Final Project - MLB Contracts

Kenny Kato

5/8/2020

Abstract

An explosion of baseball data in the past two decades has transformed everything from the top of a baseball organization to the bottom, and while much of that information has been deployed in the field to gain in-game competitive advantages, a closely related mission of baseball executives is evaluating players for the sake of accurately paying them. Using single-season and multiple-season performance data from FanGraphs.com and historical contract information from Cot's Baseball Contract, I toss the information through a couple predictive models to uncover the relationship between player performance and the ultimate payday that awaits them – though not without limitations, such as missing control information (like injury data) that I believe could sharpen the precision of these models. Nevertheless, I do find that there exists a seemingly generic quadratic relationship between the principal component analysis-derived performance metrics and the player's eventual contract average annual value (AAV). In a somewhat unexpected turn of events, I find that a linear regression model (utilizing a quadratic term) consistently outperforms a random forest regression model for both the single- and multiple-season datasets, suggesting a generally simple path relationship between performance and salary.

Introduction

Over the past two decades, Major League Baseball has experienced a transformation in the way baseball players are analyzed and evaluated. New ideas have flooded the game, ranging as far afield in their aim as Willie Mays might have once covered centerfield at the Polo Grounds. Philosophies about how to win, improve teams, keep players healthy, and more have been introduced, reinvented, or even cast into obsolescence in response to a deluge of data. Teams who used to tally up RBIs as a favorite metric have instead adopted wRC+, while teams who once used wRC+ now simply rely on sufficiently loud clubhouse trash cans and a solid A/V system. On the field, the spoils of war have increasingly gone to teams that are best able to utilize the wide variety of information now available to them, but data can be quite useful off the field as well.

Among one of the stickier issues that MLB executives and analysts are tasked with is how to value a counterfactual (impossible!): they are trying to pay a player (in this report the player is a free agent) according to their future performance, largely based on prior performance. The price of a single free agent can have sprawling consequences – on fellow free agents seeking a contract, on future free agents who are deciding whether to sign a contract extension with their current team or test free agency, on young players still having their valuations determined in the salary arbitration process, and so forth. Free agency is a highly dynamic economy and no team wants to end up being the one that signs Pablo Sandoval for $95 million.

What I am interested in finding, then, is simple: what is the nature of the relationship between salary and prior performance, and to what extent is prior performance capable of predicting salary? Theoretically, in Major League Baseball's free agent market – a ruggedly free market – players should be getting paid according to merit. While factors like individual team and personal player needs preclude us from formulaically tossing numbers into a machine and getting the correct salary spit back out at us, we should be able to estimate player salaries pretty closely simply based on performance given the meritocratic environment and the fact that players' merits as ballplayers are publicly available. With the help of principal component analysis and

a couple predictive models, I attempt to find something resembling a market price for players based on recent performance, and I add a few notes speculating why any given player might deviate from it.

Methods

Data

Fundamentally, data on the performance of individual players needs to sufficiently describe the true performance of a player to be worthwhile; with this in mind, I highly valued variety in my selection process for whose database to use, i.e. I wanted data capable of capturing the various idiosyncrasies of players. Fortunately, measuring the performance of a ballplayer is of primary concern to anyone and everyone who watches sports with semi-regularity, so data is presented pretty straightforwardly. I acquired data for both just the most recent year prior to the new contract, as well as the most recent three years, simply to see how much the model might improve with more established track records. The dataset I used carried 91 performance metrics, plus a column each for the age of the player during that season (or average age, in the case of the three-season dataset) and the team for which they played. I opted to use FanGraphs.com for my data source, and because several sources gather various types of baseball information, I think I should briefly explain why I feel that FanGraphs.com, relative to other well-known databases, offers the most accessible, valuable, and richest variety of performance metrics for predicting contract values.

While Baseball Prospectus has high-quality information and is one of the oldest and most respected organizations for analysis using advanced baseball statistics (sabermetrics), including some of their own proprietary metrics, a subscription is required to access higher-level data sorting and metrics. Baseball Savant, MLB's repository for the cutting-edge Statcast metrics that collects event-level microdata like "exit velocities" or "launch angles" on individual contact-making swings), might be my choice if I were to pursue this project in twenty years' time, but their most interesting metrics are currently only a few years old. Databases like those found on more mainstream websites like MLB.com or ESPN.com largely limit their variables to those most familiar to fans of all stripes – the fanatical and the casual – which are less nuanced and violates my "variety" condition. FanGraphs.com – which uses Retrosheet.org and Baseball Info Solutions – is free to use and hosts one of the oldest databases for sabermetrics, operating since the mid-2000s. Not only do they satisfy my variety condition, but their existence straddles the sabermetric revolution – the proliferation of newer and more descriptive statistics in the analysis of ballplayers – which suggests that their data is more likely to represent the information being used by baseball executives to make salary decisions during the time period I'm studying. Essentially, if FanGraphs.com is unable to adequately describe the performance of a player, then it simply cannot be done at this point in time.

Contract data was collected from Cot's Baseball Contracts from as early as the 2008-2009 offseason to the most recent 2019-2020 offseason. In total, there are twelve years' worth of contract data and I kept the 530 observations of position players (i.e. not pitchers) that were awarded a contract totaling at least $1 million in that timeframe. (It was originally 532, but two players were dropped – Rafael Furcal and Corey Hart – because they did not play in the majors due to injury the year prior to signing their MLB contracts.) An explanation of why I selected Cot's is not necessary, because it hosts the only historical database of contract information that I can find online. Spotrac is another popular website for information on sports contracts, but their historical databases are behind a paywall. The FanGraphs data was reduced to simply the 530 players who signed contracts between 2008 and 2020, and merged with the Cot's contract data, joined by the player. All in all, my starting datasets were 530 rows by 102 columns.

Limitations

Unfortunately, in exploring the contributions of prior performance to salary decisions, I don't get to play with a completely full deck. One dataset I wish existed publicly was a compilation of individual players' health histories. Health is of course a major influence on salary negotiations, one I would like to control for when attempting to isolate the effect of performance; some teams in recent years have even begun to invest in additional research on biomechanics, trying to optimize their ability to prevent, anticipate, manage, and recover from injuries. Understandably, the knowledge of a player's true health status is commonly restricted

to team personnel, for both personal and competitive reasons, but even the information that becomes public knowledge – like the nature of injuries that diminish playing time, body parts affected, severity of an injury, surgeries required, etc. – seems to exist only in scattered press releases and articles across the web. For now, we will simply acknowledge the limitation and trust in the ability of statistics that can be proxies for injuries – like games played or plate appearances – to get us close to the mark.

Another limitation to this project is mostly due to a personal lack of prescience, which is my failure to account for which teams sign which players. I think it is likely that salary amounts are somewhat dependent on whether a low-payroll team like the Tampa Bay Rays or Oakland A's are paying them or a high-payroll team like the New York Yankees or Los Angeles Dodgers are paying them. I can imagine both a demand- and supply-side boosting effect on salaries for above-average players: on the team side of a negotiation, a high-payroll team might in theory exercise a greater flexibility to "overpay" for a player they would like to have, while on the player side, an elite player might recognize the scarcity of comparable players available and exact something like a monopoly markup from the buying team. In light of this, then, we might actually expect to observe salaries nonlinearly boosted upwards as talent rises and becomes scarcer, and I will in fact show this relationship in the next section.

Approach

Clearly, no one wants to watch me regress a player's AAV onto 90-something variables and since many of the features of my dataset are largely correlated (sometimes even involved in each other's calculations), the size of the dataset wouldn't correspond to greater explanatory power anyway, so my first step in preprocessing the data was to run a principal component analysis (PCA) on the dataset of performance metrics to reduce the dimensions. The principal components would then effectively become my performance metrics for further analysis.

After reducing my dataset, I split it into training and testing sets (80/20) and tossed my principal components (along with a "player age" variable marking the age of the player going into the first year of the new contract) into two competing models: a random forest regression and a linear regression on the training set, with AAV as my outcome variable, to measure the quality of predictions on the testing set. The random forest model is meant to address both the potential nonlinearity for elite players that I alluded to above, as well as any other nonlinear factors I may not have considered, since I expect it to make better predictions on a dataset than the linear model. I used 8 feature samples for each tree in the 1-year set and 5 features in the 3-year set, and in both cases I bootstrapped 500 trees. I averaged the root-mean-squared error (RMSE) of each regression over 100 iterations of resampling the training and testing sets to compare the performances of each.

Results

Principal Components

Altogether, in the one-year dataset, 15 of the principal components explain about 85% of the variance, while 20 principal components explain about 90%; in the three-year dataset, 11 principal components gets 85% and 15 principal components will get us to 90%.

Several of the principal components do a solid job characterizing baseball player archetypes, many of which are addressed in the conclusion. The first principal component (PC1), for instance (which explains 25% of the variance in the 1-year set, 28% in the 3-year), appears to be a sort of negative quality index, so I guess you could say that players associated with PC1 are not used to scoring so highly. PC1 is largely the same in both the 1- and 3-year datasets. Below are the top- and bottom-eight ranked features in the 3-year.

Top:

```
##  med_pct_plus       med_pct            sh    positional soft_pct_plus
##    0.11950049    0.11236451    0.11052288    0.10991946    0.10451899
##      soft_pct           def            bu
##    0.09826271    0.09349999    0.08709827
```

Bottom:

```
##       wraa         hr        bat   wrc_plus       woba        slg
## -0.1810694 -0.1794558 -0.1790123 -0.1783494 -0.1760948 -0.1753805
##   slg_plus       re24
## -0.1752259 -0.1743781
```

Its high-ranking features are only mildly identified and appear to associate well with what baseball folk might call "journeymen," the quintessential job-seeker. They typically sign to short-term deals and so bounce around from team to team, well-acquainted with free agency. Both "positional" and "def" are defensive skill measures, while "soft_pct_plus" is how above-average the player is at making "soft contact" when hitting (not good). Its lowest-ranked features are also those that are measures of skill: both wRAA and wRC+ are measures of the amount of runs you contribute to your team above-average (run production being, of course, a fundamentally good thing in baseball).

To verify that the demand for increasing talent increases nonlinearly, I took this shiny new quality index and plotted it against AAV; I also plotted it against a well-known baseball quality index – Wins Above Replacement (WAR) – to confirm that I was correct to describe PC1 as a quality index in its own right. There is definitely a strong and well-defined relationship between the three variables:

Figure 1. PC1 and AAV

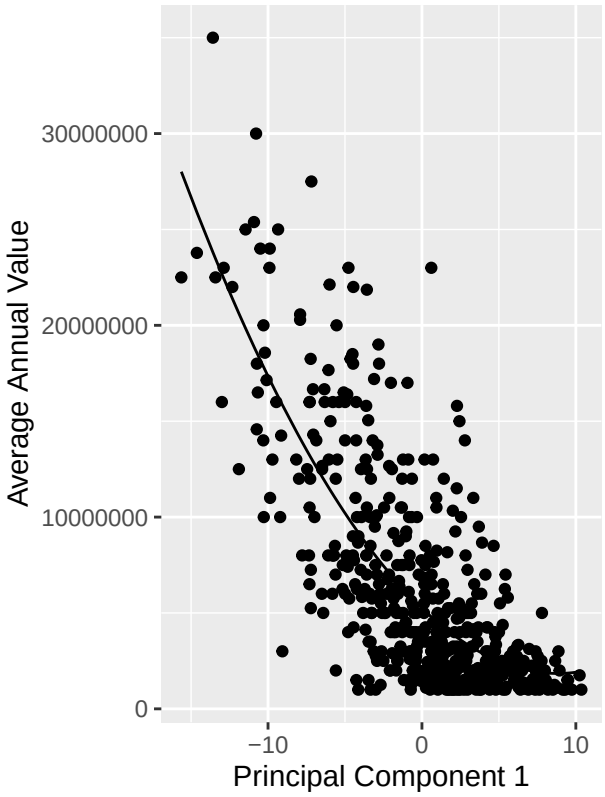
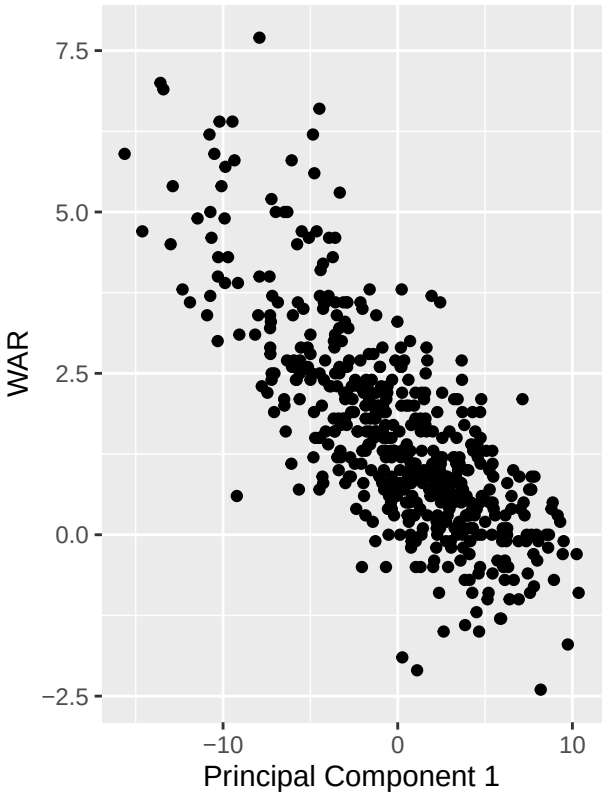


Figure 2. PC1 and WAR



Regressions

```
# Linear Model (1-Year)

# aav ~ agecontract + poly(PC1,2) + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 + PC10 +
#                      PC11 + PC12 + PC13 + PC14 + PC15 + PC16 + PC17 + PC18 + PC19 + PC20

# Random Forest Model (1-Year)

# aav ~ agecontract + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 + PC10 +
#                     PC11 + PC12 + PC13 + PC14 + PC15 + PC16 + PC17 + PC18 + PC19 + PC20
```

I use the 20 principal components as variables in the one-year regressions and 15 in the three-year regressions (90% variance explanation for each), as well as a variable for the player's age going into the first year of the contract. In the linear models, based on the relationship between AAV and PC1 that we saw above, I added the little twist of a quadratic term on the first principal component, which seemed to outperform a linear model without it and should help it compete with the random forest model. Running the two regressions over the 100 iterations presented a somewhat surprising result: the linear model (with the quadratic term) fairly convincingly outperformed the random forest predictions in terms of averaged RMSE.

Linear Model RMSE (1-Year):

```
## [1] 4083181
```

Random Forest RMSE (1-Year):

```
## [1] 4613329
```

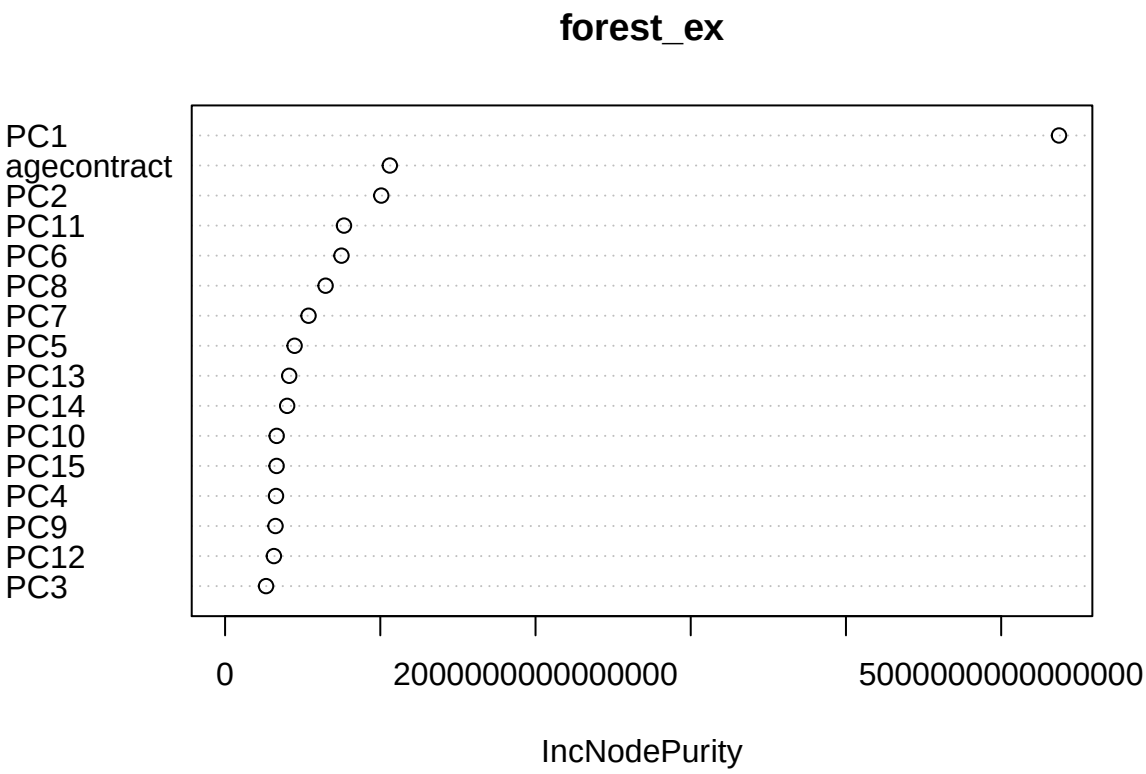
Linear Model RMSE (3-Year):

```
## [1] 3317826
```

Random Forest RMSE (3-Year):

```
## [1] 3860961
```

And here, we can see a sample variable importance plot from the random forest model:



Conclusion

In addition to PC1, we see that age plays a significant role in player evaluation – no surprise there – and PC2, PC6, PC8, and PC11 tend to remain near the top on repeated iterations. PC2 seems to represent go-big-or-go-home players, likely to strike out a lot or hit with a lot of power; PC6 appears to be productive hitters that just don't play often; PC8 players are valued highly for their defensive contributions and appear to be kind of slow – possibly catchers; and I'm not quite sure what to make of PC11, but they seem to have done well in more important game situations. The most notable loadings of the first fifteen principal components for the 3-year dataset have been included in the appendix.

Admittedly, I expected the random forest model to perform better than the linear regression. That the random forest model did not outperform the linear regression model suggests to me that at least the basic relationship between performance metrics and salaries is a fairly lazy and predictable path, as we might hope to see, and that laziness allows a lower-variance model like the linear regression outperform a higher-variance model like the random forest. We saw how important PC1 was to the model in the random forest's variable importance plot, so I think it's reasonable to think that accounting for the quadratic relationship with AAV would greatly improve the linear model.

All that said, there was still an RMSE around or in excess of $3-4 million – what drives it? I'm fairly confident in the idea that health information would significantly impact the RMSE, and perhaps someday I will get my hands on a high-quality baseball player injury dataset and redo this report. It is also true that baseball executives simply evaluate ballplayers differently, for better or worse, and much depends on individual circumstances; one ballplayer might take a discount to go to a preferred team, and a team might overpay for a player if the player might be "the missing piece" in their pursuit of a championship. Another thought that occurred to me far too late for the sake of this project is the presence of heteroskedasticity in the model that I frankly failed to really consider. I am sure there is room to refine these models in a way that

could shrink the RMSE but overall, I think we can reasonably say that the "market price" for a given player in free agency, based on their recent performance, can reliably be captured by a PCA-to-linear regression pipeline.

Appendix

Principal Component 1 (3-Year Dataset)

```
##  med_pct_plus       med_pct            sh    positional soft_pct_plus
##    0.11950049    0.11236451    0.11052288    0.10991946    0.10451899
##      soft_pct           def            bu
##    0.09826271    0.09349999    0.08709827
```

```
##       wraa         hr        bat   wrc_plus       woba        slg
## -0.1810694 -0.1794558 -0.1790123 -0.1783494 -0.1760948 -0.1753805
##   slg_plus       re24
## -0.1752259 -0.1743781
```

Principal Component 2 (3-Year Dataset)

```
##         k_pct    k_pct_plus       tto_pct   fb_pct_plus     swstr_pct
##     0.1687307     0.1684456     0.1620719     0.1519699     0.1501489
##        fb_pct pull_pct_plus         hr_fb
##     0.1471882     0.1332850     0.1145856
```

```
##          X1b           gb           ld          ifh  contact_pct
##   -0.1973673   -0.1942440   -0.1772470   -0.1687007   -0.1665111
##            h zcontact_pct          avg
##   -0.1641925   -0.1613257   -0.1592169
```

Principal Component 3 (3-Year Dataset)

```
##          iffb      iffb_pct            fb        fb_pct   fb_pct_plus
##     0.3010404     0.2259500     0.2236410     0.2048260     0.1985383
## soft_pct_plus      soft_pct pull_pct_plus
##     0.1767958     0.1641088     0.1565415
```

```
##  babip_plus       babip       gb_fb    obp_plus      ld_pct      gb_pct
##  -0.2108045  -0.2102993  -0.1815137  -0.1636485  -0.1632134  -0.1617844
## gb_pct_plus ld_pct_plus
##  -0.1588634  -0.1577585
```

Principal Component 4 (3-Year Dataset)

```
##  swing_pct oswing_pct zswing_pct        gdp  swstr_pct   avg_plus
## 0.32027439 0.31062808 0.25275096 0.19407974 0.12950723 0.12222481
##        avg         gb
## 0.10524535 0.09371323
```

```
##      bb_pct bb_pct_plus         bsr        bb_k         spd          sb
##  -0.2583826  -0.2537176  -0.2181444  -0.2028545  -0.1676627  -0.1566626
##          bb         wsb
##  -0.1562161  -0.1490189
```

Principal Component 5 (3-Year Dataset)

```
##         bb_k  contact_pct zcontact_pct ocontact_pct          ibb
##   0.24883362   0.21026375   0.18961831   0.15712190   0.12273294
##           sf      med_pct           fb
##   0.11168941   0.09379700   0.08514134
```

```
##  swstr_pct        bsr      k_pct        ubr    ifh_pct        spd
## -0.2340532 -0.2216446 -0.2136680 -0.2044374 -0.2024073 -0.1965832
##         sb k_pct_plus
## -0.1958984 -0.1868429
```

Principal Component 6 (3-Year Dataset)

```
##      avg_plus      soft_pct           avg      iffb_pct soft_pct_plus
##     0.2075974     0.2043760     0.1989076     0.1937281     0.1851167
##      slg_plus           slg  ocontact_pct
##     0.1712157     0.1644860     0.1637798
```

```
##           k           g          bb replacement          pa  k_pct_plus
##  -0.2387653  -0.1857601  -0.1732212  -0.1664236  -0.1649812  -0.1501775
##       k_pct     tto_pct
##  -0.1423064  -0.1331092
```

Principal Component 7 (3-Year Dataset)

```
##    gb_pct_plus         gb_pct        ifh_pct  soft_pct_plus            ifh
##     0.22163416     0.21656977     0.17252765     0.16486991     0.15785667
##          gb_fb       soft_pct hr_fb_pct_plus
##     0.14634570     0.13027991     0.09313689
```

```
##      ld_pct ld_pct_plus         def         fld     dollars  positional
##  -0.3742605  -0.3248925  -0.3156394  -0.2604978  -0.2221996  -0.2103973
##         war         rar
##  -0.2006586  -0.1976252
```

Principal Component 8 (3-Year Dataset)

```
##         fld         def     dollars         war         rar      gb_pct
##   0.3727233   0.3498608   0.2387733   0.2209060   0.2208577   0.2203412
## gb_pct_plus       gb_fb
##   0.2201324   0.1731289
```

```
## ld_pct_plus      ld_pct         X3b         bsr       babip  babip_plus
##  -0.2601828  -0.2044428  -0.1445110  -0.1410546  -0.1348777  -0.1346429
##      fb_pct fb_pct_plus
##  -0.1317960  -0.1203728
```

Principal Component 9 (3-Year Dataset)

```
##   zone_pct    med_pct        def positional        fld      babip
##  0.4455234  0.2188685  0.1977781  0.1605310  0.1366149  0.1360376
##        avg        rar
##  0.1278944  0.1229721
```

```
##           yr      buh_pct ocontact_pct     soft_pct       ld_pct
##   -0.4921234   -0.2220959   -0.1734661   -0.1638474   -0.1592856
##       clutch   oswing_pct  bb_pct_plus
##   -0.1505375   -0.1238766   -0.1043524
```

Principal Component 10 (3-Year Dataset)

```
##       ifh_pct cent_pct_plus           ifh        clutch            yr
##    0.28001662    0.19988622    0.15193417    0.14797438    0.11435306
##      iffb_pct          iffb           ubr
##    0.11345772    0.10284759    0.09964518
```

```
##           buh       buh_pct            bu            sh    zswing_pct
##   -0.46750352   -0.39442531   -0.38509870   -0.26403860   -0.18608783
##     swing_pct pull_pct_plus    oswing_pct
##   -0.14329061   -0.12449046   -0.08476609
```

Principal Component 11 (3-Year Dataset)

```
##        clutch oppo_pct_plus           ibb    zswing_pct           wpa
##     0.4258517     0.3062607     0.2981327     0.2934049     0.2255828
##     swing_pct soft_pct_plus cent_pct_plus
##     0.1965893     0.1826672     0.1794842

## pull_pct_plus            yr      hard_pct    positional hard_pct_plus
##   -0.29289828   -0.17142286   -0.13632370   -0.10875181   -0.09517372
##           hbp       ifh_pct  zcontact_pct
##   -0.09282878   -0.09192184   -0.09131760
```

Principal Component 12 (3-Year Dataset)

```
##           hbp       ifh_pct            sh      soft_pct soft_pct_plus
##     0.4307820     0.2642180     0.2258893     0.1915430     0.1878628
##      iffb_pct            bu    babip_plus
##     0.1843710     0.1806937     0.1774542

##           wsb            sb           X3b           spd           bsr
##    -0.3082962    -0.1795911    -0.1756812    -0.1492773    -0.1404644
## cent_pct_plus       buh_pct       med_pct
##    -0.1394366    -0.1307238    -0.1008877
```

Principal Component 13 (3-Year Dataset)

```
##        clutch           hbp pull_pct_plus           wpa           ubr
##    0.55627486    0.38355977    0.30562361    0.24273003    0.12538504
##           bsr       ifh_pct          re24
##    0.12530446    0.10325663    0.08549822

## oppo_pct_plus cent_pct_plus      soft_pct soft_pct_plus            cs
##   -0.27264715   -0.25291861   -0.13926433   -0.10756966   -0.09693135
##      iffb_pct          iffb  ocontact_pct
##   -0.09627121   -0.09486429   -0.09317161
```

Principal Component 14 (3-Year Dataset)

```
##           hbp oppo_pct_plus           wsb            sh            bu
##     0.4387953     0.1810022     0.1805819     0.1574380     0.1281625
##    positional            yr cent_pct_plus
##     0.1274094     0.1223668     0.1200429

##       buh_pct      iffb_pct pull_pct_plus         babip    babip_plus
##    -0.4711488    -0.2481293    -0.1777285    -0.1746002    -0.1710134
##          iffb           fld   ld_pct_plus
##    -0.1602261    -0.1483474    -0.1422722
```

Principal Component 15 (3-Year Dataset)

```
##         clutch            bu            sh           buh  hard_pct_plus
##      0.3522783     0.2456995     0.2307781     0.1886902     0.1850718
##            fld  cent_pct_plus hr_fb_pct_plus
##      0.1646825     0.1608300     0.1583137

##   zswing_pct      buh_pct med_pct_plus          hbp      med_pct
##   -0.2905154   -0.2858089   -0.2153062   -0.1944521   -0.1897489
##    swing_pct          ubr  bb_pct_plus
##   -0.1619493   -0.1413616   -0.1142126
```