Introduction to Data Science

(Also available in Pyret)

Students learn about Categorical and Numeric data, are introduced to Tables by way of the Animals Dataset, and consider what questions can and cannot be answered with available data.

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
Student-facing Lesson Goals	• Let's learn about data inside tables.
Materials	• PDF of all Handouts and Page • Animals Spreadsheet • Animals Starter File • Lesson Slides • Printable Lesson Plan (a PDF of this web page)
Supplemental Materials	• Additional Printable Pages for Scaffolding and Practice What's Going On in This Graph?
Preparation	• **You know your students best!** You can use the sample Opening Questions we've provided, but we recommend changing or adding your own questions that are *appropriate, relevant, and engaging* for your students. • Decide how the first activity (opening questions) will be run: will questions be printed for each student, group of students, or posted around the room? • Make sure student computers can access Animals Spreadsheet and the Animals Starter File.

Glossary

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

data :: pieces of information about a group of individuals or things

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

quantitative data :: number values for which arithmetic makes sense

sample :: a set of individuals or objects collected or selected from a statistical population by a defined procedure

Questions that Matter *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

- Look at the provided list of Opening Questions, and take one minute to select a question that grabs your attention. Arrange yourselves into groups based on the question you like, making sure that each group has between 2-5 people.

- Have each person in the group quickly share their *gut reaction*: What do you think the answer is?
- After sharing initial reactions, have each person share their reasoning.
- Does everyone in your group agree about the answers to their question?

Note: Students are VERY likely to try and explain their reasoning as soon as they give their gut answers. This can taint the answers of other students in the group - emphasize that this is about exposing our "gut reaction" or bias.

Investigate

What information would you collect to answer the question you selected? *Take 5 minutes to think about what information you would need to collect, to find the answer.*

Common 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

Data is any piece of information about a group of individual or things. In this classroom, we could collect data about student names, ages, favorite foods, and so on.

For each group…

- What were your gut reactions?
- Did the question wind up being too vague? What did you need to do to make it specific?
- What ***data*** would you gather?
- What, if anything, were you surprised about?
    - 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.
    - We can't survey every school in the world (or get data on every movie ever made, or every police action!) but we can analyze 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 people from hundreds of other different fields to ask and answer questions about data.

Optional: Which Questions *Can* we Answer?

Datasets are useful for answering questions, but they can't answer all the questions that we will wonder about for a given topic. In this activity students will look at a small dataset about a cyclist's training rides and think about how they could use the table to answer each question or why a question cannot be answered using the table.

Which of you like to ride bikes? What data might you collect about bike rides? Open to What Questions Can You Answer with the Given Data? This page includes a small dataset about a cyclist's training rides and a set of questions. The data can be used to answer some, but not all, of the questions. With your partner, read each question. If it can be answered with what we know, explain how you could use the table to answer it. If it can't be answered using the table, explain why not.

Meet the Animals! *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 is a widely-used best practice in Math-Ed, 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**?". These phases should be done in groups or as a whole class, with ample time given to both Notice and Wonder.

Launch

Open the Animals Spreadsheet in a browser tab, or turn to The Animals Dataset.

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.

Notice?

Wonder?

- Turn to Questions and Column Descriptions. 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 "Answered by Dataset" — circle "Yes" if your Wonder can be answered by the dataset or "No" if it can't.

Have students share back their noticings (statements) and wonderings (questions), and write them on the board. Ask the class if each Wonder can be answered by the data, making sure that they have a few questions that *can* be answered, and a few that *can't*. Also ask if some of their wonderings are about a group as a whole, rather than just individuals.

- If you look at the bottom of the Animals Spreadsheet, 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. Sometimes it can be tricky to figure out if data is categorical or quantitative, because it depends on *how that data is being used!*

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

Categorical Data is used to *classify*, not measure. The laws of arithmetic do not make sense when it comes to categorical data.

- "Species" is a categorical variable, because we can ask questions like "which species does Mittens belong to?"
- 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, because ZIP codes identify locations, not amounts.

- What are some other categorical variables you see in this table?
  - *Name, Sex, and Fixed*

Quantitative Data - sometimes referred to as Numeric Data - is used to measure an amount of something, or to compare two pieces of data to see how much *less or more* one is compared to the other.

- "Pounds" is a quantitative variable, because we can talk about how much more one animal weighs more than another or ask what the average weight of animals in the shelter is.

- If we want to ask “how much” or “which is most”, we’re talking about Quantitative Data.

- What are some other quantitative variables in this table?
  - *Age, Legs, Weight, and Time to Adoption*

Complete Categorical or Quantitative?. Be sure to discuss your answers with your partner or group!

Synthesize

Data Science is all about making educated guesses about an entire group (called the population) based on data about a subset of that group (called the ***sample***). 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.