Baseball Top Home Run Performers Characteristics

by Arturo Parrales Salinas

September 3, 2018

Summary

This project is aimed to understand baseball top performer players' characteristics. We compare the handedness player numbers, their BMI and where do top performers fall in such characteristics. We measure the performance using home runs (HR) and the average batting rate (BA). These two variables show a logarithm relationship that help us understand where players stand respect to each other. Finally, we show the best home runners to prove their performance is at the very end of the graph and to show their BMI.

BMI – Body Mass Index. It is the Weight/(Height)^2

Dataset

The original dataset contains information of 1,157 baseball players. Among the information we have:

- **Name**. The player name
- **Handedness**. Player batting preference:
  - R for right
  - L for left
  - B for both
- **Height** (in inches)
- **Weight** (in pounds)
- **avg**. Average batting rate achieved by the player
- **HR**. Number of home runs achieved by the player

During our visualization we created new variables:

- **Height** (in meters)
- **Weight** (in kilograms)
- **BMI** (kg/m^2)

Note: In Tableau other variables were created to be able to display ranges, since Tableau needs Dimensions and Measures. We needed to make dimensions out of many of previously listed variables.

Visualization Design

The first design consideration made was to use a color blindness friendly visualization since we are aiming to anyone whole likes baseball and it is expected to have color blinded people interested in the findings.

Then, we started to consider the data and best ways to display it, while not making it tedious to look at them. We created a lot of visualizations only to explore and some of them were made leaner to show a specific finding that could be used in our main explanatory visualization set.

The first attempt of the visualization was a dashboard that combined different visualization to show the headedness groups, top home runners, BMI, height in meters, weight in kilograms versus average home run ratio to batting, and a scatter plot of height versus weight. However, this dashboard alone could not clearly convey the message of the top performers characteristics, how they relate to non-top performers and their BMI distribution. Thus, many improvements were done such as using a histogram of BMI instead of scatter plot of height versus weight, change the average home run ratio batting to average batting in the graphs, introduce a plot of average batting versus home runs, and create a story instead of a single dashboard. These changes helped a lot to tell the story of our findings better. Thus, in here we will describe the graphs used in the story and how they evolved if it was the case.

Also, it was important to add notes on each of story slides to remark the insights found and carry the audience through the story. This actually conveyed the message to the audience better and it was added after feedback received from one reviewer of the story. Thus, all of the following visualizations in the story have now notes and in addition we introduced the dataset in our first slice just to let the audience know our initial data.

Player Handedness

The first graph is a packed bubble chart since it cleanly displays the 3 different headedness groups:

- L – Lefty
- R – Right handed
- B – Both hands (ambidextrous)

It also gives an easier idea of number of players in each group using the size of the circle. The other option was to create a bar chart, but it will have many numbers and it complicated the main message to show the viewers the handedness groups. In fact, if you look at the graph you can easily see that R players group is the biggest and B players are the smallest. However, in order to make this more obvious, I decided to add proportion percentages of the total to the label of the bubbles.

For the next graphs we had computed a value called the average home run ratio batting, which actually was not a useful variable, so graphs that used it were actually changed to use the average batting rate later on.

BMI of Players

The intention of this visualization is to display the BMI of players. We first used the average home run ratio batting and the BMI to show the distribution of players, but this actually was not very promising. After the feedback, using a histogram made more sense and an additional big improvement was to use the headedness to show each group among the histogram.

The histogram is definitively a better option to see the distribution of a variable, and in this case it help us see that the BMI is a bimodal distribution, and that each headedness group has a similar distribution. This was very helpful to indicate that the player's performance could be measured in a similar manner.

Players Performance

This visualization was not created in the first attempt, but it replaced the weight versus height scatterplot on the first dashboard. The idea of it came after a comment in the feedback and after the finding that the calculated average home run ratio batting variable was not useful. Then, we were forced to find a way to show players performance in a different way. The result was this visualization which is one of the most insightful graphs.

This visualization shows that the performance of the players in a logarithmic model of home runs versus average batting rate. We used a scatter plot to show the relationship of our chosen variables to measure performance. We used color to show the three handedness groups and also the BMI to add size to the circles in the plot. All this together, confirms that the R players are more, and that the performance of all groups is measured similar since they share similar distributions. It also was helpful to add the BMI to later on combine in the dashboard the interaction of this plot with the BMI histogram.

Top Players BMI

This visualization is actually a dashboard that uses a **BMI of Players** (without handedness colors) and the **Players Performance** visualizations. It has filtered the data of the top performer players, which means it only shows the histogram of the BMI of the filtered top players. From here, we wanted to show that top performers players are more common in the 23.5, 24, 25, and 25.5 BMI ranges. This clearly shows the bimodal distribution and it gives us a clue that BMI might be a good characteristic in common for top performer players.

This visualization was created after the feedback since it combines two previous graphs and it was used as an intermediate step to transition to the end. This actually helps convey the idea of the players BMI distribution showing the peaks of the modes for top performers. This makes it clearer the BMI ranges of the top performer players and then reaffirm this idea in the last slide which is the modified original dashboard.

After another review this dashboard included the packed bubble **Player Handedness** to allow the audience to see how the proportions of handedness didn't changed for top performers.

Home Runners Stats

This dashboard is the original idea to show the findings, however, it was not very clear in the first attempt. This lead to a lot of enhancements while discussing how to improve it with a friend. The first dashboard included elements that were changed or removed after feedback:

- Treemap of the top 20 home runners with the name of the players in each square colored by the handedness. It also served as a filter of the stats of the player clicked.

- The packed bubbled plot **Player Handedness** to show the handedness groups.
- Weight and height scatter plot to show where the player stands. This plot was removed since it just seemed to cause confusion and the other plots were more useful.
- BMI versus average home run ratio batting bar chart, which was removed since the latter variable was not useful.
- The height in meters versus average home run ratio batting bar chart, which was kept, but the average home run ratio batting bar chart was changed by average batting rate.
- The weight in meters versus average home run ratio batting bar chart, which was kept, but the average home run ratio batting bar chart was changed by average batting rate.

After the feedback, the dashboard had the following main elements:

- Treemap of top 20 players (that can be filtered on number of home runs). This is the heart of the dashboard. It was used to display the name in a square colored by handedness, it also uses the home runs(HR) to choose the top players, and the size of the squares using different features. This also allowed us to highlight and filter data when we hover over a player or click it, respectively. You can also select multiple players at once to filter them. The treemap plot is perfect to display text data associated to dimensions instead of a word cloud and that is the main reason to use it here to make use of the names of players and use them wisely.
- The packed bubbled plot **Player Handedness** remained the same and its purpose is still to show the handedness groups.
- The **Players Performance** visualization to show where the player stand respect others in performance. It also helps to visualize the logarithmic function of player performance.
- The **BMI of Players** (without handedness colors) to show the histogram of the players filtered.
- The height in meters versus average batting rate and weight in kg versus average batting rate bar charts, which show the height and weight groups of the filtered players, respectively. This helps understand more characteristics associated to the BMI and the players. I used bar charts to take advantage of the length of bar to allow people read this data faster and see differences better. It is important to show that there are several groups for the heigh and weight, but the BMI shows only to main mode ranges.

This whole story helps us convey the idea that BMI is an important characteristic that can help group players that are top performers and at the same time we could find that the performance of player is a logarithmic model between average batting rate and home runs. While we can see that there are more R players which translates in more R top performers. Nonetheless, out top performer is an L player who seems to had the potential to even performed better if he had perfectly followed the logarithmic model. All in all, the findings are introduced slowly in the story and the final dashboard wraps it all. On top of that it adds interactivity which can be used to explore beyond the dataset with the filters and even using the visualizations as filters itself.

Feedback

During the project, the first challenge was to get some insight with the data and try to explain it in simple visualizations. The feedback was very helpful to understand how my audience perceived my ideas and it helped me to enhance the visualizations.

Initial Design Feedback

The first attempt to show some insight was a simple dashboard with a lot of information, that was somehow redundant and the insight was obscure:
https://public.tableau.com/profile/arturo.parrales.salinas#!/vizhome/BestPlayersHomeRuns/Dashboard1?publish=yes

The main feedback was to make not only one dashboard, but to split it or make a story with simple visualizations and build until the dashboard which could show the all the findings at once to reaffirm the insight. That made me think I should think how to convey more info in every plot I had in the dashboard, so I asked for more feedback on each graph.

The packed bubbled plot **Player Handedness** received good compliments as it was nice and it was not as boring to look at and understand as a bar chart would be. This definitively was a good visualization to convey a simple idea.

Then, the weight and height scatter with the handedness bubbles was extremely criticized since I already had heigh and weight bar charts, and the previous Player Handedness bubble plot. This was considered as junk in the dashboard since it was adding no value. I decided to remove it and better find a scatter plot that actually show a relationship of the performance of players. That's how I started to explore the average batting versus the home runs, which becomes one of the main plots to explain players performance against other players.

The next plots in the first dashboard attempt were using the average home run ratio batting variable which I was asked what it actually meant. I had to think for a while if I actually had done the right thing after I explained that it was my variable to merge batting rate and home runs. However, I realized that was not a good decision, so after that question I decided to eliminate that variable and use the average batting rate instead. That decision affected the treemap since I had to use home runs instead.

After that the average home run ratio batting variable was removed, my friend suggested to replace the bar chart of the BMI with a BMI's histogram since it can convey how the players distribution looks like. Thus, I agree it can easily show what is the most common BMI among the filtered players based on the treemap.

The last two plots of height and weight versus the average home run ratio batting, I just needed to substitute the latter for average batting rate.

The dashboard right away looked so much more better, but it was time to create a story which I first explain to my friend and he agreed with me. The story is basically described in the design section above. However, as a quick summary, it starts with the size of handed groups. Then, it shows how each handed group has a bimodal BMI distribution. Then, we prove the logarithmic performance of players using average batting rate and home runs. After that, we show how top performer players still have a bimodal BMI between 23.5 and 24, and 25 and 25.5 kg/m^2. Finally, we show the top performers players and their statistics, which basically show that the previous BMI ranges are actually common among top performers. Thus, BMI can be a good metric and players can build muscle or lose it to fit in such BMI as see if they perform better. All in all, the story was made and few details to the final dashboard were added in the final design.

Pre-Final Design Feedback

After all the feedback I had a clean story, but I decided to add features to the final dashboard in the story. I basically added a new filter to filter top players and not only stay with the top 20 home runners. Also, I added a highlight linking all graphs in the dashboard, so that when a user hovers over a player, its data is highlighted in the other visualizations. This gave a better interaction and made the insight easier to perceive. One can find this dashboard and the story in the following link:
https://public.tableau.com/profile/arturo.parrales.salinas#!/vizhome/HomeRunnersStatsDashboard/HomeRunPlayers?publish=yes

Final Design Feedback

After one review by my audience, I got more feedback on how to better convey the findings in the story. Basically, I was asked to add notes to each slide to highlight my insights. Also, there was a suggestion to introduce the dataset in my very first slice.
All this feedback was taken into account and while adding it to my story, I had an idea on how to improve the **Player Handedness** packed bubble visualization. I added the percentage of the total as a label and it was nice to see the proportions of the bubbles with numbers as well. After that I noticed that the **Top Players BMI** visualization would be greatly enhanced to convey the insight on BMI and handedness of top players if I added the **Player Handedness** visualization to it. Thus, I added it.
It all came together nicely and now the story is easier to follow and understand the insight.
Once can find this final story here:
https://public.tableau.com/profile/arturo.parrales.salinas#!/vizhome/BaseballPlayerCharacteristics/HomeRunPlayers?publish=yes

Resources

- Tableau documentation. https://www.tableau.com/support/help
- BMI website. https://www.nhs.uk/common-health-questions/lifestyle/how-can-i-work-out-my-body-mass-index-bmi/