Link to first version of data visualization:
https://public.tableau.com/profile/robert.hickey#!/vizhome/Tableau_Baseball_Final/BaseballStory_v1

Link to final version of data visualization:
https://public.tableau.com/profile/robert.hickey#!/vizhome/Tableau_Baseball_Final_Viz/BaseballStory_Final

1. ***Summary:*** *in no more than 4 sentences, briefly introduce your data visualization and add any context that can help readers understand it*

   My data visualization shows performance characteristics of over 1,000 baseball players because, while the data set includes 1,157 data points, some of the points are duplicates, so the total number of players is slightly lower. My visualization shows the batting average of each player versus the number of homeruns for each player on a scatterplot, colored by handedness (there was a slight positive correlation). Another visualization also shows the expected positive relationship between height and weight, colored by whether body mass index (BMI) was above or below 25 (above which people are classified as overweight, obese, severely obese, or morbidly obese), and sized by number of homeruns, which showed there may be some slight positive relationship between BMI and batting average and homeruns. Digging into this a bit further, I used a histogram to bin the data for every .5 of BMI and look at the homeruns and batting average in each bin (which indicated there is possibly a positive relationship between BMI and homeruns and batting average), and since I saw that some of the taller players had small circles on the height versus weight scatterplot, I wanted to see if there was an inverse relationship between height and homeruns and batting average, which appeared to be possible.

2. ***Design:*** *explain any design choices you made including changes to the visualization after collecting feedback*

   When I first created an Avg vs. HR over the Name dimension, I found two players with very high batting averages: Bobby Mitchell and Dave Roberts. When I looked at the baseball_data.csv, I found that there were two rows for each of these names. When I looked at the data in the rows, I found that there were not duplicates of the same data since the data in each row with the same name were different. I thought this could be an issue when I started since human names are not great to use as primary keys in datasets. To correct this, I added a (1) next to one of the names and a (2) next to the other name for these player names. Then I refreshed the data sources, which took care of the problem in this case.

   To check to see if other names were duplicated, I looked at height versus weight in a scatterplot and found that Mike Brown was over 12 feet tall and weighed 390 pounds. Since the tallest man in recorded history was under 9 feet tall, I found this data very dubious. I also found similar troubling entries for Jim Wright, Mel Stottlemyre, and Dave Stapleton.

   Looking at the data, I found that the entry for Mike Brown was duplicated, so I deleted one of these from the CSV. There were two entries with different data for Jim Wright, Mel Stottlemyre, and Dave Stapleton so I did as I did above with the (1) and (2).

   In the height versus weight visualization in both story version 1 and the final story, I filtered the data to remove players who did not have any batting average since it is likely that these players either did not play at all or did not play enough to be a meaningful statistic for my purposes.

Please see section three for the changes I made to the visualization after collecting feedback.

3. **Feedback:** *include all feedback you received from others on your visualization from the first sketch to the final visualization*

After collecting feedback from my coworker, who is left handed, I decided to color my batter average versus homerun average by handedness to see if there were any interesting phenomena there. I found that there is some evidence that those players who are ambidextrous may hit fewer homeruns than those who are right or left handed.

After some feedback from my wife, I decided to dig into the height versus weight chart a bit more. She said she was interested in seeing whether stouter players tended to hit more homeruns. Following this, I decided to color the dots on the height versus weight scatterplot by BMI and size it by homeruns. While there may be some evidence for this, I wasn't totally sure just from looking at this scatterplot, so I decided to create a histogram that looked at homeruns and batting average for different BMIs, and I saw a bit clearer that there is some evidence that there is a positive relationship between BMI and homeruns although this relationship does not seem as strong for BMI and batting average. One this that I did notice unexpectedly from this histogram is that players with a very low BMI of around 19.5 had a very high batting average. However, since the sample size in each bin at the extremes is a bit small, we can really say anything definitive about this.

My wife also notices "a lot of small dots" on the right side of the height versus weight histogram. Based on this observation I wanted to see if height was inversely correlated with homeruns, and I also added batting average to this line graph just to see what it looked like. It turns out there may be a slight inverse relationship between these variables.

4. **Resources:** *list any sources you consulted to create your visualization*

To create this visualization, I used the lessons from the Udacity module, "Data Visualization in Tableau."

I also read through the Udacity forums here to get a sense of how my classmates were thinking about this project: https://discussions.udacity.com/c/nd002-data-visualization-with-tableau/nd002-p-create-a-tableau-story

I also consulted the online help pages on the Tableau website: https://onlinehelp.tableau.com/current/pro/desktop/en-us/publish_workbooks_tableaupublic.html

I got some ideas for binning from this student's project: https://public.tableau.com/profile/sidharth.suman#!/vizhome/FinalData_Viz_Project/Story1