Make an Effective Data Visualization

Udacity Nanodegree Project 5

- Data Visualization with D3.js
- Intro to HTML and CSS
- Javascript Basics

By: Joe Nyzio

# Make an Effective Data Visualization

Udacity Nanodegree

By Joe Nyzio

SUMMARY

This plot ranks the top 50 home run hitters from a list of over 1100 MLB players. The axes indicate height and weight, and the size of each bubble represents the number of homeruns hit by a particular player. These points are colored by player BMI's and the legend can be used to toggle between them. Scroll over each point to show the Name, Home Run total(career), Height(inches) and Weight(lbs) of each player.

DESIGN

I originally chose a scatter plot because it reminded me of the strike zone and pitch type counters used in the MLB like the image below.
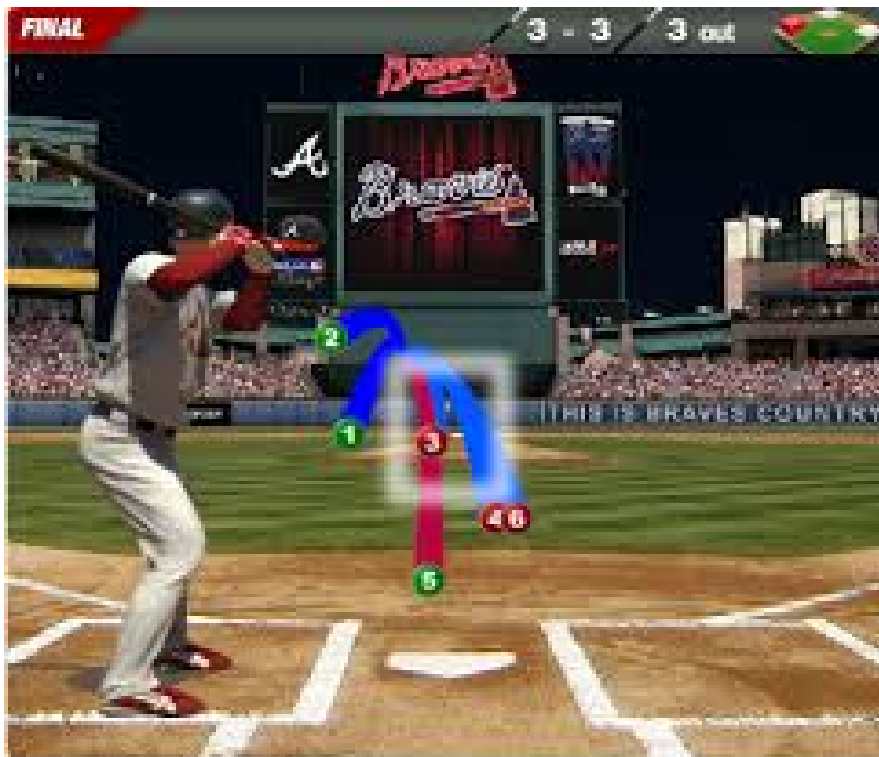
Another advantage of the scatterplot is the ability to allow the reader to get further information about specific players with the tooltip. My final plot used this idea but changed slightly to a bubble chart when I realized I needed to map multiple dimensions.

In this plot I'm trying to show that great hitters tend to have a certain body type. What I found is that there is about a 20 pound/3 inch range that the best hitters tend to reside in. Since the strike zone changes for each player based on their posture and height, the angle the ball is pitched will be defined by these parameters. As an extreme example, it's unlikely someone who is 15 feet tall would be able to hit a homerun off of the steep rise the pitch would come in at. The same goes for someone who is 15 inches tall trying to hit the ball thats coming straight toward the ground. The real game is obviously more subtle but this idea is what originally made me think there would be a general height and weight that helps a hitter most easily maximize their potential to hit the sweet spot.

Since the focus of this plot is the size of the player I chose to map variables relating to size to the x and y axis. Before making the plot I imagined being on a field and lining these players up. I'd immediately be able to see their relative heights, would order them by weight, and could see how many home runs they hit by the size of their heads. This plot is a snapshot of that lineup. Readers can get a sense of relative heights on the y axis as easily as if they were looking at this crowd. Putting weight on the x axis wasn't as visually intuitive. I thought about putting weight on the z axis so that fat guys were represented with fat dots but that erased any relationships I was hoping to show. I put homeruns on the z axis so I could get a relative sense of home run totals without scrolling over each one individually. Starting the axes at 0,0 made the differences in the players almost impossible to differentiate so I started it at 66,150 which is just slightly lower than the smallest players in the group.

I calculated the BMI and colored players accordingly. These follow an expected pattern that groups players of relatively similar body types together. This helped me notice patterns between players I had originally overlooked. For example Willie Stargell and Dave Kingman didn't appear similar at all before coloring by BMI. With this plot I can see that though Dave Kingman was larger than Willie Stargell he had the same relative proportions. Noticing that pattern I looked closer and can see they had a very similar career home run total as well. It's possible their similar home run totals have something to do with their similar overall athleticism.

Feedback

Index_1 Feedback

At this point I had just started getting the hang dimple and the surrounding workflow. My neighbor is a friend of mine I took it to him hoping for some direction.

He just sort of looked at it and didn't really care. I got him to scroll around and at least pretend he had a question he wanted to answer. It was clear I hadn't really given enough background or reason for the graph, and also that my friend was drunk and didn't feel like helping me. It was a little frustrating not being able to get any feedback but what I took from it is that it's hard to keep people's attention with a graph.

Index_1 Design Changes

I was mostly on my own as far as having things to take from this review. Things I noticed was that having no interaction didn't give him anything to do. I liked that I could partition the legend to Lefty, Righty, and Switch but thought I should use this as a way to clean the plot up a bit. I decided to make the legend clickable and interactive. While working on it I also learned to get the variables I need in order in response to a mouse hover.

Index_2 Feedback

After making the few changes I took the plot to my dad. I thought he'd be able to help since he's a baseball fan and an engineer. He scrolled around a bit looking at the players names. I didn't sense that he understood what I was going for and I felt I had to walk him through what the purpose of the graph was. He didn't click the legend until I mentioned that he could. Even after did click it he didn't use it again. He said that I shouldn't consider myself as being from Philadelphia since I don't know who Mike Schmidt is. That seems a little harsh.

I realized here that grouping by handedness probably wasn't that interesting of a way to sort the players. I also felt that 100 players was still too many people to expect someone to interact with without giving any direction.

My dad mentioned that he wished there was a way he could read the names of the players without having to scroll over them individually. He said maybe if there was numbers in the circles relating to player numbers on the side it would be easier to understand.

Index_2 Design Changes

I had learned to make an interactive legend but the way I used it didn't end up adding to the usefulness of the plot. I had too many players and not enough reasons for people to want to look at them. I decided to reduce the players to the top 10 home run hitters and color my legend by player names so there was a way to tell who is who like my dad said. I thought the idea of having numbers in the circles would look a bit cluttered but I tried to recreate the idea by having the players sorted by home run totals in the plot. I also finally realized that my axis had been sorted by categories and were poor representations of the data. Changing them to a measureAxis fixed this problem, added plot lines, and fixed an animation glitch I had in previous versions of the plot.

Index_3 Feedback

I happened to be seeing my mom around the time I finished the next version so I showed here where I was at. She doesn't know much about baseball or computers but designs bags/quilts and homes so I thought she'd be able to help with the graph from an aesthetic view.

She could see the names on the side but reached for her glasses when she moused over the points so she could read the extra info. She thought the colors were fine which I guess is why they were chosen for the default. She found it easy to understand what was going on and was double clicking the legend feature to show herself which bubble belonged to who. I had thought of using the legend for adding and subtracting players but using it as a locator tool was interesting. She didn't know that the bubble size and legend were ordered by home run totals because I had not provided any way she could have known that in the graph. This was the first time someone actually used the legend and animation for a useful purpose.

Index_3 Design Changes

This was the first time someone actually took time to explore a little without being asked to. Maybe it's because it was my mom but I think it has something to do with the interaction feature actually being useful. I noticed that graphing the legend to player names made interacting so simple that narrowing the chart all the way down to ten names was a bit too much.

In the final plot I pieced together what I learned from all of the reviews. I went with the 50 top home run hitters from the original csv file. I added numbers to the legend names so you could see the name and where they rank on the list at the same time,

then use the animation to see them in the plot. I thought this was a good middle ground between what each of my parents had said. I added a more descriptive title so that people can hopefully now tell that the 1 to 50 stands for the ranking of career home runs of each player.

Index_Final Design

The top 50 career home run hitters from my list organized by height, weight, and total home runs. Descriptive title I made slategray because black was taking some focus off of the visualization. The axes fonts are set to tangerine because the font was sharp and easily readable. I set the legend to the bottom and included all players with their home run ranking and left the default color scheme. This scheme repeated every 12 colors so I set the legend to go 6 wide which makes this repeat clear and keeps it from being overwhelming. The legend interaction can be used to clear out the graph to players of interest or locate players more easily. This visualization is taken from a csv file I made (rank50.csv) using the players with the top 50 home run totals from a list of over 1100 players originally named baseball_data.csv.

Udacity Feedback/Design

After receiving feedback from Udacity I made a few changes to my design. I needed to do a better job showing a trend in the data I had noticed. Instead of coloring individual players I grouped them by BMI to help show players relative proportions. The point I'm trying to get across is that there is a general body type that the best hitters have. Almost all the best hitters fall between 180-200 lbs, 71-75 inches, and 24-26 BMI. There is variation from this trend in the graph but there is a definite cluster around this area. This coloring and some additional text is all I really changed visually about the graph, but in doing this I had to come up with a reason for having the graph at all. It provided a contextual focus for the plot. To do this I also had to create a function in python that calculated the BMI's for each player and appended it to the dataset. I made a few changes to the size and spacing of the graph. I have a 2 screen setup with a macbook and a larger screen above. I made the original graph with the webbrowser open on the larger screen ignoring the fact that the graph was unresponsive to browser size. The new layout fits above the fold on the smaller screen size of my macbook.

There are 2 changes that were suggested to me that I was unable to complete. I haven't mapped the z variables to the square root of the home runs, and I didn't show explicitly with a legend that this z variable is mapped to the home run totals. It was said that there was no dimple js documentation on how to do so, that it wouldn't be held against me, and I wasn't given any idea of how to attempt it, so I left this part as it was.

In the future as I learn more about D3 or other graphing libraries I hope to be more capable of doing this.