# **Choosing Your Dataset**

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

| **Prerequisites** | [Grouped Samples](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-grouped-samples/index.shtml) |
| --- | --- |
| **Relevant Standards** | *Common Core Math Standards*  **6.SP.A.1**  Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers.  *Next-Gen Science Standards*  **HS-SEP1-3**  Ask questions to determine relationships, including quantitative relationships, between independent and dependent variables. |
| **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](https://docs.google.com/presentation/d/1SRZ9xk69gfUrKvreqX7hf8BiYEmr7ZyuUHIk779fieU/) * [Animals Starter File](https://code.pyret.org/editor#share=1ZupMVPWvVUOM0HCWyA7cRBghSLKxPWv1) * [Research Paper template](https://docs.google.com/document/d/1_ZEIgM4zvxI7JizViVFZojnpd3Yr2rYe8puPk8pjOcs/copy) * [My Dataset (Page 45)](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/pages/my-dataset.html) * [Samples from My Dataset (Page 46)](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/pages/samples-from-my-dataset.html) * [The Design Recipe (Page 47)](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/pages/design-recipe-helper-funs.html) * [What’s on your mind? (Page 49)](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/pages/whats-on-your-mind.html) |
| **Preparation** | * Make sure all materials have been gathered. * Decide how students will be grouped in pairs. * Computer for each student (or pair), with access to the internet * [Student workbook](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science//workbook/workbook.pdf), and something to write with * All students should log into [CPO](https://code.pyret.org/) 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](https://code.pyret.org/editor#share=1ZupMVPWvVUOM0HCWyA7cRBghSLKxPWv1). |
| **Supplemental Resources** | [Jump to Datasets and Starter files](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/index.shtml#Datasets) |
| **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

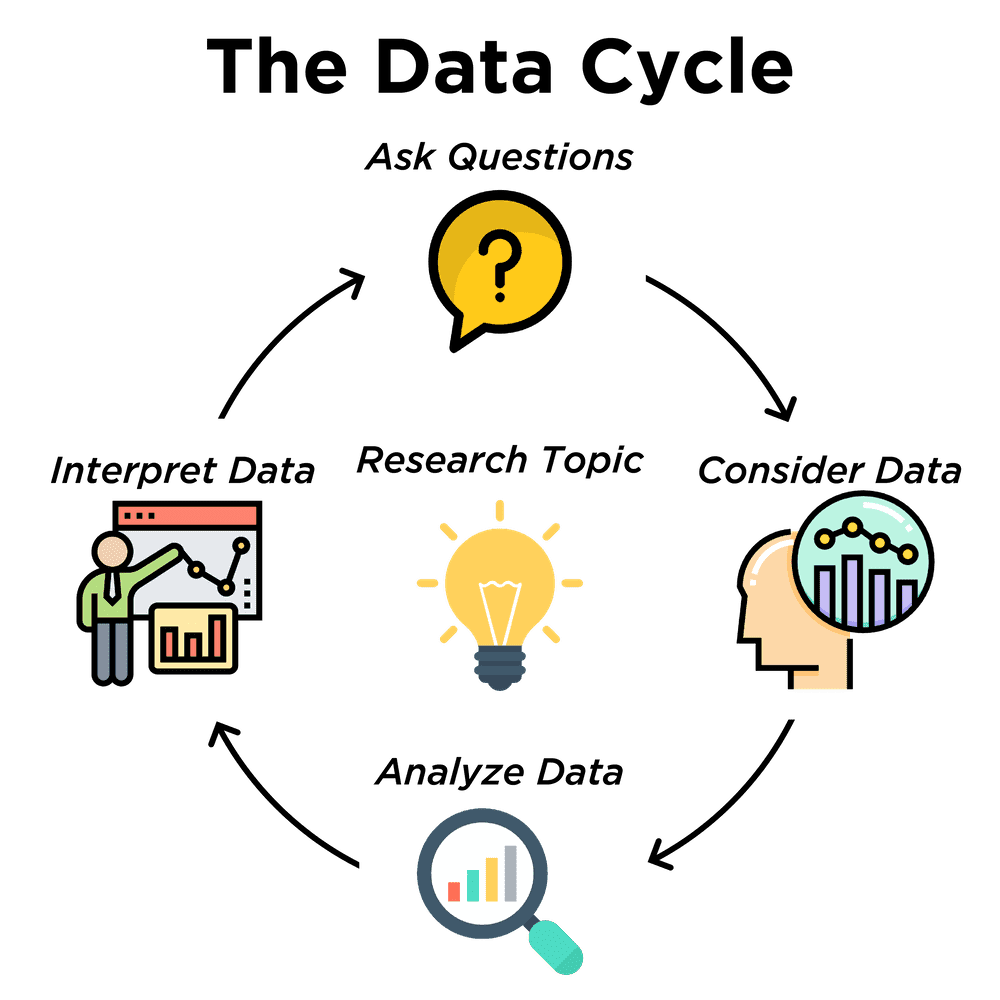
## 20 minutes

### *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](https://code.pyret.org/editor#share=12v7kzbxHt2LaSe2uI2d_OnssxtTNF0A8). See this [tutorial video](https://youtu.be/K4n9hTSqcyw) 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*

Have students choose a dataset that is interesting to them! They should have at least two questions that the dataset can help them answer, and write them on [What’s on your mind? (Page 49)](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/pages/whats-on-your-mind.html).

**World Cities' Proximity to the Ocean**

[Dataset](https://docs.google.com/spreadsheets/d/166F2V0uPtAIiU4BkITu8pDmU2hnPIWJaM3yDoOHyon0/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1zfTxSO8wGMGbXbLi5CZWEDxl5vtzqobz&v=1599623)

**NYPD Stop, Search & Frisk 2019**

[Dataset](https://docs.google.com/spreadsheets/d/1XrJPOeIJCavRjP1nHbWYKcGo79dO-x_XidLplJabGko/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1koQ_PSQgkm7ZqFxHWE47ZzGlwmF039iU&v=1599623)

**Marijuana Laws & Arrests by State 2018**

[Dataset](https://docs.google.com/spreadsheets/d/1TB0Lj88IH1bdAAKAxwsFMW-x-B-JtnLate4ArmSWDzk/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1NN8p5TncYXKcXFE1xtNZ1YSwSc0y0vxm&v=1599623)

**Police Traffic Stops, Durham, NC, 2002-2013**

[Dataset](https://docs.google.com/spreadsheets/d/1SruzpUa7zXLL49sOnbjs5dj6vjCmWnop1lk_Ejw8L6c/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1uq312iiygxSObuIRV-nJCFeXxvAipvpV&v=8c4da7d)

**College Majors**

[Dataset](https://docs.google.com/spreadsheets/d/1wIpbDIBQMjAwaLGNK-e5snmzui0DuWFdrkoSNVv9m5c/) [Starter File](https://code.pyret.org/editor#share=1qXgUV3UNjv-fSPJrZGrz9_PcFrg0Oc6e&v=1599623)

**US Jobs**

[Dataset](https://docs.google.com/spreadsheets/d/1fAzyoVgtSMl9ja-JMpou_Y5RRyoTOPh2umR_mkJYQyU/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1-zjSzBXoYgTgIqIofzZxfYiUhYiO1y2z&v=1599623)

**Refugees 2018**

[Dataset](https://docs.google.com/spreadsheets/d/1mDSr7CcpSO4aL-iV2oEfYLecssvis2Je6hN3vsomYuc/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1AgA3NuWLBjXoan6_hBVLBlvoDOjihsdB&v=1599623)

**R.I. Schools**

[Dataset](https://docs.google.com/spreadsheets/d/1XeeyAuF_mtpeCw2HVCKjvwW1rreNvztoQ3WeBlEaDl0/) [Starter File](https://code.pyret.org/editor#share=1Xu0ket-eDCeU-Uuxl82OCCGJQRW_ghWV)

**Movies**

[Dataset](https://docs.google.com/spreadsheets/d/1p2T_pHKQYubV5h4RC9CFH0W_z6ZMl_sBpGxhyKxglZo/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1GFXwGm9GZzTqzpDKYk5mdaBbZy4llpTC&v=8c4da7d)

**Fast Food Nutrition**

[Dataset](https://docs.google.com/spreadsheets/d/12yhGT-p1yMXXig27pvMEkC_E5a3tLRlXA1sXopHgwzI/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1NWAcVKR1qi2vOTyAFlclMu1tsb9tYSmS&v=1599623)

**Beverages Nutrition**

[Dataset](https://docs.google.com/spreadsheets/d/1QcPosMRFMrgayav6W3SfRjdtCn5oF_CSvoJPMmA2fJM/) [Starter File](https://code.pyret.org/editor#share=1Hb7AS7v5Nh8k4WWU2eS1iChfKt-tbD8Q&v=1599623)

**North American Pipe Organs**

[Dataset](https://docs.google.com/spreadsheets/d/1IlR9rvo4gQzcynhj4rjf_6mm2gwt5bl34j-__PgL7L0/) [Starter File](https://code.pyret.org/editor#share=1jssxWzykC98fyZtDTfnFWiw59gJI9LUF&v=1599623)

**International Exhibition of Modern Art**

[Dataset](https://docs.google.com/spreadsheets/d/108ZnfCy3V2WkpSmjYpane6gmogM3EtBLL1sD-9h0Acc/) [Starter File](https://code.pyret.org/editor#share=1UKMJekyVOkgCDNUHB0TWsQqYiJ5xGfzQ&v=1599623)

**Esports Earnings**

[Dataset](https://docs.google.com/spreadsheets/d/1SqK3BP-RKrN9adFQz2XvZUMHCAPsm530XT1hC96L6w0/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1_85l5tpS7eBt8g2eVMxHo4F9AwvQjlWT&v=1599623)

**MLB Hitting Stats**

[Dataset](https://docs.google.com/spreadsheets/d/1xjC1XZWACvQtfwHdGk_BlE2jm4aleMADHTt6PEocCjg) [Starter File](https://code.pyret.org/editor#share=1X80kXcexm4kriNFTJXLVr_8m2TZOvhzr)

**NBA Players**

[Dataset](https://docs.google.com/spreadsheets/d/1NQbLNeV5ezuAJnLNirVOb9ls8Firk4YXLzksG5RaUgA/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1a6eoAs3QD4uuffUG9P8izfyO1WnwnR2n&v=1599623)

**NFL Passing**

[Dataset](https://docs.google.com/spreadsheets/d/1tpHZdUZQ0Fzuy1G1qqPPoKS0p6PkG3hb_P_013kcVIo/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1jNSpaWNCak-C3-MLslW8RaWn0rFwS5jb&v=1599623)

**NFL Rushing**

[Dataset](https://docs.google.com/spreadsheets/d/1o8d0k46L8mkIIIpgYHXtMLxDzjXE6l1epN5tWLHyG6o/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1CFYvepEsI9cGcCn-GuZxJiljtT7KswPx&v=1599623)

**Wrestler Earnings**

[Dataset](https://docs.google.com/spreadsheets/d/10kf5ZTwXAHGMBv2n8xbgIwpnsOLAL-RtLqPi83L7s8s/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1AwXrQ7jqZfl8gokxmFDH72mlIF2BYxPh&v=1599623)

**U.S. Voter Turnout Rates 1986-2018**

[Dataset](https://docs.google.com/spreadsheets/d/1wr65EkLFAIuTHpP_TWMNjFUZ-UOF-Yvo9xKpxwioegE/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#share=1RlFah6-PiX6yOW99TWLEbueQtaKxi3_Q&v=1599623)

**State Demographics**

[Dataset](https://docs.google.com/spreadsheets/d/14fRmjAiCqGNI5Zr5kGSDjMavln7ILqhSX3N3zGd8CAU/edit) [Starter File](https://code.pyret.org/editor#share=1IzvOGVQdhi-QMKJdl7IBpM1vGSjaV_P_)

**Countries of the World**

[Dataset](https://docs.google.com/spreadsheets/d/19VoYxPw0tmuSViN1qFIkyUoepjNSRsuQCe0TZZDmrZs) [Starter File](https://code.pyret.org/editor#share=1wEHqx0f9VOxmRUVqDjnPCECOmvPWyQdS)

**U.S. Income**

[Dataset](https://docs.google.com/spreadsheets/d/1cIxBSQebGejWK7S_Iy6cDFSIpD-60x8oG7IvrfCtHbw) [Starter File](https://code.pyret.org/editor#share=1qSK5KX7cfwM4c6XtJFg5gPcVp9OBSbOU)

**U.S. Presidents**

[Dataset](https://docs.google.com/spreadsheets/d/1Frt37-nBHHxvJVBKzKLRD3kbjPLhc8CYUaIlNeNWl94) [Starter File](https://code.pyret.org/editor#share=1prkKxnGrGJJAab7XBXTqiysRli0xFEbL)

**Music**

[Dataset](https://docs.google.com/spreadsheets/d/1Yzo8GuGhMDVNyAI5OacmKZ53xJplZbXF5FT6Lwitp0w) [Starter File](https://code.pyret.org/editor#share=1UIYszUXHA0gx7jmFhEitvKxqa7MDmcxO)

**Summer Olympic Medals**

[Dataset](https://docs.google.com/spreadsheets/d/1ee30kHpV35zAO5MNQKk_nXP6iym2mX-bv_cgt-8q_oo) [Starter File](https://code.pyret.org/editor#share=1bkI_WiWZvLlS5x3A0vx6C-m-fJonIPW4)

**Winter Olympic Medals**

[Dataset](https://docs.google.com/spreadsheets/d/1-xYW4C0IRB7cDI2K8dMyVTlsQjFmB_Z4XBsHsB-TAbs) [Starter File](https://code.pyret.org/editor#share=1kFV_BmSDTSAbNDdm-IFZrGdI1I6K4-aL)

**Pokemon Characters**

[Dataset](https://docs.google.com/spreadsheets/d/1S8jf4Qf94TJKGLCcTA-Fqn4YXE7dGf_PIxv5MUeUPVo) [Starter File](https://code.pyret.org/editor#share=1Nxhiw46EFfNhmMBvUQK0NosbFnaBCQJy)

**IGN Video Game Reviews**

[Dataset](https://docs.google.com/spreadsheets/d/1Ss221kjz2WJUsTlxK7TcnsXLPoSbnfUKv-JP8gCiGRw) [Starter File](https://code.pyret.org/editor#share=1IVnp6-NCuvbk1cQH6CRly_6I6-OEBOVI)

**2016 U.S. Presidential Elections**

[Dataset](https://docs.google.com/spreadsheets/d/1WP-aE8jLa6fAWntG-qukev1nGAe3xO1RX7AMUZWOXEY/edit?usp=sharing) [Starter File](https://code.pyret.org/editor#program=1jhMJusTzXfRaVDwdl8u2o6st_1NgbWn3)

**U.S. Cancer Rates**

[Dataset](https://docs.google.com/spreadsheets/d/1deL-6Xh9WUN2fTcV95V7RLseAcplGNA6BHTnpHCMgME) [Starter File](https://code.pyret.org/editor#share=1qJxOMxULihXhtrIVaLPboFtYpUiLnFct)

**Sodas**

[Dataset](https://docs.google.com/spreadsheets/d/15n0dLqBWffE2JNOmYHcvavqMwvHXpy5_UyZfT3Q7pfs) [Starter File](https://code.pyret.org/editor#share=12AFxlJuNjmxYTUV0FWa7z3D5CHRYBXoG)

**Cereals**

[Dataset](https://docs.google.com/spreadsheets/d/1y3AoywSnyGpu-QmmEwKvW-xstZ6B9JhH5gTUx5XYTo4) [Starter File](https://code.pyret.org/editor#share=1syDC8qoFjO4EdqycRAfWlKMHanzrqTSH)

Open the [Research Paper template](https://docs.google.com/document/d/1_ZEIgM4zvxI7JizViVFZojnpd3Yr2rYe8puPk8pjOcs/copy), 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.

We have also compiled some [notes on these datasets](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/resources/pages/datasets-overview.html), which we recommend for all teachers before having their students choose a dataset.

### *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 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)](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/pages/my-dataset.html), 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)](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/pages/samples-from-my-dataset.html).
* Students should fill in the My [Dataset](https://docs.google.com/document/d/1_ZEIgM4zvxI7JizViVFZojnpd3Yr2rYe8puPk8pjOcs/edit#heading=h.6cy9t2stox4e) portion of their Research Paper.
* Students should fill in the Categorical [Visualizations](https://docs.google.com/document/d/1_ZEIgM4zvxI7JizViVFZojnpd3Yr2rYe8puPk8pjOcs/edit#heading=h.h6dvbsrt7r00) 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)](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/pages/design-recipe-helper-funs.html), and use the Design Recipe to write the filter functions that you planned out on [Samples from My Dataset (Page 46)](https://bootstrapworld.org/materials/spring2021/en-us/courses/data-science/lessons/ds-choosing-your-dataset/pages/samples-from-my-dataset.html). 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](http://introdatascience.org/) and [GAISE](https://www.amstat.org/asa/files/pdfs/GAISE/GAISEPreK12_Intro.pdf)