# Data visualization – What makes a good baseball player?

By Michael Bong

## INTRODUCTION

I have created a data visualization aimed at answering the question: 'What makes a good baseball player?'. We have chosen Tableau as our visualisation tool as it is well suited for both exploratory and explanatory analysis.

The links to the first and final visualisations hosted on Tableau Public can be found below:

- **First visualisation**:
  - https://public.tableau.com/profile/michael.bong#!/vizhome/First_DataViz/Story1
- **Final visualisation** *(Best viewed in full screen)*:
  - https://public.tableau.com/profile/michael.bong#!/vizhome/Final_DataViz/Whatmakesagoodbaseballplayer

## SUMMARY

The data visualization was created with the aim of answering the question: 'What makes a good baseball player?'. We define a 'good baseball player' as those who excel at putting points on the board, and come in one of 2 forms:

1. 'Specialist' with good batting average and/or good home runs,
2. 'All-rounder' with both good batting average and good home runs – someone who can do everything well

Multiple relationships between a player's physical attribute (height, weight, handedness) and their performance (batting average, home runs) are explored to discover insights.

*In conclusion, we find that the best players, also known as the 'All-rounders' have a height range between 70-76 inches (72.7 inches on average), a weight range of 160-220 pounds (187.3 pounds on average), and right handedness.*

## DESIGN

Both the initial and final (post feedback) design decisions are explained below. The focus is on visualizations that made it to the story.

### INITIAL DESIGN

Quite a few visualizations were created initially and added to the story.

Sets are first created. A set of top 100 batters and the top 100 home runners are created. Players in either and both these sets are labelled as 'Specialists'. Players who are in both these sets are labelled as 'All-rounders'.

In all these visualizations, the maximum point is labelled. The color scheme used is color-blindness friendly and uses the blue-orange palette.

Bar charts/Histograms were used to show the distribution of height, weight and handedness. Average batting and home runs are also added as bar charts under the distribution plots. 'Color' is used to show the high and low points for average batting and home runs.

Scatter plots are then used to explore the relationship between player height and weight, and between batting average and home runs. 'Color' encoding is used to show the different handedness of players. Linear trend lines are added to show the direction of the relationship. Average lines have been added to enable us to identify above average/below average players in terms of size and performance.

Line charts are created to show the relationship between performance and height and weight. Linear trend lines are added to show the direction of the relationship.

The dashboard was created to enable users to interact and view multiple measures on a single sheet.

A story was created to walk readers through the process of answering the question 'What makes a good baseball player?'.

## FINAL DESIGN

The feedback provided in the 'Feedback' section below has been taken on board and multiple design changes have been made.

Color encoding was made consistent across the board. Specifically, color encoding has been introduced to clearly show which players fall into which player groupings for all visuals except for distribution plots.

Data ink ratio is increased by removing the player name labels from the scatter plots as they do not add value.

A table called player stats is added to enable users to quickly find the top batters and top home runners and see which grouping they fall under.

The player averages table is added to provide users with a quick gauge of the average height, weight and performance of each player grouping.

More user-friendly player groupings have been created as follows. A set of top 100 batters and the top 100 home runners are created. Players in either and both these sets are labelled as 'Specialists' (Orange). Players who are in both these sets are labelled as 'All-rounders' (Blue). This grouping helps us to focus on these top performing players to determine what physical attributes to look out for when recruiting the next top player.

Bins were created to aggregate all height and weight records as using the raw height and weight measures resulted in volatile trends. Height is aggregated in bins of 2 inches while weight is aggregated in bins of 10 pounds. This has resulted in a smoothing of performance charts against height and weight.

In the dashboard, multiple changes were made. Firstly, the dashboard was a given a descriptive name. Secondly, the visuals on the dashboard are rearranged and only the relevant visuals that add value in deriving answers to the question 'What makes a good baseball player?' remains. Only the relevant filters are kept in the dashboard. Additional plots to address missing information such as performance by handedness are introduced to the dashboard.

In the story, descriptive narratives of each step are now provided to help readers quickly hone in on findings. Only the steps/findings that add value remain.

# FEEDBACK

A list of feedback provided by a trusted partner was recorded and the actions taken can be found below:

1. Are name labels needed in scatter plot?
    a. Name labels have been removed from the scatter plot (showing relationship between batting average and home runs)
2. Do we need distribution information in the dashboard?
    a. Distribution information provides good insights into the makeup of top player groupings we are trying to replicate/recruit in the future. Therefore, distribution information is kept in the dashboard, with changes made to only show the relevant bits of information (resulting in removal of performance information).
3. How is the performance by handedness represented in the dashboard?
    a. It was not represented previously. It is now represented with a plot on the dashboard.
4. Do we need that many filters in the dashboard?
    a. There were many irrelevant filters. Only 2 filters, 1 to filter by 'Specialists' and another to filter by 'All-rounders' remain in the dashboard.
5. In the scatter plot showing relationship between batting average and home runs, how do I know which points (players) fall into which set (type of player group)?
    a. Color encoding has been introduced to clearly show which players fall into which player groupings for all visuals except for distribution plots.
6. How do I know what this story/dashboard is covering without a proper title?
    a. The story and dashboard have been provided descriptive titles.
7. Can we please change the narrations in the story to be more descriptive and focus on explaining insights?
    a. The narrations in the story have been updated to be descriptive and highlight the insights that have been derived.
8. Is the relationship between players height and weight relevant in this analysis?
    a. This relationship is irrelevant to determining what is the makeup of a baseball player with high performance, and thus has been removed.
9. For both height and weight, the performance charts have too much fluctuations, can we smooth it?
    a. The performance charts for both height and weight has been smoothed using bins. Height is aggregated in bins of 2 inches while weight is aggregated in bins of 10 pounds.
10. The sets and color encoding are quite confusing and needs to be consistent.
    a. The color encoding is now consistent across all visuals and plots.

# RESOURCES


https://interworks.com/blog/rcurtis/2016/10/31/tableau-deep-dive-sets-computed-sets/

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

https://docs.google.com/document/d/1w7KhqotVi5eoKE3I_AZHbsxdr-NmcWsLTIiZrpxWx4w/pub?embedded=true