Baseball Data Project Writeup

LINKS TO PROJECT ITERATIONS:

00:https://public.tableau.com/profile/andrew.phan#!/vizhome/proj00/Sheet1

01: https://public.tableau.com/profile/andrew.phan#!/vizhome/proj01/Story1

02: https://public.tableau.com/profile/andrew.phan#!/vizhome/proj02/Story1

03: https://public.tableau.com/profile/andrew.phan#!/vizhome/proj03/Sheet9

FINAL: https://public.tableau.com/profile/andrew.phan#!/vizhome/Project_Final_2/Story1

SUMMARY:

I chose to explore the baseball data set, which contains some information and statistics about 1,157 baseball players, including name, number of home runs, batting average, weight (in pounds), height (inches), and handedness (right, left, or both). I searched for possible relationships between number of home runs, a statistic associated with performance in baseball, and the other variables in the data set.

DESIGN:

I originally started off just throwing together a bunch of scatterplots with the performance related variables, batting average, and number of home runs, on the y-axis. After getting some feedback, I found it was a good idea to have a certain question in mind I wanted to answer, and then creating plots based off that question. I decided to focus on the question, “Are there any significant relationships between the number of home runs hit and the other variables in the data set?” As a result, I decided to cut out the plots I had with batting average on the y-axis, and only keep the ones with home runs hit.

I chose to go with a story, as I wanted to examine relationships including number of home runs hit and the other variables in the baseball data set one at a time. I felt like because the plots I had in mind, other than the histogram, were all scatterplots, it would look a little repetitive/unappealing to have them all on one dashboard; a story made more sense.

Moving forward with the feedback I received, I began with a histogram to display the distribution of home runs hit, as it was my variable of interest, and a look at the distribution served as a great introduction. I used the number of bins suggested by Tableau.

I then polished off the remaining scatterplots. In terms of color choices, for the first three plots I simply chose different colors for each plot since there was only place for one color in each of them. Since I did have an interesting group of players in the bottom right of the height versus home runs hit and weight versus home runs hit plots (the tallest/heaviest players), I decided to make groups for them and have them colored differently. However, after some feedback about people possibly interpreting these the wrong way, I decided to remove the separate color for the group, as it wasn’t very interesting or significant. Also, seeing as I had weight and height of the baseball players, and had plots having to do with each of them, I eventually decided to create BMI as a calculated field, as it is calculated from weight and height and it seemed like a BMI versus number of home runs hit visualization would intuitively follow after the weight versus number of home runs hit and height versus number of home runs hit scatterplots.

I included a filter for handedness in my plot of batting average versus number of home runs to emphasize that the point being illustrated by the visualization held for all handedness categories. After reading through the rubric, I also decided to include a handedness filter in my other scatterplots, as to include more interactiveness for viewers. Because I included a filter for handedness, there were three colors that could be customized. I chose to go with the color blind color palette, and chose colors I believed would be distinctive in the plot: yellow, grey, and red. Choosing the color blind palette, and then choosing distinctive colors allows for my visualization to be seen and understood easily by color blind and non color blind viewers. I also adjusted the opacity to be lower for all of my scatterplots after receiving some feedback about the large number of overlapping points, especially in my plot of batting average versus number of home runs.

One of the most important things I had to work on, based on feedback, was conveying my message better. My overall takeaway message was that you can not use height, weight, BMI, or even batting average to single handedly predict number of home runs. However, this was not conveyed very well until I included filters on each scatterplot that allow you to narrow down the range of values for the variable on the x-axis. This allows for viewers to see that even within a very small interval of heights, weights, or batting averages, the spread for number of home runs hit is fairly high, indicating that there was not much of a significant relationship between number of home runs hit and these variables

The notes I included were originally fairly long and wordy. After some feedback, I cut them down to allow for the visualizations to speak for themselves more.

FEEDBACK:

AFTER SUBMITTING proj00 + proj01:

It is better if you start your exploration with a question, and make a story to answer the question. For example, what factors that may highly be correlated with HR? To answer this you may arrange your viz like this.

1. An intro, univariate, bivariate viz for highest or lowest HR, compare their other features.
2. Bivariate, compare all scatterplot of HR
3. Something like your batting average vs. home runs plot to conclude

AFTER SUBMITTING proj02

The text you have provided is a bit long, it is better if the vizzes more self-explanatory or you add an explanation about them in two or three sentences

For the last (the multivariate viz), it is better to reduce the opacity of each points, because you have a lot of overlapping points.

AFTER SUBMITTING proj03

Be careful with the group of separate color, as it may confuse people as it has different color coding; people may interpret it as a different feature

How can you convey your message better?

DATA FILES

https://www.google.com/url?q=https://s3.amazonaws.com/udacity-hosted-downloads/ud507/baseball_data.csv&sa=D&ust=1527015626610000

RESOURCES

http://blog.usabilla.com/how-to-design-for-color-blindness/