**A Multivariate Analysis of Baseball Batting Statistics and Salaries**

**ISQS 6350**

**Group 5: Daniel Brauckman, Jordan Gussett, Bryan Loeffler and Elijah Stahr**

**Introduction**

In a non-salary capped sport, professional baseball teams are often competing on unlevel grounds. Each team is looking for their own, distinct, competitive advantages. For some, like the Yankees, Red Sox, and Dodgers, their advantage is financial - as in they can buy the best players. For other smaller market teams, like the Tampa Bay Rays and Arizona Diamondbacks, they must identify trends and performance indicators to surface value in players to make them competitive that they would not be able to afford otherwise. A Sports Illustrated article details one example of the Rays acquiring a pitcher based on his spin rate, a key indicator the organization looks for in development of pitching, and dropped his batting average against from .222 with the Pirates to .091 with the Rays. (Verducci). This is just a small window into the types of edges these organizations look for to compete on a yearly basis with those that can overpay for multiple star players with obvious talent and production.

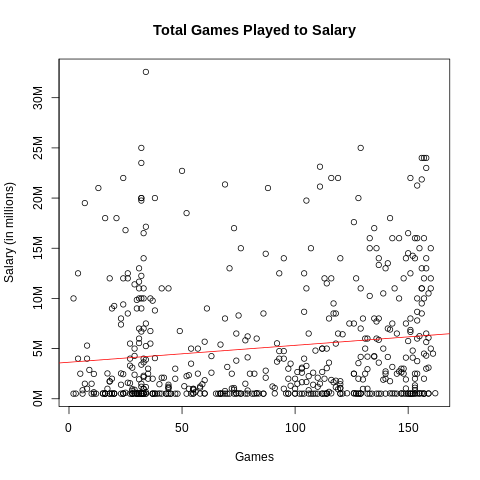
From general managers of professional sports franchises to general managers of a fantasy baseball team played amongst 10 to 12 friends, multivariate analysis can assist in identifying trends and patterns that can uncover value or competitive advantages. Anyone can spot a good player but through team building the strength of your supporting cast is what makes the most difference.

This project is based around doing multivariate analysis on baseball statistics. We used a publicly available dataset on Kaggle, involving baseball players names, various raw counting batting statistics, and salary information. Each observation (row) corresponds to a unique player in the MLB in the 2015 season, and each variable (column) corresponds to one of the various statistics recorded for each player. Our goal is to analyze the pay of athletes, and see if salary can be predicted or understood by various batting statistics. We also seek to understand if a player may be overvalued or undervalued based on their batting statistics relative to their salary. All of these statistical explorations have motivations in a better understanding of the game of baseball, and could even possibly be applied to more efficiently manage baseball teams. For example, if a rookie player is found to be clustered with many highly paid, established players, it may be beneficial for a team to trade for the player while their salary is still relatively low.

**Data Cleaning and Visualization**

We decided to use data from 2015, as it is the most current and impactful on modern salaries. We also removed data with 3 or less at bats. This is to remove any outliers due to low data. By the nature of our merge, we only retain players with recorded salaries. This is acceptable, because we plan on analyzing the salaries anyway. Additionally, the batting dataset required some cleaning before merging, as playerIDs (our planned column for join) were not unique due to mid-season trades. We account for this by summing statistics between the teams, as we care about the season statistics for the player without regard for their team. The final step in data cleaning is to replace the abbreviated column names with understandable column names, which will help analysis in the future. See appendix for additional detail on data cleaning specifics.

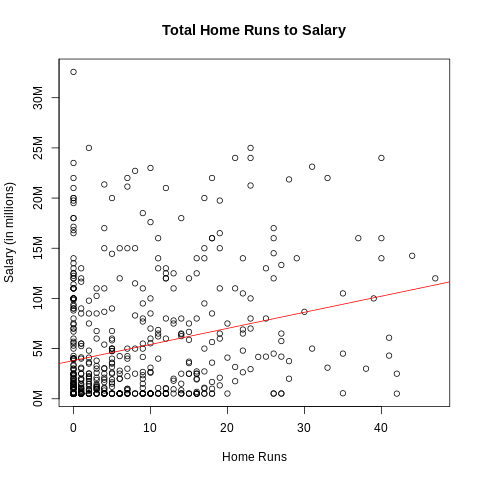
To gather a preliminary baseline, we gathered a subset of preconceived valuable metrics that were present in our dataset and plotted them against each player's salary to see if any variables presented interesting insight or unexpected correlation(s) going into analysis.



Graph 1 (Total Games Played to Salary): See Appendix section 1.1



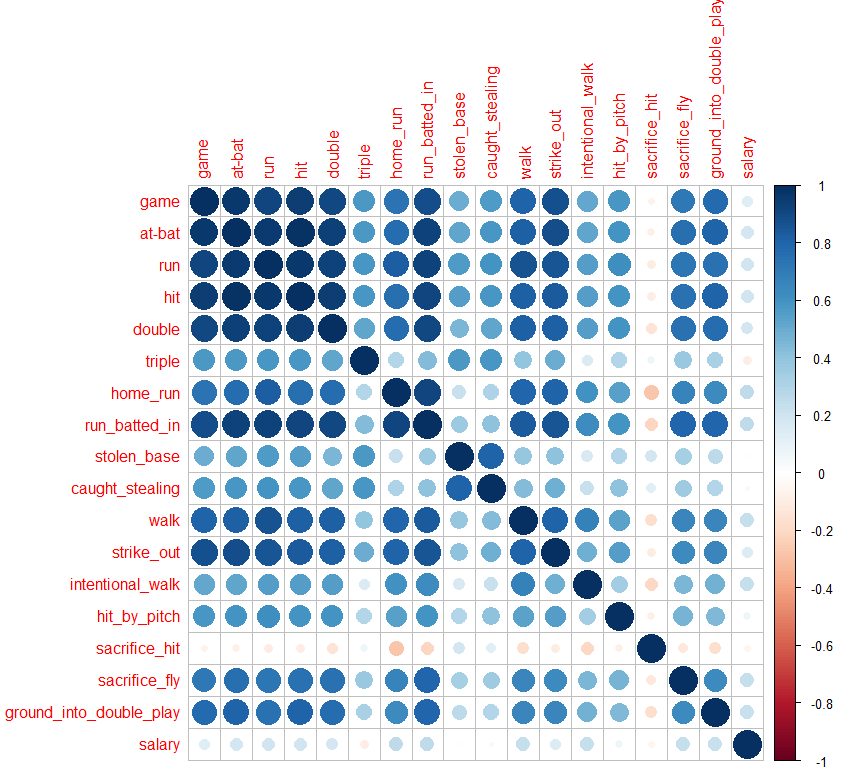
Graph 2 (Total Hits to Salary): See Appendix section 1.2



Graph 3 (Total Home Runs to Salary): See Appendix section 1.3



Graph 4 (Total RBI to Salary): See Appendix section 1.4

****

Graph 5 (Correlation Heatmap): See Appendix section 1.5

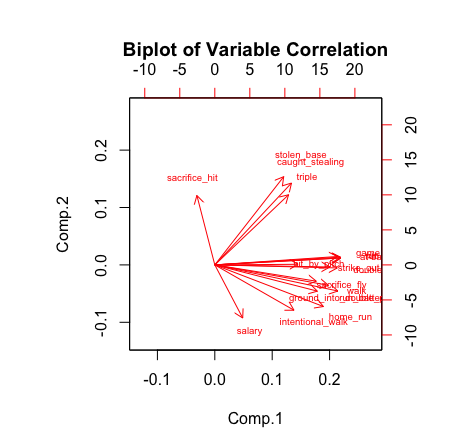
**Dimensionality Reduction**

In this section, we seek to reduce the dimensionality of our data in order to gain a better understanding of how our variables interact, and which variables introduce the most variability in our dataset.

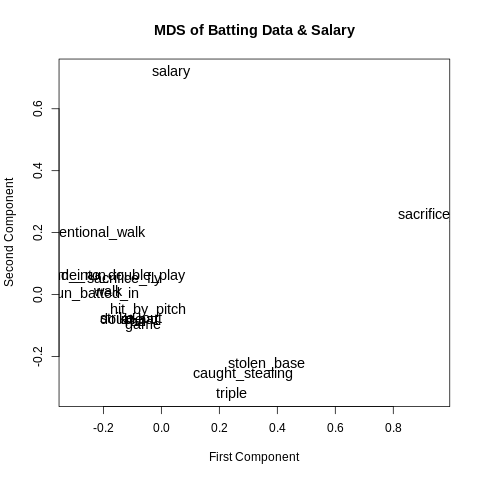
With 18 variables available in the dataset, dimension reduction is important to gather an understanding of which variables can be grouped together to develop insights about the data. Types of hits, for example, like single, double, etc. can present noise regarding the overall number of hits since the two are inherently related. Reduction eliminates some redundancy identifying relevant variables.

The first four components in PCA account for approximately 81% of the variance in the dataset. (See Appendix section 2: Table 1 for loadings)

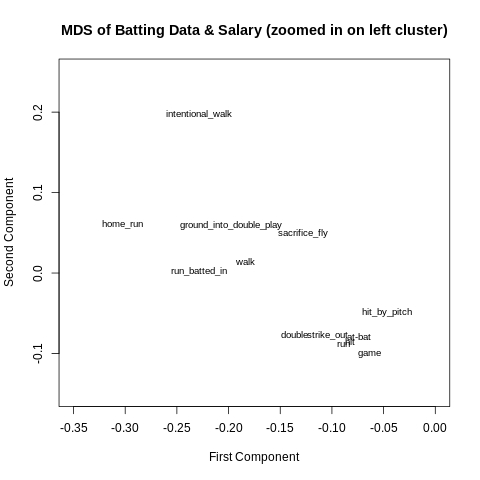
The first principal component covers the variance that exists in all hitting categories - it encompasses almost all types of players from power, to speed, to average hitters. The second component indicates a faster player. Triples and stolen bases are often attributed to speed archetypes, where power metrics often relate to players that aren't as fast. The loading values indicate high relative values to speed, such as triples, steals, and caught stealing (more attempts) and a contrasting value with power that's often not associated with a player that plays with speed. The third component seems to be a player with fewer at bats than an every day starter but enough to register relevant statistics, like a prevalent pinch hitter. Pinch hitters typically aren't utilized to bunt, hence the contrasting value of sacrifice hits. The fourth component appears to be pitchers that hit in the National League, albeit something no longer relevant as the designated hitter became universal in 2022. Given pitchers don't hit but once every 5 days, they're often used to bunt (sacrifice hit) and this component has a high loadings value in that category to support. It's also rare a pitcher gets on base explaining the contrasting values for steals categories.



Graph 6 (Biplot of Variable Correlation): See Appendix section 2.1



Graph 7 (MDS of Battling Data & Salary): See Appendix section 2.2

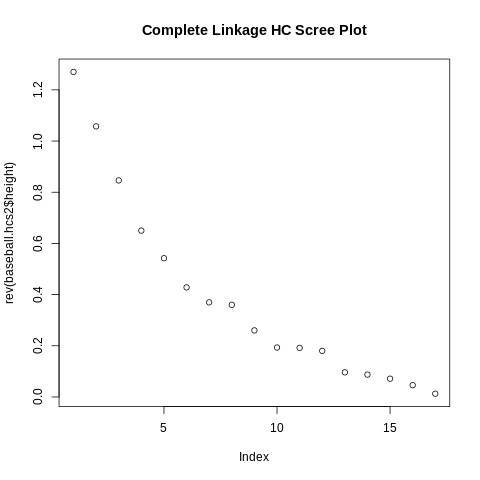


Graph 8 (MDS of Battling Data & Salary (Zoomed in on left cluster)): See Appendix section 2.3

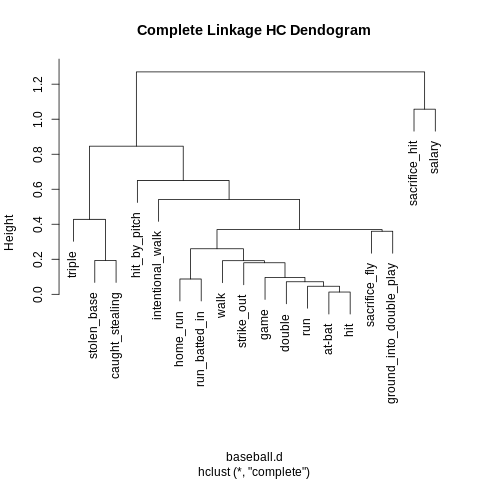
We can see various clusters of statistics in this multidimensional scaling. In the zoomed out graph, we see salary and sacrifice hit are distanced from other clusters. We can see the bottom cluster has stolen base, caught stealing and triples. The stealing statistics being close together makes sense, as more steal attempts leads to more successful and unsuccessful steals. In the zoomed in graph we have two clusters. The bottom right cluster is made up of many of the variables with high correlations in the correlation matrix. The left cluster has both positive statistics (like home runs) and negative statistics (like ground into double play). We will have to look into these associates further in the report. We explore the associations between variables seen here further in cluster analysis.

**Cluster Analysis**

In this section we create data clusters with the various algorithms we learned over the course of this class. The motivation behind clustering the data is to see if we are able to find and understand outliers within the clusters, or to see if clusters form at all. If clusters do form, they could give us insight into archetypes of batters, if such a thing exists.



Graph 9 (Complete Linkage HC Scree Plot): See Appendix section 3.1



Graph 10 (Complete Linkage HC Dendogram): See Appendix section 3.2

We utilized k-means clustering in an attempt to support the interpretation of PCA and hierarchical clustering. We encountered overlap in hierarchical clustering so the k-means allowed us to provide some additional context to the scaled data for further analysis.

In an effort to classify characteristics of each cluster, we utilized averaging each scaled variable per cluster to interpret relationships that could assist in determining what type of player fit into each cluster. (See Appendix section 3: Table 1 for cluster totals)

Cluster 1 - Replacement Level Players: 146 players. Almost every variable has a z-score very close to 0. This cluster could be classified as the average, every day player. The group averages indicate that there is no exceptional category for these players. They gather enough playing time to accumulate an average number of statistics but aren’t in any way valuable above another replacement level player.

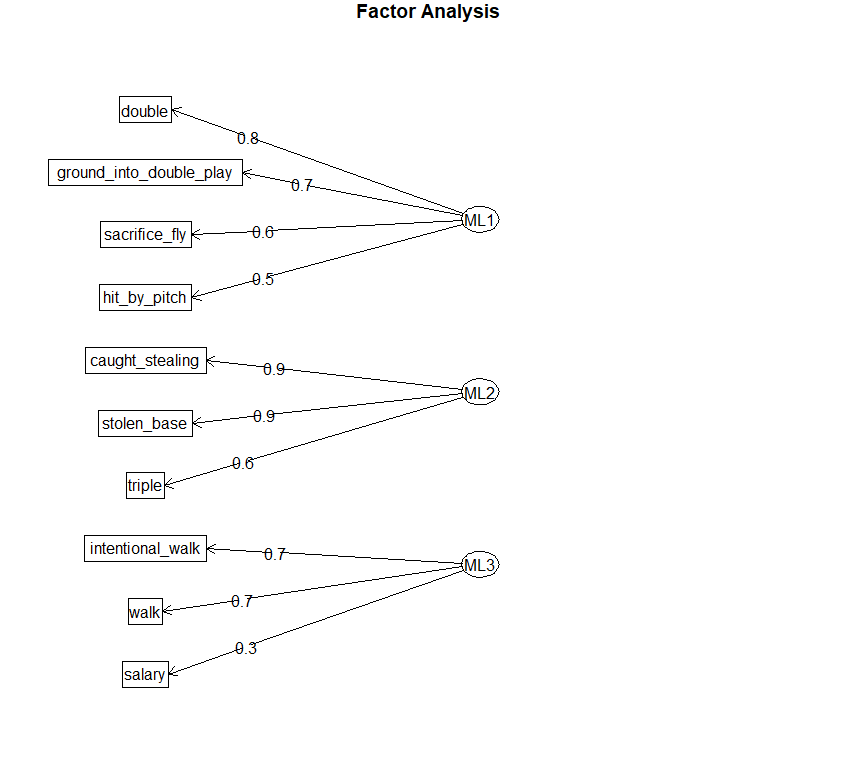
Cluster 2 - Benchwarmers: 228 players. The value that stands out is that this group is 1.02 standard deviations below the mean for games played. Subsequently, all other categories fall below the mean as each variable is a cumulative measure. Every single variable’s average is below the mean besides sacrifice hits. This grouping seems to be filled with bench players that received spot starts, pinch hit at-bats, or players that were qualified due to the number of at-bats but got injured at some point.

Cluster 3 - Table Setters: 67 players. This group has z-scores above 1 standard deviation for several valuable categories, among them stolen bases (1.96), triples (1.52), runs (1.16), and hits (1.22). This group could be categorized as speed or contact players - players that don’t typically hit for a lot of power but are considered “table setters,” or players that get on base for other run producing players to drive in. Notably, this group is slightly below average for salary (-0.16) so depending on how an organization is building a team (small ball, run manufacturing) this is a category where metrics can be used to identify value.

Cluster 4 - Run Producers: 90 players. With average z-scores of 1.62 in homers, 1.60 in RBI, 1.35 in doubles, and 1.36 in intentional walks, this cluster seems to clearly be the run producing or power driven players. Given the close to average values in stolen bases and triples this cluster doesn’t play with a lot of speed. Other values to note, 1.23 strike outs and 1.31 sacrifice fly also indicate an all or nothing group of hitters. The salary z-score is 0.85, meaning the league will typically pay above average price for this archetype of hitter.

**Confirmatory Factor Analysis**

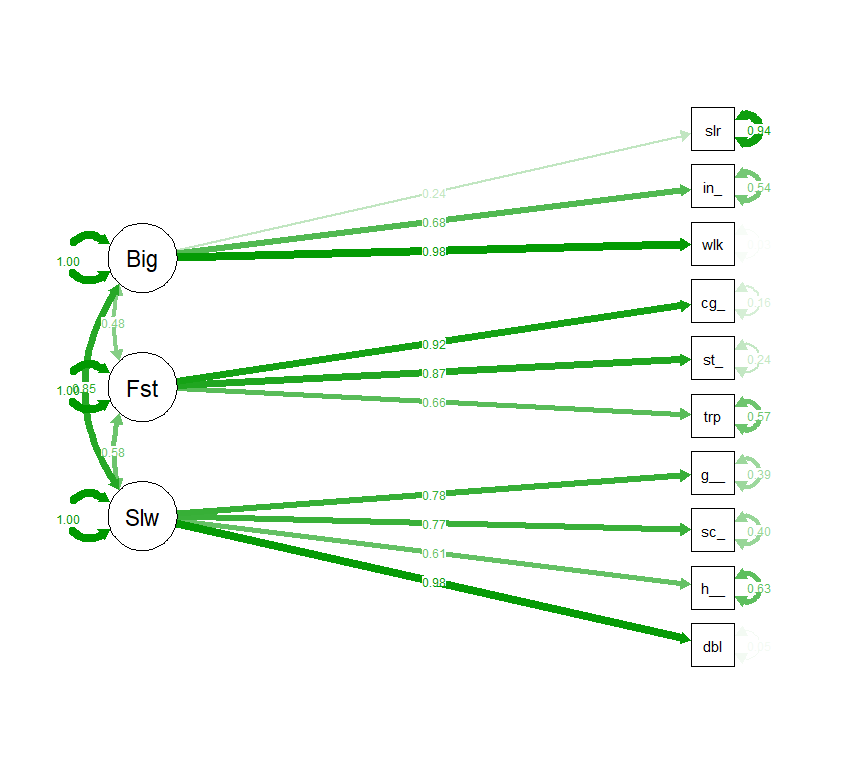
In this section, we first do an exploratory factor analysis (EFA). This helps assist us in understanding the relations between variables. We then create a model with latent variables, in an attempt to verify the integrity of the model generated by our EFA model. We chose to remove columns in order to optimize our p-value and RMSR values, as to ensure a well-fitted model. This is mostly due to the fact that several of the original columns in the dataset are too highly correlated to be included in an effective EFA.



Graph 11 (Factor Analysis): See Appendix section 4.1

The above model describes our three factors that are composed of our manifest variables. We use these groupings to form our three latent variables. We chose to characterize the first factor as slow/utility players, the second as fast players, and the final factor as big hitters. (See Appendix section 4: Table 1 for EFA loadings)

Our original exploratory factor analysis gave us interpretable factors that aligned with our previous understanding of the dataset. Knowing this, we chose to use the model to continue into CFA.



Graph 12: See Appendix section 4.2

After enacting the CFA, the model is shown to have acceptable goodness of fit indices (GFI = 0.94, AGFI = 0.89 and RMSR = 0.06). Note that the estimated variance of the disturbance for the variable "walk" is very low, and we fail to show significance with a p-value of 0.4 for this parameter. However, every other disturbance term passes a 0.05 significance test. (see appendix)

Considering the latent variables, we can reach some interesting conclusions. Salary being included in the latent variable with home runs and intentional walks makes sense, as star players are more likely to hit home runs, and are more likely to be intentionally walked out of fear of a strong performance at the plate. Hit by pitch is also included in this group. Two possible hypotheses are that pitchers may intentionally hit better batters, similar to an intentional walk. It is also possible that the star/stronger batters may be more likely to "crowd" the plate, getting nearer to the strike zone for better positioning on outside pitches, and increasing the chances of a walk by being hit by a pitch. Finally, walk is also included in this group, but with a low p-value and similar loadings across all three factors in EFA, it is likely not worth considering too heavily.

The latent variable we describe as "fast", encompasses manifest variables caught stealing, stolen base, and triple. Stolen base and triple make sense to be associated, as they both require speed and base running knowledge. Caught stealing might seem counterintuitive for a fast player, but it is important to note that in baseball, slower players are unlikely to even attempt to steal a base. Thus, faster players having more steal attempts means they inevitably will have more times caught stealing. This connection is interesting to see here, as it is the same grouping as found in our complete linkage hierarchical clustering earlier.

The final latent variable is harder to directly characterize. In the model I describe it as "slow", but the variable may better be described as utility, or simply weaker batters. There are three manifest variables comprising this latent variable (double, ground into double play, and sacrifice fly). Ground into double play is a negative hitting statistics, where the batter's weak hit causes two outs for their team. A sacrifice fly is where a batter will, either intentionally or unintentionally, hit a pop fly and get out while allowing an on-base runner to score. This statistic is generally positive, although many stronger athletes may not be asked by a coach to hit a sacrifice fly, as they may have confidence in them to get on base and score the RBI without getting out. These three statistics may indicate that the batter is not as strong.

**Conclusion**

In any statistical analysis, there will be some difficulties encountered. This project was no different. One example of this in our project is how the number of at-bats/games played is by far the biggest influence on the amount of every other statistic. This leads to difficulties in the analysis. For example, take the variable ground into double play (GIDP). You would expect this statistic to be a powerful way to identify weaker hitters, or at least have a negative correlation with something like home runs. Instead, we see a positive correlation between GIDP and home runs, hits, and at-bats. As a real world example, Albert Pujols, one of the greatest hitters of all time, also has by far the greatest number of GIDPs. The positive correlations between one statistic considered positive and one statistic considered negative make it difficult to analyze batters effectively (at least with our methodology).

For future endeavors we may be better served performing analysis on ratios of certain statistics that might give a better indicator of performance or examining fielding and pitching in relation to get a more comprehensive view of a player’s contributions. For example, a better hitter with more at-bats would almost certainly have more strikeouts, but more homeruns, than a replacement level player that didn’t play everyday. It would be beneficial to see if there was a stronger indicator with relation to salary for calculated values or ratios like Slugging Percentage, K/BB or K/HR ratio, clutch hitting metrics, or defensive runs saved. That would better differentiate a good hitter from a bad one than raw counting statistics, or give context to a bad hitter being more valuable as a good defender than a good hitter that’s a liability defensively. Additionally, analysis on year over year performance could provide insight into the direction the league as a whole is headed to get a head start on where to identify value. If homeruns are down and steals are up year over year an organization can look to counteract the market when it comes to pricing of players that fit the archetypes at the time. A good example is the emphasis currently on home runs and a disregard for strikeouts but with rule changes implemented teams may sway back to a speed and contact based approach that hasn’t been present in baseball for years.

Despite the difficulties, we were able to gain some of the desired knowledge from our data and open avenues to explore more advanced data that either further supports our preliminary findings or uncovers new insights.

**Appendix: see attached RMarkdown compiled Docx file**

**Citations**

Verducci, Tom. “Rays’ Success in Fixing Pitchers Is the Envy of the Rest of MLB ...” *Sports Illustrated*, August 8, 2023, www.si.com/mlb/2023/08/08/rays-pitching-development-kyle-snyder.