Choosing Your Dataset

Students summarize their dataset by exploring the data and identifying categorical and quantitative columns, datatypes, and more. They also define a few sample rows, random subsets, and logical subsets.

Prerequisites	Grouped Samples
Relevant Standards OK K12CS CSTA NGSS	*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 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.
Materials	• Lesson Slides (Google Slides) • Computer for each student (or pair), with access to the internet • Student workbook, and something to write with
Preparation	• Make sure all materials have been gathered • Decide how students will be grouped in pairs • All students should log into CPO and open the "Animals Starter File" they saved from the prior lesson. If they don't have the file, they can open a new one
Supplemental Resources	

Language Table

Types	Functions	Values
Number	`num-sqrt, num-sqr`	`4, -1.2, 2/3`
String	`string-repeat, string-contains`	`"hello", "91"`
Boolean	`==, <, <=, >=, string-equal`	`true, false`
Image	`triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized`	
Table	`count, .row-n, .order-by, .filter, .build-column, random-rows`	

The Data Cycle

20 minutes

Overview

Students learn about the *Data Cycle*, which helps them get situated in the process of analyzing the datasets they will select in this lesson. They browse through the library of provided datasets, and choose one they want to work with. NOTE: the

selection process can also be done as a homework assignment, if all students have internet access at home.

Launch

Zoom out a little and help students reflect on what they've done so far. Students began by exploring the Animals Dataset, formulating questions and exploring them with data displays. This led to further questions, making subsets, and asking more questions.

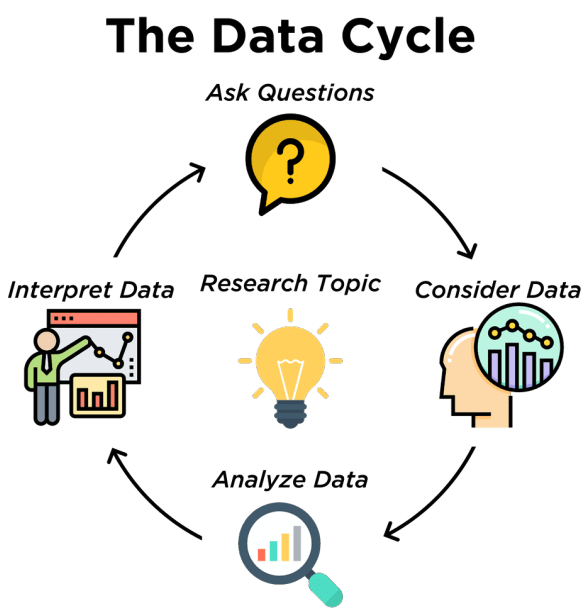
The Data Cycle[*] is a *roadmap*, which helps guide us in the process of data analysis.

(Step 1) We start by **Asking Questions** - statistical questions that can be answered with data.

(Step 2) Then we **Consider Data**. This could be done by conducting a survey, observing and recording data, or finding a dataset that meets our needs.

(Step 3) Then it's on to **Analyzing the Data**, in which we produce data displays and new tables of filtered or transformed data in order to identify patterns and relationships.

(Step 4) Finally, we **Interpret the Data**, in which we answer our questions and summarize the results. As we've already seen from the Animals Dataset, these interpretations often lead to *new* questions....and the cycle begins again.



Explain to students that they will now select a dataset for them to work with for the remainder of the course. Make sure they understand that it genuinely has to be something they are interested in - their engagement with the data is critical to engaging with the class.

Students can also find their own dataset, and use this (Blank Starter file). See this tutorial video for help importing your own data into Pyret.

Students must have at least 2 questions that are both *interesting* and *answerable* using their dataset.

Investigate

Choose a dataset that is interesting to you! You should have at least two questions that the dataset can help you answer. Write these questions down on What's on your mind? (Page 49).

Movies	[Dataset Starter File]
Schools	[Dataset Starter File]
US Income	[Dataset Starter File]
US Presidents	[Dataset Starter File]
Countries of the World	[Dataset Starter File]
Music	[Dataset Starter File]
NYC Restaurant Health Inspections	[Dataset Starter File]
Pokemon Characters	[Dataset Starter File]
IGN Video Game Reviews	[Dataset Starter File]
2016 Presidential Primary Election	[Dataset Starter File]
US Cancer Rates	[Dataset Starter File]
US State Demographics	[Dataset Starter File]
Sodas	[Dataset Starter File]
Cereals	[Dataset Starter File]
Summer Olympic Medals	[Dataset Starter File]
Winter Olympic Medals	[Dataset Starter File]
MLB Hitting Stats	[Dataset Starter File]
Spotify Top Songs	[Dataset Starter File]

Open the Research Paper template, and save a copy.

- Students fill in their first and last name(s), the teacher name on the first page of the Research Paper.
- Students should also copy the link to the dataset (spreadsheet), and paste it into the first page of the Research Paper.
- Students should click "Publish" in their Pyret Starter File, then copy/paste the resulting link into the first page of the Research Paper.

Synthesize

Have students share their datasets and their questions.
For the rest of this course, students will be learning new programming and Data Science skills, practicing them with the Animals Dataset and then applying them to their own data.

Exploring Your Dataset

flexible

Overview

Students apply what they've learned about describing and making subsets from the Animals Dataset to their own dataset.
Note: this activity can be done briefly as a homework assignment, but we recommend giving students an *additional class period* to work on this.

Launch

By now you've already learned what to do when you approach a new dataset. With the Animals Dataset, you first read the data itself, and wrote down your Notice and Wonders. You described the columns in the Animals Dataset, identifying which were categorical and which were quantitative, and whether they were Numbers, Strings, Booleans, etc. Finally, you used the Design Recipe and table methods to make random and logical subsets.
Now, you're doing to do the same thing *with your own dataset*.

Investigate

- Have students look at the spreadsheet for their dataset. What do they **Notice**? What do they **Wonder**? Have them complete My Dataset (Page 45), making sure to have at least two Lookup Questions, two Compute Questions, and two Relate Questions.
- In the Definitions Area, students use `random-rows` to define **at least three** tables of different sizes: `tiny-sample`, `small-sample`, and `medium-sample`.
- In the Definitions Area, students use `.row-n` to define **at least three** values, representing different rows in your table.
- Have students think about subsets that might be useful for their dataset. Name these subsets and write the Pyret code to test an individual row from your dataset on Samples from My Dataset (Page 46).
- Students should fill in My Dataset portion of their Research Paper.
- Students should fill in Categorical Visualizations portion of their Research Paper, by generating pie and bar charts for their dataset and explaining what they show.

Turn to The Design Recipe (Page 47), and use the Design Recipe to write the filter functions that you planned out on Samples from My Dataset (Page 46). When the teacher has checked your work, type them into the Definitions Area and use the `.filter` method to define your new sample tables.

Choose one categorical column from your dataset, and try making a bar or pie-chart for the whole table. Now try making the same display for each of your subsets. Which is most representative of the entire column in the table?

Synthesize

Have students share which subsets they created for their datasets.
[*] From the Mobilizing IDS project and GAISE