What Factors Contribute to Home Runs?

First draft of Tableau story

Second draft

Final draft

Summary

When I first explored the data of 1,157 baseball players, I wanted to see what might lead to a high number of home runs, so I examined the relationship of that variable with others. While the number of right-handed players was more than double that of left-handed players, left-handed batters hit more home runs on average. I created heat maps of home runs per weight and height, and found that those who weighed around 190 pounds or were 72-74 inches tall had the most home runs. Finally, I found the strongest relationship with batting average, since low home run hitters had a mean batting average of 0.236 while high home run hitters had a mean batting average of 0.262 (which is a big difference in baseball).

Design

First Draft

For the graphs of handedness for the full dataset and for the home runs variable, I created a bar chart, because this was the simplest way to compare the values in a way that the viewer can easily see. The names “Left-Handed”, “Right-Handed” and “Ambidextrous” were used instead of “L”, “R”, and “B”, since these are easier for the viewer to interpret. I aggregated average number of home runs instead of sum because there were a lot of right-handed players in the dataset, so comparing the sums did not make sense. I did not order the bars based on height so that the names of each group could be aligned vertically for comparison between the charts. I decided to use a heat map for the height and weight variables because I thought it was interesting to see individual heights and weights that had a lot of home runs, sometimes significantly higher than a weight or height of only a one-unit difference. The default color scheme was kept since these are continuous variables. The legend was kept in order for viewers to understand what different

colors meant on that scale. For home runs by player and handedness, this hierarchy made sense, because the first thing the viewer should know is the player, since that is what is being compared. The handedness was placed before the value in the hierarchy since it would not make sense to have the bar before a text value. The bars were sorted in descending order so that the viewers could see the top players right away. Additionally, a filter was created for the user, in case someone wanted to see the difference made when choosing a type of handedness. Packed bubbles were used to visualize the same variables, but in a way that was easier for the viewer to see them grouped by handedness. The number of circles in each group combined with the size of the circles helped the viewer see what proportions of each group were high home run hitters. The default color scheme was kept because the colors chosen helped denote categorical differences in the handedness variable. The legend was kept to help the user clearly know which color pertained to which group. The final plot in the original story was a scatterplot since it showed the relationship between two quantitative variables. A line graph was not used because it would look too messy, due to shape of the graph. The axes were oriented in that manner because it made it easier to compare batting averages for different home run counts when home runs were on the horizontal axis. I did not use any aggregations for this plot because I wanted to see relationships among all players. Any axis name change in this story was done in order to create more clarity for the viewer. The different colors picked for the bar charts and scatterplot (and later histograms) were only used to create some visual variety for the viewer. Ironically, the original Tableau story I created did not have much of a story behind it, and really only presented some interesting visualizations centered around home runs. This became clear to me after receiving feedback from my younger brother, who seemed confused as to the purpose of certain plots in the narrative.

Second Draft

For my second version of the story, I added, removed and rearranged plots in order to create a clearer narrative: that there is not really one variable that you can say contributes to more home runs hit. The packed bubbles plot was removed because it was confusing to my viewer and did not add anything to the narrative. I moved the top home run hitters bar graph to the second position, and changed the focus of the graph from the players themselves to their handedness by revising the language in the story point. This change was made so that the graph could act as proof as to why left-handedness does not mean you will be a top home run hitter, contributing to the narrative of lack of relationships between variables. For the heat maps, I realized that my language in the story point created confusion as to what they were trying to convey, when I said, “players who hit the most home runs fall within

specific weight and height ranges." I edited the story point to clarify their purpose. After focusing my narrative on this lack of relationships, it came to my attention that the heat map was just as misleading as the first bar graph, because it was implying a relationship that was not there. Just because certain widths and heights of players have high numbers of home runs, this does not mean that having those specific heights and widths led the player to hit more home runs. Because I had created histograms of the widths and heights in the exploratory phase of the visualization, I was able to quickly confirm this suspicion. I found out that these width and height values at which there were a lot of home runs also happened to be the areas of the distribution where most players fell. This finding meant that maybe there were just a lot of home runs at these values because they were common values. I decided to keep the bar graph and heat map in the story even though they were misleading, in order to create a sub narrative. The sub narrative was that even accurate data visualization can mislead viewers to infer a relationship that is not there, and that it is important to verify those relationships with other plots. The caption for the final plot was kept, since it actually contributed to the story. A final story point was added with no plot in order to summarize the narrative for the viewer. In this draft, I also changed the tooltips to have variable names that were clearer for the viewer.

Final Draft

After receiving feedback from a Udacity reviewer, I changed the narrative again: that all variables that I explored contributed in some way to the number of home runs hit, some variables contributing more than others. I made this change after the reviewer pointed out that my findings that I thought were disproven by subsequent plots actually still held up and I had inferred the wrong information from some plots. The first change I made, based on a suggestion of the reviewer, was to add an introduction slide that summarizes the dataset and what the viewer should expect out of this analysis. The next change was to better convey the difference between the top and bottom bar graphs on the first slide. I created a calculated field for the top graph called "Percentage of Players" that would show the percent of players in each category of handedness. The axis for the graph was changed to "Percentage of Players" as well. This did not change the heights of the bar graphs but allowed me to point out those differences in percentages in the caption. Seeing percentages instead of numbers of players helps the viewer more easily conceptualize the difference between right-handed, left-handed, and ambidextrous players for the full dataset. Knowing the actual percentages also makes the difference in average home runs between the groups more surprising. I also reordered the x-axis in terms of height so that people could see which categories

came on top right away. I removed the bar graph of top home run hitters by name and handedness because it did not contradict my finding from the first slide like I thought it did. Instead, it only showed that there were majority right-handed players in the top 20, which made sense since the dataset is majority right-handed. This graph does not change the fact that left-handed hitters hit more home runs on average, so it was removed for the sake of having a clear story. I also removed the histograms for height and weight, because they too did not contradict the findings in the graph before them. Instead, they only made the narrative more confusing because I presented them as if they were a contradiction. If the point of the story is to show some interesting findings that came up when exploring home runs, then it makes no sense to try to contradict the findings, even if the findings are small. The reviewer also mentioned that the caption of the home runs vs. batting average scatterplot was confusing, in that it claimed that players with the most home runs have about the same batting average as most players. Not only was it hard to know from the graph what the batting average of most players is, it was also hard to tell what a high number of home runs is. The reviewer pointed out with histograms that many players had a zero batting average and zero home runs, which weakened the claim further. I therefore added in a histogram of home runs before the home runs vs. batting averages plot in order to show the high number of zeros in the dataset. I then changed the scatterplot to a bar chart to better show the difference between low and high home run hitters. Instead of using batting average, I compared mean batting average among groups. In this bar chart, I put the players with zero home runs in their own category and split the hitters between high and low based on the mean of non-zero home run hitters (see Baseball Project Calculation.Rmd file for calculations). I put the zero home run hitters in their own category because it did not make sense to put them in the comparison between low and high hitters, especially if there were so many of them. Additionally, if a player has zero home runs, there might be a reason behind that (maybe they are a pitcher) so they probably have a batting average that will bring the mean of the low hitters down significantly. This splitting of the number of home runs was done with a calculated field. The labels of the categories read, "HR = 0", "0 < HR < 60" and "HR ≥ 60," 60 being the mean of the non-zero home run values. This was meant to show the viewer exactly what group each player fell into. The y-axis for the graph read "Mean Batting Average" to make the aggregation type clear. With this new bar chart, the difference between low and high home run hitters in terms of batting average became more pronounced. After doing some research, it turns out that a 0.236 batting average, or that of the low home run hitters is significantly worse than 0.262, or that of the high hitters. The tick marks on the y-axis were more spread out in order to emphasize this difference. The color of the graphs on the first slides were changed to green, in order to differentiate themselves in that they

graphed handedness on the x-axis, rather than number of home runs. All other graphs used a blue shade since they all dealt with the same variable. Finally, I changed the summary slide to reflect this new narrative and how the findings contributed to that narrative.

Feedback

First Draft

I showed the first draft of the story to my younger brother, and the second draft was created based on his feedback and on changes I realized should be made after the fact. Most of the visualization was clear to him, except for the packed bubbles plot and the heat map. He could not understand what either plot was trying to convey. It took a bit of verbal explanation for my brother to understand the groupings in the packed bubble plot. After explaining it, I realized that if he was confused, other people would be too, and since it was not adding anything new to the story, I decided to remove it. For the heat map, he thought that there were intervals that ran vertically on the plot, and could not understand why the first two intervals were of five and why the later ones were different. While I realized the importance of this plot in my narrative and decided to keep it, I changed the language in the story point to more clearly explain how each box in the heat map functions.

Second Draft

A reviewer from Udacity looked over the second draft and gave feedback that shaped the final draft. Their first piece of advice was to create an introductory slide with no graph, in order to introduce the dataset and what the reader might find in the story. The second comment they made was to change the top graph on the first slide to reflect percentage of players, rather than count, and then to highlight the difference in percentages in the caption. For the bar graph of top twenty home run hitters, the reviewer suggested two changes. The first was to change the color of the bar graph to match the color of those on the first slide, since it dealt with home run counts by handedness too. The second was to either change the caption or get rid of the graph. Their reasoning was that the top 20 hitters were majority right-handed because the rest of the dataset was that way, but that did not change the fact that left-handed hitters had a higher average number of home runs. I decided to get rid of the plot in order to take away unnecessary explanation for the viewer. The next piece of advice was to redo the plots in the slide after the heat map and change

the caption. The reasoning behind this request was that these slides in their current state contradicted the findings from the heat map. The reviewer proposed something along the lines of creating a histogram on top of the current ones to show average home runs at individual heights and weights. These additional plots would accompany the original histograms to give the viewer a better understanding of how the average home runs for heights and weights differ from the values of the full distribution. Instead, I decided to get rid of that slide entirely, since the reviewer appreciated the heat maps slide and adding another histogram would still obscure the insights that the heat maps provided. Finally, the reviewer said that the language of "the players with the most home runs have the same batting average as most players" was confusing. To support this comment, the reviewer showcased a histogram of home runs, in which 709 of the 1,157 players had zero home runs. They also showed a histogram of batting averages, in which the peaks occurred at 0 and 0.24. Therefore, it was hard to tell what the batting average of "most players" was or what was considered a high home run number. In order to take this change into account, I showed the same histogram of home runs in my new story, and highlighted how many players had zero home runs. This would set the stage for how I would divide the number of home runs in the next graph. I then created a bar graph instead of a scatterplot that divided the variable for number of home runs into zero home runs, low home runs, and high home runs. This change allowed me to make better comparisons among players, as well as realize that even a difference in batting averages that appears trivial might actually be significant. The final piece of advice was to change the caption in the summary slide to restate the findings, which I did, now that I changed my narrative once again. Overall, these changes helped me present the interesting findings of my story in a much clearer way.

References

https://community.tableau.com/thread/124852

http://m.mlb.com/glossary/standard-stats/batting-average