Summary

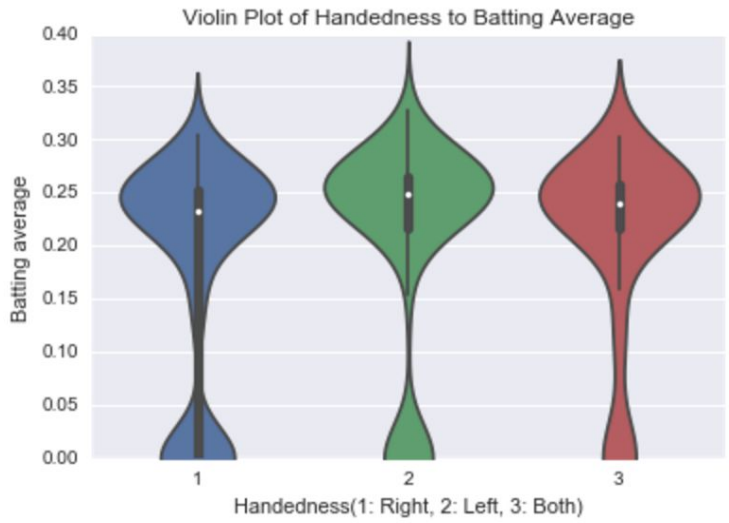
I have explored the baseball data set to check if there is a correlation between the handedness of players and their batting performance. The dataset I used had the following structure:

	name	handedness	height	weight	avg	HR
0	Tom Brown	R	73	170	0.000	0
1	Denny Lemaster	R	73	182	0.130	4
2	Joe Nolan	L	71	175	0.263	27
3	Denny Doyle	L	69	175	0.250	16
4	Jose Cardenal	R	70	150	0.275	138

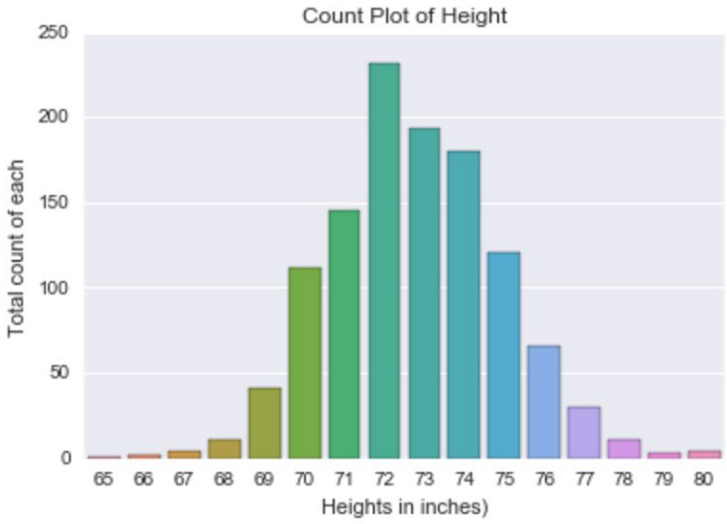
I explored the dataset in my iPython Notebook and upon my investigation, it did become pretty clear that **left handed players tend to perform better than right handed players** overall. The walkthrough of my design and methodology and how I incorporated feedback is as follows:

Data Exploration:

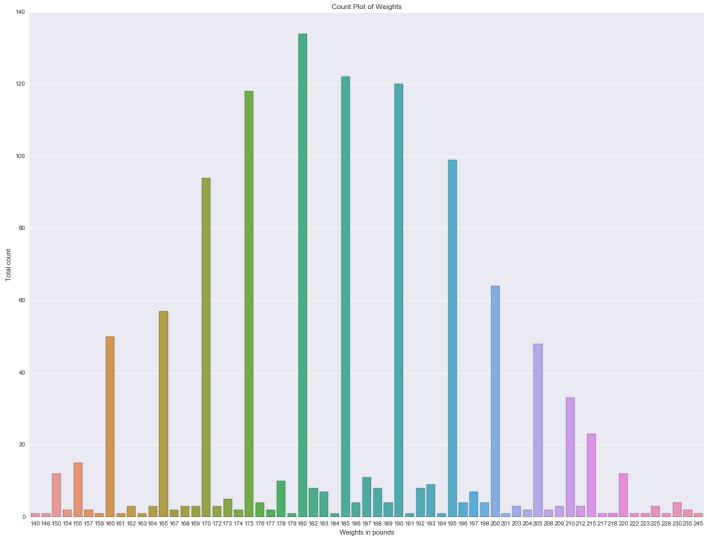
I explored the dataset using a few bivariate plots and could see that left handed players tended to have slightly higher averages than right handed players as can be seen below:



After this, I explored the heights and weights of the players. I noticed that this data was heavily skewed at certain values, presumably because people tend to round these values when entering them into systems. The heights values looked as follows:

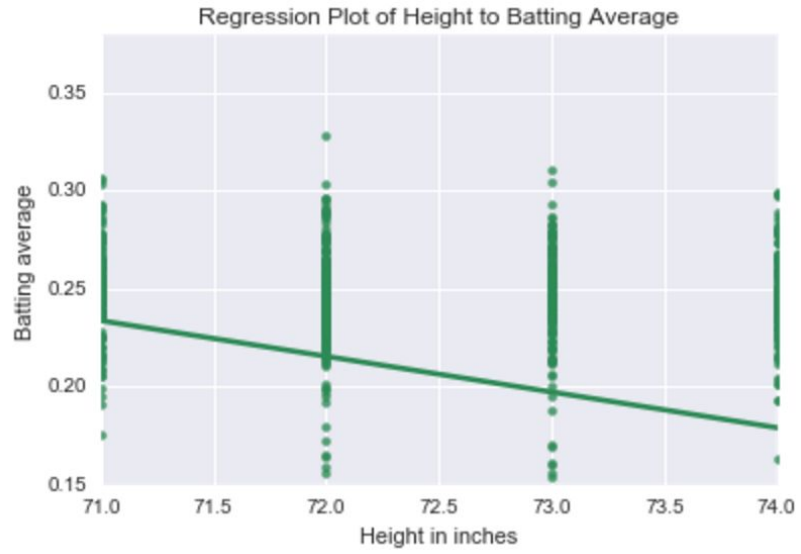


Similarly for weights, the values which were most common tended to be multiples of 5(the headings and axis labels are not clearly visible here but can be viewed in the notebook attached with this submission, for reference, the x axis has the height values and the y axis has the count of the number of players for each of those heights).



My next step in analyzing the heights and weights was to plot regression lines for each to check for any correlation. My result for heights was as follows:

I limited the the weight values to the middle 3 quadrants of values(175 pounds to 195 pounds) and it was interesting to the slope of the line, but it wasn't significant enough to cause much alarm. My similar result for height is as follows:



After this initial data exploration, I dove into dimplejs to plot by first bit of data -> Home Runs vs Batting Average faceted by Handedness.

Design

My visualizations progressed in the following order dimple1.html -> dimple2.html -> dimple3.html-> dimpl4.html -> dimple5.html ->dimple6.html.

Dimple6.html is my result plot to convey the overall message of my experiment.

Dimple1.html

Initially I thought of what are the features I'd like to use and what would be the best way to represent them. I played around with a basic histogram to show batting ranges to start with. But the sheer number of averages makes this unfeasible. I then thought of a scatter plot as a suitable alternative as it translates well when we have a high number of indexed values to represent. Therefore to show the range of batting averages of the players vs the number of home runs they hit I used a scatter plot.
The legend is clickable and you can see the clustering of the home runs by averages for all the players based on their handedness.

Feedback Part 1:
'If you are trying to make a point about which players have a better batting performance based on their handedness I'm not sure this actually conveys that message. The clustering does a great job of separating the players out and I like the fact that you can click on the legend to see that but it's hard to decipher who is actually better.'

Dimple2.html

To do a better job in showing the actual performance of these players, the clustering was not really helping as there were too many values to decipher. I therefore grouped the batting average of players into specific categories as follows:

```
def col_update3(row):
    val = 0
    if row['avg'] >= 0.0 and row['avg'] <= 0.1:
        val = 0.05
    elif row['avg'] > 0.1 and row['avg'] <= 0.125:
        val = 0.1125
    elif row['avg'] > 0.125 and row['avg'] <= 0.15:
        val = 0.1375
    elif row['avg'] > 0.15 and row['avg'] <= 0.175:
        val = 0.1625
    elif row['avg'] > 0.175 and row['avg'] <= 0.2:
```

```
        val = 0.1875
    elif row['avg'] > 0.2 and row['avg'] <= 0.225:
        val = 0.2125
    elif row['avg'] > 0.225 and row['avg'] <= 0.25:
        val = 0.2375
    elif row['avg'] > 0.25 and row['avg'] <= 0.275:
        val = 0.2625
    elif row['avg'] > 0.275 and row['avg'] <= 0.3:
        val = 0.2875
    elif row['avg'] > 0.3 and row['avg'] <= 0.325:
        val = 0.3125
    elif row['avg'] > 0.325 and row['avg'] <= 0.35:
        val = 0.3375
    elif row['avg'] > 0.35 and row['avg'] <= 0.375:
        val = 0.3625
    elif row['avg'] > 0.375 and row['avg'] <= 0.4:
        val = 0.3875
    return val
baseball['grouped_Avg'] = baseball.apply(col_update3, axis=1)
```

I then applied a count value to each row so that I would be able to get the number of players in each category. I used a scatter plot to represent these values and I was able get a much clearer picture.

Feedback Part 2:

‘The scatter plot looks good and I can tell there are more players who are left handed in the 0.3 and above batting average category. Just for continuity sake you may want to connect the plots on the graph to clearly show that lefties are greater on the right side of the graph’

Dimple3.html

I improved upon my previous graph by adding a line plot to connect the points on the scatter plot to show the trend that lefties have a higher batting average.

Dimple4.html

I went back to the drawing board to do some analysis of the mean and median batting averages of lefties and righties. I did the same analysis on for the home runs as well. I generated the following dataframe to show my results:

	handedness	mean_HR	mean_avg	median_HR	median_avg
0	B	32.144231	0.205048	13.0	0.2405
1	L	56.148734	0.204513	23.5	0.2480
2	R	42.598372	0.176620	14.0	0.2330

Clearly, the mean and median batting averages for lefties is higher than righties. I plotted the median values in the form of a bar chart with the legend showing the actual median scores for players in either handedness. I chose the median as it is less affected by outliers as compared to the mean scores. I chose to use a bar plot as I am representing categorical data with my handedness and it is the cleanest way to do so.

Dimple5.html

Similar to dimple4.html I plotted the median home runs hit by players of either handedness in a bar plot. It clearly shows that lefties hit more home runs in general than righties. I again considered the median in this case so that the effect of outliers is minimized and chose a bar plot as this is a representation of categorical data and bar plots are the cleanest way to represent the same.

Feedback Part 3:(paraphrased from the reviewers)

‘Firstly, it is better to have 1 final plot rather than 2 with all the consolidated analysis you would like to portray. The title needs some work, you could have a shorter title followed by a short description of what the plots are portraying. Your code needs to be neatly formatted with comments to show what you are doing. Your first 4 files are formatted but not your fifth plot. Also, your last two files lack enough comments. The use of the tooltip is not necessary as it conveys no new information. Also, the order of the legend is switched and has different colors, so you might want to improve that as well. Apart from this, your plots do a good job of conveying the final message, which is lefties come out on top.’

Dimple6.html

For my final plot I took the feedback and worked on each of them to ensure I can make my final result as clear as possible. First, I consolidated my 2 final plots into 1. So now the final result is viewable in one window and the end user doesn't have to toggle between files. The tooltip was not conveying any new information so I did away with it. Also, the legend is unnecessary as the plots are self explanatory in that respect. I updated the title of the of file to show what I think should be the take away from this project, that lefties come out on top. I added a short 1 line description for each plot below the heading for the end user to get a more explicit idea of what the plots are conveying. On the code front, I cleaned and formatted the code and removed all the code that I had initially commented out. So as a result of all these changes, I now have 1 file which shows the consolidated findings from my entire experiment.

Resources

http://dimplejs.org/
https://www.sitepoint.com/create-data-visualizations-javascript-dimple-d3/