Create a Tableau Story

Udacity Data Analyst Nanodegree: Project 6

Submission by Ravi Dayabhai

- Final (V3.0) Tableau Visualization[1]
- Draft (V1.0) Tableau Visualization
- Final Data Source (including custom fields)

Summary

From the prescribed baseball data source, I wanted to investigate how the physical attributes of professional baseball players (e.g., height, weight) related to their hitting styles (e.g., handedness, home runs). I begin the investigation getting a sense of the overall distribution of players and then drilled down to discover how switch hitters' (i.e., those that bat both left- and right-handed) difference in physical composition related to hitting characteristics. Switch hitters tended to be, on average, more "wiry" (as measured by a height/weight ratio) than dedicated right- or left-handed hitters and also tended to hit fewer HRs. This informs a hypothesis that could be subsequently explored: that the "lankier," perhaps less powerful hitters generally fill the top of the batting order (to get on base for more powerful "clean up" hitters) and tend to play positions that require speed or quickness (e.g., second base, shortstop vs. catcher, pitcher, first or third baseman).

Design

V1.0

In general, I wanted to stick to an orange and blue palette so that orange served as a stark contrast (and was used to lead the viewer's eye). In the first slide, I chose a line type for the histogram because of the density of data and to show how many weights were likely simply rounded to the nearest multiple of 5 lbs. (i.e., a clear separation between orange and blue lines). The joint distribution used color and size to describe counts in order to give the viewer a sense of where the dense parts of the distribution were (mimicking three-dimensions; gives sense of the joint distribution's "density" at specified coordinates).

The box-and-whisker plots when conditioning on handedness in the subsequent slide were chosen to reduce the overwhelming number of data points (given the data had to be disaggregated). The orange conditional averages and 95% confidence intervals are meant to lead the reader to make

[1] I accidentally overwrote first submission (sidebar: saving work with Tableau Public is an infuriating experience), so the "final" version has all accumulated changes incorporated.

the natural comparison between handedness categories. I left most of the descriptions in the tooltips to cut down on the chart clutter; alpha was adjusted to below 50% for intervals so as not to distract from the main point of the visual.

V3.0

Changes following Peer Feedback

No major changes aside from 1) changing the histogram in the first panel from a line type to a bar type and 2) removing the average line in the first "handedness" chart to avoid confusing the reader as to why the average is not 0 (basically, the average of ratios is not the same as the ratio of averages). I also (as per **Feedback**) opted for a jitter plot instead of box-and-whisker since the latter was distracting from the main point (differences in conditional means). For the jitter, I chose empty circles (with lower alpha/higher transparency) so each one could be distinguished on the canvas and because overlap in points show up "darker," again indicating an area that is more "dense."

Changes following Submission Review 1

Following the first review, I also added a home run count tooltip (with names) to [literally!] connect the dots between the "power" hitting and player size (see: second slide, first chart). This point regarding intuitive tooltips was well taken, since the interaction a user might have with the charts on the second slide would be 1) to identify individual players and 2) further condition the data based on home runs (e.g., "How are big-time home run hitters compare to average or not as prolific power hitters in terms of size [stratified by handedness]?"). For the second chart, I did the same, but included an "indexed player size" (read: a more intuitive label for the "relative height-to-weight versus average height-to-weight" measure) filter; this second chart provides similar information, but allows filtering by size to see home run production to give the user maximum investigative flexibility. Both charts included better labeling of handedness values; dimensions (e.g., pounds, inches) were included in field names to better describe physical characteristics.

Another change is the added inclusion of the batting average metric in the analysis (slide 2) to provide a more robust picture of the "hitting" story. It mimics the style of the original home run (by handedness) chart, and both filter by the indexed player size measure to allow the reader to drill into hitting characteristics when conditioning on size.

Finally, a third slide was added in order to get a sense of variability of hitting by player handedness over different player size ranges, especially with an eye toward differing frames of switch hitters. Balls (colors encoding handedness, size encoding frequency) were chosen since bars used too much ink and overwhelmed the primary comparison to be made: position relative to each other and average standard deviation for all data.

Feedback

I asked a co-worker (data engineer) for feedback on the initial version of my visualization. A summary of their commentary is provided below (see full transcript in **Appendix**); highlighted feedback was incorporated into the final version before submission.

- Change the style of the histogram from lines (suggests continuous values) to bars (to bring to better align with the point being made -- that "multiple of 5" weights dominate the dataset in an unusual way.
- This does not necessarily suggest changes in measurement *over time*. A general statement that can be made is that there is clear *heterogeneity* in data collection.
- Remove the average dotted line in the first "handedness" chart because the fact that the average across all data does not equal zero basically distracts the reader from the main point (i.e., switch hitters are above average when it comes to Height/Weight ratios).
- Jitter plots in lieu of box-and-whisker plots.

Resources

- Udacity DAND: Data Set Options
- Tableau Help: Build a Box Plot
- Information Lab: Show Me How: Box-and-whisker Plot
- Evolytics: Tableau 201: How and Why to Make Customizable Jitter Plots

Appendix

Colin F.'s Feedback

> *Very cool! Some thoughts:*
> *- the "weight is a multiple of 5" thing:*
> *- I found the line graph a little confusing because every value on the X axis falls into one of the two categories, but the line graph appears as if every value is represented in both the "multiple of 5" category and the "not multiple of 5 category". I think maybe there could be a better way to visualize this. (I wonder if it should be separate from what you're trying to show with the description of the distribution?)*
> *- I also didn't quite understand your text that says "might indicate differences in data collection methods over time"… that didn't seem to go with the visualization because there's no time dimension in anything you're showing.*
> *- I really like the joint height/weight distribution graph… packs a lot of information into a small space while still being clear. (Near the edges of the graph, I think the light blue is a little too light…*

it took a while for me to find the data in the 140 and 240 categories since there are no points in the darker color.)
- handedness graphs: what is the black dashed line? My guess is it's the mean across the whole dataset? If so, I think you can remove it in the top graph, as it it a bit confusing at first that it's not 0, and then once you think about it for a while, you realize it's because the mean of ratios is not the same as the ratio of means. But then people looking at the graph have ended up spending a bunch of time thinking about a property of addition and fractions that is unrelated to the point you're trying to make.
- maybe a jitter chart to get rid of box plots (not main point of graphic)? Tableau doesn't have a happy medium between disaggregated and aggregated data (box plots does this semi-summary)