Summary:

I explored the baseball dataset to figure out how various variables such as height, weight, batting average and home runs are distributed, and I tried to figure out how the home run scored by the players are related to the other variables. I also created a new variable height/weight as the ratio of height and weight of the players, to figure out how it is related to the players' performance parameters; home run and average score.

The key finding of the data visualization exercise is that in the height range 70-76 inches and weight range 170-220 there are a few heavy weight star players with impressive home runs and averages and most of them are left handers.

Design choices:

- Bar chart, with color by handedness, worked well for players as categorical variables.

  Reason for choosing bar chart:

  As I wanted to plot player's name against the home runs scored by them, bar plots make the perfect sense, because, names of the players are categorical variables. Further, I chose to color based on the handedness of the players. It enhanced the visibility of the data and finally I sorted the bars in decreasing order of home runs.
  For such a large number of players, no other designs such as pie chart or box plots would have made sense.

  Observations:

  - Reggie Jackson (left handed) is the top most player with 563 home runs to his credit, followed by Mike Schmidt (right handed) with 548 home runs.
  - In the top five players, three are left handed.
  - The most successful ambidextrous player is Reggie Smith with 314 home runs.

- Histograms for height and weight of the players, faceted by handedness.

  Reason for choosing histogram:

  Since, height and weight are numerical data, a histogram is the best way to depict the distribution of numerical data.

  Observations:

  - Height is normally distributed with mean around 72 inches and weight has a bimodal distribution with 184 lb being the most popular weight followed by 168 lb.
  - Here most striking part is the proportion of left, right and both handed players remain almost constant across the distribution.

- Circles to compare the total number of home runs and average runs scored by the left handed, right handed and ambidextrous players.

  Reason for choosing circles:

  I could have chosen pie chart to depict the relative proportion of home run scored by the players of different handedness. But, I chose circles for two reasons: (i) aesthetics, circles look better than the pies, that is just my opinion. (ii) circles gives us total size of each as well as a basis for comparison.

  Observations:

  - When it comes to total home runs, right handed players have highest total home runs, followed by left handed and both handed ones.
  - But, for average runs, left handed players lead the pack followed by right handed and both handed players.

- Box plot to display the actual distribution of data points for the total home runs and average runs.

  Reason for choosing box-plot:

  When it comes to looking at the actual data, no design rivals box plot, thus we can see the three quartiles, outliers as well as median. Thus, box plot provides vital insights in the pattern of the distribution of data. I could have chosen histograms and faceted it for handedness, but it would have fallen short of what box plot can do.

  Observations:

  - Median home run and 75% percentile home run is in the order: left handed > right handed > both handed.
  - When it comes to number of outliers at the higher end, the order is: right handed > left handed > both.
  - Average runs and 75% percentile average runs show the pattern similar to that of home run, the order is: left handed > right handed > both handed.
  - When it comes to number of outliers, there are just a few outliers that too at the lower end, the order is: both handed > left handed, there is no outlier in the distribution of right handed players.
  - Further, there is a very wide range for right handed players, more than twice that for the left-handed players.

Bubble plots

Reason for choosing bubble plots:

Bubble plots provide excellent visualization across 3rd, 4th, 5th and so on dimensions of the data that is not possible by any other visualization design. For example, I deployed color saturation, size, shape, details to capture the various layers of the data.
Bubble plots are very helpful in identifying outliers (top performers) here and determining the salient features of the outliers.

1.Bubble plots between height and total home runs and between weight and total home runs, both faceted by color saturation and size.

Observations:

- There is a sweet spot for both height and weight, in which range lie the most players with good number of home runs. For weight it is, 175-210 lb and for height it is, 70-76 inches.
- In the height range 70-76 inches and weight range 170-220 there are a few heavy weight star players with total home run over 200.

2.Bubble plots between height/weight and total home runs and between height/weight and average runs, both faceted by color saturation and size.

Observations:

- For the height/weight range between 0.36-0.44, there is a bunch a very high caliber left handed baseball players that such as Reggie Jackson, Willie Stargell, Rod Carew, Lyman Bostock rule the game.

Feedback:

1. I posted my tableau workbook on Udacity's slack channel, but all I could got was one feedback from a fellow student to make the names of the worksheet more intuitive. I made the necessary changes in half of the worksheets to make them more intuitive.

URL to the slack channel:

https://udacitydatascience.slack.com/messages/C72AP9J3Y/convo/C72AP9J3Y-1519486558.000009/

2. I asked for feedback from a colleague, he advised me to consider a box plot for the average run too, and yes, these box plots were markedly different from that for the home runs.

URL to the older work book:

https://public.tableau.com/profile/ranjan.kumar5496#!/vizhome/MyDANDviz/Story

URL to the new work book:

https://public.tableau.com/profile/ranjan.kumar5496#!/vizhome/MynewDANDviz/Story

Changes made after the review:

1. Renamed the legends and labels to make them more intuitive.
2. Changed the colors to make the visualizations colorblindness friendly.
3. Created a few more visualizations; worksheets and dashboard.

URL to the latest work book:

https://public.tableau.com/profile/ranjan.kumar5496#!/vizhome/MyDANDVizlatest/Story?publish=yes

Resources:

I didn’t use any resource outside the course video and datasets to create this visualization.