Histograms

Students explore new visualizations in Pyret, this time focusing on the distribution in a quantitative dataset. Students are introduced to Histograms by comparing them to bar charts, and learn to construct them by hand and in Pyret.

Prerequisites	Defining Table Functions
Relevant Standards K12CS CSTA NGSS CC-Math	*Select one or more standards from the menu on the left (⌘-click on Mac, Ctrl-click elsewhere).*
Lesson Goals	Students will be able to… • create histograms using the Animals Dataset • visualizations of frequency using their chosen dataset, and write up their findings
Student-facing Lesson Goals	• Let's use histograms to talk about data.
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`	

Glossary

bar chart :: a display of categorical data that uses bars positioned over category values; each bar's height reflects the count or percentage of data values in that category

frequency :: how often a particular value appears in a data set

histogram :: a display of quantitative data that uses vertical bars positioned over bins (sub-intervals); each bar's height reflects the count or percentage of data values in that bin.

shape :: The aspect of a dataset that tells which values are more or less common

Review

Have students open their Animals Starter File, and click "Run". (If they do not have this file, or if something has happened to it, they can always make a new copy.)

- Turn to The Design Recipe (Page 51), and write the functions you see there. When you're ready, type the contracts, purpose statements, examples and definitions into the Definitions Area.

- Use the `.build-column` method to add a new column to the animals table, showing the weight of every animal in kilograms.

- Use the `image-scatter-plot` function to plot all of the animals, putting `age` on the x-axis, number of `weeks` in the shelter on the y-axis, and `smart-dot` as our function.

Introducing Histograms

Overview

Students look at a bar chart and a histogram, compare/contrast them, and make observations about what they have in common and how they are different. Then they learn a more formal explanation of histograms.

Launch

Have students complete Summarizing Columns (Page 52).

The display on the left side of that page is a *Bar chart*.

- The x-axis lists the values of a categorical variable (`species`).

- The y-axis shows the *frequency* of categorical values in the dataset.

- This chart happens to show the categorical values in alphabetical order from left to right, but it would be fine to re-order them any way we wish. The bar for "dogs" could have been drawn before the one for "cats", without changing the meaning of the display. *It never makes sense to talk about the "shape" of a categorical data set* , since that shape holds no meaning.

The display on the right side is called a *histogram*.

- Histograms show the distribution of *quantitative* data.

- Since quantitative data must follow a natural order, these bars *cannot* be re-ordered.

- Histograms allow us to see the shape of a data set.

Investigate

To build a histogram, we start by sorting all of the numbers in our column from smallest to largest, marking our x-axis from the smallest value (or a bit below) to the largest value (or a bit above) and dividing into equally-sized intervals, or "bins". For example, if our values ranged from 3 to 53 we might mark our x-axis from 0 to 60 and divide it into bins of width 10. If they range from 22 to 41 we might mark our x-axis from 20 to 45 and divide it into bins of width 5. Once we have our bins, we put each value in our dataset into the bin where it belongs, and then count how many values fall in each bin. This count determines the height of the bars on our y-axis.

Turn to Making Histograms (Page 53), and try drawing a histogram from a dataset.

Possible Misconceptions

Note that intervals on this display include the left endpoint but not the right. If we included the right endpoint and someone had 0 teeth, we'd have to add on a bar from -5 to 0, which would be awfully strange!

Synthesize

Review: How are histograms and bar charts different?

Choosing the Right Bin Size

Overview

Students make histograms from the animals-dataset, and explore different bin sizes.

Launch

The size of the bins matters a lot! Bins that are too small will hide the shape of the data by breaking it into too many short bars. Bins that are too large will hide the shape by squeezing the data into just a few tall bars. In this workbook exercise, the bins were provided for you. But how do you choose a good bin-size?

Investigate

A display of how long it takes animals to get adopted can make it easier to get an idea of what adoption times were most common, and if there were any unusually long or short times that it took for an animal to be adopted.

Suppose we want to know how long it takes for animals from the shelter to be adopted.

- Find the contract for the `histogram` function.

- Make a histogram for the `"weeks"` column in the `animals-table`, using a bin size of 10.

- How many took between 0 and 10 weeks? Between 10 and 20?

- Try some other bin sizes (be sure to experiment with bigger and smaller bins!) - what shapes emerge? What bin size gives you the best picture of the distribution?

Look at the histogram and count how many animals took between 0 and 5 weeks to be adopted. How many took between 5 and 10 weeks? What else do you Notice? What do you Wonder?

Some observations you can share with the class, to get them started:

- We see most of the histogram's area under the two bars between 0 and 10 weeks, so we can say it was most common for an animal to be adopted in 10 weeks or less.

- We see a small amount of the histogram's area trailing out to unusually high values, so we can say that a couple of animals took an unusually long time to be adopted: one took even more than 30 weeks.

- More than half of the animals (17 out of 31) took just 5 weeks or less to be adopted. But those few unusually long adoption times pulled the average up to 5.8 weeks. We'll talk more about Shape of a histogram in the next lesson, and about its effect on average (the mean) in the lesson after that.

If someone asked what was a typical adoption time, we could say: "Almost all of the animals were adopted in 10 weeks or less, but a couple of animals took an unusually long time to be adopted—even more than 20 or 30 weeks!" Without looking at the histogram's shape, we could not have drawn this conclusion.

What would the histogram look like if most of the animals took more than 20 weeks to be adopted, but a couple of them were adopted in fewer than 5 weeks?

Synthesize

Have students talk about the bin sizes the tried. Encourage open discussion as much as possible here, so that students can make their own meaning about bin sizes before moving on to the next point.

Rule of thumb: a histogram should have between 5–10 bins.

Histograms are a powerful way to display a data set and assess its *shape*. Choosing the right bin size for a column has a lot to do with how data is distributed between the smallest and largest values in that column! With the right bin size, we can see the *shape* of a quantitative column. But how do we talk about or describe that shape, and what does the shape actually tell is? The next lesson addresses all of these.