Zachary Allen

5/7/2018

Final Project Write Up

My statistic is used to measure the efficiency of players based on time in the league and measured against other players during a similar time period. The statistic looks at players who played between the 2005 and 2017 who played for at least 5 years. My code allows this range to be changed with selection of start year, end year, and needed duration for eligibility. From these inputs, the code returns the necessary mySQL query to be used to grab the debut and finale dates of the players within the range.

This was where I ran into my first major difficulty with mySQL. The debut and finale dates were given in string format so, I tried to convert the first for numbers (the year) into an unsigned integer. This worked well for single select queries but, when I tried to store these in their own table, I ran into a can’t convert “” to integer error. I thought it had something to do with the conversion and after a couple hours of troubleshooting, I realized that some of the debut or finale dates were left empty which prevented the code from converting blanks into integer values. Once I finally figured this out, I could apply most of the code I had developed for the first submission.

The next thing I did was developed code that would rank the players based on decided weights of their RBI’s, 2B, 3B, HR, CS, and GIDP over the years they played. I required the batters have at least 200 AB’s for the year they played. This value is adjustable in the code. This provided a baseline to analyze the batters relative to each other. Once again, I ran into a problem with the mySQL. When I ran these mySQL scripts, I decided to run them and save them for each year individually. This was the easiest way I found to do this query. However, when I went to save I ran into multiple issues. First off, the --secure-file-priv setting needed to be reset to “” and the computer rebooted. Then, the file location needed to be chosen. So, I devised a way to find the current path and add a directory data to hold all the Excel files. This would have worked perfectly and made creating the data super easy. Unfortunately, choosing your own path for mySQL isn’t accepted. They require you to use a temporary location which seemed like more work to figure out than was necessary. Therefore, I just performed that data transfer to excel files manually. This was a painful process and made my code rather rigid in the end if that was the only way to make alterations to the AB’s required to be considered a viable candidate.

The advantage of my code is, if you can get past these lame security restrictions, my code pumps out all the necessary mySQL queries to recreate my entire experiment for any values you choose. In addition, once the data has been created, the python script allows the weights of the things considered important to change to the user’s own idea of what would work best, making it a malleable system that can be adapted and used for many people’s ideas and theories.

I think this is a good statistic because of my weighting scheme. I was focused mostly on putting your team in a better position to either score runs or have runners in scoring positions. So, I started with RBI at a +.4 weight. This was my highest weight. I figured that this statistic was better than hitting either a 2B or a 3B because those don’t guarantee runs. I weighed it higher than HRs because the RBI would be double counted in many HR scenarios. Next, the 2B and 3B at +.3 because these put runners in scoring position. Finally, on the positive side, HRs at +.3. Next, I considered negative values for caught stealing (CS) of -.2 because it often cost the offensive team an out and eats into the ability to score. After that, ground into double play (GIDP) at -.5. I considered this the worst statistic on the offensive side of the ball. This player needed to be penalized for being responsible for two outs during the same play.

My code allows the user to view this data as plots for each individual player and the next step would be to perform more statistical analysis on the data that has been received and regularizing so that players who played different amounts of years could be treated equally or at least analyzed on seperate spectrums.

This statistic is similar to Bill James’ approach to characterize players with relevance to other hall of fame players. Except, my program allows you to perform this analysis over more years which could be used to help understand why these players performed so successfully during certain periods of their career. Additionally, PECOTA performed a similar statistic to this to understand when a player transitioned from a young player to old player or why their careers lasted as long as they did.

My statistic fills the need of making the process of creating the data from Lahman’s and transferring it over to a python environment extremely easy and extremely manipulatable. This type of code is difficult to come by and the ease of use of my code is an incredible advantage to people looking to get into the data analysis and by-pass the work of having to understand the complexities of dealing with mySQL and making very adaptable code that can be changed and rerun with little to no errors. However, my statistic doesn’t include any original statistics so that would be a really good way to improve upon my statistic. The ease of doing this would be very simple and would only require changing one section of the python code to be able to produce all this data in one continuous executable. Additionally, adding this new statistic and weighting it into the function wouldn’t be too difficult.

This code can also be used to create and evaluate pitchers with minimal changes to the code. I think the same script for creating the viable players in a time range could be reused and then the selected database would transition over to the pitcher statistics. My idea for weightings used in both starting and relief pitchers are included in the mySQL script but, I abandoned these analysis when I hit so many setbacks with creating new tables and being able to store the data to excel files without having to do this by hand. That essentially eliminated my ability to quickly process the data and would have caused me extensive time to make data sets for pitchers. Currently the 2005-2017 excel data is strictly for batting.