










Visualizing the “Shape” of Data

Students explore the concept of "shape", using histograms to determine whether a dataset has skewness, and what the direction of the skewness means. They apply this knowledge to the Animals Dataset, and then to their own.

Prerequisites	Choosing Your Dataset																		
Relevant Standards <div><div>OK K12CS CSTA NGSS CC-Math</div></div>	Select one or more standards from the menu on the left (⌘-click on Mac, Ctrl-click elsewhere).																		
Lesson Goals	Students will be able to... <ul style="list-style-type: none">• Create histograms for variables in the Animals Dataset• Create visualizations of frequency using their chosen dataset, and write up their findings																		
Student-facing Lesson Goals	<ul style="list-style-type: none">• Let’s investigate what the shape of a histogram can tell us about the data.																		
Materials	<ul style="list-style-type: none">• Lesson Slides (Google Slides)• Computer for each student (or pair), with access to the internet• Student workbook, and something to write with																		
Preparation	<ul style="list-style-type: none">• 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	<table><tr><th>Types</th><th>Functions</th><th>Values</th></tr><tr><td>Number</td><td>num-sqrt, num-sqr</td><td>4, -1.2, 2/3</td></tr><tr><td>String</td><td>string-repeat, string-contains</td><td>"hello", "91"</td></tr><tr><td>Boolean</td><td>==, <, <=, >=, string-equal</td><td>true, false</td></tr><tr><td>Image</td><td>triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized, histogram</td><td></td></tr><tr><td>Table</td><td>count, .row-n, order-by, .filter, .build-column, random-rows</td><td></td></tr></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, histogram	  	Table	count, .row-n, order-by, .filter, .build-column, random-rows	
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, histogram	  																	
Table	count, .row-n, order-by, .filter, .build-column, random-rows																		

Glossary

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

skewed left :: A distribution is skewed left if there are a few values that are fairly low compared to the bulk of data values. A display of the data will show a longer tail to the left.

skewed right :: A distribution is skewed right if there are a few values that are fairly high compared to the bulk of data values. A display of the data will show a longer tail to the right.

symmetric :: A symmetric distribution has a balanced shape, showing that it's just as likely for the variable to take lower values as higher values.

Review

15 minutes

Have students turn to [Reading Histograms \(Page 54\)](#), and complete the matching activity there.

Describing Shape

20 minutes

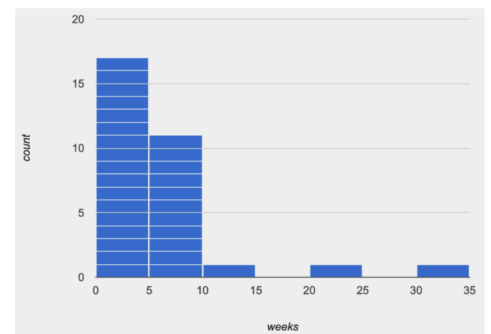
Overview

This activity focuses on *describing shape* based on a histogram. Students learn about "left skewed", "right skewed", and "symmetric" data, and what those descriptions tell us about a dataset.

Launch

Shape is one way to *summarize* information in a dataset, to quickly describe what values are more or less common.

Consider the image on the right: most of the data points are clustered on the left side, and it contains a few unusually high values way off to the right. We might describe this histogram by saying that it is "*skewed right, or has high outliers.*"

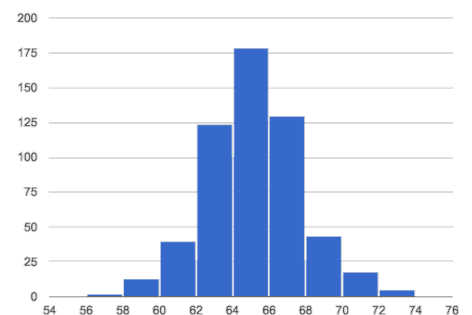


Here are the most common shapes that we see for real-world data sets:

Symmetric: values are balanced on either side of the middle.

In a *symmetric* distribution, it's just as likely for the variable to take a value a certain distance below the middle as it is to take a value that same distance above the middle. Examples:

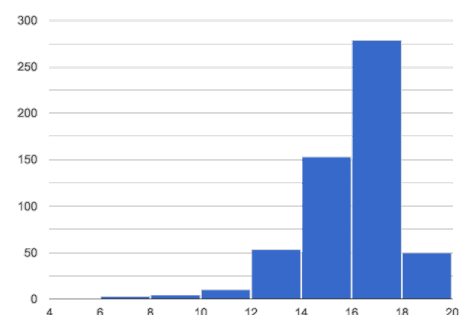
- Heights of 12-year-olds would have a symmetric shape. It's just as likely for a 12-year-old to be a certain number of inches below average height as it is to be that number of inches above average height.
- In a standardized test, most students score fairly close to what's average. Also, we see just as many students scoring a certain number of points above average as we see scoring that same number of points below average. The shape is symmetric (and bulges in the middle because most students score fairly close to what's average).



Skewed left, or low outliers.

In a distribution that is *skewed left*, values are clumped around what's typical, but they trail off to the left with a few unusually low values. Examples:

- Number of teeth that adults have in their mouths would be skewed left or have low outliers. Most adults will have close to a full set of 32 teeth, but a few of them with serious dental problems would have a very small number of teeth. We won't get anyone in our data set who has 10 or 20 *extra* teeth in their mouths!
- If most students did pretty well on an exam, but a few students performed very badly, then we'd see a shape that has left skewness and/or low outliers.

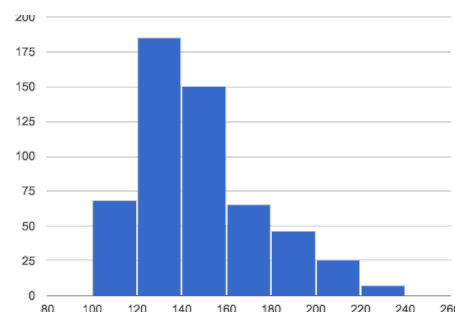


Skewed right, or high outliers.

In a distribution that is *skewed right* values are clumped around what's typical

in a distribution that is **skewed right**, values are clumped around what's typical, but they trail off to the right with a few unusually high values. We see this shape often in the real world, because there are many variables — like “income” or “time spent on the phone” — for which a few individuals have unusually high values, which aren’t balanced out by unusually low values (things like “income” and “phone time” can’t be less than zero). Examples:

- Age when a woman in the U.S. gives birth would be skewed right or have high outliers. A few women would be unusually old (40+ years), above the average age of 26 (check the tabloids!), but none of them could be even close to 40 years below average to balance things out!
- A data set of earnings almost always shows right skewness or high outliers, because there are usually a few values that are so far above average, they can’t be balanced out by any values that are so far below average. (Earnings can’t be negative.)



Investigate

- Make a histogram for the pounds column in the animals table, sorting the animals into 20-pound bins:
- Would you describe the shape of your histogram as being skewed left, skewed right, or symmetric?
- Which one of these statements is justified by the histogram’s shape?
 1. A few of the animals were unusually light.
 2. A few of the animals were unusually heavy.
 3. It was just as likely for an animal to be a certain amount below or above average weight.
- Try bins of 1-pound intervals, then 100-pound intervals. Which of these three histograms best satisfies our rule of thumb?
- On [Identifying Shape \(Page 55\)](#), describe the shape of the histograms you see there.
- On [The Shape of the Animals Dataset \(Page 56\)](#), describe the pounds histogram and another one you make yourself. When writing down what you notice, try to use the language Data Scientists use, discussing both skew and outliers.

Challenge Questions: - Compare histograms for the `pounds` column of both cats and dogs in the dataset. Are their shapes different? How much overlap is there? - Compare histograms for the `age` column of both cats and dogs in the dataset. Are their shapes different? How much overlap is there? - Can you explain why the amount of overlap between these two distributions different?

Synthesize

Discuss as a class, making sure students agree on the description of the shape.

Your Analysis

flexible

Overview

Students repeat the previous activity, this time applying it to their own dataset and interpreting their own results. **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

Now it’s time to try looking at the shape of your own dataset! Pick one quantitative column in your dataset, and hypothesize whether you think it will be skewed right, skewed left, or symmetric. What do you think?

Investigate

- How is your dataset distributed? Choose two quantitative variables and display them with histograms. Explain what you learn by looking at these displays. If you’re looking at a particular subset of the data, make sure you write that up in your findings on [The Shape of My Dataset \(Page 57\)](#).

- Students should fill in the **Quantitative Visualizations** portion of their Research Paper, using histograms they've constructed for their dataset and explaining what they show.

Synthesize

Have students share their findings.

Closing

Histograms are a powerful way to display a data set and see its *shape*. But shape is just one of three key aspects that tell us what's going on with a quantitative data set. In the next unit, we'll explore the other two: center and spread.