Introduction to Computational Data Science

Students are introduced to the Animals Dataset, learn about Tables, Categorical and Quantitative data, and consider the kinds of questions that can be asked about a dataset.

Prerequisites	None
Relevant Standards K12CS CSTA CC-ELA	*Select one or more standards from the menu on the left (⌘-click on Mac, Ctrl-click elsewhere).*
Lesson Goals	Students will be able to... • Explain the difference between Categorical and Quantitative data • Identify whether a variable in a dataset is Categorical or Quantitative • Identify the Header Row and Identifier Column of a Table
Student-facing Lesson Goals	• Let's learn about data inside tables.
Materials	• Lesson Slides (Google Slides) • Computer for each student (or pair), with access to the internet • Each student (or pair of students) should have a Google Account. • Student workbook, and something to write with • Opening questions printed for each student, group of students, or posted around the room. **Note:** these are just ideas to get you started. Use questions that you know will interest *your* students!
Preparation	• Make sure student computers can access the Animals Spreadsheet and the Animals Starter File. • Make sure all materials have been gathered • Decide how students will be grouped in pairs • Decide how the first activity (opening questions) will be run
Supplemental Resources	
Language Table	No language features in this lesson

Glossary

categorical data :: data whose values are qualities that are not subject to the laws of arithmetic.

data science :: the science of collecting, organizing, and drawing general conclusions from data, with the help of computers

programming language :: a set of rules for writing code that a computer can evaluate

quantitative data :: number values for which arithmetic makes sense

---

Introduction

20 minutes

Overview

Students look at opening questions, either at their desks or in a walk around the room. They select a question they are personally interested in, and think about the data required to answer that question. This process draws a direct line between answering questions they care about and the basics of data science.

Launch

- Give students 2 minutes to choose a question that grabs their attention, and group themselves by question. Ideally, no student will be the only one interested in that question.
- Have students spend 2 minutes coming up with a hypothesis about what the answer is, and explaining why. Does every student in a single question-grouping have the same answer?

Investigate

- *What information would you collect to answer this question?* Give students 5 minutes to think about what information they would need to collect, to find the answer.

Possible Misconceptions

Students may lean towards questions about *individuals*, instead of questions about what's true for a *group of individuals* who vary from one to another. For example, instead of wondering what movie gets the highest rating, they should ask what's the typical rating for movies in a list, or how much those ratings tend to vary.

Synthesize

Have students share back the different data they would gather to answer their questions. For each question, students would likely have to gather many different kinds of data. If we wanted to find out if small schools are better than big schools, for example, we might want to gather data on SAT scores, college acceptance, etc. Each of these is a **variable** in our dataset: any two schools we look at could *vary* by each of them.

What's the greatest movie of all time? Is Climate Change real? Who is the best quarterback? Is Stop-and-Frisk racially biased? We can't survey every school in the world, get data on every movie ever made, or every police action - but we can do an analysis for a *sample* of them, and try to infer something about all of them as a whole. These questions quickly turn into a discussion about data—how you assess it, how you interpret the results, and what you can *infer* from those results.

The process of learning from data is called ***Data Science***. Data science techniques are used by scientists, business people, politicians, sports analysts, and hundreds of other different fields to ask and answer questions about data.

We'll use a ***programming language*** to investigate these questions. Just like any human language, programming languages have their own vocabulary and grammar that you will need to learn. The language you'll be learning for data science is called Pyret.

The Animals Dataset

25 minutes

Overview

Students explore the Animals Dataset, sharing observations and familiarizing themselves with the idiosyncrasies and patterns in the data. In the process, they learn about Categorical and Quantitative data.

Notice and Wonder Pedagogy

This pedagogy has a rich grounding in literature, and is used throughout this course. In the "Notice" phase, students are asked to crowd-source their observations. No observation is too small or too silly! Students may notice that the animals table has corners, or that it's printed in black ink. But by listening to other students' observations, students may find themselves taking a closer look at the dataset to begin with. The "Wonder" phase involves students raising questions, but they must also explain the context for those questions. Sharon Hessney (moderator for the NYTimes excellent

What's going on in this Graph? activity) sometimes calls this "what do you wonder...and **why**?". Both of these phases should be done in groups or as a whole class, with time given to each.

Launch

Have students open the Animals Spreadsheet in a browser tab, or turn to The Animals Dataset (Page 2) in their Student Workbooks.

Investigate

This table contains data from an animal shelter, listing animals that have been adopted. We'll be analyzing this table as an example throughout the course, but you'll be applying what you learn to *a dataset you choose* as well.

- Turn to ../../lessons/ds-intro/pages/exploring-animals-dataset.adoc in your Student Workbook. What do you _Notice about this dataset? Write down your observations in the first column.
- Sometimes, looking at data sparks questions. What do you *Wonder* about this dataset, and why? Write down your questions in the second column.
- There's a third column, called "Question Type" — we're going to return to that later, so you can ignore it for now.
- If you look at the bottom of the spreadsheet file, you'll see that this document contains multiple sheets. One is called `"pets"` and the other is called `"README"`. Which sheet are we looking at?
- Each sheet contains a table. For our purposes, we only care about the animals table on the `"pets"` sheet.

Any two animals in our dataset may have different ages, weights, etc. Each of these is called a **variable** in the dataset. Data Scientists work with two broad kinds of data: Categorical Data and Quantitative Data. ***Categorical Data*** is used to *classify*, not measure. Categories aren't subject to the laws of arithmetic. For example, we couldn't ask if "cat is more than lizard", and it doesn't make sense to "find the average ZIP code" in a list of addresses. "Species" is a categorical variable, because we can ask questions like "which species does Mittens belong to?"

What are some other categorical variables you see in this table?

Quantitative Data is used to measure an amount of something, or to compare two pieces of data to see which is *less or more*. If we want to ask "how much" or "which is most", we're talking about Quantitative Data. "Pounds" is a quantitative variable, because we can talk about whether one animal weighs more than another or ask what the average weight of animals in the shelter is.

We use ***Categorical Data*** to answer "what kind?", and ***Quantitative Data*** to answer "how much?".

- Turn to page Categorical or Quantitative? (Page 3), and answer the questions 1-5
- Sometimes it can be tricky to figure out if data is categorical or quantitative, because it depends on *how that data is being used!*
- On ../../lessons/ds-intro/pages/exploring-animals-dataset.adoc in your Student Workbook, fill in the blanks for questions 8-13.

Synthesize

Have students share back their noticings (statements) and wonderings (questions), and write them on the board.
Data Science is all about using a smaller sample of data to make predictions about a larger population. It's important to remember that tables are only a *sample* of a larger population: this table describes some animals, but obviously it isn't every animal in the world! Still, if we took the average age of the animals from this particular shelter, it might tell us something about the average age of animals from other shelters.

Question Types

10 minutes

Overview

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

Launch

Once we have a dataset, we can start asking questions! But 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. Most questions can be broken down into one of four categories:

- **Lookup questions** — These can be 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 "is Sunflower fixed?" or "How many legs does Felix have?"
- **Compute questions** — These can be answered by computing an answer across a single column. Examples of computing questions might be "how much does the heaviest animal weigh?" or "What is the average age of animals from the shelter?"
- **Relate questions** — These ones take the most work, because they require looking for relationships between multiple columns. Examples of analysis questions might be "Do cats tend to be adopted faster than dogs?" or "Are older animals heavier than young ones?"
- **Can't answer** — These are questions that just can't be answered based on the available data. We might ask "are cats or dogs better for elderly owners?", but the Animals Dataset doesn't have information that we can use to answer it.

Investigate

- Come up with examples for each type of question.
- Look back at the Wonders you wrote on ../../lessons/ds-intro/pages/exploring-animals-dataset.adoc. Are any of these Lookup, Compute, or Relate questions? Circle the question type that's appropriate. Can you come up with additional examples for each type of question?

Synthesize

Have students share their questions with the class. Allow time for discussion!

Closing

Debrief with the class, and have students reflect on what they learned by writing on What's on your mind? (Page 5). Some prompts that may be helpful:

- What new vocabulary did you learn?
- What question was exciting to you, and what data would you need to answer it? Is that data Qualitative or Quantitative?
- What do you hope to learn in the next lesson?

Additional Exercises:

- What Questions Can You Answer?