

Project 5 for Udacity data analysis nanodegree

Effective Visualization of Correlation between Variables of Baseball Players' Performance

Zhongyuan Zhou



Related courses

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

Summary

An effective visualization of correlation between variables of career baseball players' performance is created by batting average and logarithm of home runs (log home runs) in a scatter chart. The process how to effectively visualize the correlation is illustrated. It is shown that the batting average is not linearly correlated to the home runs but nearly linearly and positively correlated to the log home runs in which the correlation coefficient is 0.59. Visualizations of correlation in different coordinates and using different variables are created and can be accessed by clicking the sidebars.

Design

The data set of career baseball players' performance is chosen to be the data of my project. The data set contains 1157 career baseball players' information: Name, handedness (righty, lefty, or switch), height, weight, batting average, and home runs. In this data set, there are quite a few entries that have zero batting average. These players may probably be pitchers who never went to bat and thus will be removed in this project [1]. In this data set, there are also some entries that have zero home runs. These data will also be removed in this project because a new variable, the logarithm of home runs (log home runs), will be introduced for effective visualization. After removal of these zero data entries, only 871 career baseball players left in the data set, which will be used in this project.

In the data set, each player is characterized by three variables: height, weight, and handedness, and each player's performance is measured by two variables: batting average and home runs. It may be interesting to show the relationship between the variables of baseball players' performance. In this project, I will present an effective visualization of correlation between the variables of baseball players' performance as well as how to effectively visualize the correlation.

Correlation is one of the statistical tools to show the connection of two or more variables. The strength of the connection is characterized by correlation coefficient. The most common correlation coefficient is Pearson product-moment correlation coefficient, which measures the strength of the linear relation between variables. Correlation coefficient ranges from -1 to 1. The correlation coefficient is negative if one variable goes up when the other goes down. The correlation coefficient is positive if both variables rise or fall together.

In data visualization it is very important to select a right chart because an inappropriate chart may lead to a poor judgment and wrong conclusion whereas an appropriate chart can lead to right decision and conclusion [2]. One of the proper charts to visualize the correlation of two variables is scatter chart. A scatter chart provides a visual representation of correlation between two variables by displaying data points on a two-dimensional graph. The variable that is considered as an explanatory variable is plotted on the x axis, and the response variable is plotted on the y axis [3]. In this project, I treat the home runs as an explanatory variable and batting average as a response variable. They are plotted on the x axis and y axis, respectively. In order to visualize the relation between the correlation and handedness, color scatter charts are employed to color the third variable, handedness.

Revised version

In this revised version, I have taken all the reviewer's suggestions and comments in his/her report [1] into account and I have made great changes in the visualization both on the theme and chart types. In order to create effective visualizations and tell a clear and interesting story for the variables of career baseball players' performance, I only focus on the correlation between the variables of career baseball players' performance. I will present an effective visualization for the correlation and illustrate the process how to effectively visualize the correlation.

1. Batting average versus home runs in linear-scale coordinates

I plot the batting average versus the home runs in linear-scale coordinates. In this coordinates the batting average is nonlinearly dependent of the home runs and increases nearly logarithmically with the home runs. The correlation between the batting average and home runs is low. The (Pearson product-moment) correlation coefficient obtained from my computation is 0.40 [4]. The correlation coefficients for lefty, righty, and switch baseball players' are 0.33, 0.44, and 0.43, respectively. The correlation is slightly dependent of handedness.

2. Batting average versus home runs in logarithmic-scale coordinates

Due to the logarithmic dependence of batting average on home runs, I plot the batting average versus the home runs in the coordinates of logarithmic scales in the home runs axis. In this coordinates, the batting average increases nearly linearly with the logarithmic home runs. However, the correlation between them is still nonlinear and the correlation coefficient is still 0.4. The correlation is dependent of handedness. The dependence of the correlation on the handedness is in the exactly same way as that in the linear-scale coordinates.

3. Batting average versus log home runs in linear-scale coordinates

Due to nearly linear dependence of the batting average on the logarithmic home runs I introduce a new variable, $\log \text{home runs} = \log(\text{home runs})$. Then I plot the batting average versus the log home runs in linear-scale coordinates. In this case, the batting average is nearly linearly and positively correlated to the log home runs with a higher correlation. The correlation coefficient of the batting average and log home runs is 0.59. The correlation coefficients for lefty, righty, and switch baseball players' are 0.44, 0.60, and 0.46, respectively. This indicates that the correlation is dependent of handedness.

4. Conclusion

The batting average is not linearly related to the home runs but it is nearly linearly correlated to the log home runs. Thus the correlation between the variables of career baseball players' performance can be effectively visualized by the batting average and log home runs. The correlation coefficient is 0.59. The correlation is dependent of handedness.

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 feedback 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 feedback 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 68% hitters are righty, 30% hitters are lefty, and only about 10% hitters are switch.

Based on the readers' feedback 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 date 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*, *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 the readers raised in the feedback, I added an explanatory text under the chart of *index_v2.html*, “the Players are represented by the circles, sized by the homeruns and colored by the handedness”, to illustrate the relationship between the circle sizes and the 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 [5]

```
\begin{equation}
! [equation] (BMI = \frac{\text{mass}(\text{kg})}{\text{height}^2(\text{m})} = 703 \frac{\text{mass}(\text{lb})}{\text{height}^2(\text{in})}) .
\end{equation}
```

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

Feedback from the reviewer: Reviewer Report

This submission has a lot of interesting charts. I'd say there are actually too many, so it's difficult to tell what exactly the visualization is communicating to the reader.

Because there are so many charts showing so many different aspects of the data, the project is coming across more as exploratory rather than explanatory. It's showing the reader a lot of the data all at once rather than a clear, interesting finding or relationship in the data.

In the opening chart, the main takeaway is that height and weight are positively correlated; however, height and weight being correlated is something that people would probably already know. It's not necessarily a clear finding in terms of baseball players. If height, weight or BMI were correlated to batting average or home runs, that would be an interesting finding; from the current encodings, I can't really see if there is a correlation or not. Also, whatever variables are put on the x and y axes, those will be the ones given the most importance visually. Encoding home runs to circle radius makes it difficult to tell if home runs generally increases with height and weight or decreases. When including circle radius in a chart, it's also good practice to include a legend for circle radius although d3.js unfortunately doesn't do this for you, so I won't require it.

The README file also mentioned that there are more right handed players in the top 100 than left handed players. There are indeed more right handed players in the top 100, but then again, there are many more right handed people in the world than left handed.

The top 100 player charts showing BMI vs. home runs and Batting Average vs. home runs aren't really showing a clear pattern either. In terms of the quantum pattern that emerges, I'm also a little bit skeptical. The quantum pattern could be due to how height and weight were measured (only in whole numbers) or any other rounding that was done rather than something intrinsic about baseball players. I'd be a little bit skeptical about it.

The top 100 player ranking chart is more exploratory than explanatory. Adding in rank by color as well as radius size doesn't really give a clear insight or correlation. It would be interesting to some baseball fans to click around the visualization and explore the data, but that makes it a bit more exploratory rather than explanatory.

Another aspect about the data set to consider is that there are quite a few entries that have zero batting

averages. These players were probably pitchers who never went to bat, and they should probably be removed before drawing conclusions about performance. Unless of course the difference between pitchers and other players is investigated and something interesting comes out of it.

What I'm looking for to pass this section is a really clear narrative that has an interesting finding about baseball players. For example:

- is batting average and the number of home runs correlated?
- do left handed players have an advantage over right handed players?
- does body size or body type (BMI) correlate to more career home runs or better batting averages?

So to pass this section, try focusing on one really compelling finding and make that the focus of the submission. This current submission is focusing on many things at once and hasn't really found an interesting correlation yet, so it's hard to follow the submission. A compelling finding would be something specifically found in the data set related to baseball player performance, and it would be something that might not be intuitive unless somebody had done extensive analysis on the data set first.

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 [6].

1. What do you notice in the visualization?
2. What questions do you have about the data?
3. What relationships do you notice?
4. What do you think is the main takeaway from this visualization?
5. 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

1. 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.
2. Most of top career homerun hitters have weights from 180 lbs to 200 lbs with height from 70 inches to 76 inches.
3. What is the relation between among different circle areas?
4. No, no any needs to be taken away.
5. No. I can understand everything showed in the graphic.

Feedback 2

1. The concentration seems fall onto those players who are weight is between 180 to 200 pounds and height is between 70 to 76 inches.
2. No questions.
3. Concentration of left-handed and Concentration of right-handed are quite similar.

4. Interactive data visualization techniques made the switching views much easier.
5. No, it is straight forward.

Feedback 3

1. 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.
2. I would like to know what kind of the feature that the diameters represent for baseball players.
3. I can see from the chart that more homerun hitters are weight between 175-200 lbs. And their heights are between 70-76 inches.
4. 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.
5. 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

1. This figure contains lots of information.
2. No.
3. 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.
4. 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.
5. I think I understood it correctly but I might be wrong.

Feedback 6

1. Player performances vs. body parameter (such as handedness, weight and height) distributions.
2. Why the data visualization is more focused on the body parameter distribution instead of performance?

3. Most players have weight ranging from 180 to 200lb, height 70-75 inches.
4. Performance is not explicitly displayed on the first glance.
5. No.

Feedback 7

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

Resources

- [1] See reviewer's "Reviewer Report" in Section "Feedback"
- [2] <http://chandoo.org/wp/2010/04/19/chart-selection-process/>
- [3] <http://www.netmba.com/statistics/plot/scatter/>
- [4] <http://www.sthda.com/english/rsthda/correlation-matrix.php>
- [5] https://en.wikipedia.org/wiki/Body_mass_index
- [6] <https://www.udacity.com/course/viewer#!/c-nd002/l-3184238632/m-3181478727>