Choosing Your Dataset

(Also available in CODAP)

Students practice making a variety of chart types and then begin to investigate a real world dataset, which they will continue to work with for the remainder of the course.

Lesson Goals	Students will be able to… • Explain why they chose their dataset • Describe their dataset • Make subsets from their dataset
Student-facing Lesson Goals	• Let’s all choose an interesting dataset to investigate.
Prerequisites	• Simple Data Types • Contracts: Making Tables and Displays
Materials	• PDF of all Handouts and Page • Data Exploration Project Slide Template • Lesson Slides • Printable Lesson Plan (a PDF of this web page)
Supplemental Materials	• *Global Food Supply & Production Starter File* • *Blank Dataset Starter File for Bootstrap:Data Science* • *Tutorial Video: Importing Your Own Data into Pyret*

Preparation

- Decide how much choice you’re ready to offer your students before you begin. Research shows that choice increases student engagement! But focusing the whole class on a single dataset is also an option.
  - Would focusing your students on a single dataset make this doable for you? Because you teach younger students who might need more scaffolding? Or because you are new to teaching data science and managing fewer moving parts would increase your confidence? We recommend focusing on Global Food Supply & Production Starter File.
  - Are you ready to jump straight into supporting your students in working on a wide range of topics of their choosing? We have a full dataset library!
  - Want to give students choice from a shorter curated list…to shorten the decision-making process, focus on topics related to curriculur goals, or just to have fewer options to manage during class? We’ve assembled descriptions of individual datasets here. For those looking for a precurated shorter list, we’ve starred a few of them for you.
  - **If you have time**, you may want to complete all of the lessons with everyone getting extra practice analyzing Global Food Supply & Production Starter Fileand then have your students choose a dataset to analyze for their culminating research papers!

Glossary

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

dataset :: a collection of related information that is composed of separate elements, but can be manipulated as a unit by a computer

random sample :: a subset of individuals chosen from a larger set, such that each individual has an equal probability of being chosen

statistical inference :: using information from a sample to draw conclusions about the larger population from which the sample was taken

Review: Consider Data *20 minutes*

Overview

Students practice making lots of chart types, focusing specifically on the "Consider Data" step in the Data Cycle and how it can be used alongside Contracts to help go from questions to code.

Launch

The Data Cycle is a roadmap that guides us in the process of data analysis. You’ve learned that the Data Cycle includes four steps. Let’s review what those steps entail.

- In the **Ask Questions** phase of the Data Cycle, what are some of the different types of questions we can ask?
  - *Lookup, arithmetic, and statistical questions.*
- What’s the difference between an arithmetic question question and a statistical question?
  - *A statistical question does not specify a particular arithmetic process, while an arithmetic question does.*
- What does the **Consider Data** phase entail?
  - *We need to ask two questions: "What rows should we investigate?" and "What columns do we need?"*
- During the **Analyze Data** phase of the Data Cycle, we choose what kind of display we’ll need to answer our question. Which two displays work with categorical data? Why might you choose one over the other?
  - *Bar and pie charts workwith categorical data. A pie chart only makes sense when you have the full picture, whereas a bar chart shows the count. .*
- In your own words, what happens during the **Interpret the Data** phase?
  - *We answer questions and summarize results, which often leads to new questions.*

Investigate

In this lesson, we’re going to get some practice with the second step of the cycle - Consider Data. This entails isolating the Rows and Columns needed to answer various questions, and using our knowledge of Contracts to help turn those questions into working code!

Complete Consider and Analyze.

Be sure to review student answers.

Synthesize

- What strategies did you use to determine which columns to isolate?
- Why do the contracts for some displays require more arguments than others?

Choosing a Dataset *30 minutes*

Overview

Students select a dataset that interests them, and do some thinking about *why* it interests them, *what questions they'd like to answer* and *what hypotheses they have*. They'll be analyzing this data for a long time, so it's critical to ensure a high degree of buy-in before signing off on a student's choice!

Launch

Note: **If you are opting to focus your whole class on a single dataset, we recommend skipping this section of the lesson.** *You'll instead want to jump to "Dataset Exploration Project.")*

Data Science: it's all about YOU!

What data matters to ***you?*** What questions do ***you*** care about? We live in a world filled with data, gathered about almost every subject you can imagine.

- Climate sensors are gathering data on temperature, humidity, oxygen and more…practically everywhere on the globe.
- Census data tracks the number of different groups of people, as well as their education, income level, and more.
- Companies like Facebook, Amazon, and Google gather massive amounts of data on the websites you visit, what you chat about online, what you purchase, etc.

This data is used to set public policy, draw voting districts, approve drugs, calculate school funding, decide which advertisements you see, and more.

- Where else do you see data being gathered?
- What are some other ways data is used in the world around you?

What follows is a list of every ***dataset*** already provided to students, with a corresponding Starter File that instantly imports the (cleaned) data into Pyret. We suggest giving students a direct link to this page, which lists all of the relevant links found in the lesson plan.

Students can also find their own dataset, and use this Blank Dataset Starter File for Bootstrap:Data Science. For help, see this Tutorial Video: Importing Your Own Data into Pyret.

For teachers using a single dataset, we recommend using Global Food Supply & Production Starter File. This dataset focuses on global food supply and production through environmental / geographic / cultural lenses and the variables were carefully selected to make sure it lends itself well for all kinds of data displays and discussions. You can, of course, opt to choose any dataset you'd like, from our library or otherwise.

NOTE: *We have compiled some Notes on our provided datasets, to help you decide which might be most useful in your classroom.*

Investigate

Have students choose a dataset that is interesting to them and save a copy of it in their programs!

Looking for a shorter list? We've starred a few good beginner datasets.

The Environment & Health

Global Waste by Country 2019	[Dataset Starter File]
World Cities' Proximity to the Ocean	[Dataset Starter File]
Earthquakes	[Dataset Starter File]
Air Quality, Pollution Sources & Health in the U.S.	[Dataset Starter File]
Health by U.S. County	[Dataset Starter File]
COVID in the U.S. by County	[Dataset Starter File]
Arctic Sea Ice	[Dataset Starter File]

Politics

Countries of the World	[Dataset Starter File]
Gerrymandering	[Dataset Starter File]
Marijuana Laws & Arrests by State 2018	[Dataset Starter File]
LAPD Arrests 2010-2019	[Dataset Starter File]
NYPD Stop, Search & Frisk 2019	[Dataset Starter File]
Refugees 2018	[Dataset Starter File]
State Demographics	[Dataset Starter File]
U.S. Income	[Dataset Starter File]
U.S. Jobs	[Dataset Starter File]
U.S. Voter Turnout 2016	[Dataset Starter File]

Sports

Esports Earnings	[Dataset Starter File]
MLB Hitting Stats	[Dataset Starter File]
NBA Players	[Dataset Starter File]
NFL Passing	[Dataset Starter File]
NFL Rushing	[Dataset Starter File]

Entertainment

★Movies	[Dataset Starter File]
IGN video game Reviews	[Dataset Starter File]
International Exhibition of Modern Art	[Dataset Starter File]
North American Pipe Organs	[Dataset Starter File]

Pokemon	[Dataset Starter File]
Music	[Dataset Starter File]

Education

College Majors	[Dataset Starter File]
U.S. Colleges 2019-2020	[Dataset Starter File]
★R.I. Schools	[Dataset Starter File]
Evolution of College Admissions in California	[Dataset Starter File]

Nutrition

Soda, Coffee & Other Drinks	[Dataset Starter File]
Fast Food Nutrition	[Dataset Starter File]

Would you like to contribute a dataset of your own, or is there something you'd like to change about one of ours?

Synthesize

- What did you select, and why?
- What questions did you come up with?

For the rest of this course, you'll be learning new programming and ***Data Science*** skills, practicing them with the Animals Dataset and then applying them to you own data.

Dataset Exploration Project *flexible*

Overview

Students are introduced to the Dataset Exploration Project. They will apply what they have learned to add four items to their Data Exploration Project Slide Template: (1) a description their dataset, including its source, structure, and relevance, (2) at least one bar chart, (3) at least one pie chart, and (4) any interesting questions they develop. To learn more about the sequence and scope of the exploration project, visit Project: Dataset Exploration.)

Launch

Today, we are going to start digging into the datasets we've chosen to study at length. Each time we learn about a new data science concept in this class, we will add displays, questions, and analyses to the Data Exploration Project Slide Template.

- Open the Data Exploration Project Slide Template.
- Create and save your own copy of the slide deck.
- Let's take a look! Peruse the slides to get a sense of what this cumulative project includes.
- What do you Notice? What do you Wonder?
  - *Students will likely notice that many displays they are unfamiliar with are referenced. They may wonder how there is going to be so much analysis on just one dataset!*

Encourage students to familiarize themselves with the template, highlighting some important features:

- Blue text is included to provide examples.
- Slides can be duplicated if students want to add additional displays or interpretations.

Investigate

By now you've already learned what to do when you approach a new dataset. Think back to your first exposure to the Animals Dataset. You read the data and wrote down your Notices and Wonders. You described the columns. You even took some ***random samples*** of the dataset to explore ***inference*** and probability.

Now, you're doing to do the same thing *with your own dataset.*

- Open your chosen dataset starter file in Pyret.

- Look at the spreadsheet or table for your dataset. What do you **Notice**? What do you **Wonder**?
- Complete My Dataset, making sure to include at least two questions that *can* be answered by your dataset and one that *cannot*.
- Save a copy of your starter file. In the Definitions Area, use `random-rows` to define **at least three** tables of different sizes: `tiny-sample`, `small-sample`, and `medium-sample`.

Today we will begin adding to our Data Exploration Project Slide Template. First, we are going to describe our dataset.

- **It's time to add to your Data Exploration Project Slide Template.**
- Complete all of the slides you see in the "About this Dataset" portion of the slide deck. It may be helpful to refer to My Dataset.

Ensure that students have thoughtfully described their datasets. Then, explain that they are going to add bar and pie charts, along with their interpretations of them.

- Choose one categorical column from your dataset that you will represent with a bar chart.
- What question does your display answer?
- Now, write down that question in the top section of Data Cycle: Categorical Data.
- Complete the rest of the data cycle, recording how you considered, analyzed, and interpreted the question.
- Repeat this process for at least one more categorical column - but this time, create a pie chart.

Once students have at least one bar and pie chart, it's time to add their findings to the Data Exploration Project Slide Template.

Copy/paste at least one bar chart and one pie chart into your slide deck. Be sure to also add any interesting questions that you developed while making and thinking about these displays.

You may need to help students locate the "Bar Charts" section and the "Pie Charts" section. The "My Questions" slide is at the end of the template.

Synthesize

Share your findings with the class!

Did you discover anything surprising or interesting about your dataset?

What questions did the bar and pie charts inspire raise?

Did other students make any discoveries that were surprising or interesting to you?