Baseball Performance Data Analysis
Udacity DAND Final Project – Data Visualization, Storytelling, and Tableau

By Javier Soto

Introduction

This analysis presents the explanatory data visualization steps used and details the findings and conclusion on the main determinants of baseball players' performance. The analysis was executed using Tableau Public's data visualization tools and a Udacity provided data set containing 1,157 baseball players with statistics on handedness (right versus left-handed and ambidextrous), height (in inches), weight (in pounds), batting average, and home runs (including average home runs). It focused on the relationship between the performance attributes of batting average and home runs on the one hand and the physical attributes of weight, height, and handedness on the other.

Link to Initial Version:
https://public.tableau.com/profile/javier7547#!/vizhome/Story2_45/PerformanceinBaseball?publish=yes

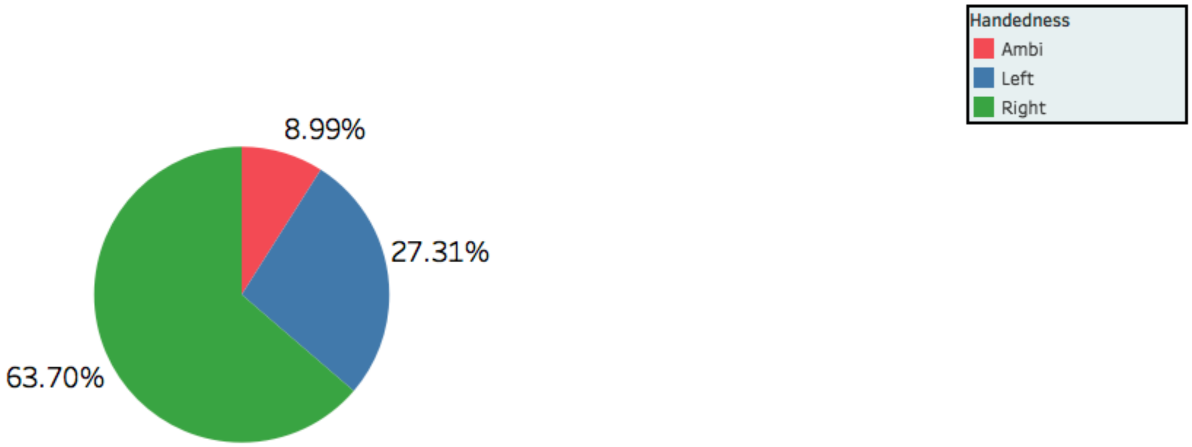
Link to Final Version:
https://public.tableau.com/profile/javier7547#!/vizhome/PerformanceinBaseballAnalysis-FinalVisualization/PerformanceinBaseball

Summary

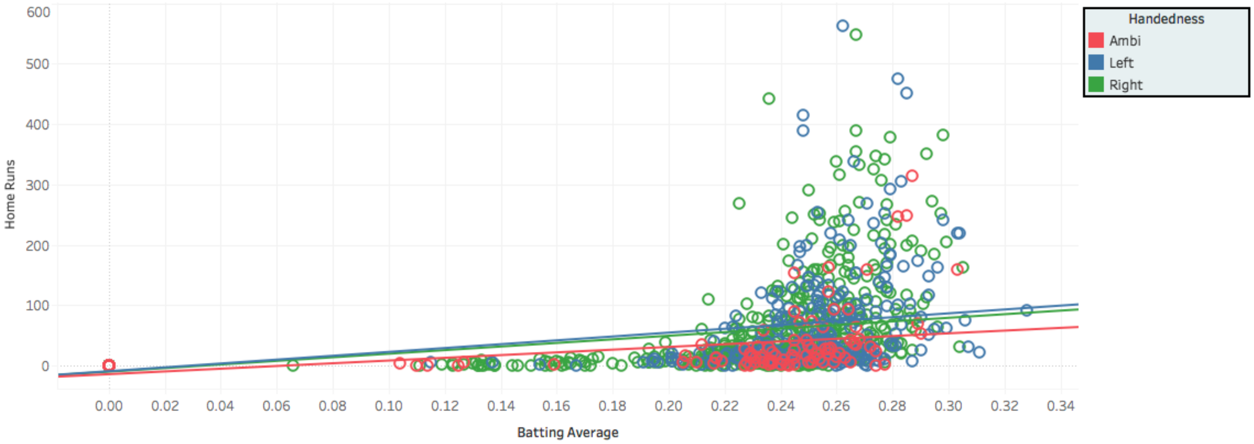
This project employed various visualization techniques to explore the baseball data observations. This included analysis on the relationships between handedness, batting average, height, weight and home runs.

Design

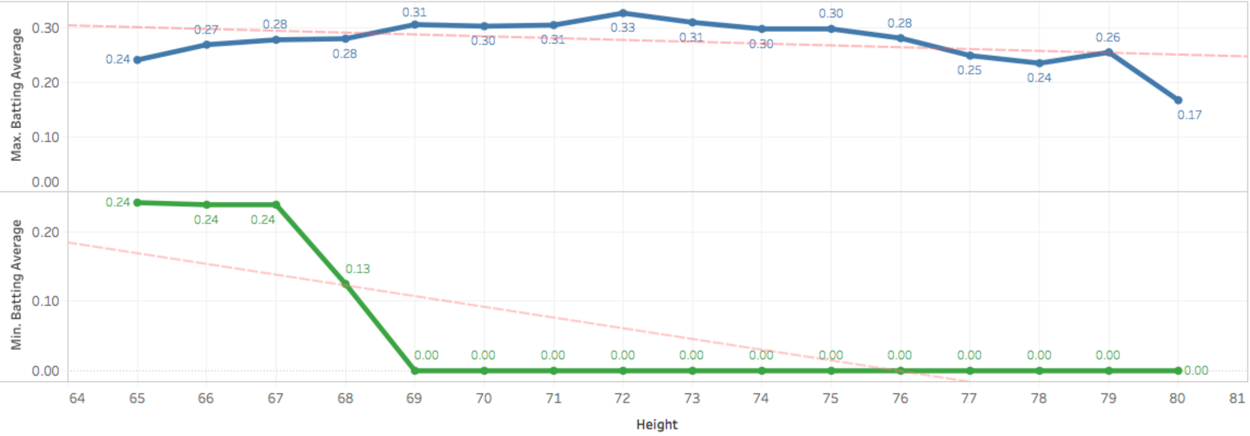
First, I started with a pie chart detailing the percentage of players who are right handed, left handed, or ambidextrous. Pie Charts are ideal for giving the reader a quick idea of the proportional distribution of the data, particularly with small values. Most of the baseball players are right-handed, according to our data. 64% of the batters are right-handed versus 27% left-handed. Only 9% are ambidextrous switch hitters.



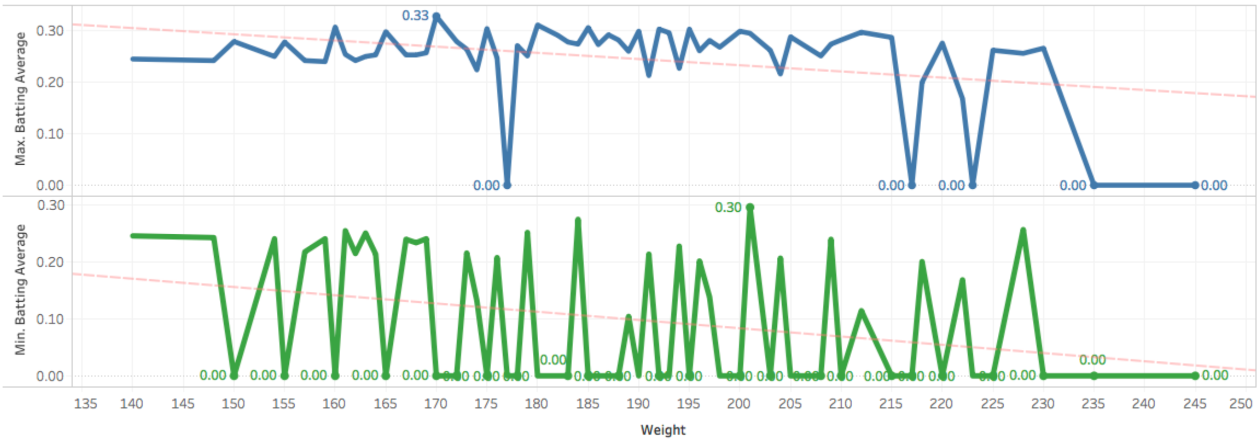
Next, I generated a scatterplot of each player's number of home runs (on the vertical y-axis) against their batting average (on the horizontal x-axis) and color coded for handedness. Scatterplots are ideal to show the relationship between two variables. It is also considered the best method to show non-linear patterns. Also, the range of data flow, i.e. maximum and minimum values, can be easily determined and reading them are straightforward. I also added trendlines to this plot. A linear trendline is a best-fit straight line that is used with simple linear data sets. It can show whether there is an increase or decrease at a steady rate. The added trend lines (categorized across handedness) suggests there is a slight positive relationship between the number of home runs hit and a player's batting average. Surprisingly, the left-handed players' trend line has the highest slope.



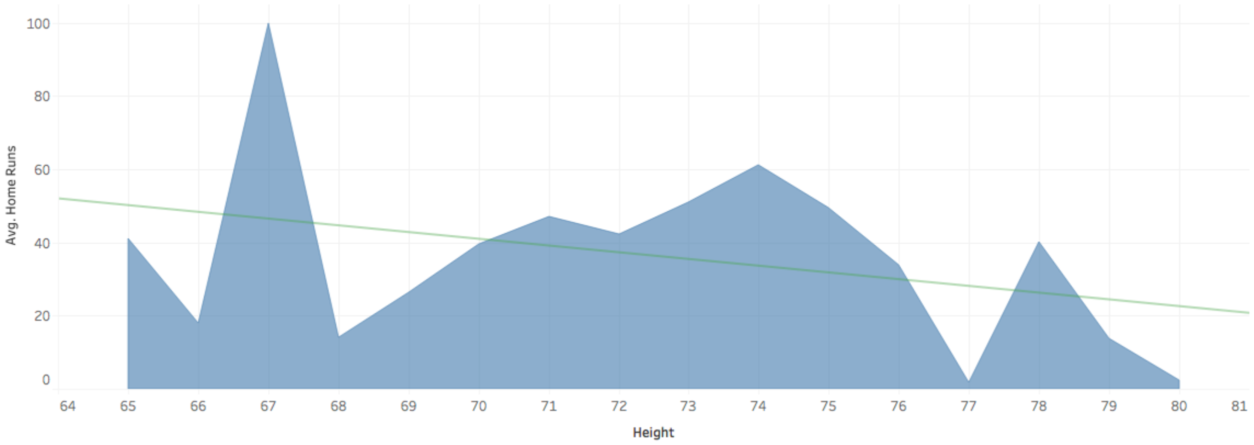
This was followed by a plot to determine whether height was a predictive factor for a player's batting average and to visually capture the maximum and minimum batting average range for any given height. The trend lines suggest a slight negative relationship; however, it is important to note that the data does not appear to be complete, with many players missing statistics on batting averages.



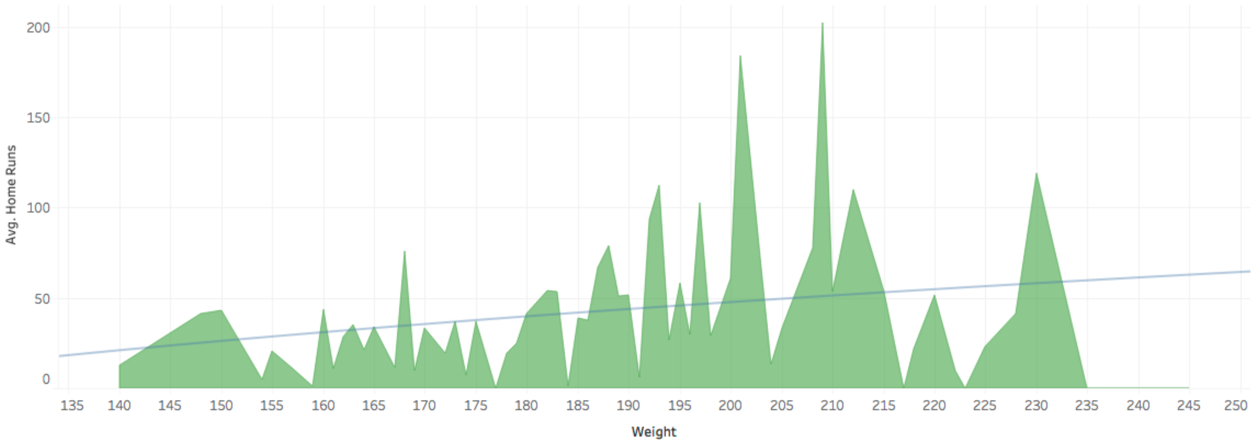
In addition, another plot was produced to determine whether weight was a predictive factor for a player's batting average and to visually capture the maximum and minimum batting average range for any given weight. Here the trend lines again suggest a slight negative relationship; however again, it is important to note that the data does not appear to be complete, with many players missing statistics on batting averages.



This was followed by an area chart (with a trend line) to plot the average home runs against height. Area charts displays quantitative data. It is based on the line chart. The area between the axis and line are commonly emphasized with colors, textures and hatchings. Surprisingly the trend line here suggests a slight negative relationship, indicating shorter players, on average, hit more home runs.



Another area chart (with a trend line) was plotted for the average home runs against height. Here the trend line suggests a slight positive relationship, indicating heavier players, on average, hit more home runs. This result is not surprising.



Feedback

Feedback was received from friends, family, peers, and udacity.com reviewers. I received comments in regards to the color schema, my choice of graphs, and readability of the presentation. In addition to aesthetic changes to the Tableau presentation, the following action was taken to address this feedback.

A full re-work of the analysis offered in the captions was done to correct midnight-hour inaccuracies. Also, added clarifications on what minimum and maximum batting average are. For the relationships between homeruns and height/weight used the mean instead of the sum. The pie chart was resized to increase the dimensions so that it takes up more of the story point frame. Also, suppressed the tabs and used the story point captions as the principal navigation tool.

It was pointed out that charts are more user-friendly when they include titles. The reason my charts were missing titles in the story points presentation is attributable to a quirk in Tableau. Single worksheets that are added to a story point end up with deactivated titles. To keep the titles, place charts on a dashboard first, then into the story points. As a work around I used the annotation feature to substitute for the titles to avoid re-doing the story. I also took some time to optimize the hover tool content to include units of measure for the height and weight variables.

Conclusion

It is concluded that the top baseball performers will be those who are left handed and shorter, with a height between 65 and 70 inches. Weight may play a factor, with heavier players hitting more home runs. Also, left-handed players may have an advantage over both right-handed and ambidextrous players.

Resources

https://stackoverflow.com/questions/tagged/tableau

https://www.tableau.com/support

https://udacity.com