DATA VISUALIZATION

UNIT – IV (PART – I)

Proportion & Percentage

- A proportion is a ratio in which the numerator is a partial amount, and the denominator is the total amount (expressed as a number between 0 and 1).
- For example, the proportion of the NYC population living in the Bronx is 0.169. A proportion is expressed as a number between 0 and 1.

- A percentage is a ratio comparing a number to 100.
- For example, 16.9% of NYC residents live in the Bronx. A percentage is generally a number between 0 and 100, but can be larger than 100 (e.g., "sales have increased by 150% year-over-year").

A step in the process

- They are most often used to communicate three different types of comparisons:
    1. Part-to-whole
    2. Current-to-historical
    3. Actual-to-target

Proportions and Percentages

1. Part to whole
2. Current to historical
3. Actual to target
4. Mean and Median

Part to whole

- A part to whole analogy is defined as a comparison between a part and a whole of one thing and how it is like a part and a whole of another thing

- For example, Fin : Fish :: Wing : Bird is a part to whole analogy and works because it demonstrates that just as a fish uses a fin to move, so too does a bird use a wing to move.
- Followers of the sport of baseball have long been fascinated with proportions of outcomes (called stats).

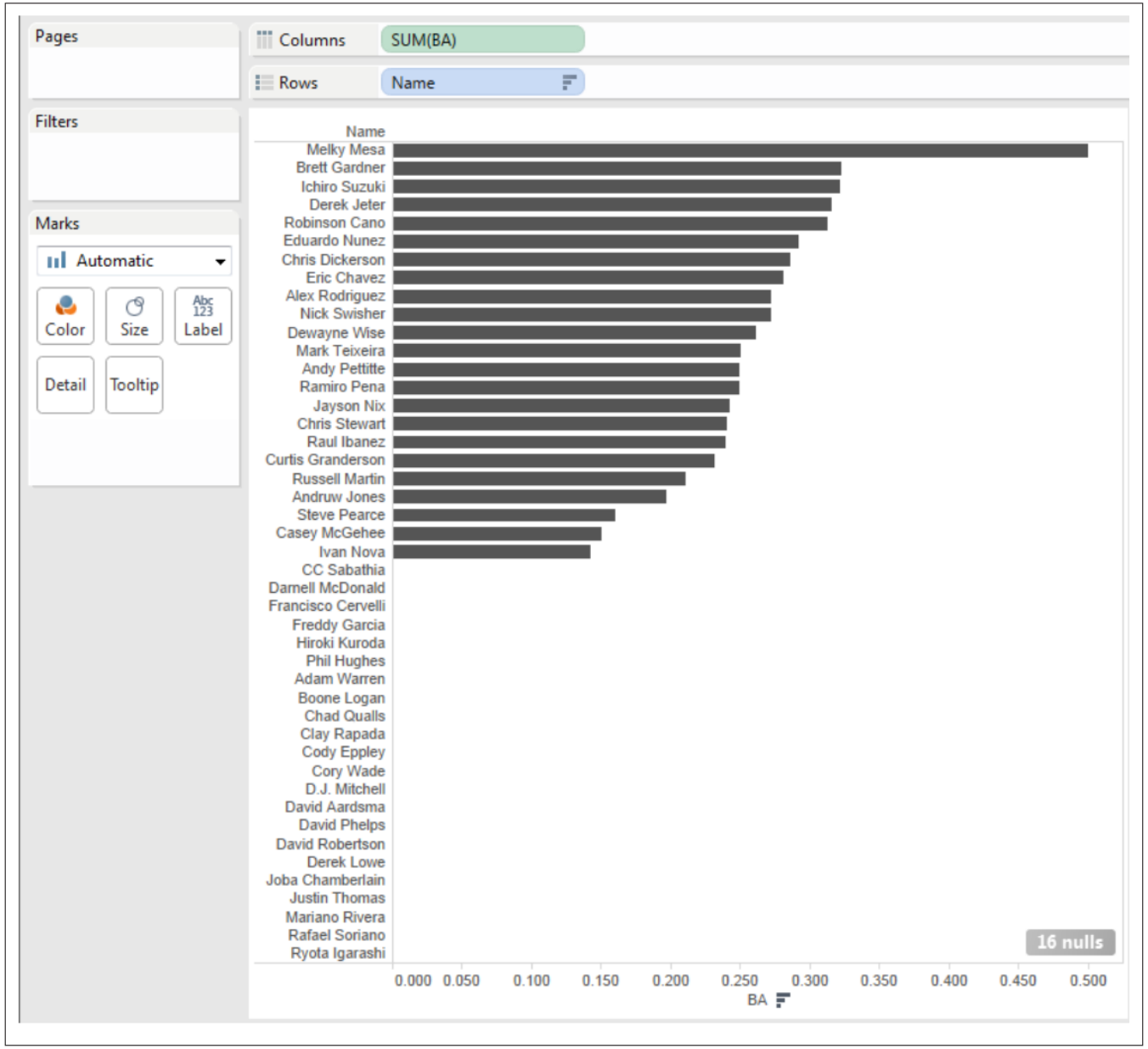


Figure 5-1. Bar chart showing BA for every Yankee with at least one at bat in 2012

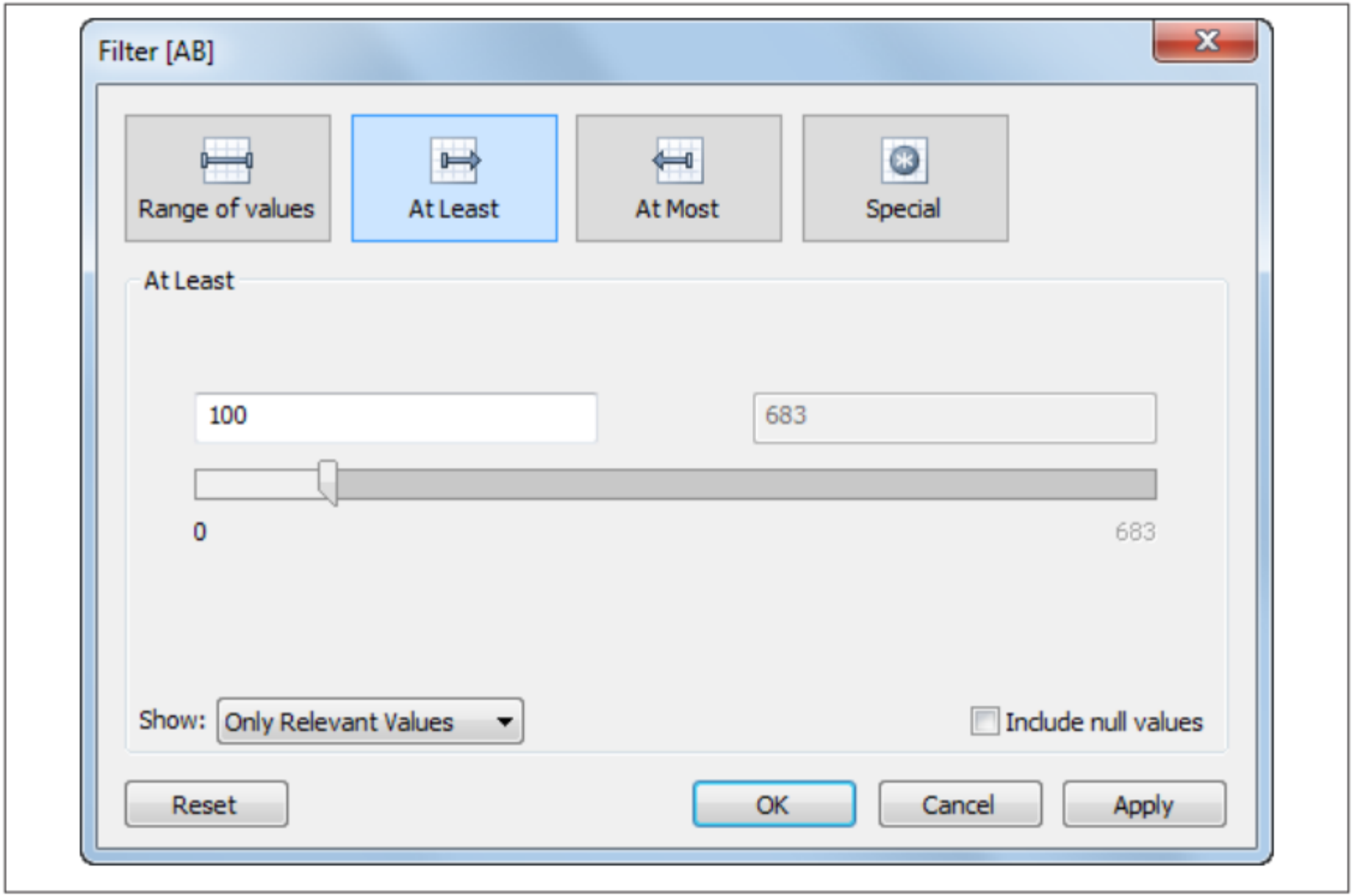


Figure 5-2. Filtering the BA bar chart to show only players with at least 100 at bats

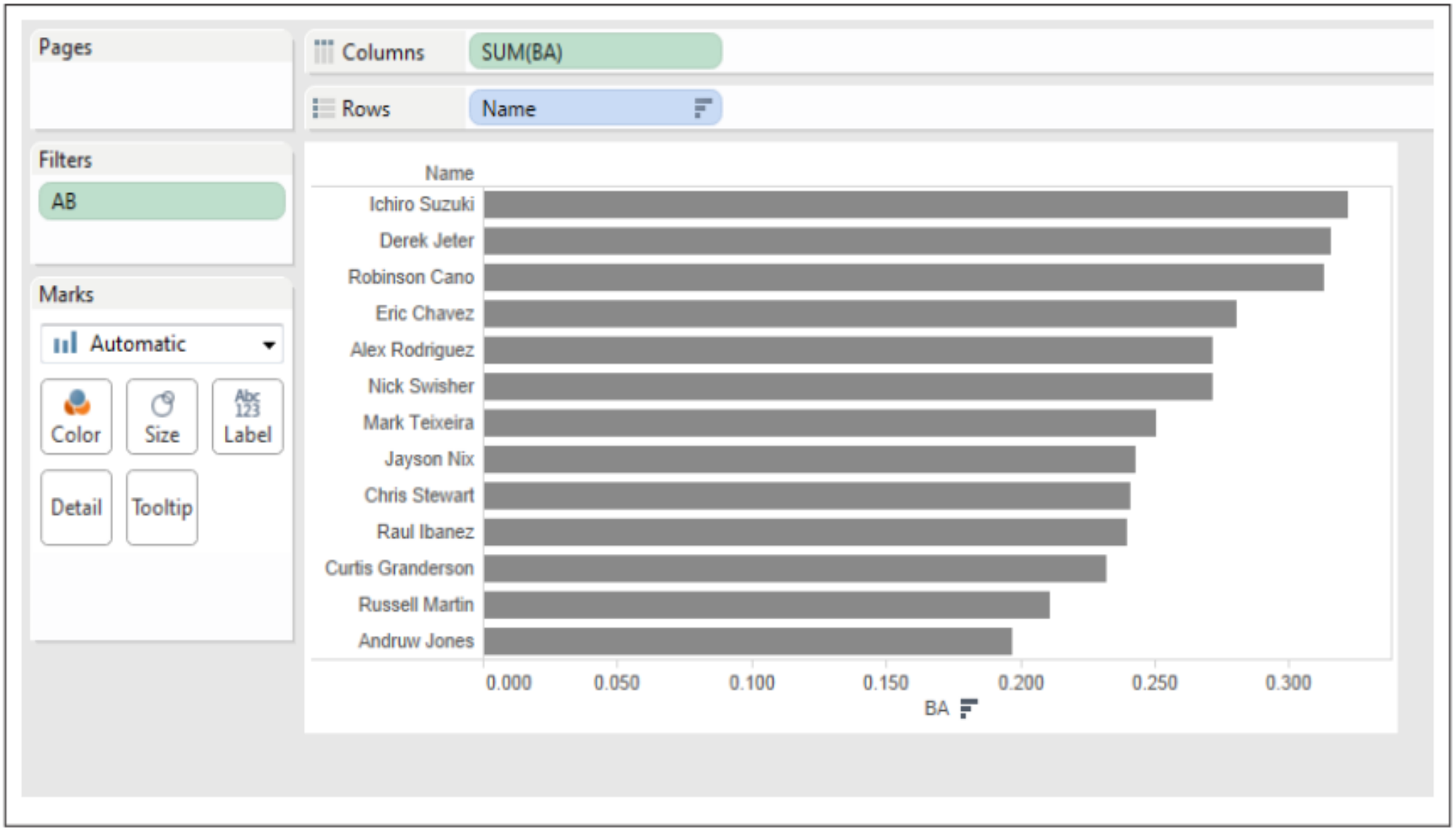


Figure 5-3. The filtered BA bar chart

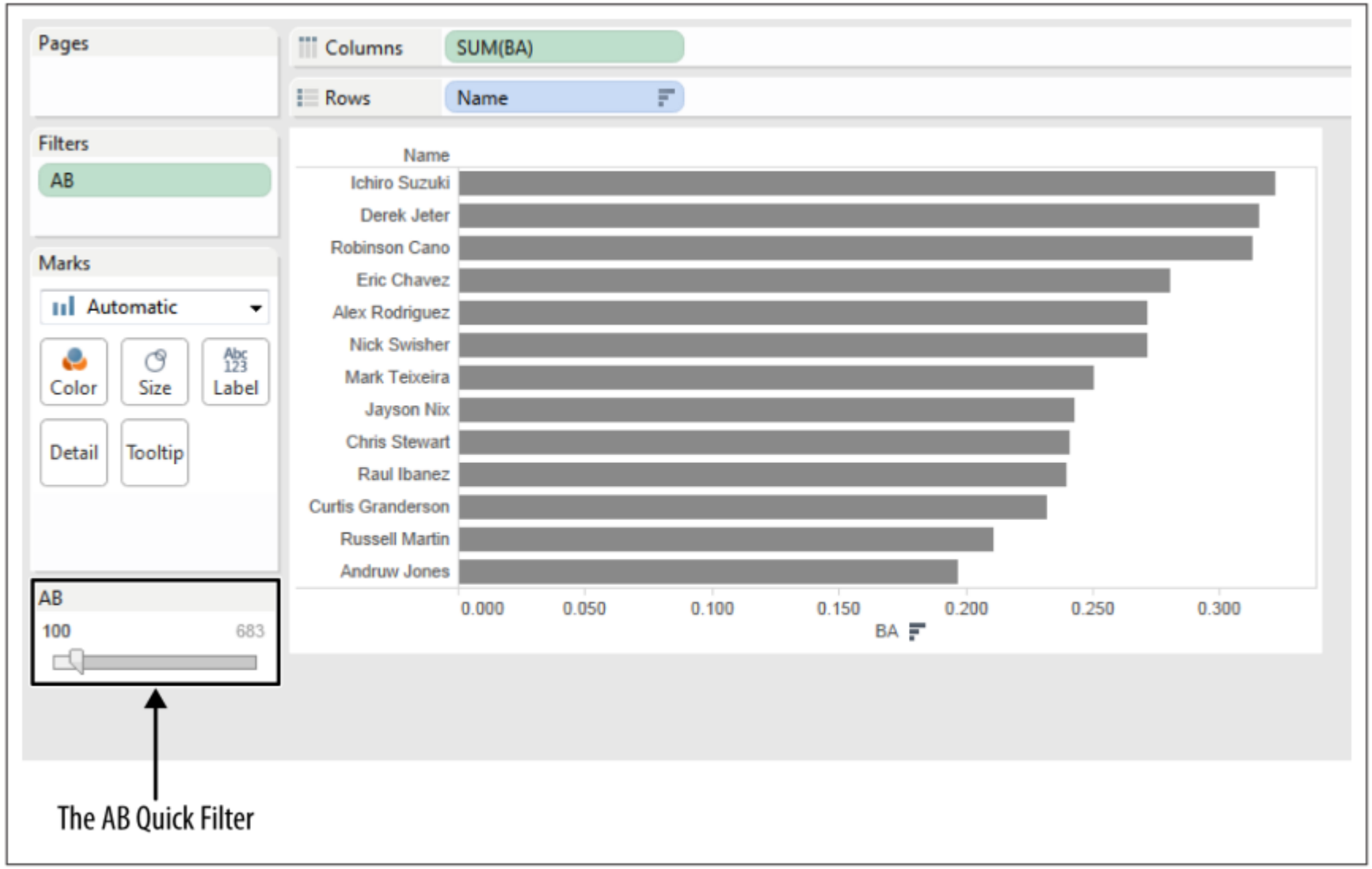


Figure 5-4. The AB Quick Filter added to the Sheet

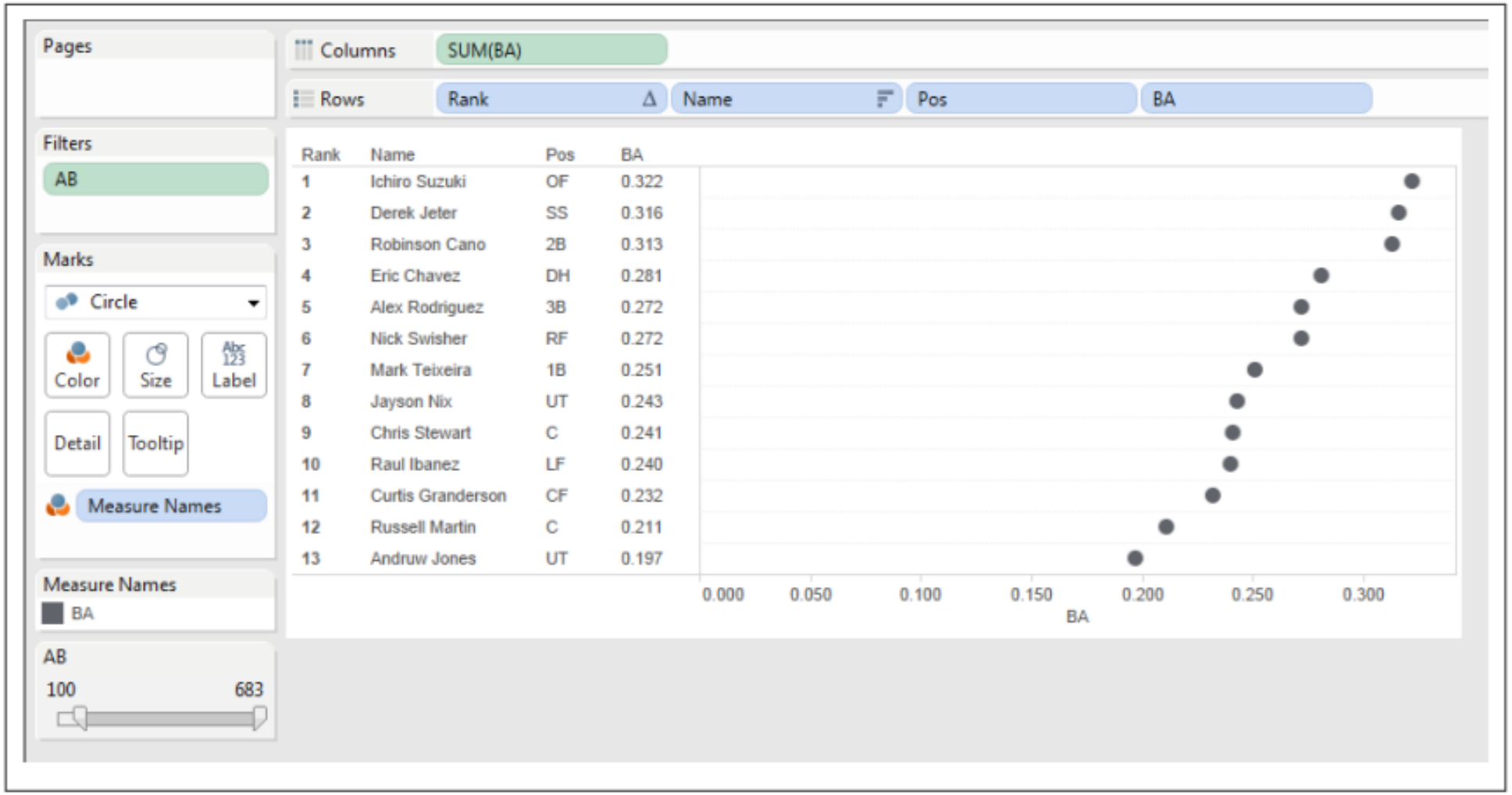


Figure 5-5. Batting average dot chart

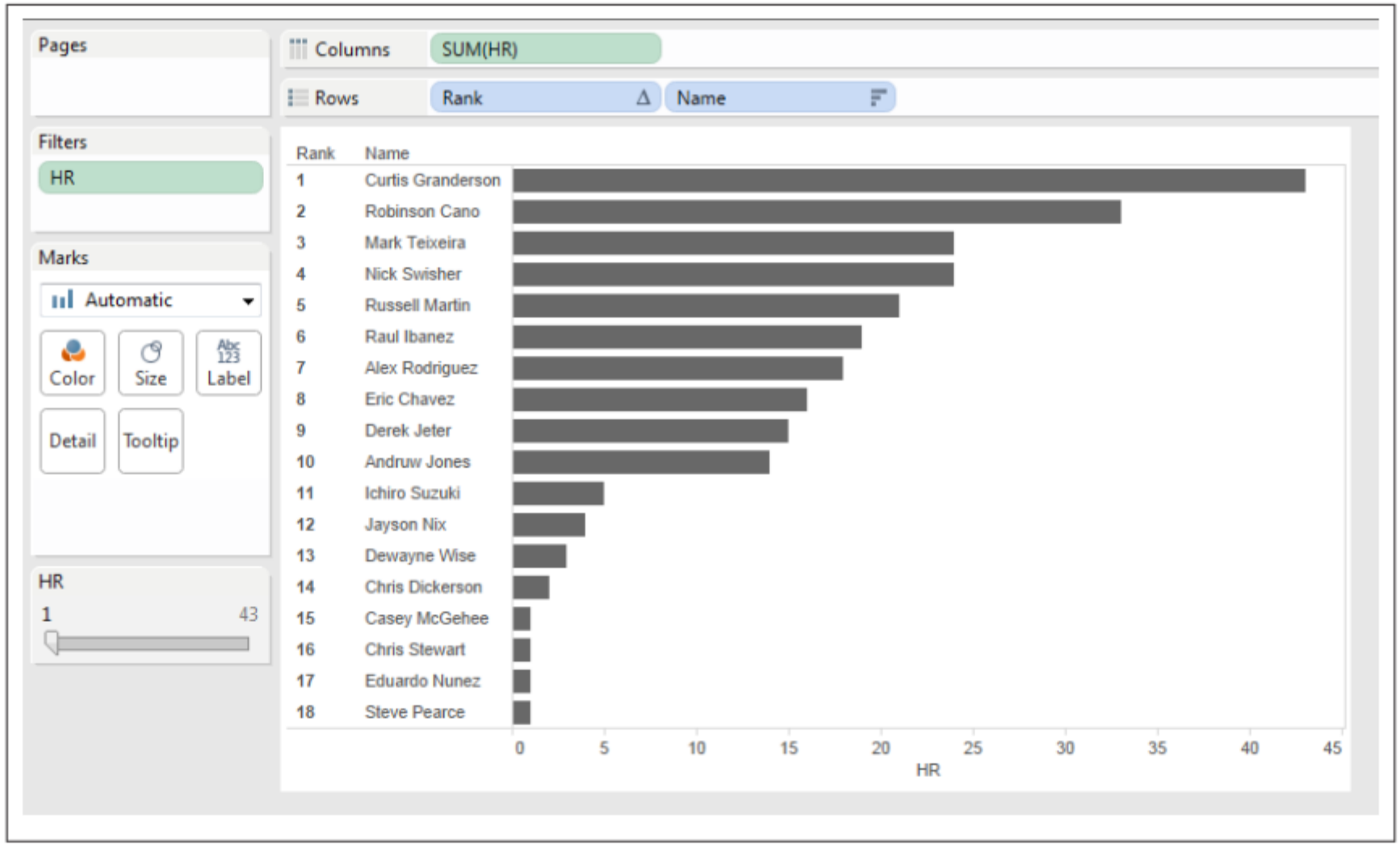


Figure 5-6. Bar chart showing Yankees in 2012 with at least one home run

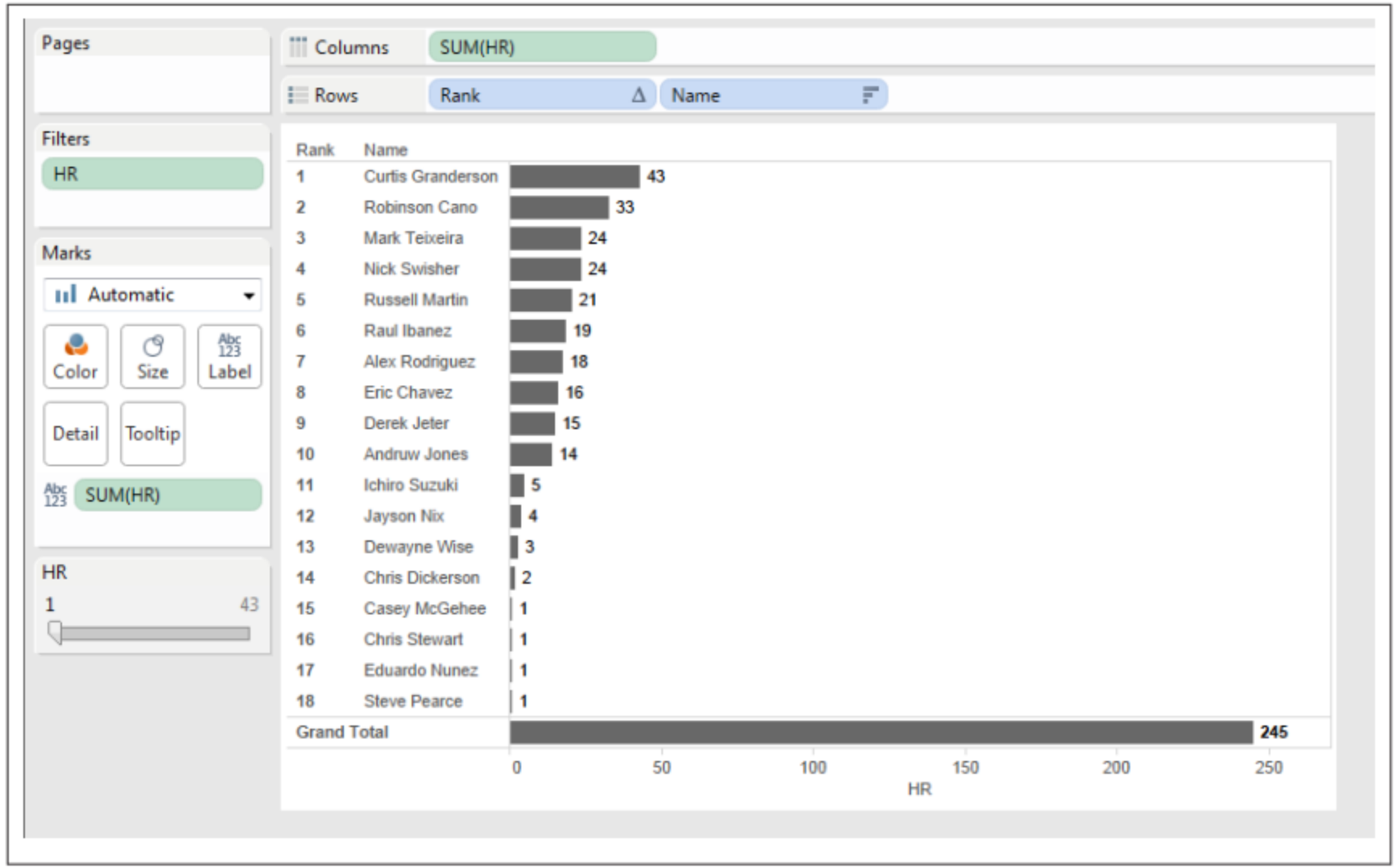


Figure 5-7. Home run bar chart with team total added

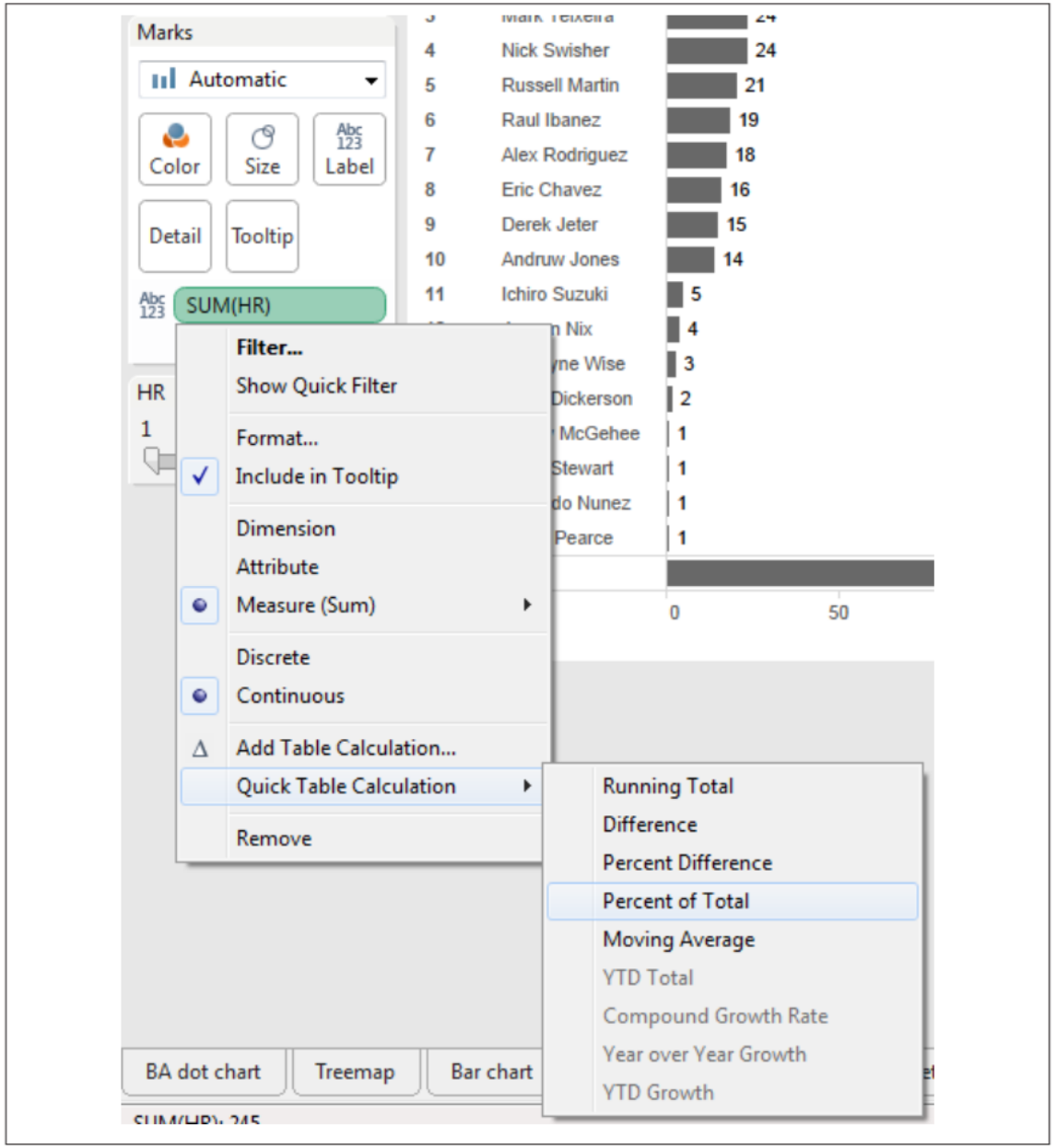


Figure 5-8. Changing the label to a "Percent of Total" table calculation

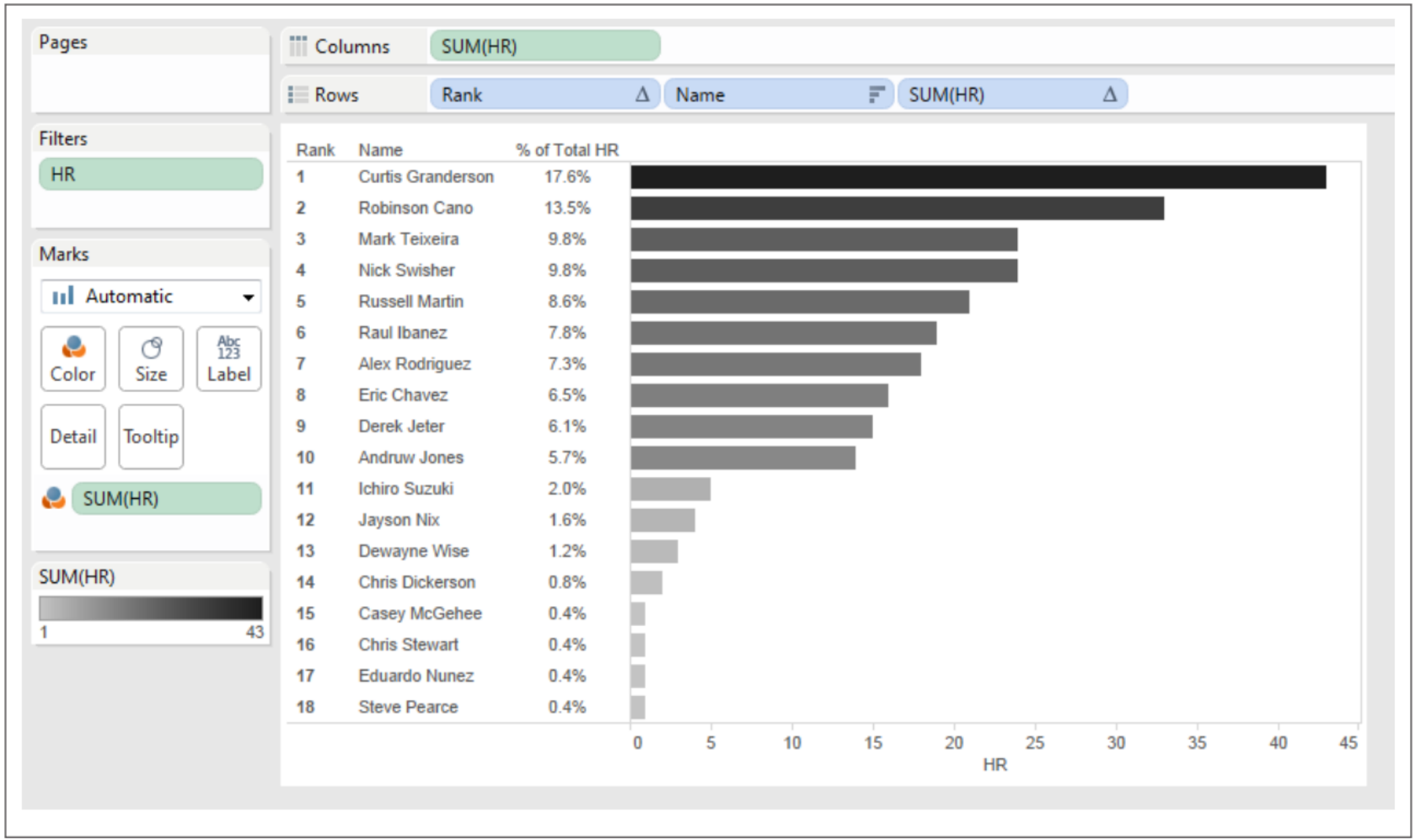


Figure 5-9. Home run tallies with labels shown as Percent of Total

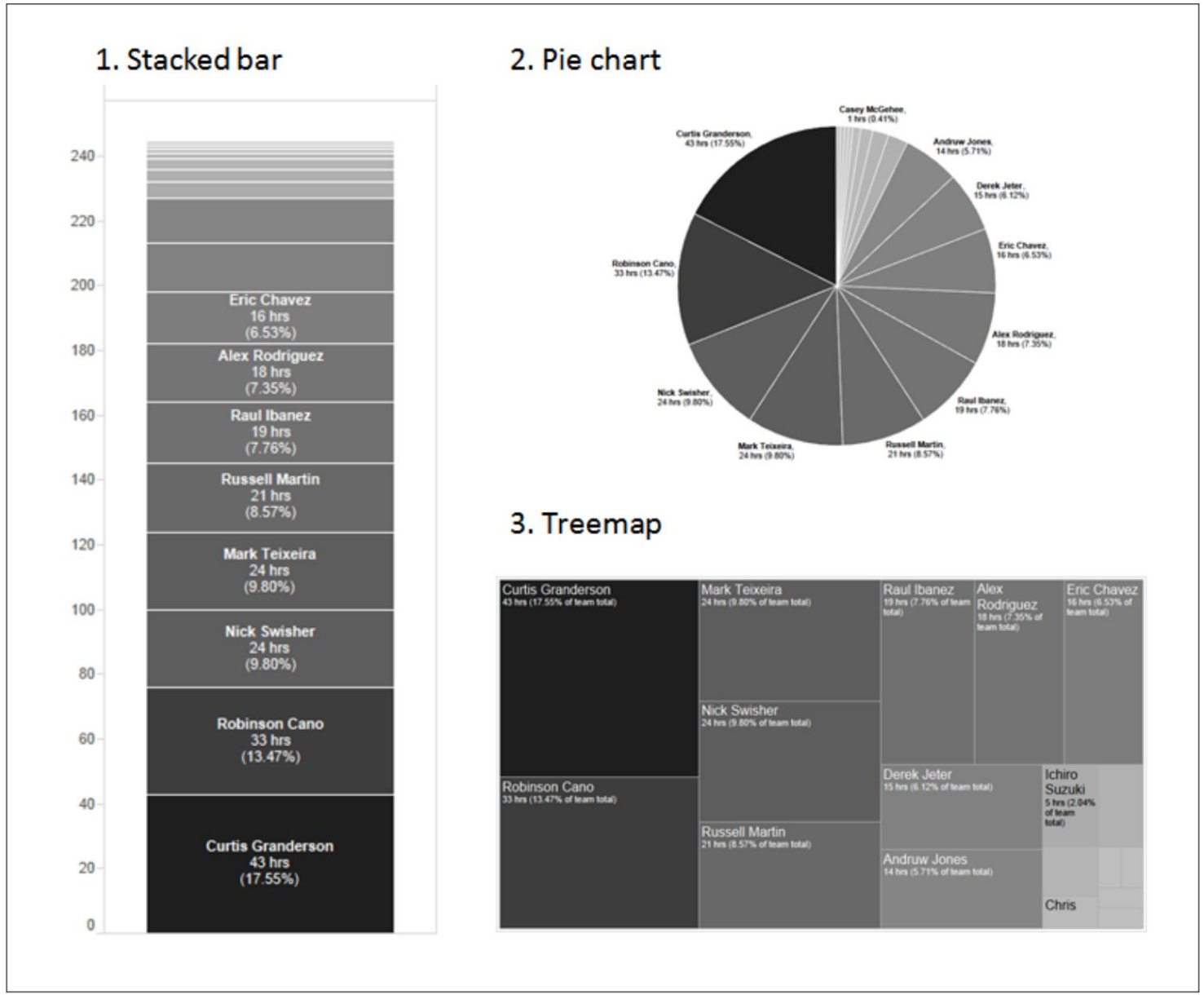


Figure 5-10. Three alternative ways to show the proportion of home runs hit by each player

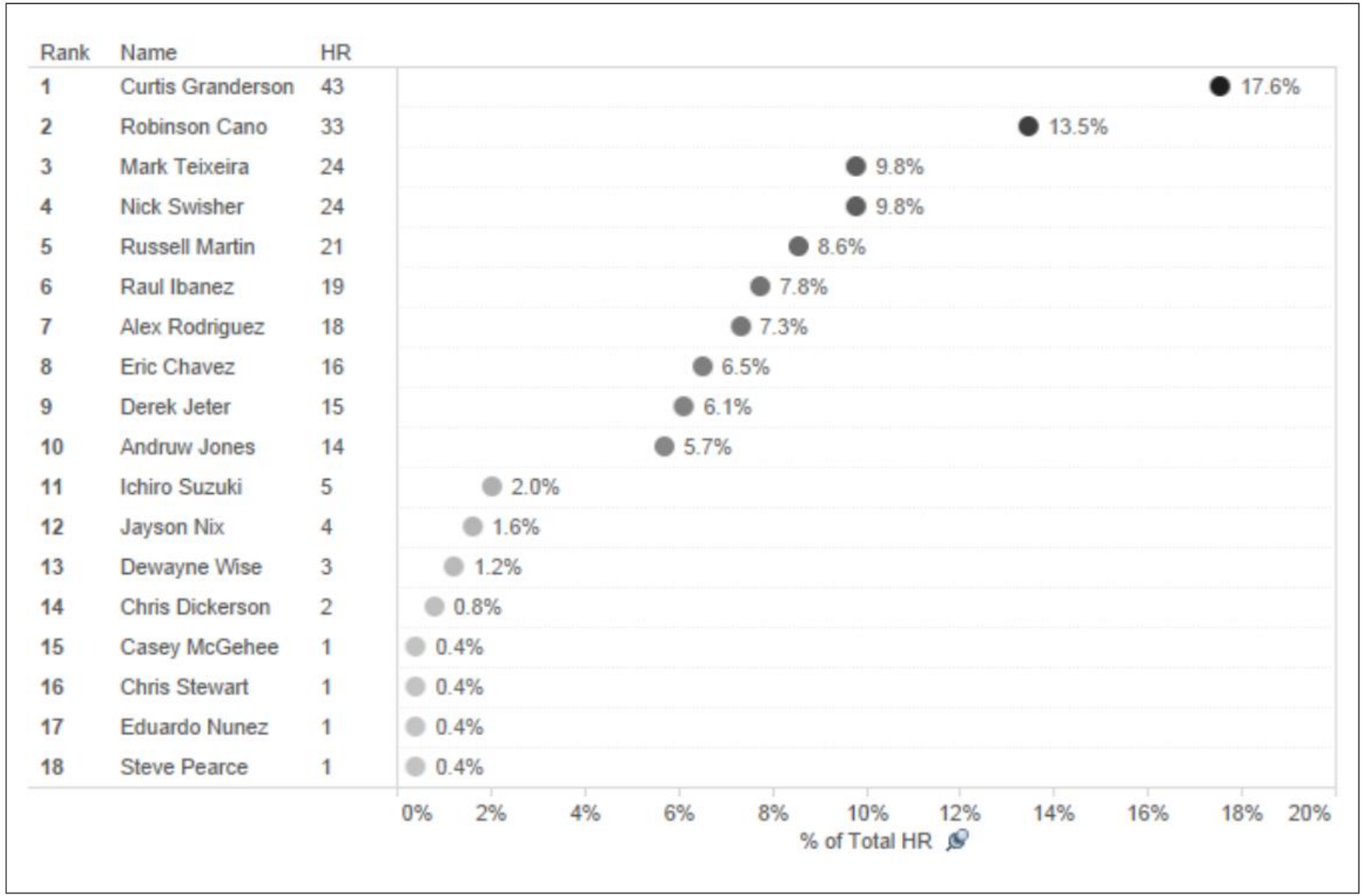


Figure 5-11. Dot chart of percentage of home runs contributed by each player

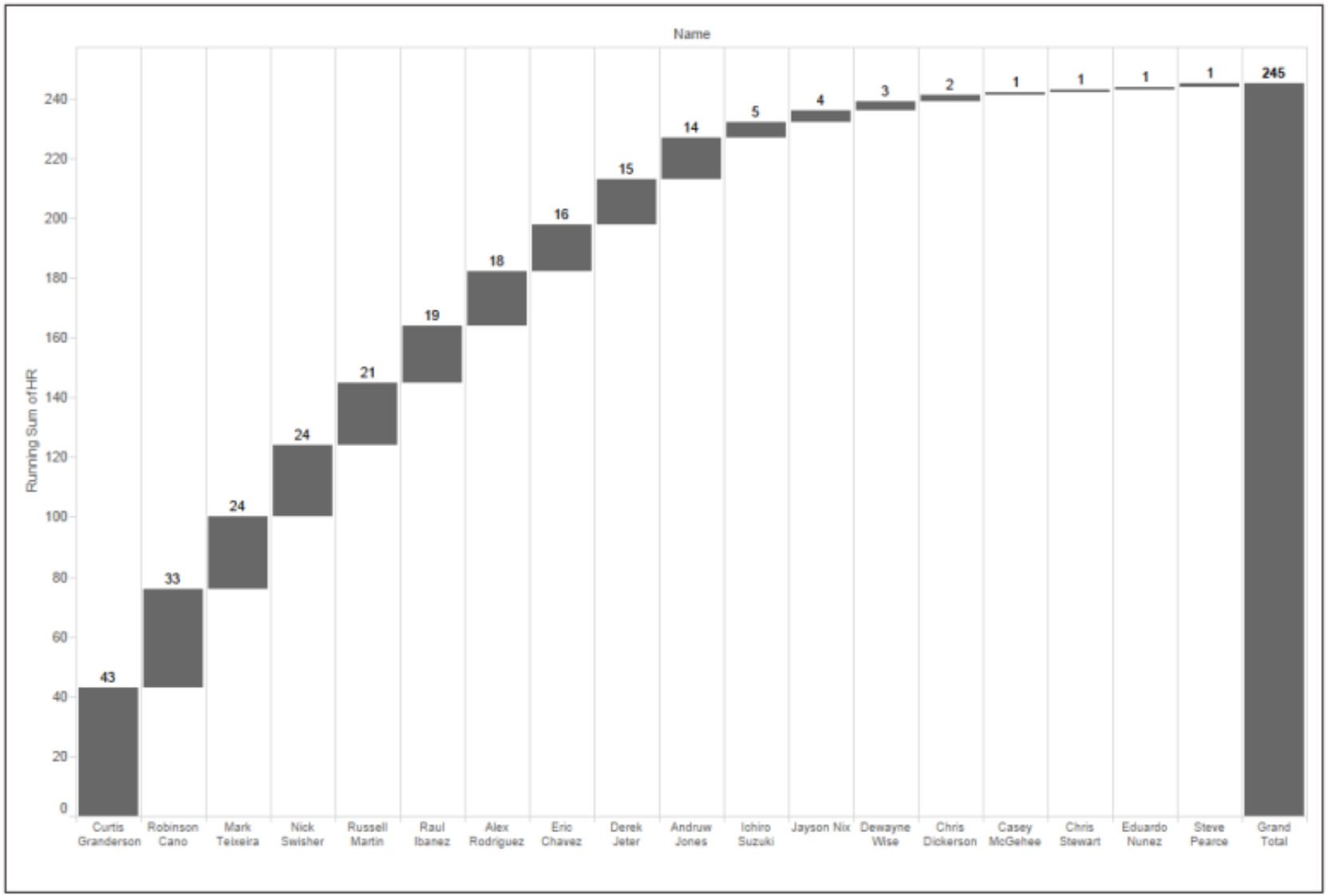


Figure 5-12. Home run data as a waterfall chart

Proportions and Percentages

1. Part to whole
2. Current to historical
3. Actual to target
4. Mean and Median

Current to historical

- New York hit 245 home runs in 2012.
- But how did that home run tally compare with other teams in their league?
- how did it compare with the amount of home runs they hit the year before?

- Was it the most in the league, the least, or somewhere in between?
- Did they improve their home run tally over 2011 or did they fail to reach the previous year's mark?
- These questions send us in search of yet more data, and we find the 2011 and 2012 league team home run totals on the Web.

	A	B	C	D	E
1	Tm	City	League	2012 HR	2011 HR
2	BAL	Baltimore	AL	214	191
3	BOS	Boston	AL	165	203
4	CHW	Chicago	AL	211	154
5	CLE	Cleveland	AL	136	154
6	DET	Detroit	AL	163	169
7	KCR	Kansas City	AL	131	95
8	LAA	Los Angeles	AL	187	129
9	MIN	Minnesota	AL	131	103
10	NYY	New York	AL	245	222
11	OAK	Oakland	AL	195	114
12	SEA	Seattle	AL	149	109
13	TBR	Tampa Bay	AL	175	172
14	TEX	Texas	AL	200	210
15	TOR	Toronto	AL	198	186

Figure 5-15. Team home run totals, 2011 and 2012

Current to historical

The Bullet Graph:

- Let's go ahead and connect to the data table and create a bullet graph by Ctrl-selecting City, 2011 HR, and 2012 HR, and then choosing bullet graph from the Show Me panel. The resulting view is shown in Figure 5-16.

- In this initial view created from the Show Me menu, the length of the blue bars is determined by the 2011 HR totals, as evidenced by the green pill in the Columns shelf. The vertical black lines are the 2012 HR totals for each team, and the bands are at 60% and 80% of the 2012 values.

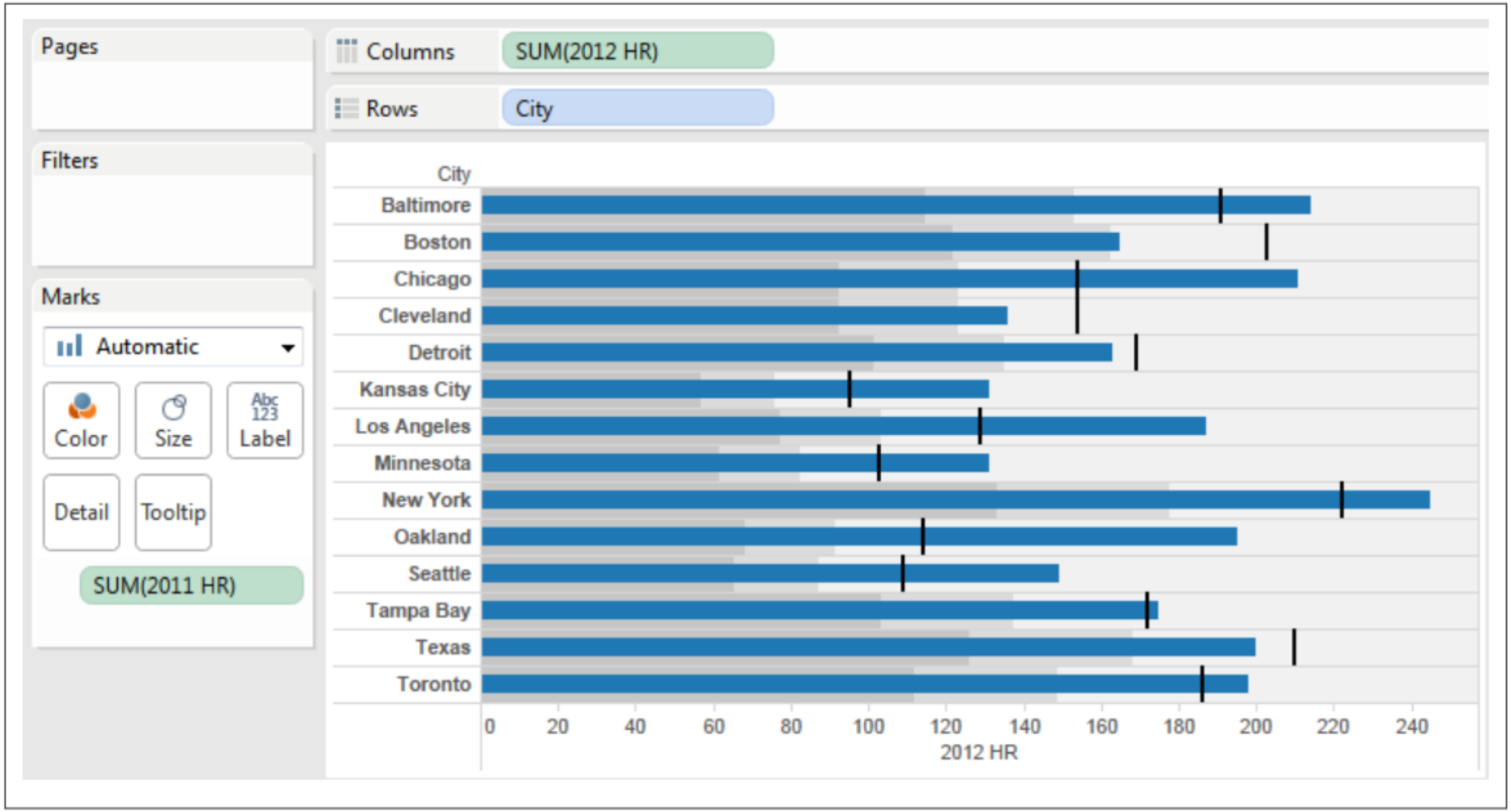


Figure 5-16. The initial bullet graph created by Show Me

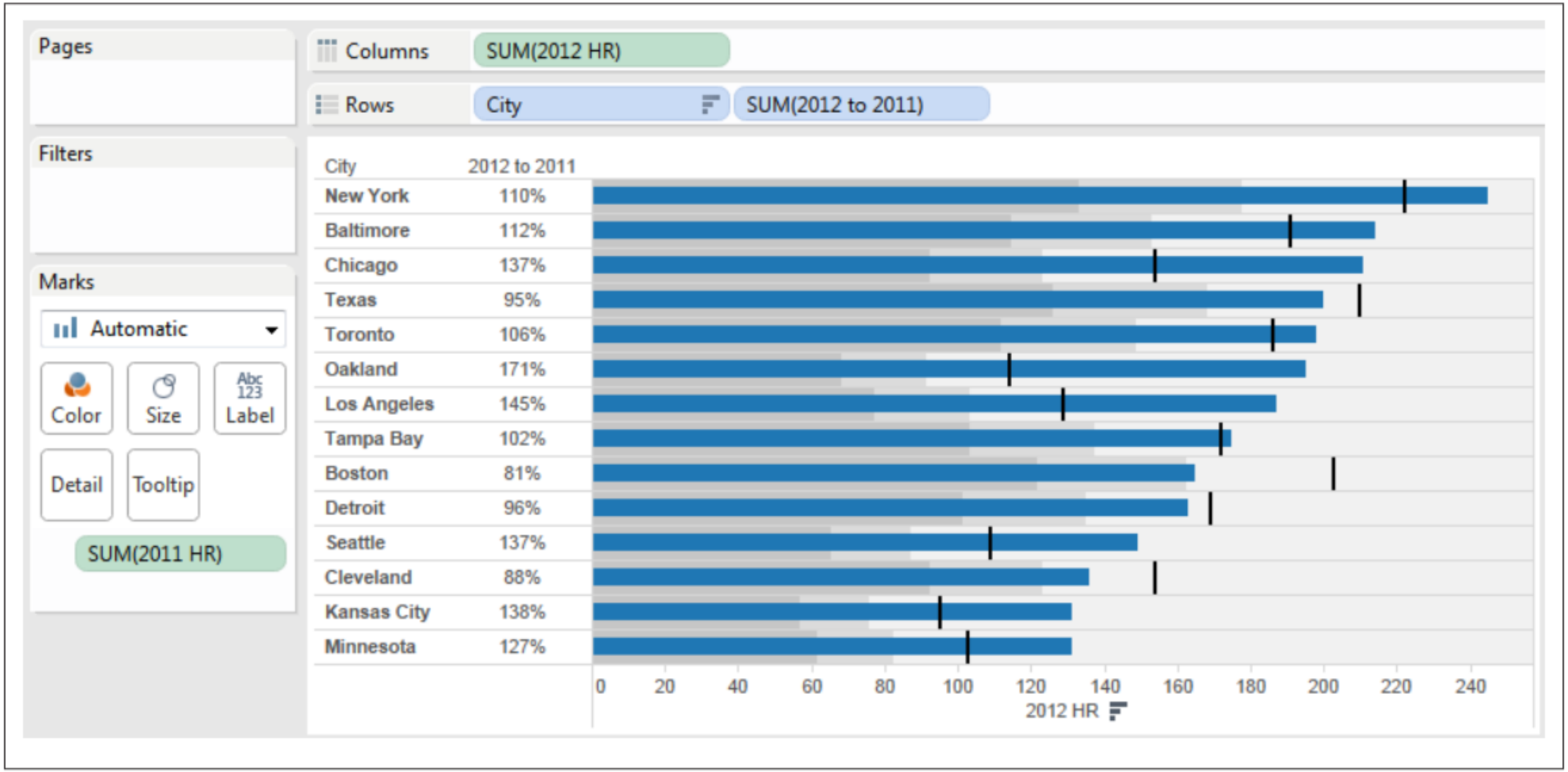


Figure 5-17. Bullet graph of AL team home runs, 2012 to 2011

Current to historical

How do we read the bullet graph?

- If the blue bar goes beyond the black line, then the team hit more home runs collectively in 2012 than they did in 2011.
- If the blue bar falls to the left of the black line, then the team hit fewer home runs, and their 2012 tally falls into one of three bands:
- The darkest gray band is for teams that hit less than 60% of their previous year total
- The next lighter gray band is for teams that hit between 60% and 80% of their previous year total
- The final band is for teams that hit between 80% and 100% of their previous year total. This time, we find

Proportions and Percentages

1. Part to whole
2. Current to historical
3. Actual to target
4. Mean and Median

Actual to target

- The world is full of people and teams with goals in the form of quotas, budgets, and performance targets. These figures are tracked and monitored religiously to determine “performance to plan.” Just listen to any monthly sales call.

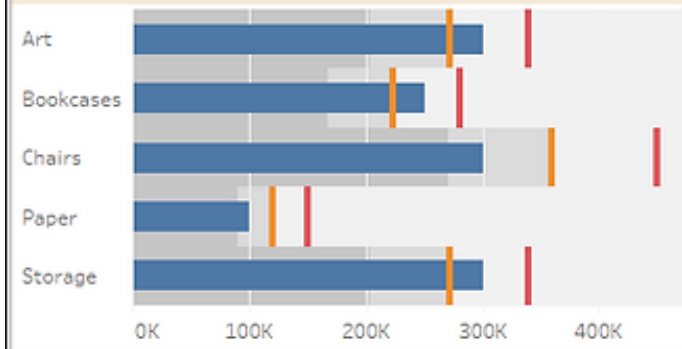
- How are we doing compared to where we want to be at the end of the year?
- How are we doing compared to where we should be right now if we want to hit our monthly goal?

CHARTING ACTUAL TO TARGET VALUES

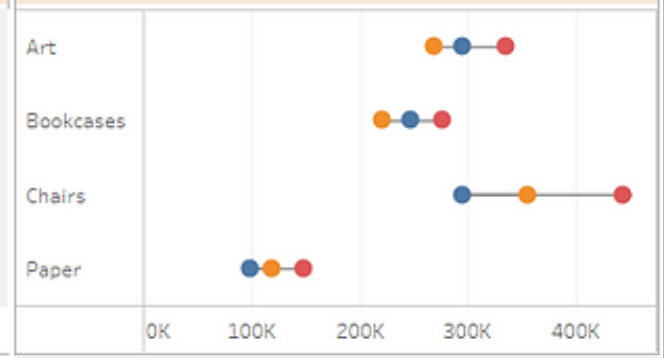
Monthly Target | Mtd Actual | Mtd Target

where Mtd = Month-to-date

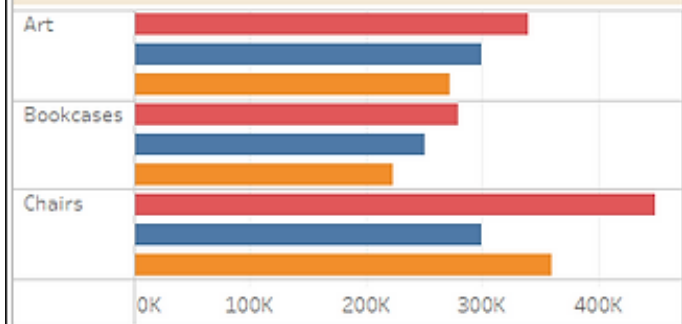
Bullet chart



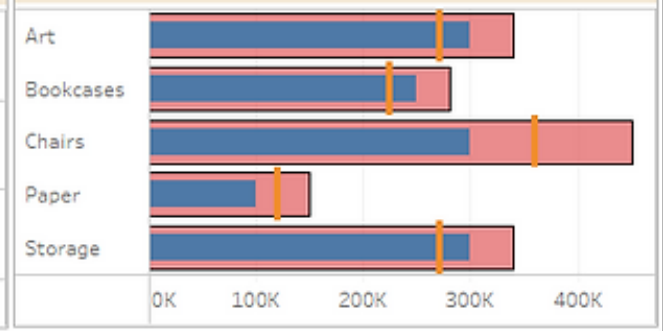
Dot plot



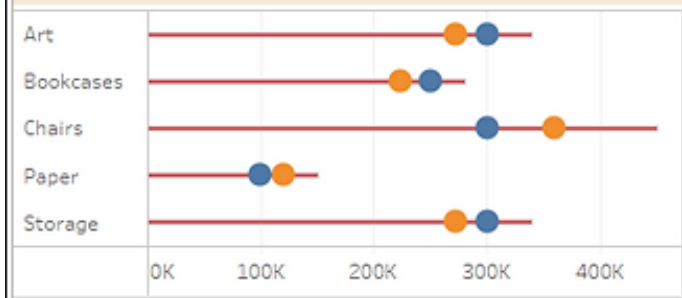
Side-by-side bars



Bar-in-bar chart + Reference line



Bar + dot plot



BANS

340,000

▲10.3% Above Mtd target

▼-11.8% Below Monthly Target

Proportions and Percentages

1. Part to whole
2. Current to historical
3. Actual to target
4. Mean and Median

Mean and Median

- The mean (or average) is determined by summing all the values in a data set and dividing by the number of values.
- The mean is considered a "representative value," meaning if you replaced each value in the data set with the mean, the overall sum wouldn't change.

- The median is the middle value in a data set in which the values have been placed in order of magnitude. Thus, half the values in the data set are less than the median, and half are greater.
- The mode is the most commonly occurring value in a data set.

Mean and Median

The Normal Distribution:

- The Gaussian, or normal, distribution, is something we've all been exposed to at some point.

- The unmistakable bell-shaped curve of the Gaussian represents a very mild style of variation, one in which the probability of a value occurring falls off dramatically the farther we move away from the mean on either side

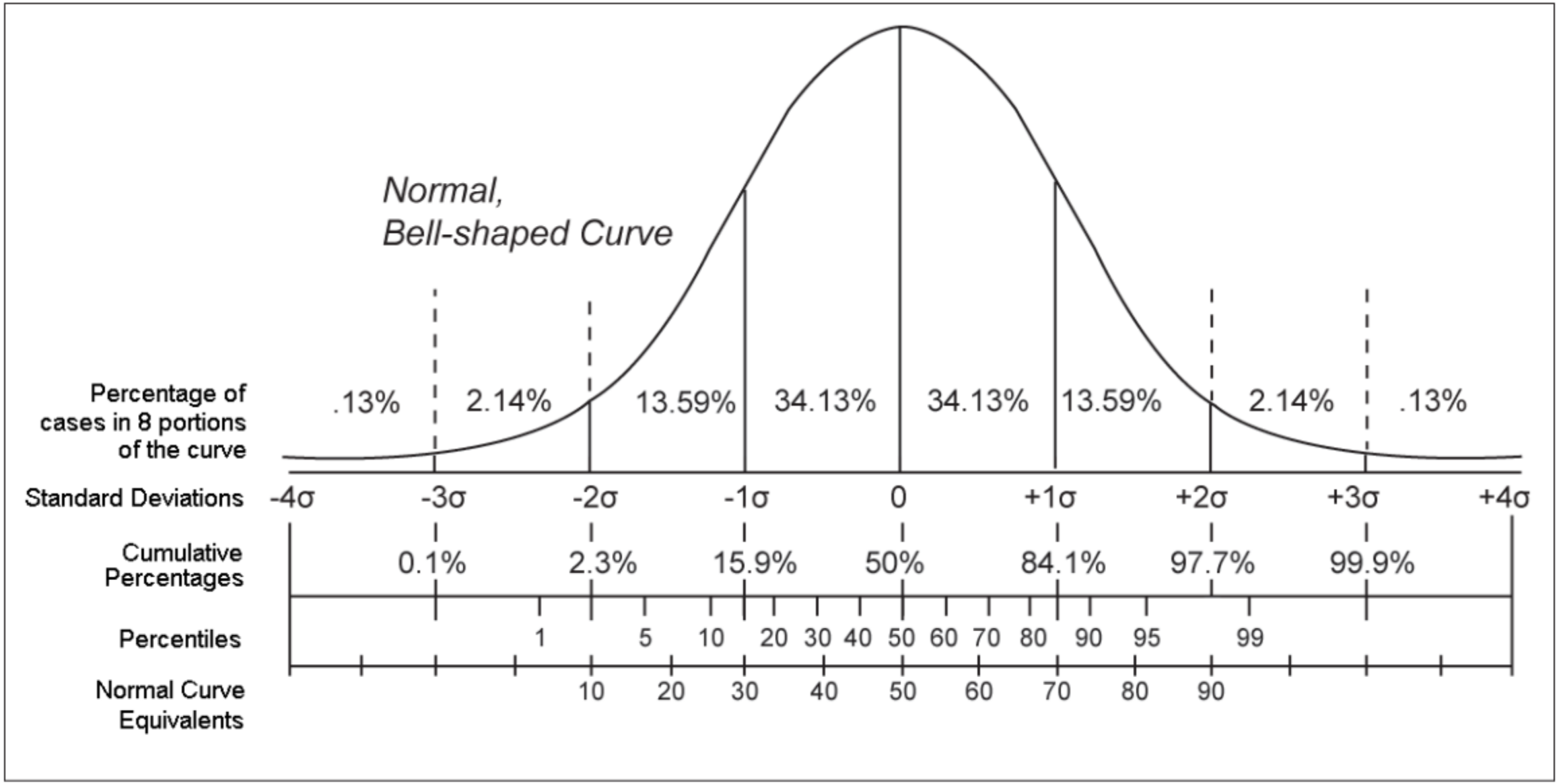


Figure 6-1. The characteristics of the normal distribution

Mean and Median

An Example of "Normal" Data:

- In baseball, a run batted in (RBI) is granted to a batter every time he enables a runner to score during his at bat. A batter can earn more than one RBI during a single at bat; a grand slam home run would result in 4 RBI

- We can create a histogram as before, to visualize the distribution of qualifying players' RBI during the 2012 season, as shown in Figure 6-2.

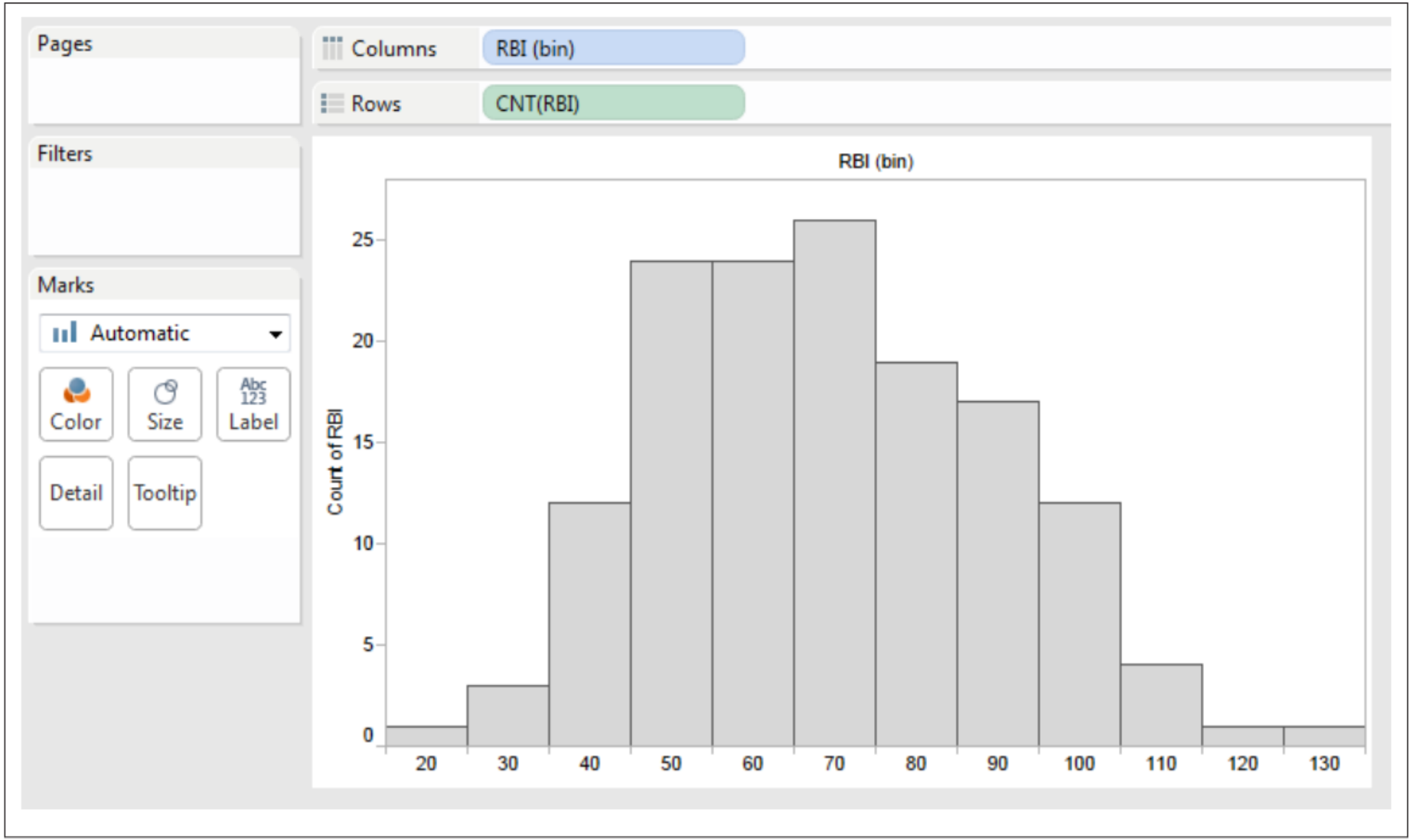


Figure 6-2. A histogram of players' RBI during the 2012 season

Mean and Median

Box Plots:

- If we Ctrl-click the Player and Pos (for "Position") Dimensions, and the RBI Measure, and then open the Show Me panel and select box-and-whisker plot, we get the chart shown in Figure 6-4.

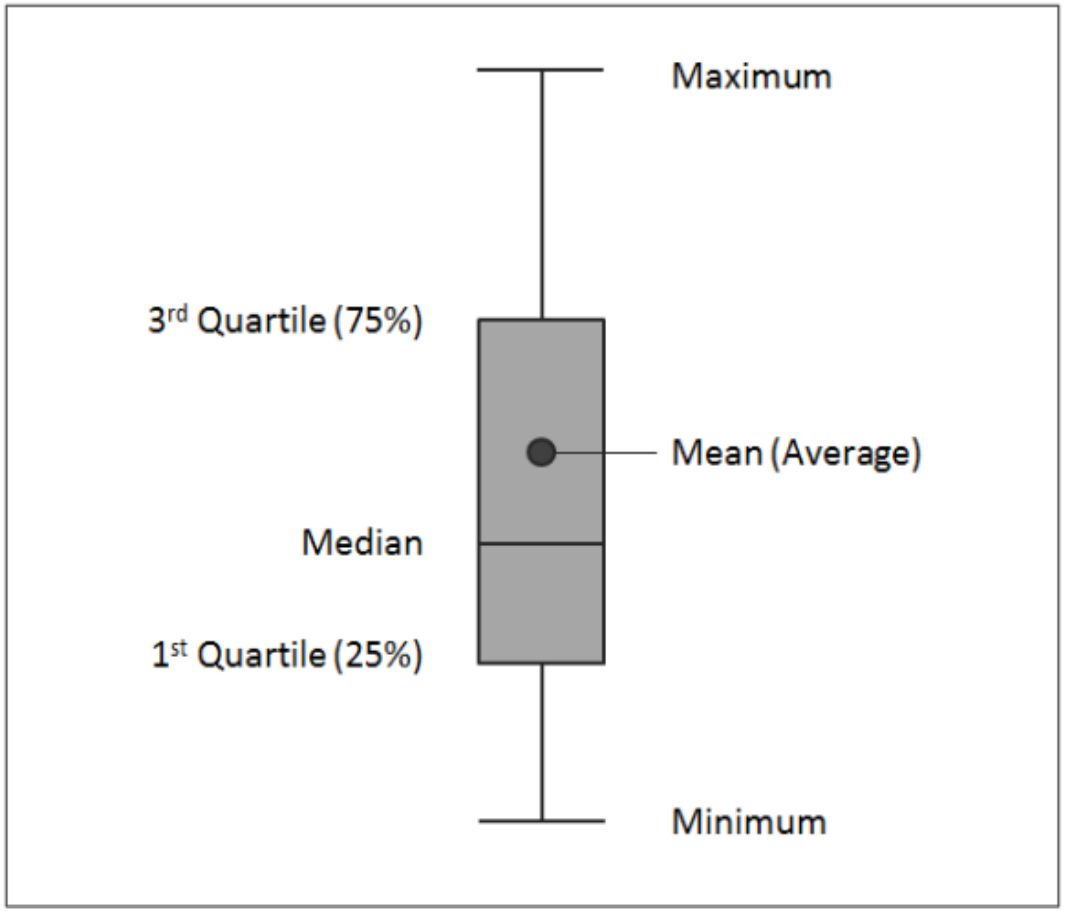


Figure 6-3. The box plot

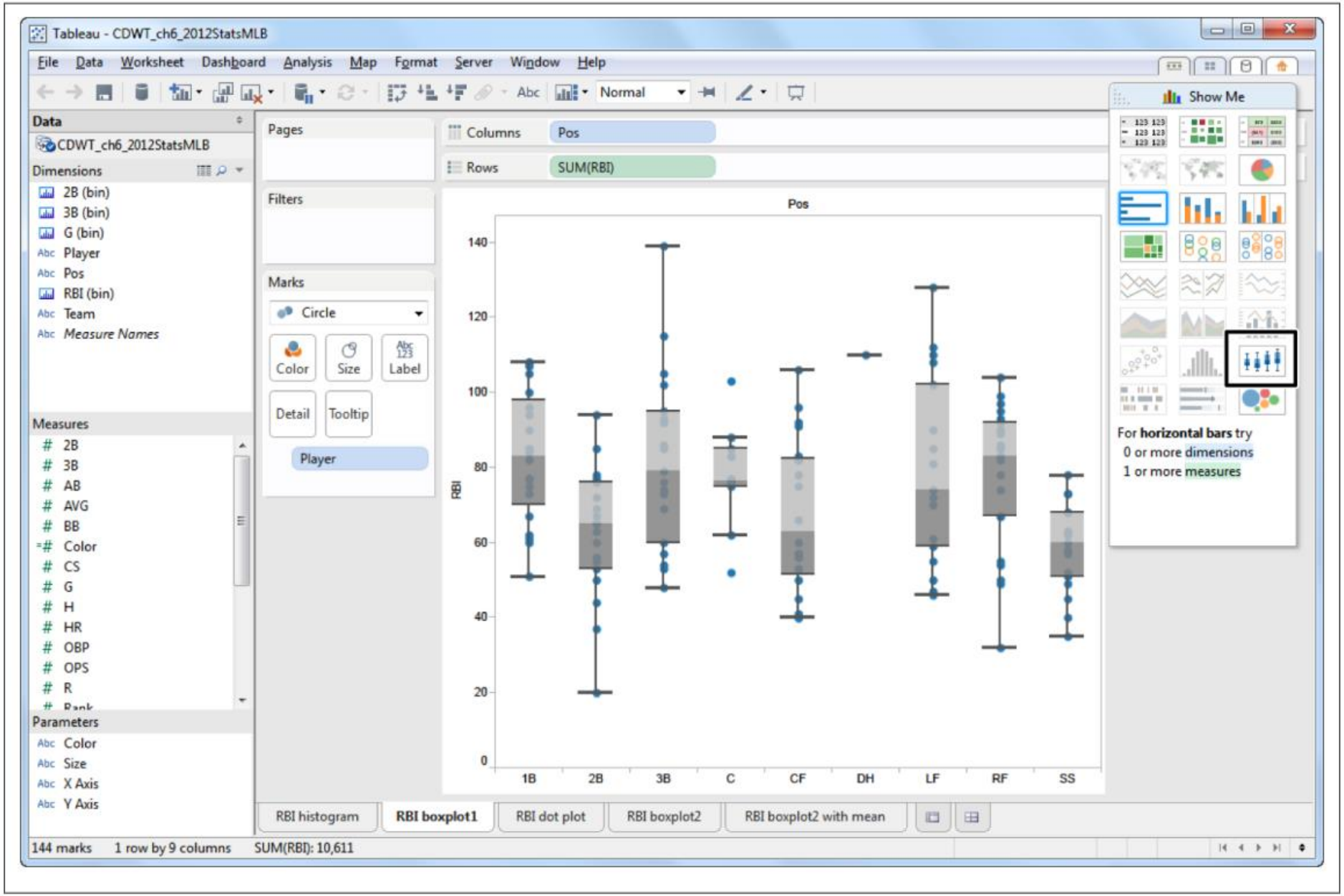


Figure 6-4. The box-and-whisker plot

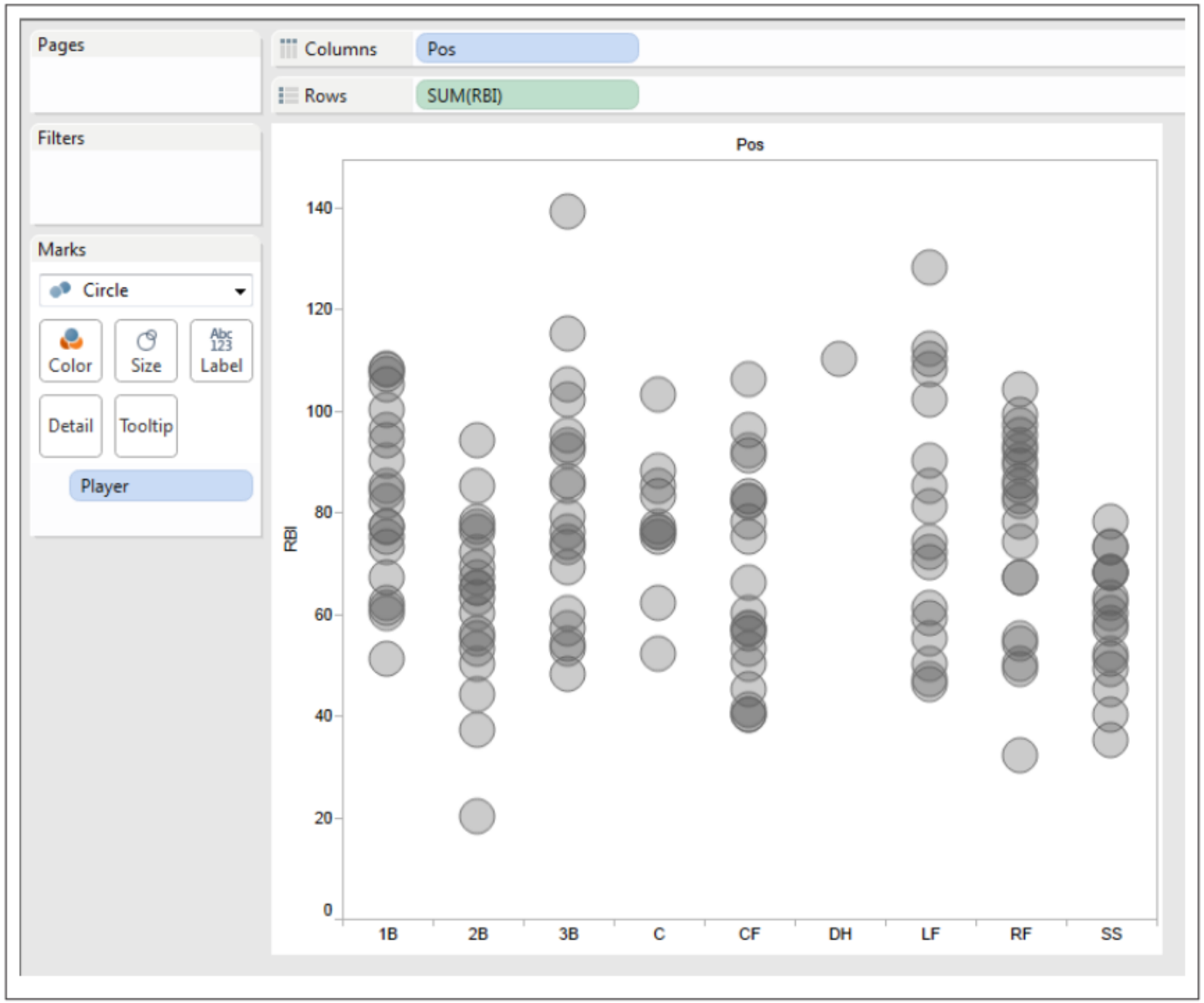


Figure 6-5. Dot plot of RBI by position

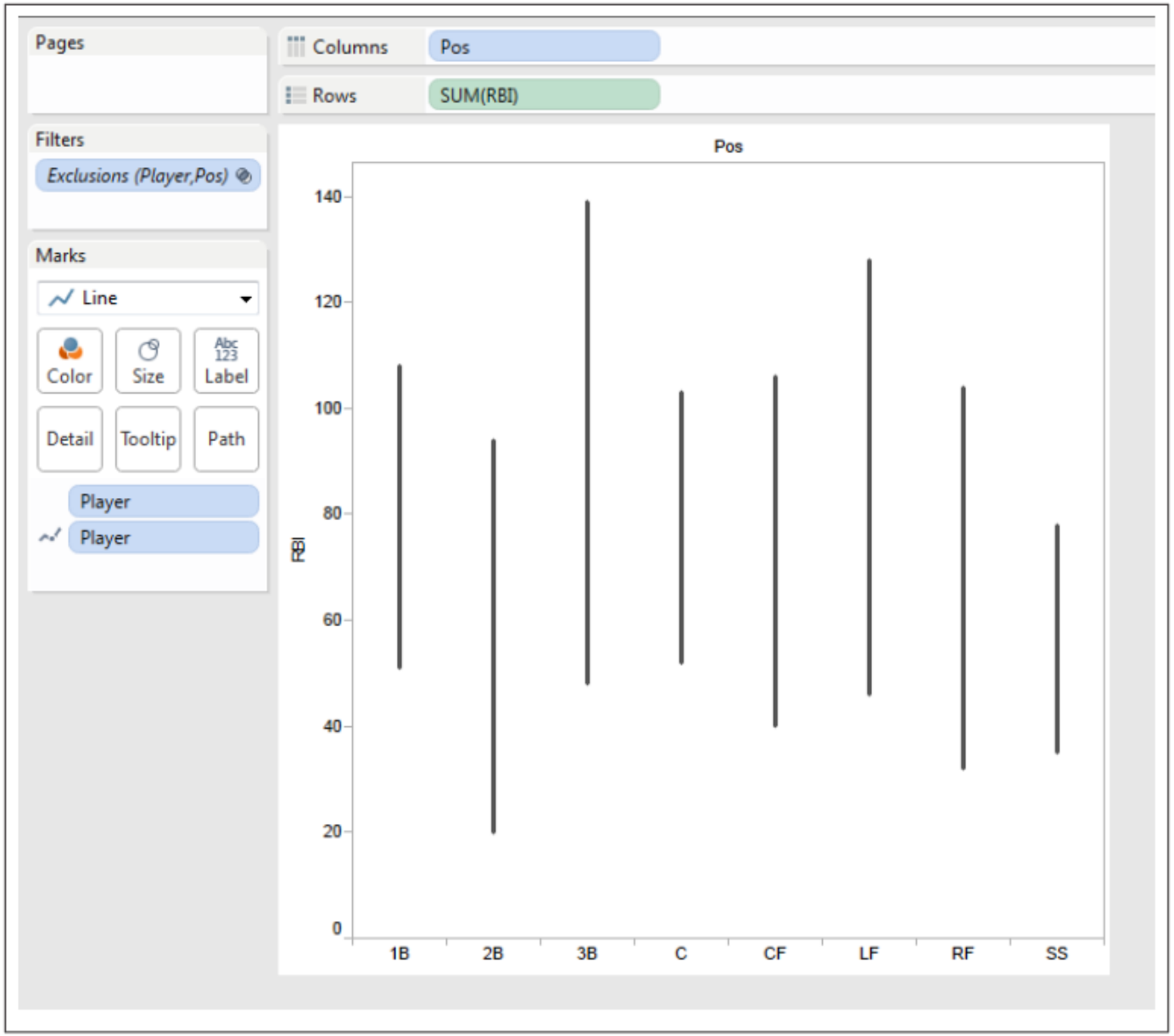


Figure 6-6. Vertical lines from the min to the max values of each position