**MOTIVATION**

I originally started thinking about this kind of data after watching the movie Moneyball with Brad Pitt. This got me interested in the myriad of sports statistics that are readily available for the public, especially surrounding Baseball. I had always wondered whether or not an experienced athlete is better than a rookie, and my hypothesis was yes, probably. So, focusing on the sport of baseball, I compared the batting averages of any given player as they progressed through their career from their rookie season to the end of their career.

**DATA SOURCES**

I obtained the batting record data from Sean Lahman’s MLB database, which contains statistical data for each player dating back to 1871. The dataset that I used, which was the batting records for these players contains over 107,000 rows of information about each player’s performance per year, per team, including at bats and hits which were the most important pieces of information to me.

**PROCESSING STEPS**

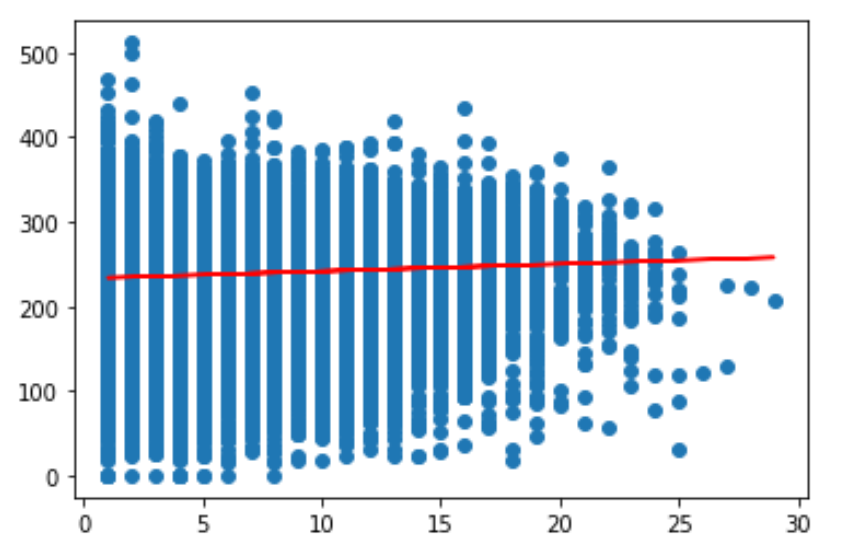
All of my data processing was done in Microsoft Excel. After downloading Lahman’s batting table, I created a SEASON# column which tracked the season that a player was currently in within their career by counting duplicates of the PLAYERID column only when the STINT column was equal to 1, meaning it was their first stint of the year. I went on to create the BATTINGAVG column which tracked a player’s batting average for any given stint from 0 to 1000 by dividing their H (hits) over their AB (at bats). Now came the process of cleaning this data, for which I removed all players who had no at bats for any given stint, in addition to removing players with less than 30 at bats in order to satisfy the statistical law of large numbers in obtaining an adequate sample size. Finally, I removed all entries from before 1970 in order to reduce the sample size to a more relevant group, and a more manageable scale. Lastly, I removed all unnecessary columns of information not relevant to my core investigative question, and exported the spreadsheet into a comma separated value list.

Later, I made the second datasheet titled CareerStats, this datasheet was built upon the BattingStats sheet I made previously, but was designed to answer a different investigative question: “Do Better Baseball Players Play Longer?”. So, I created a new column titled AVGBATAVG, which tracked the average career batting average of any given player. Furthermore, I calculated the final season for any given player in a new column titled MAXSEASON. Finally, I removed duplicate entries for the PLAYERID rows, and this left me with only the name, final season, and career batting average for every player since 1970.

(All of the processing steps have been outlined in the attached videos)

**VISUALIZATION**

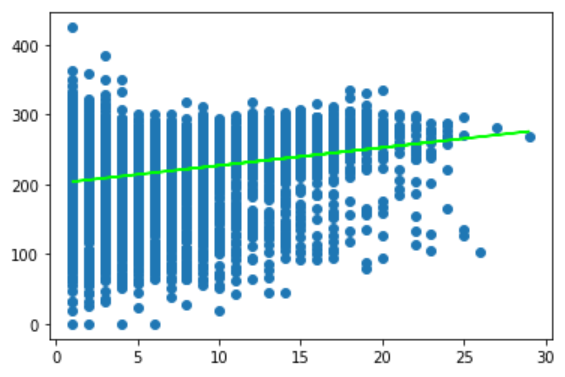
Once I had processed the data, I needed to chart it out into a more useable format, so I exported the .csv files to python and created the following graph:



This graph charts all players’ batting averages year over year with the Y axis being a player’s batting avg. from 0-1000, where the X axis charts what season the players are on in their career.

The red line of best fit charts the average player’s batting average over the course of their career. A description of this line would be Y=0.85614932X + 232.99199903

After doing some analysis, I decided to create the following graph as well:



This graph charts all players’ career batting averages, and the year which they stopped playing in the MLB. The X axis represents the final year which the player played in the MLB, and the Y axis represents the player’s career batting average.

The green line of best heat charts the average player’s predicted career batting average based off of when they drop out of the league. This line can be described as Y=2.57070625X + 201.35316972.

**ANALYSIS**

After exporting the spreadsheet into Python and discovering that a model of the data showed that players do not get any better the longer they play with the MLB, this raised a new question in my mind which was “Do Better Players Play Longer?” The reason that the question came to my mind is that by showing that better players play longer, while players do not get any better from playing longer, there would be at least some evidence to say that baseball is an entirely skill-based sport rather than one that can be improved upon based off of hard work alone. So I plotted a dot plot of my first spreadsheet which tracked players batting averages season over season and found that a large portion of players with low batting averages seem to drop off at about their 15th season. After creating this second spreadsheet, I exported that back into Python and created a second. Plot that charts all players' career batting averages in the year that they stop playing in the MLB after using a linear regression on the data, I created a line of best fit which charted the average players’ predicted career batting average based off of when they drop out of the league. This line of best fit showed that players who spend more time in the league tend to have a higher batting average career-wise than players who only stay in the league for a relatively short amount of time.

**DESCRIPTIONS OF MATERIALS:**

This Repository contains the following folders:

NOTEBOOKS: All Jupyter notebooks used for analyzing and visualizing the data

PROCESSED DATA: Final, cleaned spreadsheets and csv files

RAW DATA: Lahman’s batting csv used for this project

DATA CLEANING: Links To Walkthroughs on all data cleaning processes for this project

VISUALIZATIONS: Visualizations and graphs from the Jupyter Notebooks within this project.

**SOURCES:**

Lahman, Sean. “Baseball Archive.” SeanLahman.com, 12 June 2018,

www.seanlahman.com/baseball-archive/.

Acquired the 2019 updated batting statistical data.