FinalProj

Sean Guglietti

2025-12-09

Final Assignment - Data Visualization

Recap, Goals and Tasks

Recap

- The data I used is historical MLB batting data, retrieved from the Society for American Baseball Research (SABR). I used the data to calculate the traditional offensive statistics (batting average, on-base percentage, slugging, on-base plus-slugging/OPS), truncated the data so as to only us data from the modern era (1953-onwards), and only included records with a baseline number of at-bats. The baseline number was 600, which is a very full season. This number was not chosen completely arbitrarily, it was chosen to limit the size of the dataset from 12000 to 1200. This does of course change the interpretation of insights to pertain only to a subset of players who reached the specific number of at-bats, but it was necessary for Altair to handle the data. This, of course affects the data, primarily due to a simple fact in professional sports, if you don't play well, you don't play. In this dataset, by limiting the sample size to players with 600+ at-bats, we will only be seeing players who are good, and the outliers for bad seasons will be distributed amongst players who are exceptional defensively, such as Ozzie Smith who in 1980 had a 0.589 OPS and 0.230 AVG (very, very poor), but won the award for best defensive play, at one of the most important defensive position, shortstop.

Goals and tasks

- The goal of my visualization was to visualize the spread/distribution of performances overall, and also by season. The aim was to allow users to efficiently find top and bottom performers in a given year, while also allowing users to see the dispersion, i.e., how relatively good or bad a season was. Therefore, the main tasks, and ones most important to the design, were for users to find the best offensive season (with over 600 at-bats), to find the worst offensive season, and to find trends in the dispersion (i.e., global outliers, and trends in terms of density of good or bad seasons). Beyond that, I also wanted users to be able to gain insight on whether or not there was much of a correlation between batting average and OPS.

Key design elements

Scatter Plot

- Using a scatter plot. This is because using other marks, such as a bar graph, will not be as good as showing how the data is dispersed. Users should come away with a good idea of: what an average OPS has looked like (and how it has changed over the years), what an elite/exceptional OPS is historically, what a poor OPS is historically, and also how batting average varies with OPS.

Color coding

- Utilizing color coding for the marks. Using a gradient color coding for each point as a representation of that batters batting average allows users to gain insights on how OPS and batting average vary, and allows them to come to their own conclusions about how 'important' batting average is (at least as it pertains to the OPS-definition of success). I used 4 bins because the colors needed to stand out and having a more gradual gradient would lose the apparent-ness of the batting average-related insights I foresee users developing with this tool.

Tooltips/Interactions

- Tooltips to allow users to identify the name and year of a season, as well as the specific batting average and OPS values, which is of course vital for season-specific insights.

Final Evaluation Approach

Description

- The final evaluation approach I employed was thinkalouds with people of various baseball statistical background: one of my friends who played baseball for 15 years and is an avid follower of the MLB and advanced stats, my brother who has been interested in baseball for a few years, and is familiar with OPS, and my mother who has watched a few baseball games but is wholly unfamiliar with what a 'good' OPS or batting average would be. I hoped that my friend would be able to find some 'fun' outliers, such as Luis Gonzalez's fantastic 2001 season, Alcides Escobar's tough 2013, as well as picking out some of the more fun 'high-AVG, relatively low-OPS' seasons, such as Juan Pierre's 2003 season, or one of the many Ichiro Suzuki seasons with exceptionally high average and sort of middle-of-the-pack OPS. I also believed that my friend would pick up on trends relating to AVG and OPS, more specifically that AVG was not the greatest indicator for OPS, as many very productive seasons did not have 'exceptionally high' averages ('exceptionally high' being in the .330-.350 range). I hypothesized that my brother would be interested in finding some of the more recent players who he is familiar with, and seeing how their seasons stack up historically, while also being interested in some of the clear outliers, again such as Luis Gonzalez's 2001 season. I hoped my mother would be able to use the tool, and search for recent players that she likely knows (Shohei Ohtani, maybe Vladimir Guerrero Jr.). The procedure was simple, give my recruits the tool, ask them to verbally comment on insights and their thought process, and silently watch how they use the tool.

Results

- My experienced friend's use of the tool began by looking at the outliers, specifically noting Luis Gonzalez's 2001 season and Roger Metzger's 1972 season, the lockout years in 1994-1996, and then looking at some of the high-AVG seasons, which are indicated by the deep blue dots in the sea of light blue. He also investigated more deeply some of the more recently great and poor seasons, such as Shohei Ohtani and Anthony Volpe's contrasting 2024 seasons. He also noticed that the average OPS for the dataset appeared to be close to 0.8, which as he said, was almost a full point higher than the true league average which tends to be closer to 0.7. He then realized this was because underperforming players were not going to reach the 600 at-bat threshold, and also mentioned how surprisingly stable the average seemed to be throughout the years.
- My brother's use of the tool again began with looking at outliers. He quickly identified the outliers that have been mentioned several times, and as anticipated began to look through the recent seasons, noting Shohei Ohtani's 2024, and Vladimir Guerrero Jr.'s 2021 season. This was about the extent of the insights, as the color-coded AVG feature was not used very much, I believe this was primarily due

to the large number of points on the screen, and simply the color contrast not being sharp enough to elicit the attention.
- My mother's use of the tool allowed her to develop some insights into who some of the good and bad players were, specifically, she now believes that Luis Gonzalez is the greatest baseball player of all time. She enjoyed finding some of the players she knows, but it took her significantly longer to do so, and she spent a large amount of time just putting the cursor back and forth and seeing if she recognized any of the names. As someone with very little baseball knowledge, this was expected, and the depth of insights was not very deep, the AVG feature for example going completely unused.

Synthesis

- The scatterplot, tooltips/interactions, and color-coding allowed my experienced friend to get quite deep insights, finding the trends I had hope would be uncovered between league-wide OPS and batting average correlation, as well as global trends regarding how OPS has fluctuated year-by-year, and finding the extreme outliers. The tool was described as 'fun' by my brother and mother, but my mother is biased as she loves me very much and her opinion should be understood with that context. In the future, I think the color coding should be much more striking, to draw attention to the fact that that feature is there. I also believe I should truncate the data in a different way, i.e., finding a better sweet spot with the minimum amount of at-bats, because having a visualization trying to portray the history of OPS without Barry Bonds in it feels exceptionally wrong. I also would be interested in developing plots that display changes in seasonal averages league-wide, as well as season maximums, and these may take shape better with bar charts instead of scatter plots. Overall, the visualization largely succeeded in what I had hoped it would, and future iterations will likely build on the work as opposed to reshaping.