Project 8: Data visualization

Links to workbooks:

- Workbook before feedback
- Workbook after feedback

Summary:

In this project, I did some data visualization on the baseball dataset, in order to understand how some features, affect their performance.

So, the visualization starts by exploring the height and weight of players, then, handedness distribution, and finally, the effect of handedness on the batting rate and number of Homeruns.

Findings:

- Left-handed players have highest batting rates and number of HRs.
- Ambidextrous players have better batting average than right-handed players.
- Right-handed players have better HRs than ambidextrous.
- Height and weight don't have a linear relation with batting rate or HRs.

Design:

1st panel of the story:

- Since height and weight are continuous variables, I choose histograms to plot their distributions.
- The y-axis being the count of players for each bin, I saw no need to show it, since the values could be shown above each bin.

2nd panel:

- Handedness being a categorical variable, I choose a bar chart to plot its distribution.
- I choose a pie chart to plot this same distribution in percentage of total, because percentages are visually better in pie charts (or at least I think so).
- I used colors from the color-blind palette, so that, the color encoding makes sense to everybody.

3rd panel:

- Same as height and weight, a histogram was used to plot the batting rate distribution.
- This time though, and due to the number of bins (19), it could have been confusing to have the values of each bin above it, so I kept the y-axis.

- I also choose to keep the y-axis for the 2nd plot (batting rate average by handedness), because of the decimal values of batting rate (which I think would be bothering to read if they were put above the bins).

4th panel:

- I choose to have y-axis show as percentages to make it easy to notice how more than half players scored less than 20 HRs.
- Although there a lot of bins, I choose to show each one's value above it because most values are 1 or 2 digits long so it should not be very confusing.

5th panel:

- Player names were printed vertically to make them entirely visible since there are quite long ones.
- Same choices as the other panels were made concerning the color palette, labels and y-axis visibility.

After feedback:

- Added a scatter plot to visualize the effect of height/weight over batting rate by handedness.
- Since height/ weight and batting average are continuous variables, a scatter plot is indeed a good chart type to use.
- Handedness is used a color encoding, though, the nature of the data points makes it difficult to notice a pattern related to handedness.

Feedback:

1. From slack community:

"it is very clear and informative, and clear insights, you've used pie chart to show handedness distribution. if you need some improvements to make an update for the feedback, I think you should add anew plot type like scatter plot between avg batting and weight and color it with the type of handiness"

It is true that I only used 3 types of charts (histograms, bar charts and pie chart), which might get boring for the reader.

2. From a friend:

"The story is well sequenced"

This doesn't imply changes to the visualization, but it is good to know.

Resources:

- Change size of pie chart
- Remove grid lines