Christian Geer

Professor Bonham-Carter

CS 390: Data Analytics

November 14, 2017

Progress Report

Baseball statistics have been a part of the game for what almost feels like forever. The beginnings of baseball statistics trace back to a man named Henry Chadwick. Chadwick used his familiarity with cricket to develop basic statistics for baseball, some of which, such as batting average and runs scored, are still used today. It wasn’t until 1951 when Hy Turkin published *The Complete Encyclopedia of Baseball.* In the modern age, baseball statistics have taken a drastic turn from the rudimentary stats that had been used for so long. In recent years, aided by technological advancement, a new field called sabermetrics has emerged. Sabermetrics is a branch of statistics that uses far more advanced, and sometimes confusing, statistics that have been developed. Recently, sabermatricians have begun incorporating new data that has been captured by on field cameras, such as stat cast, giving them access to data never before available. Unfortunately, the database that I have selected only contains basic statistics, and due to various constraints, I will not be able to implement and examine the more advanced stats used in sabermetrics. The Lahman baseball database was developed and compiled by award-winning database journalist Sean Lahman. The Lahman baseball database is one of the first efforts to make a database of baseball statistics available online to the public. The database contains statistics from 1871-2016 for many professional baseball leagues within that time span. To analyze this data, I will be compiling the data into career stats for all the players, and then using that to analyze how various stats affect a player likelihood to be named to the Hall of Fame, an All-Star team, or be selected for another individual award. My analysis will mostly utilize the tidyverse r package to manipulate and plot the data, along with the dplyr and psych packages.

The biggest challenge that I have faced so far during this project is the creation of the career statistics table. In fact, at one point I had given up and intended on changing the direction of my project, as I did not think that I was going to be able to create the table; however, after making the decision I accidently stumbled upon a way to create the table, and I returned to the original project path. The major function that allowed me to do this was ‘group\_by’, which allowed me to group the observations by the ‘playerID’ variable. Another challenge that I have run into is keeping track of, and renaming, the variable names in the tables. This has been an issue sometimes because some of the functions that I am using rename the columns after they have been used. I then have to go back and rename them to what they should be so that the code works, and so that the variable names make sense. This has mostly been a nuisance more than a major issue, for it is something simple, but takes up, at the moment, 37.5% of the code that makes the career table. If I have time, I would like to research a quicker, easier way to rename the columns, but for now I will have to leave it as be, and deal with the long, tedious way I’m doing it.

Everything that I have learned so far in this project stems from the creation of the career statistics table. The major thing so far has been ‘group\_by’. This is what actually made the creation of the table possible. This allowed me sum up all of the players career statistics, and keep them in the table. Before, I believed that I was going to have to extract all of each player’s stats from each year, and then sum that up, and join all of those tables back together. Doing that with stats from over 100,000 induvial player’s seasons would have been entirely impossible with the time I have, and a poor use of time regardless. It was because of this that I had believed that I was going to have to change the direction of my project; however, thanks to ‘group\_by’, I was able to create the table and continue with what I had originally planned on doing. A couple of others things that I have picked up thus far are the use ‘summarise’, specially ‘sum()’ within it, and ‘round()’. Using ‘sum()’ within ‘summarise()’ is what allowed me to actually sum up all of the players stats after I grouped them. Round is also very important as well, for I had to calculate stats such as batting average, on base percentage, slugging percentage, etc. After calculating these, they were outputting with many decimal points, which are unnecessary for my purpose, as only 2 or 3 are needed for baseball statistics. Being able to round them allowed me to convert them to the standard format for statistics.