Visual Exploration of Sean Lahman's Baseball Database

Matthew Coker CU- Boulder

Matthew Sredojevic CU- Boulder

Taylor Gunter CU- Boulder

Tom Slesinger CU- Boulder

Abstract

Sean Lahman's baseball database was used as the source data for this project. There are four visualizations included in this project. Visualization #1 is two visualizations where the first compare the best batters to the average batters and the second compares the best pitchers to the average pitchers. Both of these visualizations are shown in terms of average and best per decade. Visualization #2 looks at the average amount of homeruns per game for each stadium since 1871 from the Parks.csv and Teams.csv files. We use *Carto* to visualize this data with a world map. Visualization #3 is a scatter plot of salary of the two leagues (American and National League) over time. Visualization #4 looks at the difference between Hall Of Fame players and rest of the league in twelve statistical categories broken out by batting and pitching, organized by year ranging from the late 1800s to early 2010s.

Introduction

This paper is looking at baseball data to try and better understand how players and statistics changed over time. We find that this is important since baseball is a growing sport throughout the world. Data analysis can help find trends in certain aspects such as salary or pitching and hitting statistics. Finding insights into these attributes can allow the average person to explore how baseball has changed over time and possibly predict where it is heading. The overall problem we found is that baseball does not have many captivating visualizations being built around this. Overall, our paper performs an exploratory analysis of many attributes of the our dataset which is based off of the Lahman Baseball Database. We've performed a holistic series of analysis and visualizations to show how baseball data can be visualized, and the benefits associated with that.

Related Work

Previous Work Related to Visualization #1

The first part of the research needed to be looked at was to determine what was the best way to determine what statistics best describes a player and most accurately shows the proficiency of a player. The first source for accuracy we chose was by Adam Houser, titled 'Which Baseball Statistic Is the Most Important When determining Team Success?' The part I was more curious about was the pitching stat because it is harder to judge a pitcher from their stats than it is a batter. The research focused on WHIP (Walks plus Hits per Innings Pitched) and K/9 (strikeouts per 9 innings pitched). The research states 'ERA is not a good measure because of the fact it is so subject to errors and WHIP is not.' The next question was what is the best batting statistic. Most everyone knows about a batting average, a batters likelyhood to get a base hit for any given at bat. There's another level up, Weighted On-Base Average (wOBA). This accounts for all the different bases a batter can get from a hit.

Where Batting average only considers if a batter got on base, ‘wOBA values certain types of hits according to how much they actually matter’ as stated by Jon Bales on his Basic Daily Fantasy Baseball Research publication, under the ‘Most Important Stats: Batting’ section

Previous Work Related to Visualization #2

There is several work related this visualization #2. Many people have pondered the question which park is the easiest to hit a homerun. Visualization #3 takes the average home runs for each park since the park was established. This information can lead to answering the question of which park is the easiest to hit a home run. If the average is high, we can see that many hitters do not have much of a problem with hitting it past the fences. What data scientist want to answer, however, is why are these parks the easiest to hit out of? According to the article “Which ballpark ranks easiest to go yard?” written by Manny Randhawa, altitude and yardage of the park plays a role in sailing the ball into the crowd. It states that due to Coor’s stadium high elevation, it makes it easier to hit a ball out of the park, leading to why the stadium has a highest home run per game average. In relation to statistics, the MLB website has also calculated the average amount of home runs hit per game in 2018, with Great American Ball Park at number 1.

Previous Work Related to Visualization #3

Only a little previous work was needed to start working on visualization #3. We first started by justifying our use of a scatterplot using the Sarikaya and Gleicher paper. This paper gives an overall verification that scatterplots are a good use to represent the kind of data that we have, and they show how easy to interpret they are. Secondly, we’ve incorporated a use of Weber’s Law. Sarikaya and Gleicher researched how Weber’s Law could be used to show that, “correlation judgment precision showed striking variation between negatively and positively correlated data” (1). This means that viewers perceive negative and positive correlations different. If we had more time, we would try to adapt this papers model of Weber’s law to visualization #3.

Previous Work Related to Visualization #4

Previous work related to Visualization four includes an analysis of why players are hitting more home runs than ever, an article that asks the question, do statistics alone still matter when considering hall of fame candidates. The first article is published by the Washington Post, and is titled “*The Statistical Revelation that has MLB hitters bombing more homeruns than the steroid era.* The article is by Neil Greenberg, and he examines the current uptick in home runs compared to the steroid era, which lasted from 1994 to 2005. During that time 11.8 percent of balls hit were home runs, but in the current timeframe 14.2 percent of balls hit are home runs. Greenberg attributes the increase to greater player knowledge that was previously unquantifiable. Specifically, players have changed the launch angle of hit balls raising it from 10 degrees to 10.8 degrees. This increase in launch angle has raised the ball’s trajectory enough to significantly increase the proportion of home runs hit in a season.

The second article was published in the New York Times by Doug Glanville. In the article, Glanville looks at 2017 inductees in Basball Hall of Fame. He waxes poetic about great moments in their careers, but concedes they are not in the hall of fame becasue of great moments. Rather, these men are inducted into the hall of fame because, “they separate themselves from the pack through consistent excellence and longevity”. Glanville supports the idea of long-term greatness as the benchmark for a hall of fame career by writing, “The scouting report fades, the lefty-righty splits are irrelevant, the weakness to his backhand side is unimportant. Greatness is now the summation of your work.”

These articles will propel our fourth visualizations to look at overall trends in statistical categories by grouping players over the years which provide a snapshot of how hall of fame players perform versus the rest of the league.

Description

Visualization #1

Visualizations one contains who visualizations. One that compares average batters and their wOBA (weighted On-Base Average) to the best batter per decade, and the other comparing average pitchers and their WHIP (Walks plus Hits per Innings Pitched) to the best pitcher per decade. The major difference to realize in the relationships of these two visualizations is that their relationships are inverse. Where the wOBA is meant to be high, the WHIP score is meant to be low. Both shows the change in average player per decade, and the best per decade, stating the explicit year that each best player is from. Both are shown as line charts to attempts to show an ever changing skillset of a player without overcrowding the data.

Visualization #2

Visualization two is a geo map showing each stadium with its average amount of homeruns per game since 1871 (since the stadium was established). When a user hovers over a marked location, a legend pops up showing the park name, park city, and the average home runs per game. This visualization was made in *Carto* using the Parks.csv file. In order to get the average home runs per game, I used Pandas in Python in Jupyter Notebook. In the AverageHomeruns.ipynb, I use both the Teams.csv and Parks.csv file. I use the Teams.csv file to get the average home runs for each game for each stadium These attributes include yearID , teamID, park, HR, and G. I then display the average home runs per game with the associated park in a table. I then input the average home runs into the Parks.csv for ease in making the visualization in *Carto.* After making the map in *Carto*, I then embed the visualization link in an html file in order to push it to the github and allow others to view it.

Visualization #3

Visualization three is a scatter plot of salary of the two leagues (American and National League) over time. Scatter plots, “ are among the most common methods for exploring and presenting data” (Sarikaya and Gleicher 1). This is because they allow the viewer to easily interpret data and find trends, clusters, or outliers in the data. This visualization shows a steady increase in salary throughout the years. After the data was laid out onto the plot, we generated two regression lines using scikit learn. These lines used the available salary data to fit a line to the points on the graph. As seen in the visualization, there is a trend that the American League is getting paid more than the National League. Linear regression is a good way to show correlation in the data for this graph. Weber's Law, " indicates that the human perception of differences in correlation and the objective differences in data correlation has a linear relationship…" (Harrison et al.). This means that there is a positive correlation for perception in scatterplots. It would be useful to apply the equation to this visualization to see whether or not if follows Weber's law. If it does, than the graph has a correct correlation and the data is being viewed by the audience correctly.

Visualization #4

Visualization four is a collection of line graphs that answer the question: Are Professional Baseball Hall of Fame Players Statistically Better Than The Average of The League? The visualization is built using Pitching.csv, Batting.csv, HallOfFame.csv, and People.csv. We merged several files to create the working data files: Pitching.csv was merged with People.csv, Batting.csv was merged with People.csv, HallOfFame.csv was merged with Pitching.csv and People.csv, and HallOfFame.csv was merged with Batting.csv, and People.csv. The non-HallOfFame.csv containing merges created data for the entire league, and the HallOfFame.csv merges created data for just Hall Of Fame inducted players. The merged datasets all used PlayerID as the primary merge key, and allowed us to look at overall statistics for all Hall OF Fame Pitchers and Batters, as well as overall statistics for the entire league. Before rendering the visualizations all data tables were grouped by year, and the average was taken for each year. The year ranges for batting and hall of fame batting are 1871 to 2010, and 1878 to 2010 for pitching. The hall of fame player data is taken during the years the players were active. For example, if a hall of fame player was active from 1980 to 1995, then his per season statistics are accounted in the hall of fame batter or pitcher data, and the overall pitcher or player data since he contributed to yearly totals.

The visualization shows twelve graphs on a single page, and compares hall of fame player statistics to overall league statistics. The categories the twelve graphs visualize are separated on batting data, and pitching data. The batting data statistics are Batting Average, Runs Batted In, Strikeouts, Doubles, Triples and Homeruns. The pitching data statistics are Wins, Losses, Earned Runs Against, Saves, Games Played, and Shutouts.

Discussion of Findings

Discussion of Findings for Visualization #1

The first thing to note is that the earlier sets of data can be uninformative. Because of missing data, the calculations are hard to create accurate representations of players from before important data was given. Before the 1900's, there was less data collected about teams and their players stats. Now days, there is more data, and more players to create more accurate data models. So as the data may change from decade to decade, it is also becoming more refined with more data. Within the batting graph, you can see that, for the most part, the best batter follows a similar pattern to the average batter, but at an increase magnitude. For one part there is significant difference between the two but as the average slightly increases, the best dramatically increases. However, in the pitching, there is less of a relationship. As the average pitchers WHIP increases in the beginning, so does the best. However as time goes on, you can see the the best pitchers WHIP going down, showing an overall increase in pitching performance as the years go on. Where the trend differs is the average pitcher WHIP. That remains to not have much or any significant change from recent past decades and only large changes can be seen from the early 1900's.

Discussion of Findings for Visualization #2

Visualization number two shows the average home runs per game since a stadium was established. This answers the question of which stadium gets the most home runs per game. Visualizing this information can lead to discussing many questions. A big one is why? Why do these stadiums have high home run averages, and why do some have low averages? In order to find the answer to these questions, a deep dive into what causes it is required. Could it be the altitude? Are some stadiums longer than others for a reason? If so, are they longer because of the altitude, and does the math check out to make sure all the stadiums are fair? These questions cannot be answered simply by a geographical representation of the data, however, it gives a good place to start in order to start looking into the reasons for why the some stadiums have high and low home run averages.

Discussion of Findings for Visualization #3

This visualization gives a quick overview of how salary has changed over time. As seen in the visualization, there is a steady increase in the amount the players in the two leagues are being paid through time. The causal mechanism for this could be a few things. First, players could be getting better over time which yields an increase in how much they are being paid. Secondly, the overall popularity of baseball may be increasing through the years which allows the leagues to pay their players more. This increase in popularity gives the leagues more revenue from game attendance as well as advertising. An interesting attribute to tweak in this graph could be normalizing the y-axis to be adjusted for inflation. If completed, we believe that the trend lines would be less steep; however, the difference in how much the teams are being paid would remain the same. Overall, this graph allows the viewer to see how salary has changed from the early years of baseball data collection to roughly present.

Discussion of Findings for Visualization #4

All twelve of the graphs show hall of fame players having better statistics than all players in the league. Notable results are that the difference between hall of fame batters, and the league is much greater than the difference between hall of pitchers and the league. The other notable result is that hall of fame trends match league trends identically. The hall of fame data is much more volatile because the sample size is smaller, but the trends match in all statistical categories. Application of these results could help the General Managers when considering signing star players. There is probably better return on signing a Hall Of Fame batter than there is a Hall Of Fame pitcher. For example, signing Bryce Harper in the 2018-2019 off season will probably provide a better return on investment than Clayton Kershaw. Even though both players are exceptionally dominant in their respective categories the visualizations show the difference between hall of fame batters and the league to be greater than hall of fame pitchers and the league.

References:

Alper Sarikaya and Michael Gleicher. Scatterplots: Tasks, Data, and Designs. IEEE Transactions on Visualization and Computer Graphics, 28(1): 402–412. 2017.

Glanville, Doug. “For Baseball's Hall of Fame, Do Statistics Alone Still Matter?” *The New York Times*, The New York Times, 21 Jan. 2017, www.nytimes.com/2017/01/20/opinion/for-baseballs-hall-of-fame-do-statistics-alone-still-matter.html.

Greenberg, Neil. “Analysis | The Statistical Revelation That Has MLB Hitters Bombing More Home Runs than the Steroid Era.” *The Washington Post*, WP Company, 1 June 2017, www.washingtonpost.com/news/fancy-stats/wp/2017/06/01/mlb-home-run-spike-shows-statcast-science-is-more-potent-than-steroids/?utm_term=.8c5b9d5ba687.

Harrison L, Yang F, Franconeri S, Chang R. Ranking Visualizations of Correlation using Weber’s Law. IEEE Trans Vis Comput Graph. 2014 Dec;20(12):1943-52. doi: 10.1109/TVCG.2014.2346979.

Houser, Adam. Which Baseball Statistic Is the Most Important When Determining Team Success? 2018, www.iwu.edu/economics/PPE13/houser.pdf.

Bales, Jon. “Most Important Stats: Batting.” RotoGrinders: The Daily Fantasy Sports Authority, 2016, rotogrinders.com/lessons/most-important-stats-batting-268627.

M. Randhawa, “Which ballpark ranks easiest to go yard?,” *MLB*, 21-Jun-2017. [Online]. Available: https://www.mlb.com/news/which-is-easiest-ballpark-to-hit-a-home-run/c-237845876. [Accessed: 05-May-2018].

“MLB Park Factors - 2018.” [Online]. Available: http://www.espn.com/mlb/stats/parkfactor/_/sort/HRFactor. [Accessed: 05-May-2018].