Exploring Baseball Dataset

A baseball dataset is being explored for understanding the relationships between statistical categories. The categories include: player's name, handedness (right handed, left handed, or both handed), height (inches), weight (pounds), batting average, and home runs. There are 1158 players in the dataset.

Research Questions

1) Does a player with higher batting average scores more home runs? What is the average above which player scores high home runs? And How does this relationship among right, left and both handed players?

Visualization: Scatter plot has been used to answer and explore this relationship. It is observed that player with an average score of 0.25 or above tend to score higher home runs. This pattern is almost similar across right, left and both handed players.   Overall, average score of all player is also around 0.25, this is clear from the histogram plot of the distributions of the average score which is centered around 0.25, though there is an outlier at 0.   This average score is almost similar across handedness except for right handed players which is in the range 0.2-0.25. Plots 1-6 are being used to answer this question.

2) How are the home run scores distributed? What is the frequency of higher Home runs? What is the relationship between average score, home runs with respect to weight and height of the players.

Visualization: Histogram is being used to analyze the distribution of the home runs. It is observed that there are few players who has score home runs above 100. Average home run score is higher for left handed players than others.   Interestingly, it is found that players within the weight range of 160-220 have higher average score with left handled players getting more score than others. And players within the height range of 69-76 have scored higher home runs and average score. Plots 7-12 are used for this analysis.

Feedback A:

“In the first plot, it says there is at least a player whose average is 0.26 and hit more than 500 HR? That is quite amazing, isn't it?”

Feedback B:

“If you want feedback on the visualization, all plot looks nice. However, the information showed is not so clear. I mean, normally you need to set a hypothesis about your data and look for relationships among the explanatory variables that confirm or reject your hypothesis.”

Feedback C:

“I can’t tell who has the most HR between the left and right-handed players. That I think you tried to show in the histogram title ‘Home runs of R, L and B handedness.’ By the way, what does the x-axis variable mean on that plot?”

Feedback D:

“Your graphs need headers or footnotes. This would provide better clarity of what conclusion you are trying to communicate to your audience.”

Feedback E:

“There’s room for improvement on your design… perhaps an animation? I’ve seen amazing tools in which labels for an item expand if you hover over them. I have also seen amazing dictionary-like definition of a highlighted word of interest.”

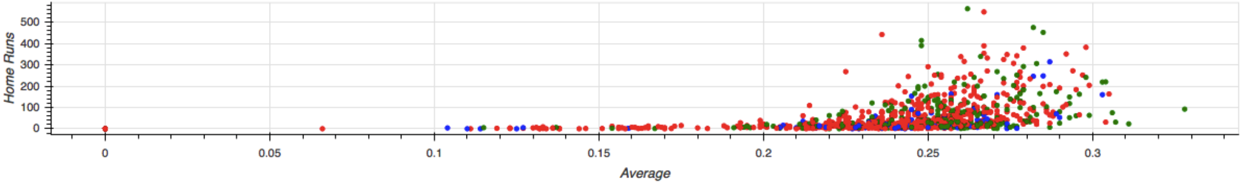
Based on the recommendations by my peers, the following changes were implemented:

- A summarization of the data to explain the relationships between the variables.

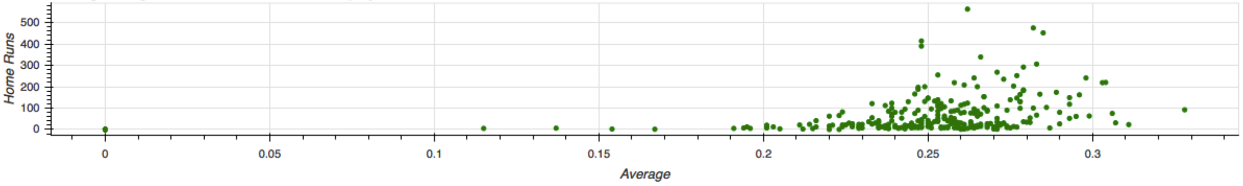
- Additional headers or footnotes for my graphs.

- Application of some sort of an animation tool.

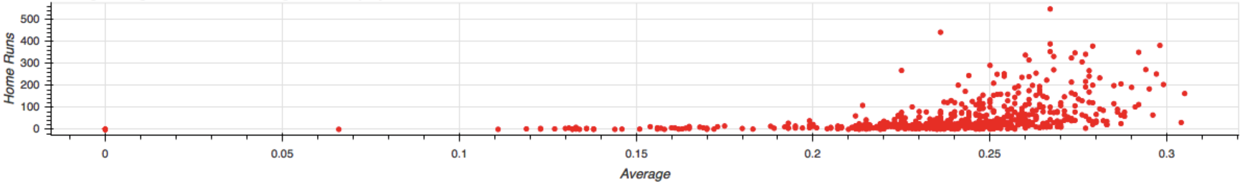
Batting Average vs Home Runs of all players



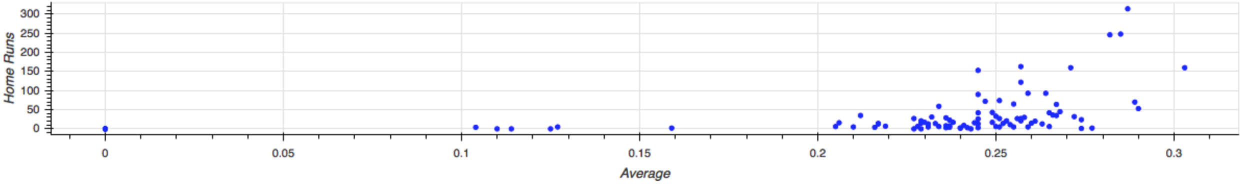
Batting Average vs Home Runs of left handed players



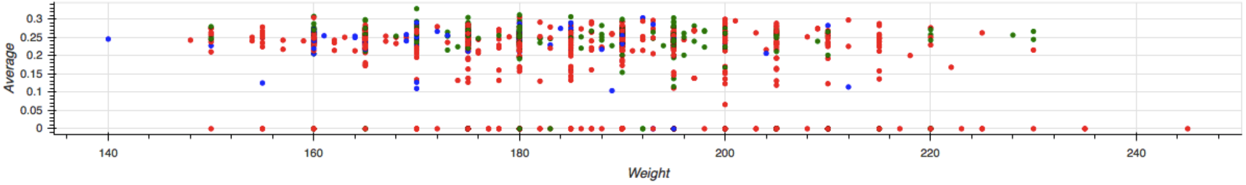
Batting Average vs Home Runs of right handed players



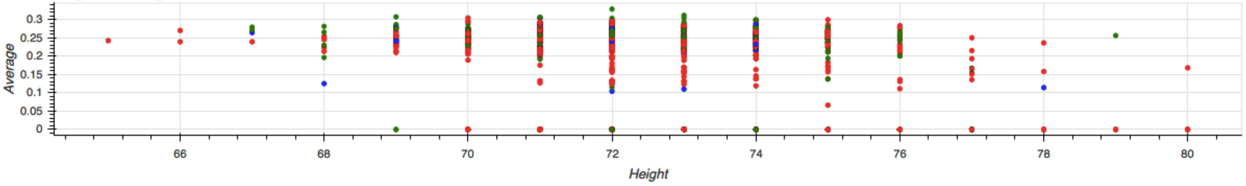
Batting Average vs Home Runs of both handed players



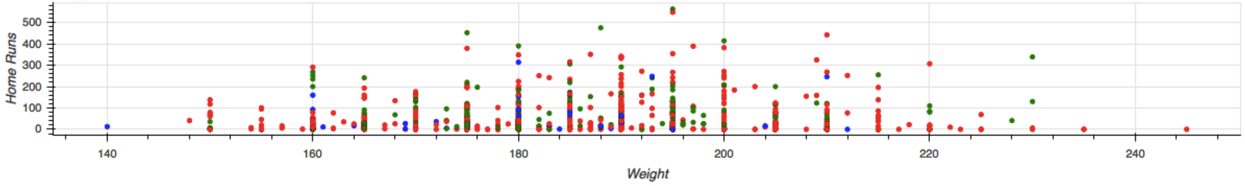
Weight vs Average



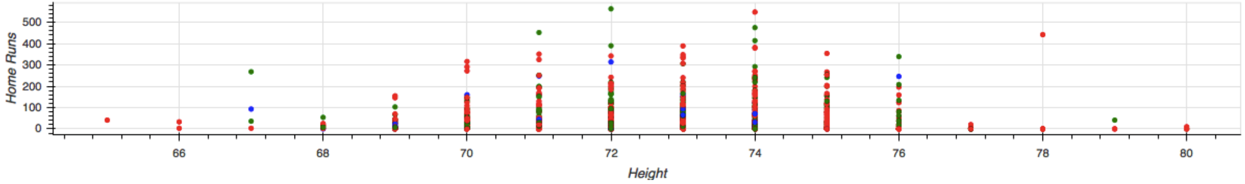
Height vs Average



Weight vs Home Runs



Height vs Home Runs



Resources:

Baseball data:
https://s3.amazonaws.com/udacity-hosted-downloads/ud507/baseball_data.csv

https://en.wikipedia.org/wiki/Baseball_scorekeeping

https://www.gamesensesports.com/knowledge/2017/3/17/righties-vs-lefties-the-importance-of-handedness-training-in-baseball-hitting

Animation tool:
https://bokeh.pydata.org/en/latest/