Christian Geer

Professor Bonham-Carter

CS 390: Data Analytics

December 11, 2017

For the final project for this class I have elected to research and analyze various questions regarding statistics in baseball. This was done that to the Lahman Baseball Database that was created by Sean Lahman and allows easy access to years of baseball statistics for the public. My personal motivation for selecting baseball statistics for my final project is a reflection of my interest in the game. Growing up baseball was, and still is, my favorite sport, and as I’ve grown with the sport so has my interest not just in the game, but also the stats that come along with it. My interest in baseball statistics reaches further than the standard batting average and home runs, for as I’ve grown older I have developed an interest in more advanced sabermetric analysis. While this type of statistical analysis is beyond the scope of this class, and my current ability and understanding, it was certainly still a motivation for me to briefly my own type of analysis on baseball statistics.

The first question that I asked for this project was which stats tended to correlate more strongly with a player’s election to the Hall of Fame. My motivation for looking in to this was that each year when the Hall of Fame elections come around it seems that everyone has a different opinion on why someone should or should not be elected to the Hall of Fame, something that has increased in recent years as sabermetrics have become more popular and ‘old school’ analyst continue to resent them in favor or the standard stats and the ‘eye test’. Because of this I wanted to see if there were any stats that tended to be more prevalent among those players that got elected to the Hall of Fame.

Before I began my analysis, I hypothesized that batting average would be the top statistic. This was because it the most widely used and understood statistics in baseball, even if you are not a baseball player or a fan of the game, there is a pretty good chance that you know what batting average is. It is also an intuitive that is easy to understand and figure out if you are not familiar with the game. However, batting average was not the statistic with the highest correlation, but rather it was runs scored with .31 – just above the .28 of batting average. The others stats that were in the top five were hits (.28), on base percentage (.22), and runs batted in (.22). Along with batting average not being first, another surprise in the results was that home runs were not in the top five, in fact, home runs were the 12th lowest correlation out of the 12 statistics measured. This was surprising because it always seems that home runs are valued so highly in the sport of baseball that they would hold more weight when it comes to election to the hall of fame. An explanation for this low correlation could be that, generally, the players that hit the most homeruns don’t do much else effectively like hit for average, and this make it hard for them to get elected.

As with most everything there are some inherent issues to the analysis that I did. One major issue with the analysis that I did is that it only accounts for batting statistics, yet you cannot truly and fairly evaluate a player without taking into account their ability in the field and in other parts of the game. Unfortunately, this is not somethings that I was able to do with the database that I was using, and is also something that would have been hard to implement with the time that I had to do it. Something that I would have like to try and add is analysis of how the correlations have changed over time. This would be interesting to see, as there have been multiple changes in the sport as far as how its player and what is valued. Another thing that I would like add is the ability to predict a player’s chance of getting into the Hall of Fame based on their stats and correlations that I found in the data.

My second question similar: what stats correlate with All-Star selections? My main motivation for researching this was to be able to compare the results to the results of the first question. This was interesting to me, as for most of the years in the dataset All-Star selections were decided by the fans and not coaches or sports writers. Because of this I thought that it might to interesting to see what stats the fans valued more or less that the Hall of Fame voters. Another thing that might affect the results is the fact that All-Star selections are based on half of a season, not a whole career, so there may be more stats that make a bigger short-term impact on fans, and would get more value for All-Star selections than Hall of Fame votes.

In my analysis of this question I found that runs score was again the top correlation (.54), with these others in the top five: runs batted in (.53), hits (.52), extra base hits (.50), and walks (.49). For this question I also believed that batting average would be the first, but ended up not even being in the top five. It seems that is a preference towards cumulative stats rather that stats of averages for All-Star selections. This could mean that seeing big number of stats such as runs scored, runs batted in, and home runs, which moved up to 6th from 12th. Something that I was very surprised about in these results was how high up walks were. Walks are a stat that are not at all glamourous and aren’t generally what fans think about when talking about the best players in the game.

Similar to the first question I would have liked to be able to see the trends in these correlations may have changed over time, and subsequently compared that to the changes in the Hall of Fame voting, to see if they have gone through similar to different changes as the game has evolved. I would also like to implement the ability to predict All-Star selections based on a player’s stats through the first part of the year.

The third question I set out to answer was who the most under and overpaid players are in the data set. A lot of times people talk about whether or not a player is playing up to his contract, and whether that player is worth the money that he is being paid by a team, and this was my main motivation for looking into this question. I wanted to see if there were any players that were/have been vastly over or under paid in their careers.

To answer this question, I used the players career total of runs scored. I chose this stat because it was the stat that had the highest correlation in both the Hall of Fame voting and All-Star voting analyses. What I found in the data was that Pete Rose was extremely underpaid for the production that he provided over his playing career. Despite the highest runs scored total in the data, Rose wasn’t even paid an average salary. One explanation for this is inflation, as Rose played in the 70s and the data includes players with salary numbers up to 2016, so players being paid what is really equivalent or less than Rose will appear to be paid more.

The major thing that I would like to change for this analysis is the effect that inflation has on the results. Unfortunately, there is no way to standardize every single year without going through and doing so for each and every year that is in the data table, which would take forever. To get around this I would most likely instead group every 5-10 years together and then do the calculations, and also probably use, for example, data back to 1950, so that there wouldn’t be an absurd amount of calculations to do. I also would’ve liked to figure out a better way to judge a player’s value than simply their runs scored, as this is going to hurt players that are not that strong in this area, but with the data that was available to me I would have had to pick something, so no matter what some players were going to get hurt by this effect.

The fourth question is similar to the last: what were the best teams for their salary? My motivation for this question was to see if there were any teams that were able to defeat the logic that spending more money on your roster is the best way to win. Throughout the history of the game the amount of money that a team has has always had a major influence on their ability to win championships. This is evident in the fact that the New York Yankees, always one of the wealthiest teams in the league along with being one of the oldest, have the most championship out of any franchise by a long shot.

The results of the analysis showed that there were in fact two teams that were able to succeed without giant payrolls. Both of these teams were surprises to see up there, for the one, the 2001 Seattle Mariners, hold the record for the most wins in a season. While the team with the biggest salary might not always win, you expect that a team that was able to accomplish something so incredible would have a lot of very good, established players, who therefore would be paid very much. The other team that was able to succeed despite a lower payroll was the 1998 New York Yankees. This is very surprising to see because the Yankees are not a team that you think of as not having a very large payroll; however, one explanation for this could be the same as what happened to Pete Rose in the last question – inflation. Due to the effect of inflation both of these teams could appear to have lower payrolls for the season, but in actuality they were still very high for the years that they were played in.

As was the case before, the major thing that I would like to change about my analysis is the lack of correction for inflation. There is no doubt that the effect of inflation is skewing the results, but with the time that was available I was not able to account for this. For this question I see two ways that I could attack the issue of inflation. First, I could do something similar to what I had proposed before, and group the years into segment of every 5-10, and do approximate calculations like that. Another option is to rank each team by how their payroll compares to the others teams in the league in the that year. I believe both of these would fix the issue just as well as the other, it just be a matter of figuring out which is easier to implement.