**Fun with Phillies Data**

# Today we’re going to be working with baseball statistics to introduce basic concepts of data science! There’s a lot to cover in just a little bit of time so we’ll skim over a lot of details, but check out the resources at the end of this guide for more.

# **You will be able to...**

1. Explain the utility of math and data science in the Phillies organization and outside of school
2. Recognize common baseball statistics
3. Describe and question visual patterns in graphs
4. Optional: perform simple joins between data tables

[**Activity 1:**](https://docs.google.com/document/d/1jTMeDhqxCkgirK3s13ttnwLIa7JqWEpCoUd9yyJcKqQ/edit#heading=h.obs3irwdjgmq) **Introductions and Data Science**

What is data science?

|  |
| --- |
|  |

List some examples of data science you see every day

|  |
| --- |
|  |

The Phillies use data science to guide a lot of their team management! We can use data science to build a roster, suggest trades and contracts, and help the coaches with in-game tactics.

[**Activity 2:**](https://docs.google.com/document/d/1jTMeDhqxCkgirK3s13ttnwLIa7JqWEpCoUd9yyJcKqQ/edit#heading=h.ipc4b0j4rxe7) **Introducing CODAP and Baseball Statistics**

You can open a blank CODAP document by navigating to <http://tiny.cc/codap>. All materials for this activity are hosted online at <https://github.com/hannah-gaudet/phillies-data-codap>.

Use this space to take notes on baseball statistics that are new to you (if any!)

|  |
| --- |
|  |

[**Activity 3:**](https://docs.google.com/document/d/1jTMeDhqxCkgirK3s13ttnwLIa7JqWEpCoUd9yyJcKqQ/edit#heading=h.8fylifh725xb) **Summary Statistics and Visualization**

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

|  |
| --- |
|  |

Which Phillies players had below-average batting averages?

|  |
| --- |
|  |

Predict what the graph will look like if we plot a batter’s BB% against their OBP.

Hint: Does walking more often mean the batter will get on base more or less?

Sketch the resulting graph

|  |  |
| --- | --- |
|  |  |

When we examine the distribution of stolen bases, the mean is much higher than the median. What does this tell us about the way the observations are distributed?

|  |
| --- |
|  |

What other statistics do you think might be correlated with one another? Try plotting them in CODAP and describing the result here.

|  |
| --- |
|  |

**Resources**

Data Science

**Codecademy**: Free online coding lessons covering a lot of different languages. They offer a specific data science track that covers a lot of basics. Python and R are good languages to get started with. <https://www.codecademy.com/>

**DataQuest**: Very similar to Codecademy, but specifically focused towards data science and associated skills.

<https://www.dataquest.io/>

**Vertabelo**: Online courses covering a wide range of topics. Specifically recommended for learning SQL.

<https://academy.vertabelo.com/>

**Kaggle**: Provides online environments for data science work. Hosts a wide range of datasets to experiment with. Kaggle also runs competitions for practicing modeling and prediction skills. <https://www.kaggle.com/>

**FiveThirtyEight**: No lessons, but covers pop culture, news, and sports with data. A great option for learning what types of questions are being asked and answered.

<https://fivethirtyeight.com/>

Baseball

**Fangraphs**: Popular site that hosts baseball statistics and is regularly publishing new data-driven baseball research. Great for deep-dives on specific players and teams as well as understanding how more advanced statistics are calculated

<https://www.fangraphs.com/>

**Baseball Reference**: Another site hosts player statistics. Keeps track of all player transactions and contracts.

<https://www.baseball-reference.com/>

**Baseball Prospectus**: Known for tracking amateur players (prospects). Also home to proprietary baseball statistics for people who want to know every stat that’s being used.

<https://www.baseballprospectus.com/>