

The Data Cycle

(Also available in [Pyret](#))

Students are introduced to the Data Cycle, a four-step scaffold for getting an answer from a dataset - and then generating the next question! Students learn to identify - and ask - statistical questions, by comparing and contrasting them with other kinds of questions.

Lesson Goals	Students will be able to... • Articulate the parts of the Data Cycle • Tell the difference between Lookup, Arithmetic, and Statistical questions • Come up with their own examples of Lookup, Arithmetic, and Statistical questions
Student-facing Lesson Goals	• Let's think about what it means to ask questions of data, and the steps to do it • Not all questions are created equal! Let's learn the difference between them
Prerequisites	• Introduction to Data Science
Materials	• PDF of all Handouts and Page • Animals Starter File • Lesson Slides • Printable Lesson Plan (a PDF of this web page)
Supplemental Materials	• Additional Printable Pages for Scaffolding and Practice

Overview

Students learn about the *Data Cycle*, which is a scaffold to support asking them in questions, thinking about how those questions relate to the data in front of them, analyzing that data, and interpreting the results.

Launch



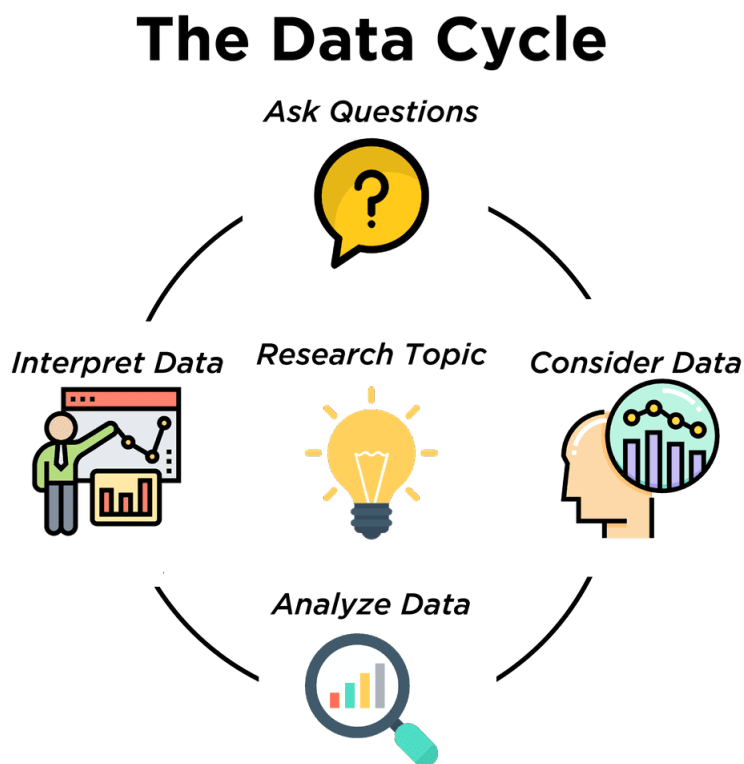
- Open your saved Animals Starter File, or [make a new copy](#).
- Working in pairs, turn to [The Animals Dataset](#), or open [the Animals Spreadsheet](#).
- You and your partner are going to answer a simple question: **are more animals fixed or unfixed?**

Data Science is all about *asking questions of data*. Sometimes the answer is easy to compute. Sometimes the answer to a question is *already in the dataset* - no computation needed. And sometimes the answer just sparks more questions!

Data Scientists ask a ton of questions, and each question adds a chapter to their **data story**. Even if a question turns out to be a dead-end, it's valuable to share what the question was and what work you did to answer it!

The Data Cycle[*] is a *roadmap*, which helps guide us in the process of data analysis.

- 1) We **Ask Questions** - which can be answered with data.
- 2) We **Consider Data**. This could be done by conducting a survey, observing and recording data, or finding a dataset.
- 3) We **Analyze the Data**, to produce data displays and new tables of filtered or transformed data, to identify patterns and relationships.



4) We **Interpret the Data**, answering questions and summarizing results. As we've already seen from the Animals Dataset, these interpretations often lead to *new* questions... and the cycle begins again.

Investigate

Are more animals fixed or unfixed? was a pretty specific question, and it was straightforward to answer. But the answers to even simple questions can lead to more interesting questions down the road!



- What other questions might come from counting the ratio of fixed to unfixed animals?
 - *Sample responses: Is there a higher percentage of fixed dogs or fixed cats? At what age do animals get fixed? Do fixed animals get adopted more quickly than unfixed animals?*

Synthesize

This is a short example of a **Data Story** - a description of how each step in the Data Cycle was used to go from a question to an answer, and then to the next question. When analyzing a real dataset, Data Scientists might explore lots of questions, resulting in many different Data Stories to tell.

Each chapter in the Data Story is valuable, and the Data Cycle represents each chapter you plan to add.

Overview

Students begin to categorize questions, sorting them into "lookup", "arithmetic", and "statistical" questions - as well as questions that simply can't be answered based on the data.

Launch

How do we know what questions to ask? There's an art to asking the right questions, and good Data Scientists think hard about what kind of questions can and can't be answered.

Ask Questions



Most questions can be broken down into one of four categories:

- **Lookup questions** - Answered simply by looking up a single value in the table and reading it out. Once you find the value, you're done! Examples of lookup questions might be "How many legs does Felix have?" or "What species is Sheba?"
- **Arithmetic questions** - Answered by computing an answer within a single column. Examples of arithmetic questions might be "How much does the heaviest animal weigh?" or "What is the average age of animals from the shelter?"
- **Statistical questions** We ask these all the time! If we ask "are dogs heavier than cats?", we know that *not every dog is heavier than every cat!* We just want to know if it is *generally* true or *generally* false! When we ask, "How old are animals at the shelter?", we know that *not every animal is the same age!* We're asking about how old they are *in general!*
- **Questions we can't answer** We might wonder where the animal shelter is located, or what time of year the data was gathered! But the data in the table won't help us answer that question, so as Data Scientists we might need to do some research *beyond the data*. And if nothing turns up, we simply recognize that there are limits to what we can analyze.



- What kind of question is "Are more animals fixed or unfixed?"?
 - It's an *arithmetic question*.
- What kind of question is "How old is Toggle?"
 - It's a *lookup question*.

Investigate



- Turn to [Which Question Type?](#), and fill out the "Type" column in the table at the bottom. For now, ignore the other columns.
- Look at the Wonders you wrote on [Questions and Column Descriptions](#). Are these Lookup, Arithmetic, or Statistical questions?
- *Optional:* For more practice, complete [Question Types: Animals](#), by coming up with examples of each type of question for the Animals Dataset.

Common Misconceptions

- Students generally struggle to make the leap into asking statistical questions. It's worth taking time on this, to support them coming up with better (and more engaging!) questions later.
- They may think that "What's the average weight of the animals?" is a statistical question, because "average" is a term that shows up in statistics. But computing the average is just pure arithmetic! A *statistical* question would be "What's the typical weight of an animal?", because it does not specify a particular arithmetic process. The answer could be the mean, the median, or even the mode! Figuring out which one to use depends on the distribution of the data, which we'll discuss more in a later lesson.

Synthesize

- How would you explain the difference between Lookup, Arithmetic, and Statistical questions?
- When you looked back at your Wonders from the Animals Dataset, were they mostly Lookup questions? Arithmetic? Statistical?
- What are some examples of statistical questions the owner of a sports team might ask? Or a researcher who is trying to see if a cancer drug is effective? Or a principal who wants to know what will help their students the most?

Consider Data

20 minutes

Overview

Students bridge from a human-language question into something more formal, by specifying the rows and columns they would need to examine. This activity stresses a hard programming skill (reading Contracts) with formal reading comprehension (identifying key portions of a statistical instruction).

Launch

Once we have our question, it's time to figure out what data we'll need to answer it!

When considering data, we ask: **Which Rows** do we need? **Which Column(s)** do we care about?

Tables are made of **Rows** and **Columns**. Each Row represents one member of our population. In the Animals Dataset, each row represents a single animal. In a dataset of temperature readings, each row might represent the temperature at a particular hour.

Columns, on the other hand, represent information *about each row*. Every animal, for example, has columns for their name, species, sex, age, weight, legs, whether they are fixed or unfixed, and how long it took to be adopted.

If we want to know which cat is the heaviest, we *only care about rows for cats*, and *we only need the pounds column*. If we want to know how many fixed animals are rabbits, *we only care about rows for fixed animals*, and *we only need the species column*.

Consider Data



- If our question is "How old is Mittens?", what rows and column(s) do we need?
 - We only need one row for Mittens, and we just need the age column
- If our question is "Which animal is the heaviest?", what rows and column(s) do we need?
 - We need to compare every row, and we only look at the pounds column
- What rows and columns did we need to answer "Are more animals fixed or unfixed?"?
 - We needed to look at all the rows, but the only column we care about is fixed.

Investigate



- Return to [Which Question Type?](#) For each question, which rows would you need to answer them? (Sometimes we need all rows, and sometimes we only need a subset.) Which columns would you look at? Write your answers in the last two columns of the table at the bottom.
- Complete [Data Cycle: Consider Data](#).

Common Misconceptions

Students often forget that questions like "Who is the oldest?" or "What is the most?" require looking at *every row* in the table.

Synthesize

Have students share their answers and discuss any questions they have about these pages.

How does asking "Which rows? Which columns?" help us figure out which configurations to use?

Overview

Students progress to the third step in the Data Cycle, by combining the "Consider Data" step with their knowledge of Contracts to help them *Analyze* that data.

Note that [blank Data Cycle worksheets](#) are available for printing as-needed!

Launch

Once we know what data we need, we can turn our attention to what we want to build with it!

- Do we need all the rows, or just some of them?
- Do we need a bar chart? A scatter plot?

What kinds of displays can help us analyze whether there are more fixed or unfixed animals? A bar chart will allow us to see the actual count ("20 fixed vs. 10 fixed") of animals in each category.

Once we've decided what to make and we know which rows and columns we're plotting, the next step is to choose the appropriate configuration.

Analyze Data



Investigate

Let's get some practice going from questions to code, and making data displays in the process!



Turn to [Data Cycle: Analyzing with Displays](#), and see if you can fill in the first 3 steps of the Data Cycle for a set of predefined questions. When you're finished, try to make the display in CODAP.

Have students share their results. What did their charts tell them?

Synthesize

In this case, we got a clear answer to our question. But perhaps that's not the end of the story! We might be curious about whether a higher percentage of dogs are spayed and neutered than cats, or whether it's even possible to "fix" a tarantula. *All of this belongs in our data story!*

Interpret Data



[*] From the [Mobilizing IDS project](#) and [GAISE](#)