**Guide to Fun with Phillies Data**

Our goal for this activity is to introduce students to asking and answering questions from data. We also want to connect their classes to the “real world” by understanding how math is used in a baseball team.

We’re providing the students with baseball statistics to help get them started.

|  |
| --- |
| What if students don’t care about baseball? If they’re interested in any sports, discuss how their sport of choice might care about measuring performance.  Outside of the realm of sports - how does Netflix know which TV shows to recommend? What could we measure to decide which social media platform is the most popular? |

# **Learning Objectives**

1. Explain the utility of math and data science in the Phillies organization and/or outside of school
2. Recognize common baseball statistics
3. Describe simple summary statistics and their relationship to graphs
4. Stretch/for students with statistics experience: perform simple joins between data tables

# **Activities**

|  |  |  |
| --- | --- | --- |
| *Activity 1* | *Introductions and Data Science* | *10 mins.* |
| *Activity 2* | *Introducing CODAP and Baseball Statistics* | *10 - 15 mins.* |
| *Activity 3* | *Summary Statistics and Visualizing Patterns* | *20 - 30 min.* |
| *Activity 4* | *Combining Data Sources* | *20 - 30 min.* |

# 

# **Index**

[**Activity 1:**](#_obs3irwdjgmq) **Introductions and Data Science** [**3**](#_obs3irwdjgmq)

[**Activity 2:**](#_ipc4b0j4rxe7) **Introducing CODAP 4**

[**Activity 3:**](#_8fylifh725xb) **Summary Statistics and Visualizing Patterns 7**

**Activity 4: Combining Data Sources 13**

# 

# **Activity 1: Introductions and Data Science** *10 mins*

*Goal: Connect with students. Explain the utility of math and data science in the Phillies organization and/or outside of school.*

*Materials needed: Worksheets optional. Leave laptops closed to encourage engagement.*

**Introductions (5 minutes)**

Instructors should give their name and pronouns (if comfortable doing so), how they use tech in their everyday life (besides using your phone!), and favorite sports team.

Students can introduce themselves with their name and pronouns (if comfortable doing so), grade/school, and your get-to-know-you question of choice. Some ideas to get started: their favorite meme (and why), the musical instrument they’re most similar to, or the super power they would want to have.

**What is data science? (5 minutes)**

The goal of this module is to explain what data science is and ground it in questions students understand and care about. The great thing about data scientists is that we’re all working on different things and have different skills!

Ask students if they know what data science is or for their best guesses. If nobody seems to be on the right track or students are hesitant to respond, break down the title. What’s data? What data might be interesting to a company? What might a general scientist do at work?

|  |
| --- |
| Here’s what Wikipedia has to say: Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is related to data mining and big data. |

Instructors can give their own definitions - if currently working in the field, how do you use data science? If not, what parts of your job might overlap with data scientists? How do you personally define it?

Once instructors have explained the concept, check in with the students. What parts of the explanation were confusing? When do they see data science in their own lives? Some examples, if they’re having trouble coming up with ideas: YouTube autoplay, Instagram explore page, facial recognition, and scheduling airplanes and trains.

What about the Phillies? We use data science to construct lineups, recommend trades, and provide in-game tactics!

**Students have learned…**

* Each other's and instructors’ names
* Common examples of data science

# **Activity 2: Introducing CODAP and Baseball Statistics** *10-15 mins*

*Goal: Understand basic CODAP functionality. Recognize common baseball statistics.*

*Materials needed: Chromebooks pre-loaded with baseball statistics with internet access. Worksheets optional.*

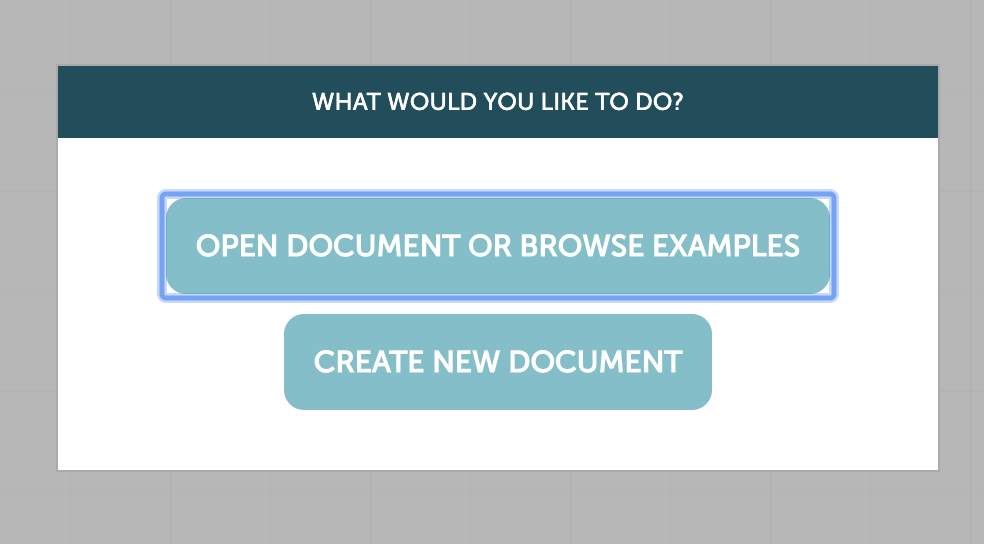
|  |
| --- |
| CODAP stands for Common Online Data Analysis Platform |

In Lesson 1, we were introduced to the types of questions data scientists are asking themselves. But before they can do that, we need to understand the data we’re working with. Some students might be more or less familiar with baseball, so we’re going to walk through the different types of statistics a data scientist might be interested in.

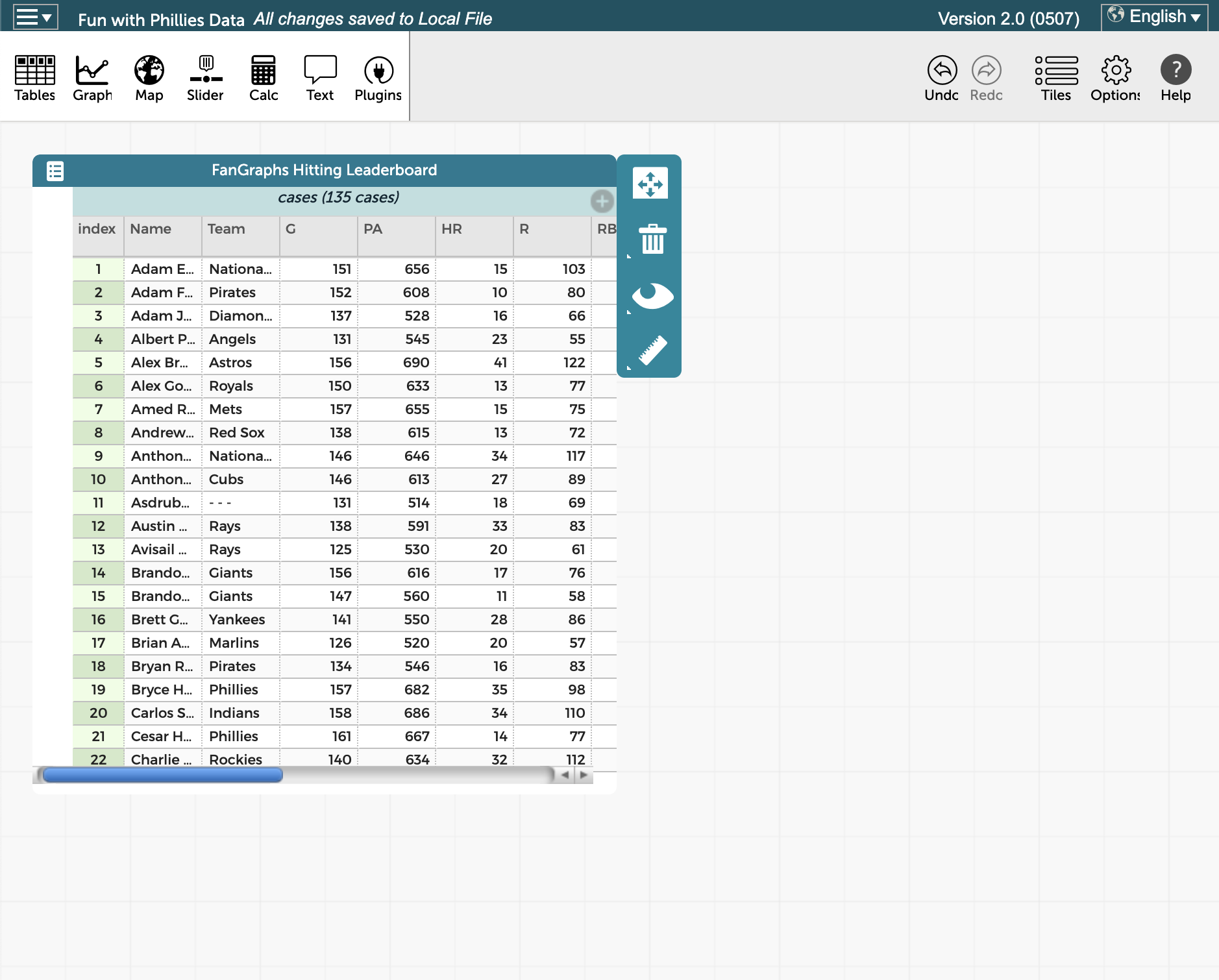
Soon we will turn you loose to explore and answer some questions, but first, let’s set up out environment…

**Setup (5 minutes)**

* Get students set up on the computers. There should be two students for everyone one laptop. To make the baseball information ramp-up easier, we can ask students to pair up with one person who’s a baseball fan and one person who knows a bit less. They may need to log on and wait for the OS to boot up. This is a good time to circulate and have smaller group conversations. What are they excited about learning today?
* Students can open up CODAP in a Chrome browser by navigating to [tiny.cc/codap](http://tiny.cc/codap). They should see the following pop-up window and select OPEN DOCUMENT OR BROWSE EXAMPLES.



* Navigate to Local File and open the phillies.codap file saved on the desktop. The opened program should look like this:



**Understanding the tables (5-10 minutes)**

These tables might seem overwhelming at first - there are a lot of abbreviations that the students might not recognize. So let’s walk through them, based on what students know. Students who are serious baseball fans will probably know all of the stats listed below and can explain them to their partners if their partner doesn’t recognize them.

Hitter Statistics

* **G** (Games Played): Number of games in which the player has appeared.
* **PA** (Plate Appearances): Number of times the player has come to the plate.
* **HR** (Home Runs): Number of home runs.
* **R** (Runs Scored): Number of runs scored.
* **RBI** (Runs Batted In): Number of times a run scores as a result of a batter’s plate appearance.
* **SB** (Stolen Bases): Number of stolen bases.
* **BB%** [(Walk Percentage)](http://www.fangraphs.com/library/offense/rate-stats/): Frequency with which the batter has walked, calculated as walks divided by plate appearances.
* **K%** [(Strikeout Percentage)](http://www.fangraphs.com/library/offense/rate-stats/): Frequency with which the batter has struck out, calculated as strikeouts divided by plate appearances.
* **AVG** (Batting Average): Rate of hits per at bat, calculated as H/AB.
* **OBP** [(On Base Percentage)](http://www.fangraphs.com/library/offense/obp/): Rate at which the batter reaches base.
* **SLG** (Slugging Percentage): Average number of total bases per at bat, calculated as Total Bases/AB.
* **wOBA** [(Weighted On Base Average)](http://www.fangraphs.com/library/offense/woba/): Combines all the different aspects of hitting into one metric, weighting each of them in proportion to their actual run value.

There’s a table embedded in the CODAP file for pitcher statistics. This isn’t incorporated in the main lesson flow, but for students who are moving particularly quickly it’s available as another resource for visualization.

Pitcher Statistics

* **W** (Wins): Number of wins.
* **L** (Losses): Number of losses.
* **G** (Games): Number of games in which the pitcher appeared.
* **GS** (Games Started): Number of games the pitcher started.
* **IP** (Innings Pitched): Number of total innings pitched (.1 represents 1/3 of an inning, .2 represents 2/3 of an inning).
* **K/9** [(Strikeouts per 9 innings)](http://www.fangraphs.com/library/pitching/rate-stats/): Average number of strikeouts per 9 innings.
* **BB/9** [(Walks per 9 innings)](http://www.fangraphs.com/library/pitching/rate-stats/): Average number of walks per 9 innings.
* **HR/9** (Home Runs per 9 innings): Average number of home runs allowed per 9 innings.
* **BABIP** [(Batting Average on Balls in Play)](http://www.fangraphs.com/library/pitching/babip/): The rate at which the pitcher allows a hit when the ball is put in play, calculated as (H-HR)/(AB-K-HR+SF).
* **GB%** [(Ground Ball Percentage)](http://www.fangraphs.com/library/pitching/batted-ball/): The percentage of a pitcher’s balls in play that are ground balls, calculated as GB/BIP.
* **ERA** [(Earned Run Average)](http://www.fangraphs.com/library/pitching/era/): The average number of earned runs a pitcher allows per 9 innings. ((ER\*9)/IP)

Why do we measure so many different things?

Every player (or movie, social media post, etc.) has multiple dimensions that we might be interested in. Consider a potential player who gets on base really often but almost never hits home runs. How does this player compare to someone else who strikes out more but can bring in a lot of runs with hard-hit balls? What about strikeout pitchers compared to those who induce weak contact?

We measure almost everything we can think of that might tell us more information about a player. Every team has their own formula for success and what they find interesting in a player

**Students have learned…**

* Simple offensive baseball statistics
* The benefit of measuring many different facets of questions

[**Activity 3:**](#_8fylifh725xb) **Summary Statistics and Visualizing Patterns** *20-25 mins*

*Goal: Describe simple summary statistics and their relationship to graphs*

*Materials needed: Chromebooks pre-loaded with baseball statistics with internet access. Worksheets optional.*

**Working with summary statistics (5 minutes)**

Here is a simple question you might have, given all this data:

*What does AVG really mean? What does it mean for a player to have a “good” batting average? What could we use as a baseline for evaluation?*

Students are going to have varied background information about summary statistics. Some may have already taken a formal statistics course in school and others might be fuzzy on the exact difference between mode and median. Again, let the students guide how much active teaching needs to happen here and there’s no need to cover everything if time is short. Introducing just median and mean will give the students plenty of information to start asking questions.

Summary statistics fall into two main categories: measures of location and measures of spread.

**Measures of location**: Where in space do these statistics fall? Where are the numbers centered?

* Average/Mean: Calculated by dividing the sum of all measurements by the number of measurements. *Ex avg(1,2,2,11) = (1+2+2+11)/4 = 4*
* Mode: The most common measurement in the set. *Ex mode(1,2,2,11) = 2*
* Median: The exact middle of a dataset, half the points fall above and below this point. *Ex median(1,2,2,11) = 2*

**Measures of spread**: How spread out/varied are our samples?

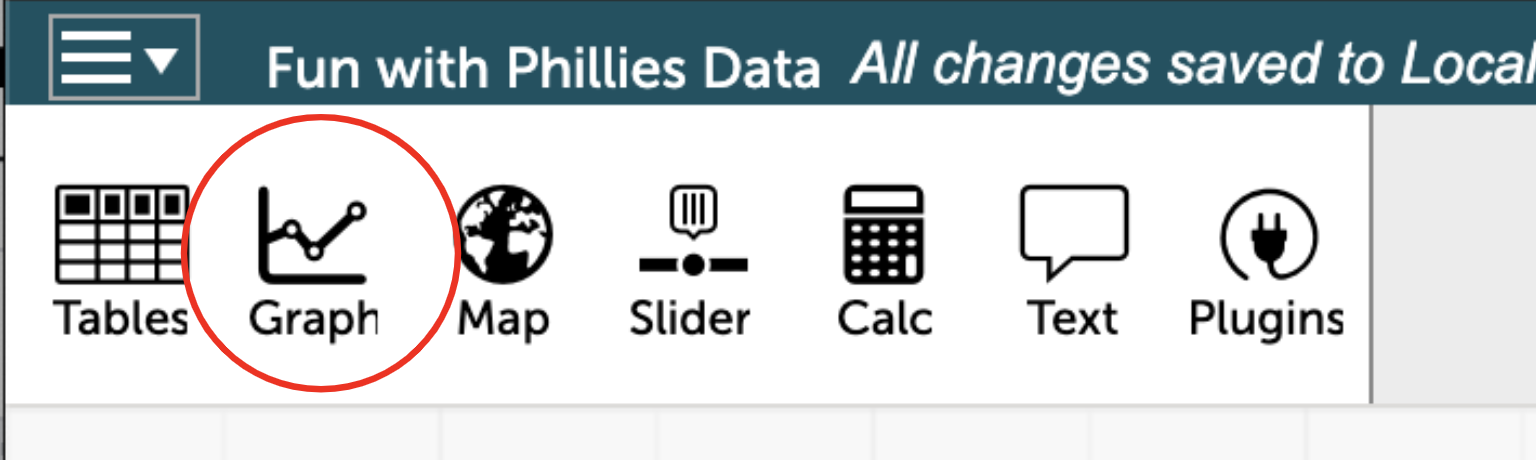
* Range: How spread out the data is. The difference between the largest and smallest measurement. *Ex range(1,2,2,11) = 11 - 1 = 10*
* Quartiles: Similar to median, measurements that indicate where the lower and upper 25% of the data fall.

**Visualizing single distributions (10 minutes)**

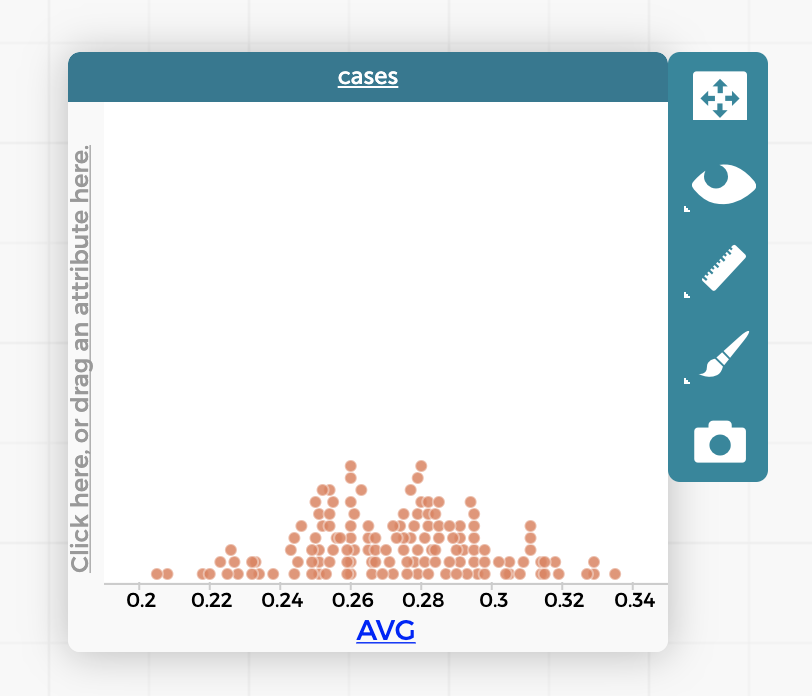
Let’s return to that question. What does it mean for a player to have a good batting average? This question is usually code for “does player x have an **above-average** batting average?”. To answer this, we can look at the underlying distribution of batting average.

CODAP makes visualization very straightforward. Building graphs is a drag-and-drop operation.

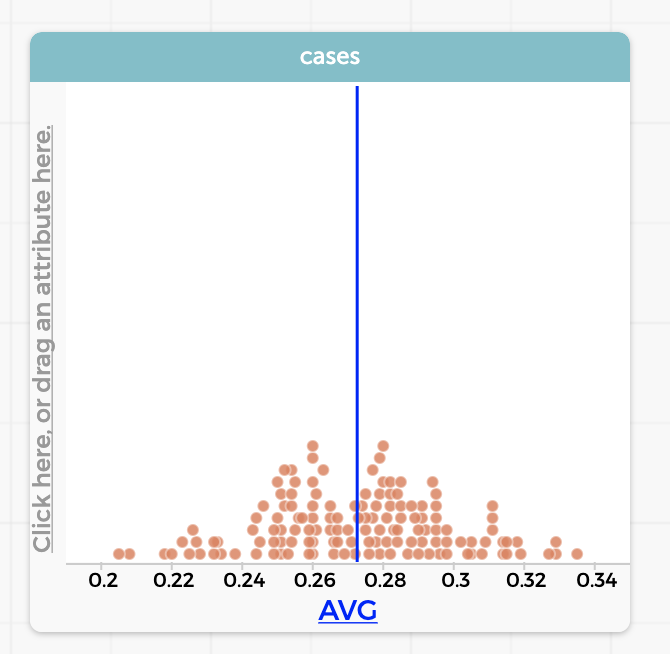
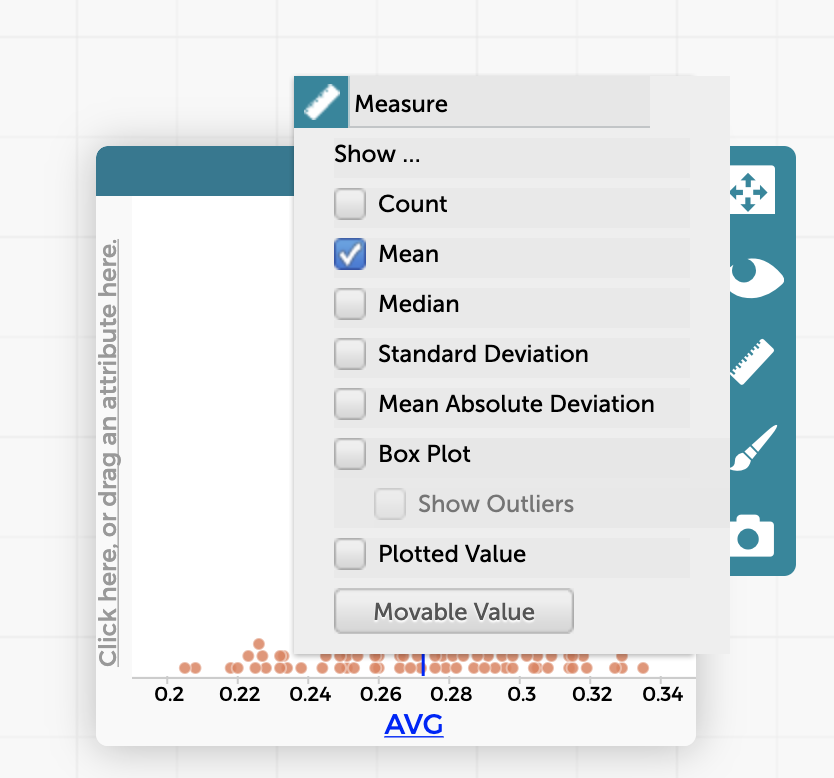
1. Have students select Graph from the left toolbar. A blank canvas will appear.



1. We’ll start by making a dot plot to visualizing the distribution of the AVG column. Selecting the column header and dragging it to the bottom of the canvas will highlight the x-axis in yellow and display Create Axis with AVG. Releasing the attribute populates the graph. Students should see this:



1. Selecting the ruler icon on the right side of the canvas displays multiple visualization options. Let’s select mean.



1. Selecting a dot on the plot will select that row of the table and selecting a table row highlights that point on the plot.

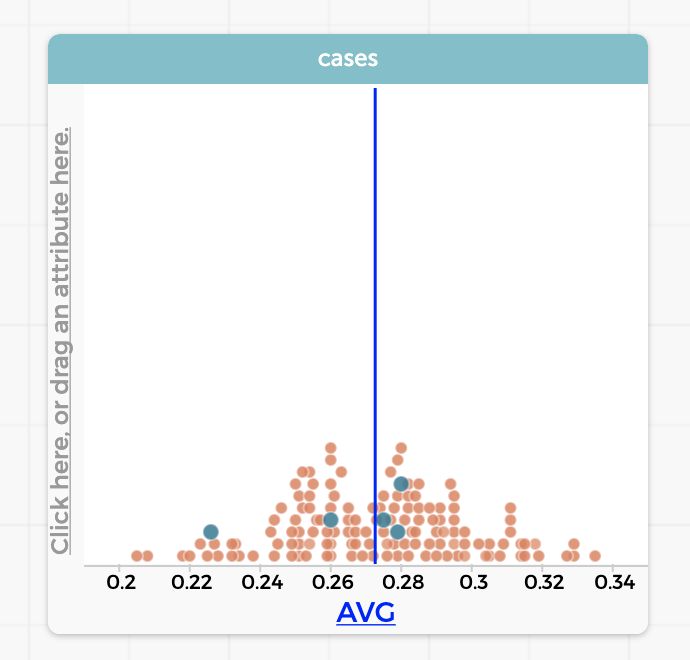
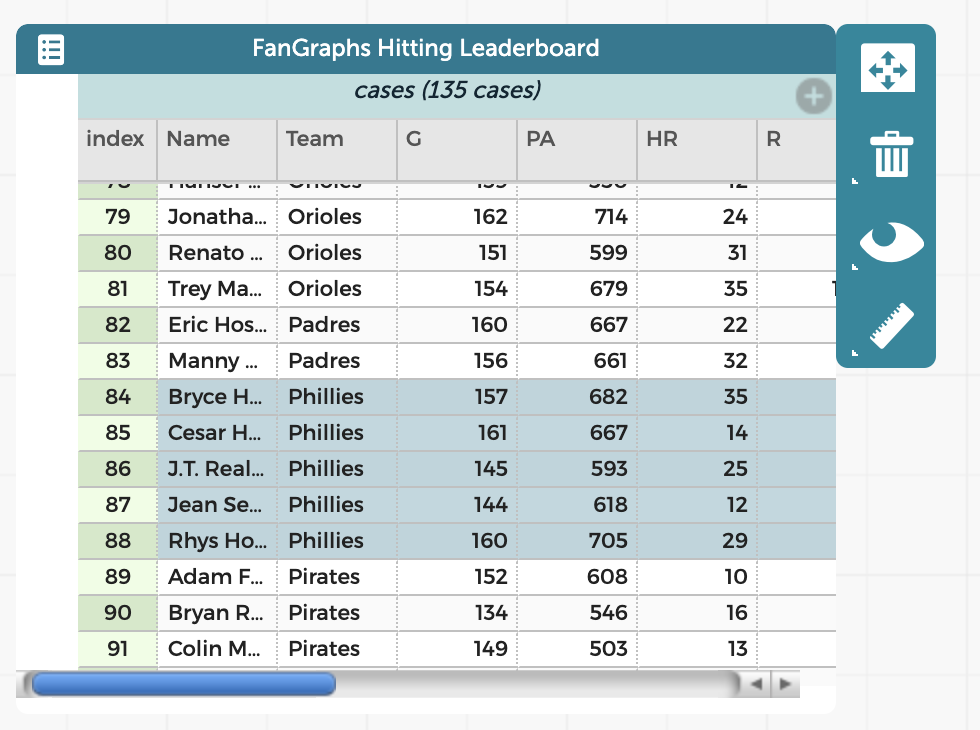
Now that students have a visual of the 2019 batting averages, let’s answer some questions. First up,

*Which Phillies players had above-average batting averages in 2019?*

1. First, we want to find out which of the dots correspond to Phillies players. How could we easily find all the Phillies players in the table? Sorting by the Team component will group players by their associated team and it’s simple to scroll and find the Phillies players.

|  |
| --- |
| Students may ask why certain players aren’t included - great question! Sample size (how many observations we have) is something data scientists care a lot about. Seeing only one coin flip doesn’t tell us much about the behavior of that coin and the same goes for baseball players. This set only includes players with over 500 plate appearances (PA) so that we trust the statistics! |

1. Select all rows in the table of Phillies players. Their dots will turn blue in our graph!



1. Selecting the players one by one will show us which have a batting average higher than (to the right of) the average. These players are Jean Segura, J.T. Realmuto, and Cesar Hernandez. If using worksheets, students can fill this answer in.

What does this mean about the other Phillies players? They have a batting average below average.

**Visualizing multiple distributions (10 minutes)**

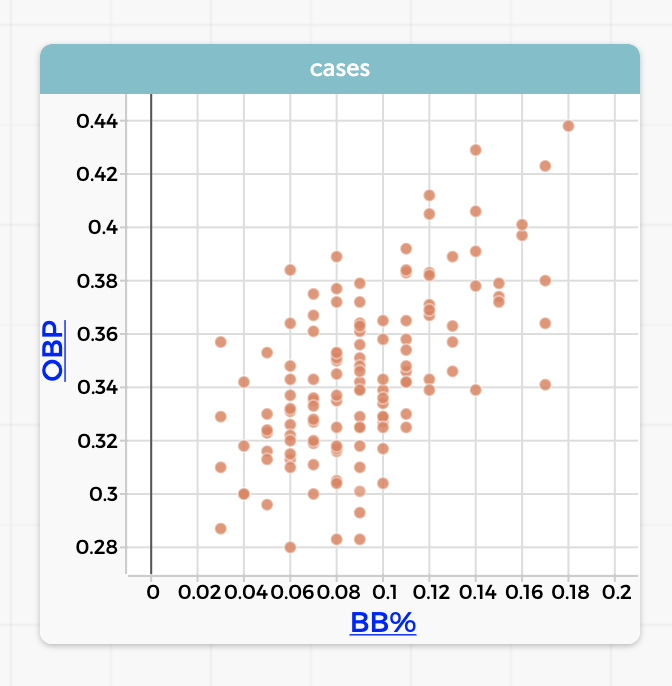
Now that students are comfortable building graphs, we can move onto 2D plots. Before they do anything in CODAP, ask them to think about the answer to this question and sketch a graph on the worksheet, if using.

*What relationship do BB% and OBP have? Would we expect OBP to increase or decrease as BB% increases?*

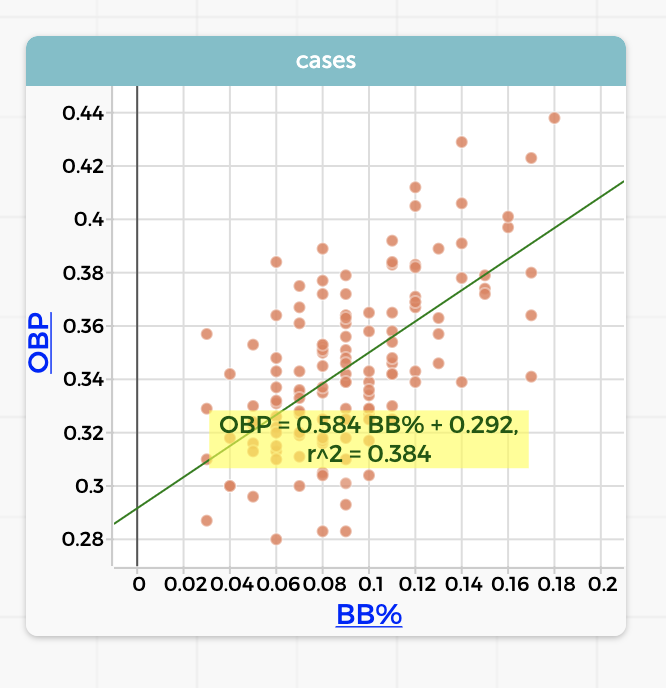
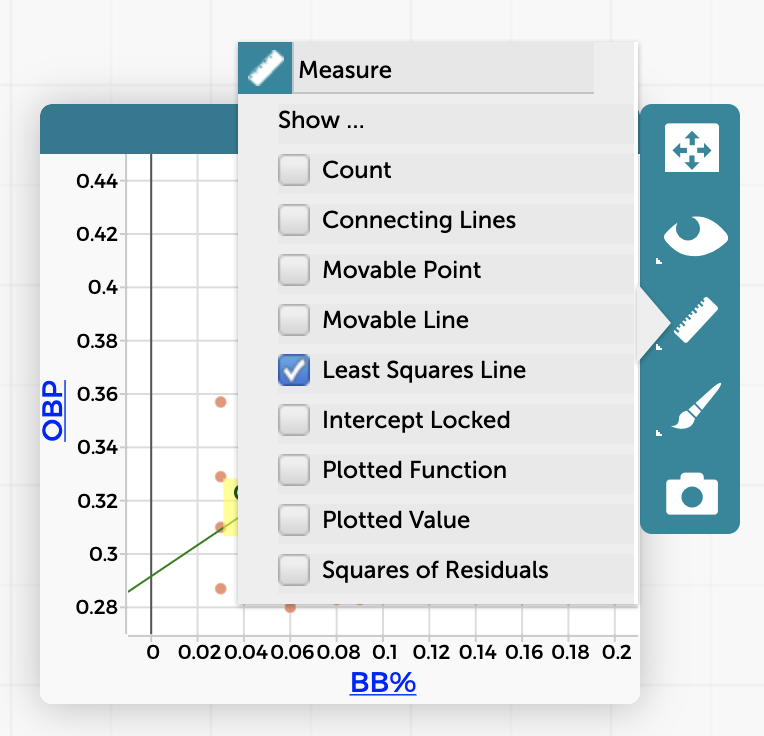
# Make sure students discuss with one another what they would expect to see. As data scientists, it’s helpful to have an idea of what we expect the data to tell us. That way we can catch mistakes in our code or recognize interesting results more easily!

If they seem stuck, walk through the meaning of these two statistics. BB% is a measurement of how often a player walks, or is given a free pass to first plate. OBP measures how frequently a player gets on base, regardless of how they do it. So, if players are walking more, they should get on base more frequently!

Plotting the two measurements against one another results in a scatter plot, which most of us have seen before. We build this graph in exactly the same way, except that both the x and y-axes will be given attributes. First, open a new graph canvas. Drag BB% over to the x-axis and OBP over to the y-axis.



We see a positive linear trend, just like we expected! Students can add a line of best fit to the graph if they’re comfortable with the concept. Just select Least Squares Line from the measurement menu. What do each of the numbers in this equation mean?

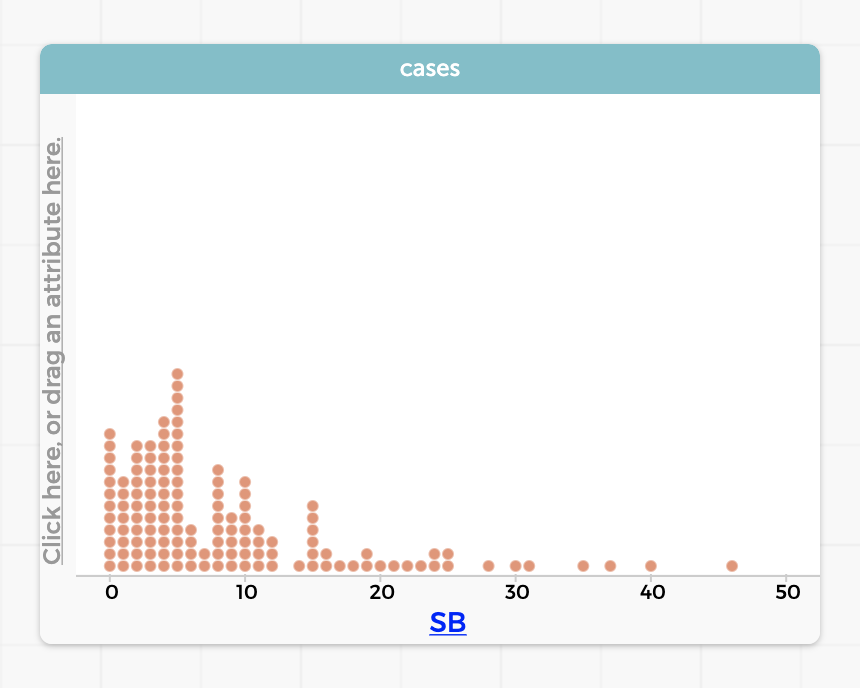


Students can pick a track here - investigating summary statistics or searching for other pairs of statistics that are correlated with one another.

**Return to summary statistics (5-10 minutes)**

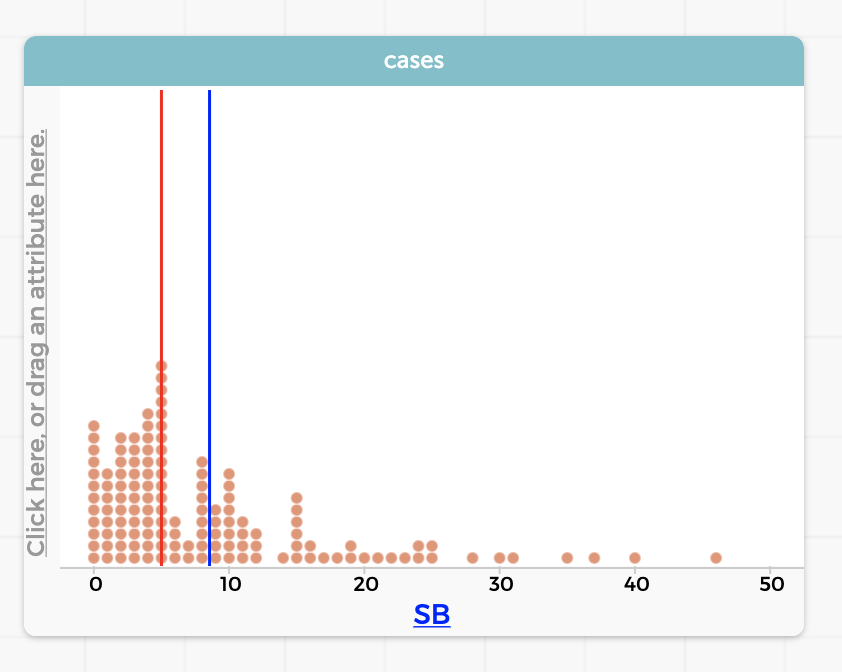
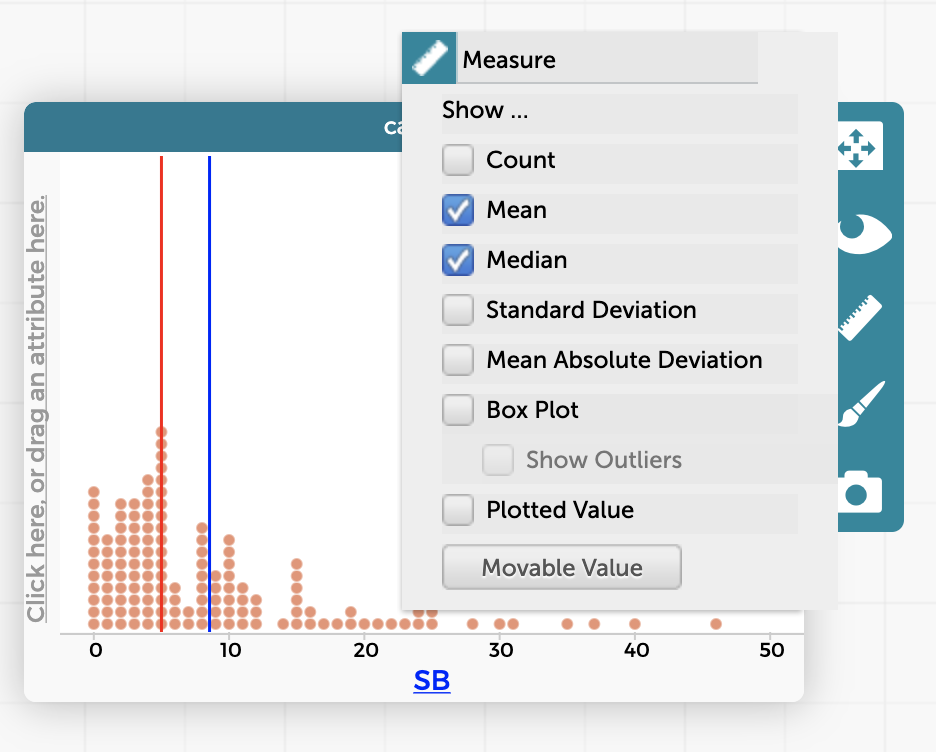
When we introduced summary statistics, we covered a lot of different types that tell us different things about the data. But why would we need so many ways to measure our data?

Visualizing stolen bases (SB) can give us a clue why. We just need a simple dot plot this time.



Before we go any further - what’s different about this plot than the AVG dot plot? Does it look like players cluster in a different way?

This time, we’re going to add *two* summary statistics to the plot: mean and median.



The median is much lower than the mean - why? Outliers (observations that are much lower or much higher than the sample mean) can pull up the average. In this case, most baseball players don’t steal that many bases. In fact, half of them steal less than 5, the median! But some baseball players steal a lot of bases, and their measurements skew the mean to be higher than the median. In fancy terms, we say this data is *right-skewed.*

The takeaway is that all summary statistics tell us something a little bit different about our data set! We can understand the population better by looking at the problem in a lot of different ways.

**Correlation scavenger hunt (5-10 minutes)**

Give students the freedom to experiment in CODAP and think of possible combinations of statistics that might show positive or negative trends. Students should feel free to open the pitcher table and experiment with that as well (data can be found under the tables button in the left toolbar). Some options, if they get stuck:

* K% vs AVG
* K/9 vs ERA

[**Activity 4:**](#_8fylifh725xb) **Combining Data Sources** *20-25 mins*

*Goal: Describe simple summary statistics and their relationship to graphs*

*Materials needed: Chromebooks pre-loaded with baseball statistics with internet access. Worksheets optional.*