Grouped Samples

(Also available in [CODAP](#))

Students practice creating grouped samples (non-random subsets) and think about why it might sometimes be useful to answer questions about a dataset through the lens of one group or another.

Lesson Goals	Students will be able to… • Make grouped samples from a population
Student-facing Lesson Goals	• Let's combine what we know about sampling and filtering with creating displays.
Prerequisites	• [Simple Data Types](#) • [Contracts: Making Tables and Displays](#) • [Contracts](#) • [Functions: Contracts, Examples & Definitions](#) • [Solving Word Problems with the Design Recipe](#) • [Row and Column Lookups](#) • [Choosing Your Dataset](#) • [Table Functions: Bringing it all Together](#)
Materials	• [PDF of all Handouts and Page](#) • [Grouped Samples Starter File](#) • [Lesson Slides](#) • [Printable Lesson Plan](#) (a PDF of this web page)

Glossary

grouped sample :: a non-random subset of individuals chosen from a larger set, where the individuals belong to a specific group

ratio :: the relative sizes of two or more values

Problems with a Single Population   *10 minutes*

Overview

This activity is all about *grouped samples*: Students make a bunch of non-random samples from the Animals Dataset, and see how each samples might answer the same question differently.

Launch

When looking at a scatter plot of animals, it looks like the amount an animal weighs may have something to do with how long it takes to be adopted.
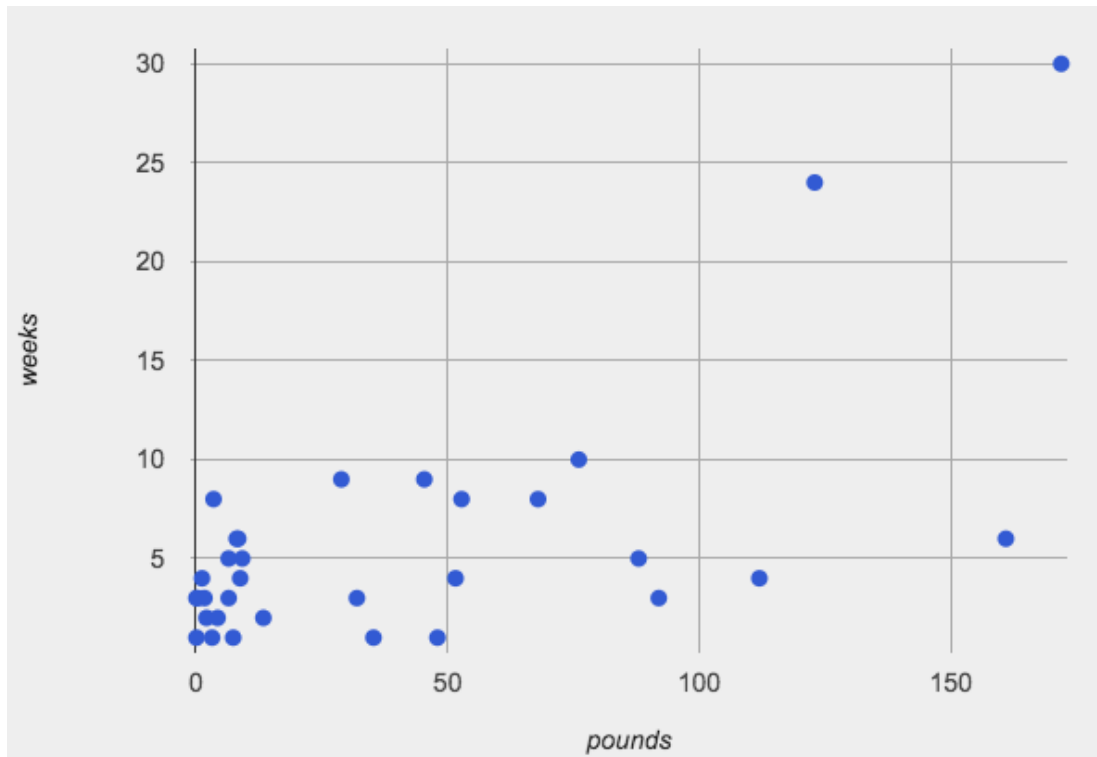
But if we label the dots by animal, we notice *every data point* after 25 pounds belongs to a dog from the shelter! The cats are all clumped together in the lower weight range, making it hard to see how weeks to adoption may relate to a cat's weight.
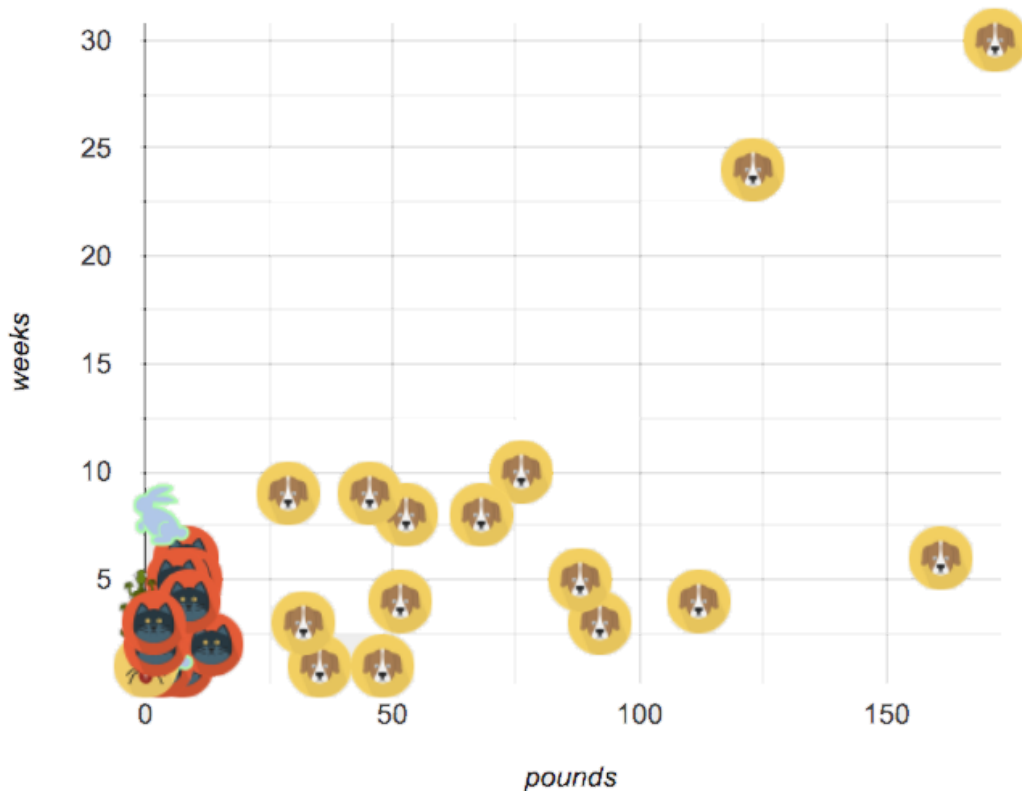
Investigate

Divide the class into groups of 3-4, with one student identified as the "reporter".

- Looking at this scatter plot (above), does it make sense to analyze all the animals together? Why or why not?
  - *No. Every data point after 25 pounds belongs to a dog from the shelter. The cats are clumped in the lower weight range.*

- Are there some questions where it would be important to break up the population into species-specific populations? What are they?
  - *Sample response: Yes. If we want to know whether dogs or cats are more likely to be fixed, we would need to look at each species separately.*

- Are there some questions where it would be important to keep the whole population together? What are they?
  - *Sample response: Yes. If we want to know if, in general, young animals are adopted more quickly, we would look at the entire population.*

- Be ready to share your answers with the class!

Have the reporters share their findings with the class.

Synthesize

You've been handed a dataset from a country where half the people have access to amazing medical care, and the other half have no healthcare.

- Why might it be important to look at a particular sample of a population?
  - *Sample response: Maybe we want to determine if emissions from a nearby factory impact the health of residents of one particular neighborhood.*
- Why is it sometimes bad to blindly take random samples?
  - *If we took a random sample of the population as a whole, we might think that they are generally middle-income and have average health. But if we ask the same question about the two groups _separately, we would discover inequality hiding in plain sight!_*

Grouped Samples   *20 minutes*

Launch

Depending on the question we're asking, sometimes it makes more sense to ask about "just the cats" or "just the dogs". Averaging every animal together will give us an answer, but it may not be a *useful* answer.

> Sometimes important facts about samples get *lost* if we mix them with the rest of the population!

Data Scientists define *grouped samples* of datasets, breaking them up into sub-groups that may be helpful in their analysis.

Earlier, you learned how to define values in Pyret. We can define Numbers, Strings, Images, and even rows:

```
name = "Flannery"
age  = 16
logo = star(50, "solid", "red")
sasha= row-n(animals-table, 0)
```

Let's use this skill to define *Tables…*

We already know how to define values, and how to filter a dataset. So let's put those skills together to define a grouped sample of the dogs in the shelter:

```
dogs = filter(animals-table, is-dog)
```

The `filter` function walks across each row in the table, and passes it to the `is-dog` function. If `is-dog` produces `true`, `filter` adds it to a new table. Otherwise, it just silently moves on to the next row. Finally, we define the name `dogs` to be the table produced by `filter`.

Investigate

A "kitten" is an animal who *is a cat* and who *is young*. How would you define a table of just kittens?

- Open the Grouped Samples Starter File, save a copy, and click "Run".
- What functions do you see defined here? What do they do?
- Turn to Grouped Samples from the Animals Dataset, and see what code define a table of only young cats.

- Can you fill in the code to define the other grouped samples?

- In the starter file, scroll to the bottom of the Definitions Area. Add your solutions from [Grouped Samples from the Animals Dataset](). Test them out!

Synthesize

- How could we make a bar chart showing the distribution of `sex` in the `kittens` sample?

- How could we filter *and* sort a table?

- How can we combine functions?

Displaying Samples

Overview

Students revisit the data display activity, now using the samples they created.

Launch

Making grouped and random samples is a powerful skill, which allows us to dig deeper than just making charts or asking questions about a whole dataset. Now that we know how to make grouped samples, we can make much more sophisticated displays!

Let's start with question: *what's the ratio of fixed to unfixed **cats** at the shelter?* Let's use the Data Cycle to get an answer, using our knowledge of grouped samples.

Ask Questions This is an Arithmetic Question. We know it's not a lookup question because there's no ratio written somewhere in the table for us to read. Instead, we'll have to count all the fixed cats and the unfixed cats, then compare the totals.

Consider Data We know that we'll need to count **only the cats!**, and can ignore everything else. And once we've picked the rows for cats, the only column we want is the `fixed` column. This is a huge hint that **we'll need to filter the dataset!**

Analyze Data We could use a bar-chart or a pie-chart to do this analysis, but since we care more about the *ratio* ("2x as many fixed as unfixed") than the count ("20 fixed vs. 10 fixed"), a pie chart is a better choice.We've decided what to make and we know which rows and columns we're plotting, so the next step is to *write the code!*

Interpret Data What did our displays tell us? In this case, we got a clear answer to our question. But perhaps that's not the end of the story! We might have *new* questions about whether a higher percentage of dogs are spayed and neutered than cats, or whether it's even possible to "fix" a tarantula. *All of this belongs in our data story!*

Investigate

- Complete Displaying Data, using what you've learned about samples to make more sophisticated data displays.

- Complete [Data Cycle: Analyzing Categorical Data](#).

Optional: : For students working on their own datasets…

If you are working on your own dataset, use [Samples from My Dataset](#) to define your own grouped samples!

Blank Design Recipes are also available [here](#).

Synthesize

- What connections do you see between the "Consider Data" and "Analyze Data" steps?
- How do we know when we need to filter? How do we know when we *don't?*