance the influence of player seasons. The logical weigthing here is the total PAs between the two seasons in question. The "corr" function from the "boot" package assists us greatly in this task.

```
> wCorrAvg <- corr(cbind(year2year$AVG.x, year2year$AVG.y),
+                  w=(as.numeric(year2year$PA.x) + as.numeric(year2year$PA.y)))
> wCorrRBI<- corr(cbind(year2year$RBI.x, year2year$RBI.y),
+                 w=(as.numeric(year2year$PA.x) + as.numeric(year2year$PA.y)))
> wCorrR <- corr(cbind(year2year$R.x, year2year$R.y),
+                w=(as.numeric(year2year$PA.x) + as.numeric(year2year$PA.y)))
> wCorrOBP<- corr(cbind(year2year$OBP.x, year2year$OBP.y),
+                 w=(as.numeric(year2year$PA.x) + as.numeric(year2year$PA.y)))
> wCorrSLG <- corr(cbind(year2year$SLG.x, year2year$SLG.y),
+                  w=(as.numeric(year2year$PA.x) + as.numeric(year2year$PA.y)))
> wCorrwRCPlus <- corr(cbind(year2year$wRCPlus.x, year2year$wRCPlus.y),
+                      w=(as.numeric(year2year$PA.x) + as.numeric(year2year$PA.y)))
> wHRPerFB<- corr(cbind(na.omit(year2year)$HR.FB.x, na.omit(year2year)$HR.FB.y),
+                 w=(as.numeric(na.omit(year2year$PA.x) + as.numeric(na.omit(year2year$PA.y)))))
> wBBPct <- corr(cbind(year2year$BBPct.x, year2year$BBPct.y),
+                w=(as.numeric(year2year$PA.x) + as.numeric(year2year$PA.y)))
> wKPct <- corr(cbind(na.omit(year2year)$KPct.x, na.omit(year2year)$KPct.y),
+               w=(as.numeric(na.omit(year2year$PA.x) + as.numeric(na.omit(year2year$PA.y)))))
> f<-data.frame(c("Avg","RBI","R","OBP","SLG","wRC+","HRPerFB","BBPct","KPct"),
+               c(wCorrAvg, wCorrRBI, wCorrR, wCorrOBP, wCorrSLG,
+                 wCorrwRCPlus, wHRPerFB, wBBPct,wKPct))
> setnames(f, c("Stat","R") )
> f[with(f, order(-R)), ]

     Stat         R
9    KPct 0.8871530
8   BBPct 0.8009087
7 HRPerFB 0.7569801
5     SLG 0.7291001
2     RBI 0.6982567
4     OBP 0.6740915
6    wRC+ 0.6673462
3       R 0.5591737
1     Avg 0.5570178
```

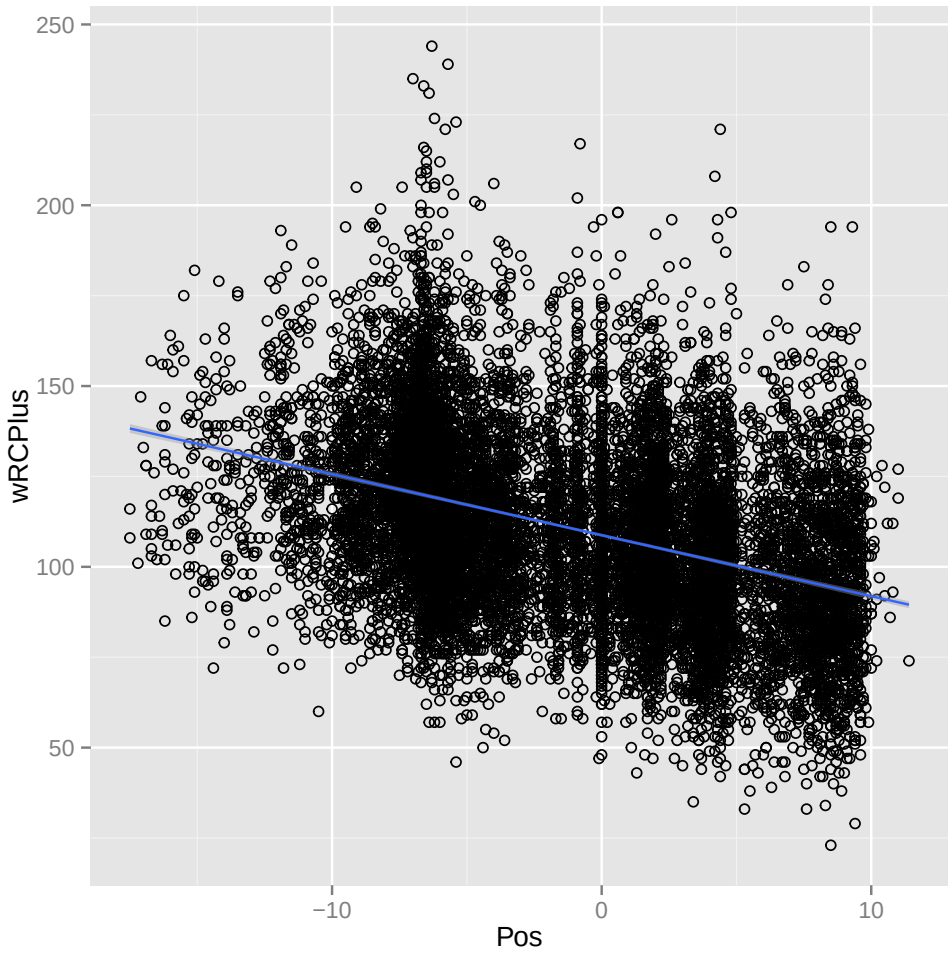
Roughly speaking, the magnitude of correlation here indicates how well a statistic acts as a measure of a specific skill. The closer we are to a skill with our measurements, the more we expect it to persist through time. It's not surprising then that modern projection systems use historical KPct and BBPct outcomes as foundational building blocks. It's also interesting that wRC+ does not have a great deal of repeatability. What this tells us is that in order to build wRC+ projections, we should look to project individual components and then build a composite wRC+. In other words, don't use past wRC+ to predict future wRC+.

Note that only a subset of the columns available were included in the year to year analysis. Any number could be added/analyzed with minor modifications to the above code.

Defensive Spectrum

It may be interesting to a semi-casual observer as to why there is no bucketing of players into positions(1B, 2B CF etc). Instead we've used the built in positional adjustments to transform those positions into numerical values. Positions that are harder to play (SS, C) are the highest values while easy to play positions that can be manned by the David Ortiz's of the world (1B, DH) are given the smallest values. An additional benefit here is that it accounts for players who play multiple positions in one season.

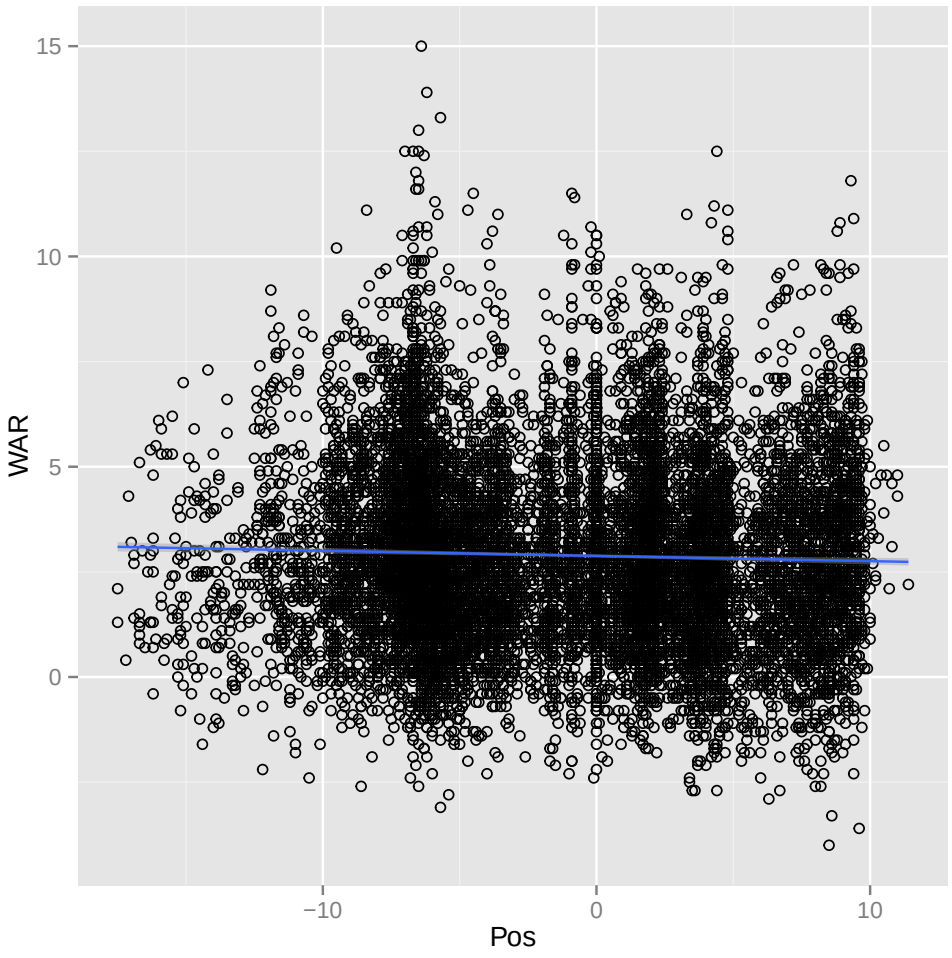
```
> defSpecToOffCorr <- corr(cbind(masterHitters$Pos, masterHitters$wRCPlus), w=(masterHitters$PA))
> p3 <- ggplot(masterHitters, aes(x=Pos, y=wRCPlus)) +
+   geom_point(shape=1) +    # Use hollow circles
+   geom_smooth(method=lm)
```



There's a clear relationship here. It seems as though players chosen to play the easy defensive positions are usually good hitters. Intuitively this makes sense as a team may be willing to take a hit on run prevention to get a good offensive player.

Similarly, we can see how the Wins Above Replacement (WAR) value model accounts for this.

```
> defSpecToWARCorr <- corr(cbind(masterHitters$Pos, masterHitters$WAR), w=(masterHitters$PA))
> p4 <- ggplot(masterHitters, aes(x=Pos, y=WAR)) +
+   geom_point(shape=1) +    # Use hollow circles
+   geom_smooth(method=lm)
```



Here we see no correlation. Again, this is expected as WAR attempts to encapsulate the premium value that should be assigned to the players able to play difficult positions. The flattening of the line of best fit here is a result stemming directly from that.

Conclusion

We've examined a number of things with this data set. It has been shown that league wide offense has been the subject of ups and downs throughout time. It appears that Colorado and the Yankees are historically more likely to put up runs than other teams but for very different reasons. Year to year player performance has strong correlation, especially for those metrics most closely measuring and underlying skill and finally the relationship between position and offensive performance has been examined (as well as a cursory glance as to how WAR takes this into account).