

Project 5 for Udacity data analysis nanodegree

Effective Data Visualization of Career Baseball Players' Performance

Zhongyuan Zhou

Related courses

- Data Visualization with D3.js
- Intro to HTML and CSS
- JavaScript Basics

Summary

Effective data visualization is created for career baseball players' performance in bubble and scatter charts. In the bubble charts, the homerun hitters are symbolized by circles in height-weight coordinates, the hitters' homeruns are characterized by circle diameters, and the handedness/ranks are represented by colors. Most homerun hitters have weights of 175 to 200 lbs and heights of 70 to 76 inches, more homerun hitters are righty, and weights correlate to heights nearly linearly. In the scatter charts, the homeruns and batting averages are plotted versus players' body mass indices (BMIs) and colored by their handedness. A "quantum pattern" emerges on the distribution of baseball players' performance along the BMIs. The players can be toggled by clicking legends, the players' information (name, weight, height, homerun etc.) can be abstracted by pointing the mouse to the corresponding circles, and different charts can be switched by clicking sidebars.

Design

I chose the data set of baseball players as the data of my project and I wanted to create data visualizations to show the baseball players' performance. The data set contains 1157 career baseball players' information: Name, handedness (righty, lefty, or switch), height, weight, batting average, and home run. Obviously the players are characterized by three variables: height, weight, and handedness, and the players' performance is measured by two variables: home run and batting average. Thus four dimensional charts are necessary in order to show the players' performance (home run or batting average).

First sketch

Chart files: index_v1.html

Initially, I proposed to visualize the homeruns of hitters' in a color bubble chart. In the chart, the weight and height are used as x and y coordinates, respectively, and the hitters are represented by circles with their positions determined by the hitters' weights and heights. The hitters' homeruns are characterized by circle diameters and the handedness is depicted by circle colors. To make the visualization visibly and clearly in a limited screen of 1300×600 pixels, I picked up the top 100 homerun hitters to visualize. Based on this idea, the first sketch was drafted and stored in index_v1.html.

After the first sketch was created I sent it to seven readers for review. I got feedbacks from them all. Although I did not include any context in the sketch to explain what the chart showed most of the readers understood what it was shown. The feedbacks are given in section "Feedback" in detail and are briefly summarized below.

What the readers notice in the visualization:

- (1) Most of the homerun hitters have weights from 175 to 200 lbs and heights from 70 to 76 inches;
- (2) The relationship between the height and weight is nearly linear.

The questions that the readers have about the visualization:

- (1) What is the relationship between the players' homeruns and the circles diameters?
- (2) Can you include batting average information into picture by making a 3-d from weight, height, and batting average?

Something that nobody notices:

- (1) Nobody notices the relationship between the homeruns and the handedness although it is clearly shown that the number of righty hitters is dominant. About 60% hitters are righty, 30% hitters are lefty, and only about 10% hitters are switch.

Based on the readers' feedbacks I revised my chart by adding some context to help readers understand the visualization. I also created more charts, both bubble and scatter charts, to visualize other variables in the data set, such as batting average. After revision and tuning, the data visualizations of the data set are greatly improved as demonstrated below.

Second version

Chart files: index_v2.html, BMI_avg_v2.html, BMI_avg_v2_All.html, BMI_Homerun_v2.html, BMI_Homerun_v2_All.html, and RANK_homerun_v2.html

In this step, I first add a subtitle to charts to illustrate what the project is about and what the purpose of the charts is.

As a response to the question (1) that readers raised in the feedbacks, I added an explanatory text under the chart of index_v2.html, "Players are represented by circles, sized by homeruns and colored by handedness", to illustrate the relationship between the circle sizes and homeruns.

As a response to the question (2), I created a scatter plot, BMI_avg_v2.html, to show the batting averages. In the first sketch, the hitters' homeruns are plotted versus weight and height in the color bubble chart. From that chart, it is hard to find any profound relation between the homeruns and players' weight and height. Thus instead of the weights and heights, I plotted the top 100 players' batting averages versus body mass index (BMI) in a scatter plot, BMI_avg_v2.html. The BMI is defined by [1]

$$\text{BMI} = \frac{\text{mass}(\text{kg})}{\text{height}^2(\text{m})} = 703 \times \frac{\text{mass}(\text{lb})}{\text{height}^2(\text{in})}.$$

From this scatter plot, it seems that a "quantum pattern" emerges in the distribution of the batting averages along the BMI. To endorse this pattern, I plotted the batting averages of all 1157 players' versus BMI in BMI_avg_v2_All.html. A "quantum pattern" is clearly shown in this chart, i.e., the players' batting averages distribute along the BMI only at some BMI values. The quanta of the BMI are 0.65. This is an interesting finding and is significant for the baseball players.

The "quantum pattern" in the distribution of batting averages arouses my interest to explore the distributions of homeruns along BMI. Thus the visualization of the top 100 hitters' homeruns was created versus the BMI in the scatter plot, BMI_Homerun_v2.html. Again, a "quantum pattern" emerges on the homerun distribution. The "quantum pattern" is endorsed by the scatter chart, BMI_Homerun_v2_All.html, where the distribution of all 1157 baseball players' homeruns was visualized versus the BMI. The quanta of the BMI are also 0.65.

In addition, another color bubble chart, RANK_homerun_v2.html, was created to show the distribution of the top 100 homerun hitters along the weight and height, and colored by the players' ranks. This chart provides a way to directly find the player information by rank.

Third version

Chart files: index_v3.html

I assemble all the charts created in the second version step and put them in a single chart file, index_v3.html, via interactive data visualization techniques. By clicking the sidebars all the charts created can be accessed via this single chart.

Final version

Chart files: index_final.html

In the final step, I smooth the charts by tuning the font size, style, color and opacity. I also explore the data visualization and craft messages or stories around the charts.

Feedback

After the first sketch was drafted I sent it to seven readers for review. I asked them to answer five questions after looking at the chart as suggested in the project overview [2].

- (a) What do you notice in the visualization?
- (b) What questions do you have about the data?
- (c) What relationships do you notice?
- (d) What do you think is the main takeaway from this visualization?
- (e) Is there something you don't understand in the graphic?

I got feedbacks from them all. The feedbacks (answers to the questions above) are given below.

Feedback 1

- (a) From the chart it is easy to visualize that the most of players are distributed in the center; the blue and red circles are almost equal.
- (b) Most of top career homerun hitters have weights from 180 lbs to 200 lbs with height from 70 inches to 76 inches.
- (c) What is the relation between among different circle areas?
- (d) No, no any needs to be taken away.
- (e) No. I can understand everything showed in the graphic.

Feedback 2

- (a) The concentration seems fall onto those players who are weight is between 180 to 200 pounds and height is between 70 to 76 inches.
- (b) No questions.
- (c) Concentration of left-handed and Concentration of right-handed are quite similar.
- (d) Interactive data visualization techniques made the switching views much easier.
- (e) No, it is straight forward.

Feedback 3

- (a) I notice that bubbles with three colors and varied diameters scatted in the chart. Most of them are located in the center of the chart.
- (b) I would like to know what kind of the feature that the diameters represent for baseball players.
- (c) I can see from the chart that more homerun hitters are weight between 175-200 lbs. And their heights are between 70-76 inches.
- (d) I think the visualization compares the number of left-handed, right-handed, and switch/both handed. As well, the chart illustrates players' weights and heights.
- (e) I think the chart is very clear about the attributes of homerun hitters, as height, weight and hand-using.

Feedback 4

Yes, the pictures are clear. You use size and color for homerun and handedness in a 2-d height and weight plane. Also every information of a player will be shown by clicking the circle. The design is very useful for people. Just some small points: can you include batting average information into picture by making a 3-d from weight, height, and batting average? Also can you use circle size for batting average in a 3-d from weight, height and homerun. This is an interesting data and many kind of analysis can be done.

Feedback 5

- (a) This figure contains lots of information.
- (b) No.
- (c) There is no significant relationship between the performances of players and their handedness. Positive linear relationship seems to exist between the height and weight of players.
- (d) In the range depicted in the figure, players can perform despite their physical characteristics, such as height, weight and handedness. Not much players are switch handed.
- (e) I think I understood it correctly but I might be wrong.

Feedback 6

- (a) Player performances vs. body parameter (such as handedness, weight and height) distributions.
- (b) Why the data visualization is more focused on the body parameter distribution instead of performance?
- (c) Most players have weight ranging from 180 to 200lb, height 70-75 inches.
- (d) Performance is not explicitly displayed on the first glance.
- (e) No.

Feedback 7

- (a) I see a nice and clear performance chart for all the baseball players.
- (b) Is the bubble size corresponding to the good or bad performance, e.g. the bigger the bubbler, the better the performance?
- (c) Players with medium weight and higher seem to do better. Regardless left or right handed.
- (d) This is more visualized way to quickly find out who are the top performers with all the related information.
- (e) Not really.

Resources

- [1] https://en.wikipedia.org/wiki/Body_mass_index
- [2] <https://www.udacity.com/course/viewer#!/c-nd002/l-3184238632/m-3181478727>