

UNDERSTANDING DISTRIBUTIONS OF DATA – PET FOOD MANUFACTURING

TEACHER VERSION

Subject Level:
Middle School Math

Grade Level:
6

Approx. Time Required:
45–60 minutes

Learning Objectives:

- Students will be able to describe a data distribution by its center, shape, spread, and outliers.
- Students will be able to compare different data formats (including graphs) of a data distribution.

Activity Description

Students will analyze raw data and graphical representations about businesses in each U.S. state that manufacture dog and cat foods. Students will use these various data formats to determine the center, shape, spread, and outliers of the data distribution.

Suggested Grade Level:
6

Approximate Time Required:
45–60 minutes

Learning Objectives:

- Students will be able to describe a data distribution by its center, shape, spread, and outliers.
- Students will be able to compare different data formats (including graphs) of a data distribution.

Topics:

- Box plots
- Center
- Dot plots
- Histograms
- Outliers
- Shape
- Spread

Skills Taught:

- Interpreting a data distribution using a graphical representation
- Comparing data distributions

Materials Required:

- The student version of this activity, 19 pages
- Calculators

Activity Items:

The following items are part of this activity. Items, their data sources, and any relevant instructions for viewing the source data online appear at the end of this teacher version.

- Item 1: Unsorted Raw Data, Dog and Cat Food Manufacturers by State, 2012
- Item 2: Sorted Raw Data, Dog and Cat Food Manufacturers by State, 2012
- Item 3: Dot Plot of Data
- Item 4: Histograms of Data With Varying Intervals
- Item 5: Box Plot of Data
- Item 6: Dog and Cat Food Manufacturers With 0 to 4 Employees, by State, 2012

For more information to help you introduce your students to the U.S. Census Bureau, read "*Census Bureau 101 for Students.*" This information sheet can be printed and passed out to your students as well.

Standards Addressed

See charts below. For more information, read "*Overview of Education Standards and Guidelines Addressed in Statistics in Schools Activities*."

Common Core State Standards for Mathematics

Standard	Domain	Cluster
CCSS.MATH.CONTENT.6.SP.A.2 Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and overall shape.	6 SP – Statistics & Probability	Develop understanding of statistical variability.

Common Core State Standards for Mathematical Practice

Standard
CCSS.MATH.PRACTICE.MP3. Construct viable arguments and critique the reasoning of others. Students will defend the advantages and disadvantages of various distribution graphs.
CCSS.MATH.PRACTICE.MP4. Model with mathematics. Students will interpret information about real-life data from graphical representations.
CCSS.MATH.PRACTICE.MP6. Attend to precision. Students will use the language of mathematics as it applies to center, shape, spread, and outliers.

National Council of Teachers of Mathematics' Principles and Standards for School Mathematics

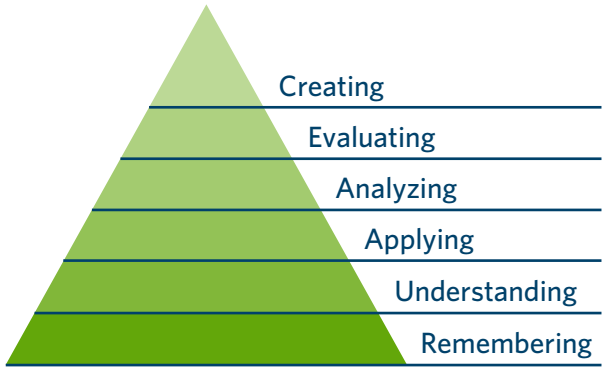
Content Standard	Students should be able to:	Expectation for Grade Band
Data Analysis and Probability	Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them.	Select, create, and use appropriate graphical representations of data, including histograms, box plots, and scatterplots.
Data Analysis and Probability	Select and use appropriate statistical methods to analyze data.	Discuss and understand the correspondence between data sets and their graphical representations, especially histograms, stem-and-leaf plots, box plots, and scatterplots.

Guidelines for Assessment and Instruction in Statistics Education

GAISE	Level A	Level B	Level C
Formulate Questions	X		
Collect Data			
Analyze Data		X	
Interpret Results	X		

Bloom's Taxonomy

Students will ***understand*** different graphical representations of data sets and will ***analyze*** a data set's center, shape, spread, and outliers.



Teacher Notes:

Before the Activity

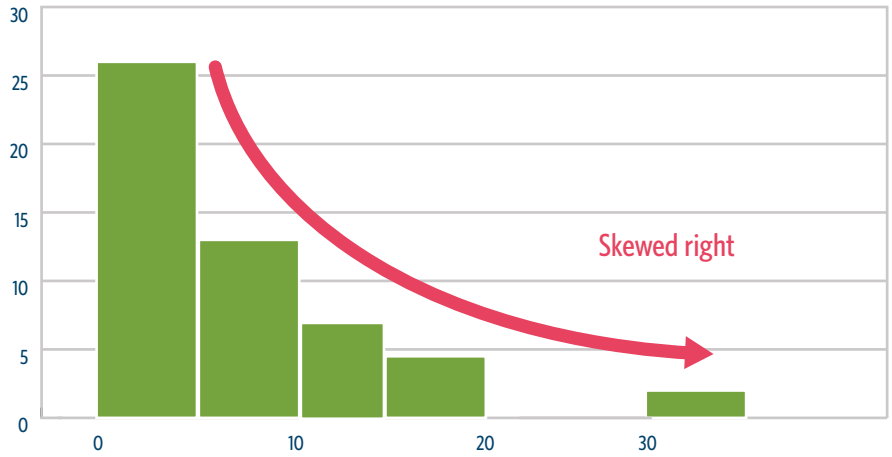
Students must understand the following key terms:

- **Center** – the middle of a data set, usually identified as the median or mean
- **Mean** – a measure of center in a set of numerical data, computed by adding the values in a list and then dividing by the number of values in the list
- **Median** – a measure of center in a set of numerical data, identified as the value appearing at the middle of a sorted version of the list (or the mean of the two middle values if the sorted list contains an even number of values)
- **Mode** – the value that appears most frequently in a data set
- **Range** – the numerical difference between a data set's maximum value and minimum value
- **Shape** – the general form of a data distribution (e.g., bell-shaped, bimodal, irregular, uniform)
- **Spread** – the amount of variability among the values in a data set
- **Distribution** – an arrangement of a particular variable's values that shows the frequency of each value's occurrence in a data set and that can be described by its center, shape, and spread
- **Histogram** – a way of displaying numeric data on a graph using horizontal or vertical bars so that the height or the length of the bars indicates frequency
- **5-number summary** – a summary that provides information about a data set and contains five values: the minimum (smallest value), first quartile, median (middle value), third quartile, and maximum (largest value)
- **First quartile (Q_1)** – also known as lower quartile, the value that divides a sorted data set into the smallest 25 percent of the data and the largest 75 percent
- **Third quartile (Q_3)** – also known as upper quartile, the value that divides a sorted data set into the largest 75 percent of the data and the smallest 25 percent
- **Interquartile range** – a measure of variability in a set of numerical data that is the distance between the first and third quartiles of the data set and that encompasses half of the data points (i.e., the difference between Q_3 and Q_1 in a 5-number summary)
- **Outlier** – an extremely high or low value that noticeably differs from the other data points in the set; it is any point that falls 1.5 times the IQR below the lower quartile or above the upper quartile

Students should have the following skills:

- Ability to read graphs and correctly identify the variables represented along the horizontal and vertical axes; for example in this activity, students should be able to see that the the horizontal axis of the histograms represents ranges of the number of dog and cat food manufacturers in a state and the vertical axis represents the frequency of states with dog and cat food manufacturers within each range.
- Ability to interpret a box plot
- A basic understanding of what it means for a graph to be skewed right or skewed left

Teachers should remind students of the difference between a histogram that is symmetrical and a histogram that is skewed right or skewed left. See an example below, where the longer "tail" goes to the right:



Teachers should note for students the general rule that histograms with between five and 12 bars provide sufficient information for an observer to draw conclusions about center, shape, spread, and outliers. They could also remind students how to count data for purposes of creating adjacent bars in a histogram (e.g., in the skewed right histogram above, a state with 10 manufacturers is counted in the bar beginning at 10, rather than the one to its left that goes from five to 10).

Teachers should also explain to students that this activity uses data from the Economic Census of the United States. For more information, go here (*www.census.gov/programs-surveys/economic-census.html*).

Teachers should ask the following questions:

- Have you ever had the responsibility of feeding a pet?
- Where does pet food come from?
- Do you know of any businesses in our state where dog or cat food is produced?
- Which states do you think produce the most dog and cat food? Why do you think so?

Teachers will divide the class into five groups, assigning each group an item from **Items 1–5** for completing the corresponding section in part 1.

During the Activity

As students work with their groups on their section of part 1, teachers should move around the room and encourage them to think about the number of data points in the data set when identifying its center, shape, and spread, rather than the various items' visual presentations of these measures, which may be misleading. For example, the center may be the "balancing point" of the data set rather than the direct middle point on a graph. Teachers should also tell students that they should be prepared to defend their answers.

After students complete their group's section, teachers will direct groups to present their answers to the class, so that all students can experience all data formats. Teachers will direct students to fill in the four remaining sections of part 1 as they listen to the other groups present. Teachers should ask students how easily, if at all, they were able to calculate the mean and median of the data set using the data format they were assigned.

For part 3, students will likely choose to put the data in numerical order if they decide to make a box plot or a histogram. Teachers could remind students to graph outliers as individual data points if they choose to create a box plot (more accurately called a "modified" box plot). Because dot plots are the easiest graphs to create, teachers should recommend this format for less advanced students. Teachers could also help students recognize that sorting the data would be helpful in locating the center and outliers for any data format.

After the Activity

Teachers should ask students to reflect on what they learned.

Extension Idea:

- Teachers could have students practice with distribution graphs using data from any subject area. For example, teachers could find data related to holidays and current events from Facts for Features here ***(www.census.gov/newsroom/facts-for-features.html)*** or related to special observations from Stats for Stories here ***(www.census.gov/newsroom/stories.html)***.

Student Activity

Click *here to download a printable version for students.*

Activity Items:

The following items are part of this activity and appear at the end of this student version.

- Item 1: Unsorted Raw Data, Dog and Cat Food Manufacturers by State, 2012
- Item 2: Sorted Raw Data, Dog and Cat Food Manufacturers by State, 2012
- Item 3: Dot Plot of Data
- Item 4: Histograms of Data With Varying Intervals
- Item 5: Box Plot of Data
- Item 6: Dog and Cat Food Manufacturers With 0 to 4 Employees, by State, 2012

Student Learning Objectives:

- I will be able to describe a data distribution by its center, shape, spread, and outliers.
- I will be able to compare different data formats (including graphs) of a data distribution.

Part 1 – Identify Variability Measures in Different Data Formats

With your group, complete the section below that matches your assigned item from the end of this activity. Your group will then share your answers with the class; as other groups present, fill in the answers to the other four sections of part 1.

Refer to **Item 1: Unsorted Raw Data, Dog and Cat Food Manufacturers by State, 2012** to answer these questions:

1. Based on what you see in this data format, can you easily identify the center of the data set? Why or why not?

   It is difficult to identify the center of the data set in this format because the data are not sorted.

   Teachers should note that students may incorrectly guess that Missouri is the center of the data because it is the middle state in the alphabetical list.

2. About how many manufacturers are in a "typical" state?

   Most states have fewer than 10 manufacturers. States with 5 manufacturers are the most common.

3. What appears to be the spread of the data in this format?

   The number of manufacturers in a state varies from 0 to 33.

4. How are the data distributed as displayed in this format?

   Student answers will vary but could include that there is not much shape in this format. States with more manufacturers are sprinkled throughout the alphabetical list of states.

5. Do there appear to be any outliers in this data set as it is presented in this format? What are they, and how do you know they are outliers?

   California is a noticeable outlier, with 33 manufacturers. There are six states with 0 manufacturers, but these data points are not smaller enough than those for the other states to be considered outliers.

Refer to Item 2: Sorted Raw Data, Dog and Cat Food Manufacturers by State, 2012 to answer these questions:

1. Based on what you see in this data format, can you easily identify the center of the data set? Why or why not?

   The number 4 is easily identified as the median because it is near the middle of this sorted list. (Some students may calculate the mean (5.98 ≈ 6), but that would not be required.)

2. About how many manufacturers are in a "typical" state?

   The mode is 5 manufacturers.

3. What appears to be the spread of the data in this format?

   The number of manufacturers in a state varies from 0 to 33.

4. How are the data distributed as displayed in this format?

   Student answers will vary but could include general observations similar to those for Item 1. Students might also notice that there are many states with 7 or fewer manufacturers and not very many with more than 15 or 20.

5. Do there appear to be any outliers in the data set as it is presented in this format? What are they, and how do you know they are outliers?

   The most extreme values are 0 and 33. California (with 33) is more of an outlier because it is much larger than the second-highest data point, for Texas (with 19), while those states with 0 manufacturers are not as remarkable because the closest data point to these is 1 manufacturer.

Refer to **Item 3: Dot Plot of Data** to answer these questions:

1. Based on what you see in this data format, what appears to be the center of the data set? Calculate the median to support your answer.

   The median is 4 manufacturers (in the middle if you count the dots).

   Teachers should note that students may incorrectly identify 16 as the median because it is in the middle of the graph's horizontal axis.

2. About how many manufacturers are in a "typical" state?

   The most common value (i.e., the mode) is 5 manufacturers, but most of the values appear to be clustered around 4.

3. What appears to be the spread of the data in this format?

   The number of manufacturers in a state varies from 0 to 33.

4. How are the data distributed as displayed in this format?

   Most states seem to have 7 or fewer manufacturers (the data are clustered between 0 and 7). There are comparatively few states with 10–16 manufacturers, and only two have more than 16.

5. Do there appear to be any outliers in this data set as it is presented in this format? What are they, and how can you tell they are outliers?

   The value of 33 (for California) is positioned much farther to the right than the other values, so it is likely an outlier.

Refer to **Item 4: Histograms of Data With Varying Intervals** to answer these questions:

1. Based on what you see in this data format, can you easily identify the center of the data set? Why or why not?

   Student answers will vary but could include: between 5 and 10 manufacturers, based on the bars in each of the three graphs. The varying widths of the bars make it difficult to estimate the median or mean from this data format.

2. Is it easy to identify how many manufacturers are in a "typical" state? Why or why not?

   No, the mode is not easy to identify because the bars do not show individual data points.

3. What appears to be the spread of the data in this format?

   The number of manufacturers in a state varies from 0 to 35 manufacturers in Graphs A and B, and from 0 to 40 in Graph C.

4. How are the data distributed as displayed in this format?

   Most of the data are clustered on the left, which means the data are skewed right.

5. Do there appear to be any outliers in this data set as it is presented in this format? What are they, and how can you tell they are outliers?

   The data between 30 and 35 manufacturers represent outliers, as they are positioned farther away from the rest of the data. It is difficult, however, to tell how many data points are in this outlier range based on these graphs.

6. Which histogram (Graph A, B, or C) represents the data best? Defend your answer.

   Student answers will vary but could include the following observations:

   - **Graph A, because it groups the data in small intervals (each bar representing a range of 5 manufacturers) that reveal the outliers and clusters well.**
   - **Graph B, because it has even smaller increments than the other two graphs. (Some students may argue, however, that this graph gives too much detail and masks the clustering.)**
   - **Graph C's grouping of all 50 states into only three bars distorts the true center, and the details of shape and clustering are lost, so students could conclude that this graph provides less information than do Graph A or B.**

Refer to **Item 5: Box Plot of Data** to answer these questions:

1. Based on what you see in this data format, can you easily identify the center of the data set? Why or why not?

   Because the vertical line of a box plot represents the median, the center is easily identifiable in this case as 4 manufacturers.

   Teachers should note that some students may say, incorrectly, that the value of 16 manufacturers is the center because it is in the middle of the graph's horizontal axis.

2. About how many manufacturers are in a "typical" state?

   Half of all data points appear to be between 3 and 7.

3. What appears to be the spread of the data in this format?

   The number of manufacturers in a state varies from 0 to 33.

4. How are the data distributed as displayed in this format?

   The data are clustered on the left, meaning they are skewed right.

5. This is considered a modified box plot, meaning that instead of including outliers in its whiskers, these values are graphed separately as individual data points. How many values are shown as outliers in this modified box plot, and are they much greater than or much less than most other values in the data set?

   There are four outlier values; all of them are much greater than the rest of the values.

6. The data inside the box (the interquartile range) represent one-half of all the values in the data set. Given this information, complete the sentence below.

   One-half of all states have between **2** and **7** dog and cat food manufacturers.

Part 2 – Compare the Utility of Data Formats for Assessing Variability

1. Looking at **Items 1** and **2**, which one helps you identify the following measures of variability best and why?
   - Center

     Item 2 is better, because it is easy to see that the 4s and 5s are more common and are in the middle of the data set.

   - Spread

     Student answers will vary. If students choose Item 1, they could say that finding the least and greatest values by inspection is easy because the data set is small. (Teachers could note that an unsorted list would not be practical for identifying the spread for larger data sets.) If students choose Item 2, they could say the least and greatest values are easy to find since they are at the beginning and end of the sorted list.

   - Shape

     Item 2 is better because similar data values are arranged together so it is easy to see where clusters are located.

   - Outliers

     Student answers will vary. If students choose Item 1, they could say the large numbers stand out because there are not very many of them. It is easy to see that 33 is the largest data value in the list. (Teachers could note again that this might not be the case for a larger data set.) If students choose Item 2, they could say the least and greatest data values are easy to find because they at the beginning and end of the sorted list.

2. What is a major advantage of the data format in **Item 1** over that in **Item 2** in terms of assessing variability?

   Student answers will vary but could include that it is easier to find the number of manufacturers in a specific state (such as a student's home state).

3. What is a major advantage of the data format in **Item 2** over that in **Item 1** in terms of assessing variability?

   Student answers will vary but could include that it is easier to see the smallest and largest data values, as well as the most popular ones (mode).

4. Now consider **Items 3–5**. Which of these graphical representations helps you identify the following measures of variability best and why?
   - Center

     Student answers will vary. If students choose Item 3, they could say that this dot plot lets them find the middle dot (i.e., the median) or balancing point (i.e., the mean) easily. If students choose Item 4, they could say the histograms let them imagine where the balancing point is easily. If students choose Item 5, they could say that this box plot lets them see the median easily because it is the middle line in the box.

- Spread

  Student answers will vary. If students choose Item 3, they could say this dot plot lets them see individual data values and that they are in order. If students choose Item 4, they could say the histograms let them estimate the spread from the intervals on the horizontal axis (though it is impossible to be exact). If students choose Item 5, they could say the box plot lets them see that the least and greatest data values are at the extremes of the graph.

- Shape

  Student answers will vary. If students choose Item 3, they could say this dot plot lets them see data clusters easily because the dots stack up. If students choose Item 4, they could say the heights of the bars in the histograms easily display the clusters of data. If students choose Item 5, they could say the box plot lets them easily see where the IQR is because of the box.

- Outliers

  Student answers will vary. If students choose Item 3, they could say this dot plot lets them see the first and last dots (the extremes) and the distances between individual dots. If students choose Item 4, they could say that a small bar as the first or last bar of a histogram, with a gap between it and the other bars, might represent an outlier. If students choose Item 5, they could say the box plot, because it is modified, easily shows outliers as dots instead of as part of whiskers.

5. List one major advantage to each type of graphical representation (Items 3–5) in terms of assessing variability:

   Student answers will vary but could include the following observations:

   - Dot plot:
     - **Displays individual data values**
     - **Shape, data clusters, and outliers are easily visible.**
   - Histogram:
     - **Very practical for large data sets**
     - **Data clusters and shape are easily visible.**
   - Box plot:
     - **Practical for large data sets**
     - **Outliers, IQR, and median (i.e., spread) are easily visible.**

Part 3 – Apply Your Knowledge: What Would You Do?

A new method of manufacturing is being developed that will speed up the production of dog and cat food. It has been recommended that this method be adopted by any manufacturer that employs fewer than five people. Fortunately, the Economic Census included businesses with 0–4 employees in its data collection on dog and cat food manufacturers (see **Item 6: Dog and Cat Food Manufacturers With 0 to 4 Employees, by State, 2012**). Examine **Item 6** to answer the following questions:

1. What graphical representation would you choose to display this distribution? Explain your reasoning, including why your choice is better than others for finding the center, shape, and spread.

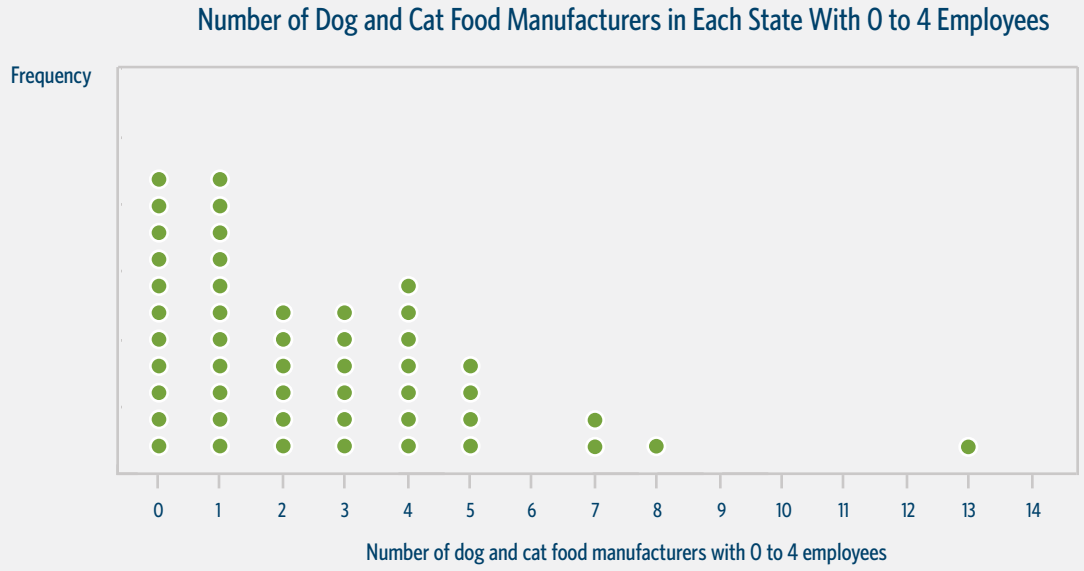
   Student answers will vary but could include either of these observations:

   - **Because the numerical range of data is small, a dot plot is reasonable and will accurately display the center, shape, and spread, since every point will be easily visible.**
   - **Because the number of data points is somewhat large, a box plot or histogram would be less tedious to make, and it would still accurately show the center, shape, and spread of the data.**

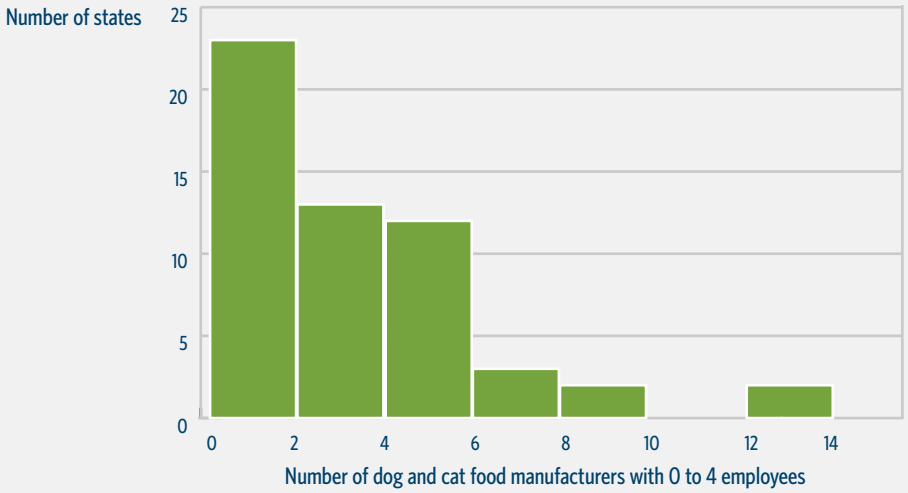
2. Create the graphical representation you chose in question 1 to display the distribution. Be sure to label your axes and give your graph a title. To determine if a value is an outlier, check to see if it is more than 1.5 times the IQR above the third quartile, or 1.5 times the IQR below the first quartile (show your work).

   Using the 1.5 IQR rule, 13 qualifies as an outlier in this data set. The IQR = $Q_3 - Q_1 = 4 - 1 = 3$, meaning that (1.5*IQR) = 4.5. Any value greater than $Q_3 + (1.5*IQR) = 4 + 4.5 = 9.5$ is therefore a high outlier. The mean is 2.53 business and the median is 2 businesses.

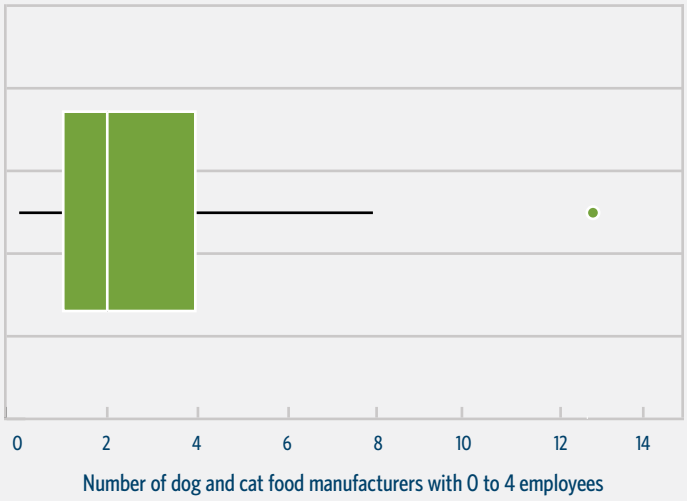
   Sample student graphs follow:



Number of Dog and Cat Food Manufacturers in Each State With 0 to 4 Employees



Number of Dog and Cat Food Manufacturers in Each State With 0 to 4 Employees



Item 1: Unsorted Raw Data, Dog and Cat Food Manufacturers by State: 2012

State	Number of Businesses Manufacturing Dog and Cat Food
Alabama	4
Alaska	2
Arizona	5
Arkansas	3
California	33
Colorado	7
Connecticut	1
Delaware	2
Florida	5
Georgia	5
Hawaii	0
Idaho	1
Illinois	16
Indiana	11
Iowa	10
Kansas	12
Kentucky	4
Louisiana	0
Maine	0
Maryland	3
Massachusetts	2
Michigan	6
Minnesota	5
Mississippi	2
Missouri	16
Montana	0

Item 1: Unsorted Raw Data, Dog and Cat Food Manufacturers by State: 2012 (Continued)

State	Number of Businesses Manufacturing Dog and Cat Food
Nebraska	10
Nevada	1
New Hampshire	3
New Jersey	5
New Mexico	1
New York	7
North Carolina	7
North Dakota	0
Ohio	12
Oklahoma	7
Oregon	4
Pennsylvania	14
Rhode Island	1
South Carolina	4
South Dakota	5
Tennessee	4
Texas	19
Utah	3
Vermont	5
Virginia	3
Washington	13
West Virginia	1
Wisconsin	15
Wyoming	0

Source: U.S. Census Bureau 2012 Economic Census

Item 2: Sorted Raw Data, Dog and Cat Food Manufacturers by State: 2012

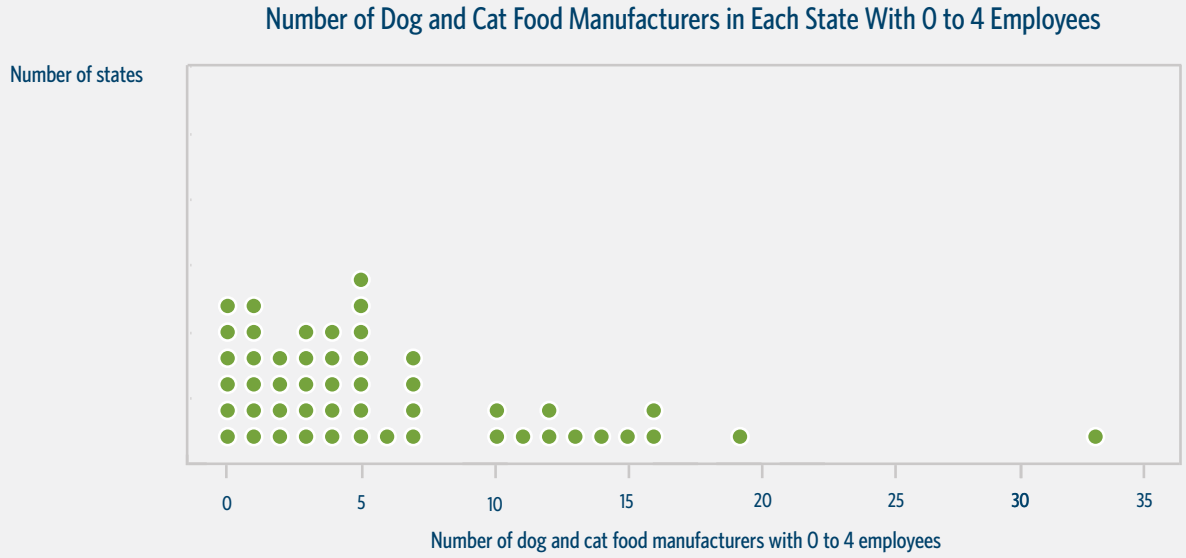
State	Number of Businesses Manufacturing Dog and Cat Food
Hawaii	0
Louisiana	0
Maine	0
Montana	0
North Dakota	0
Wyoming	0
Connecticut	1
Idaho	1
Nevada	1
New Mexico	1
Rhode Island	1
West Virginia	1
Alaska	2
Delaware	2
Massachusetts	2
Mississippi	2
Arkansas	3
Maryland	3
New Hampshire	3
Utah	3
Virginia	3
Alabama	4
Kentucky	4
Oregon	4
South Carolina	4
Tennessee	4

Item 2: Sorted Raw Data, Dog and Cat Food Manufacturers by State: 2012 (Continued)

State	Number of Businesses Manufacturing Dog and Cat Food
Arizona	5
Florida	5
Georgia	5
Minnesota	5
New Jersey	5
South Dakota	5
Vermont	5
Michigan	6
Colorado	7
New York	7
North Carolina	7
Oklahoma	7
Iowa	10
Nebraska	10
Indiana	11
Kansas	12
Ohio	12
Washington	13
Pennsylvania	14
Wisconsin	15
Illinois	16
Missouri	16
Texas	19
California	33

Source: U.S. Census Bureau 2012 Economic Census

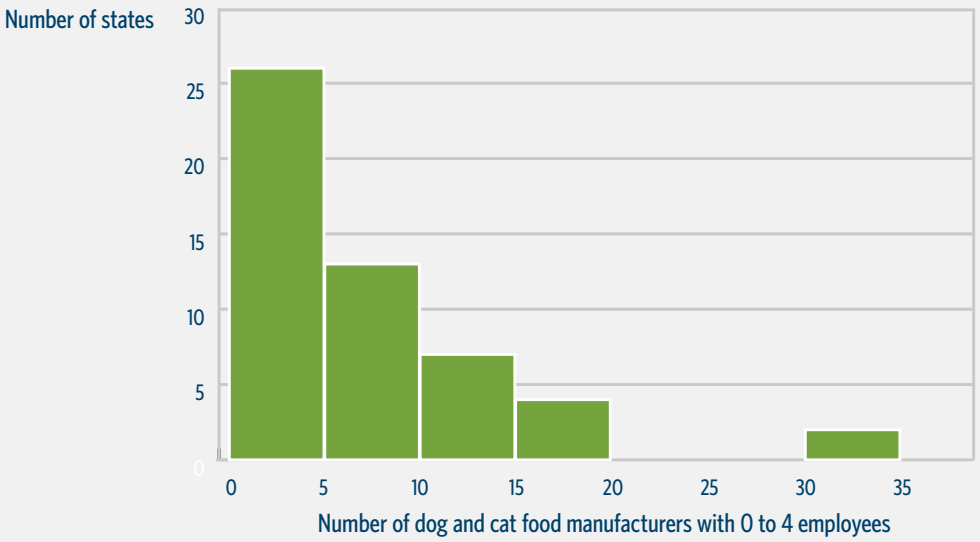
Item 3: Dot Plot of Data



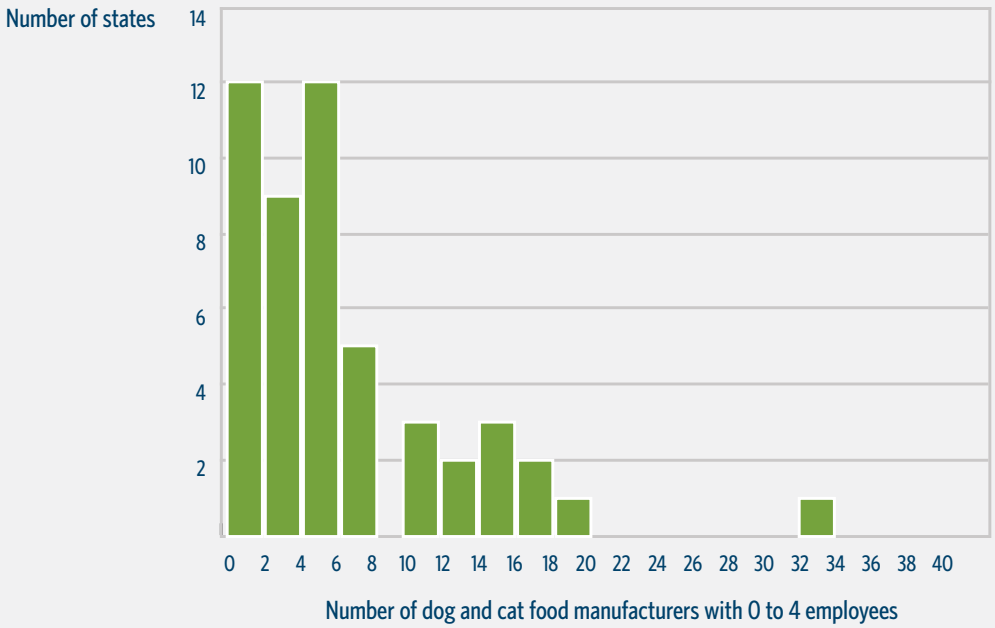
Source: U.S. Census Bureau 2012 Economic Census

Item 4: Histograms of Data With Varying Intervals

Graph A: Number of Dog and Cat Food Manufacturers With 0 to 4 Employees in Each State (Intervals of 5)

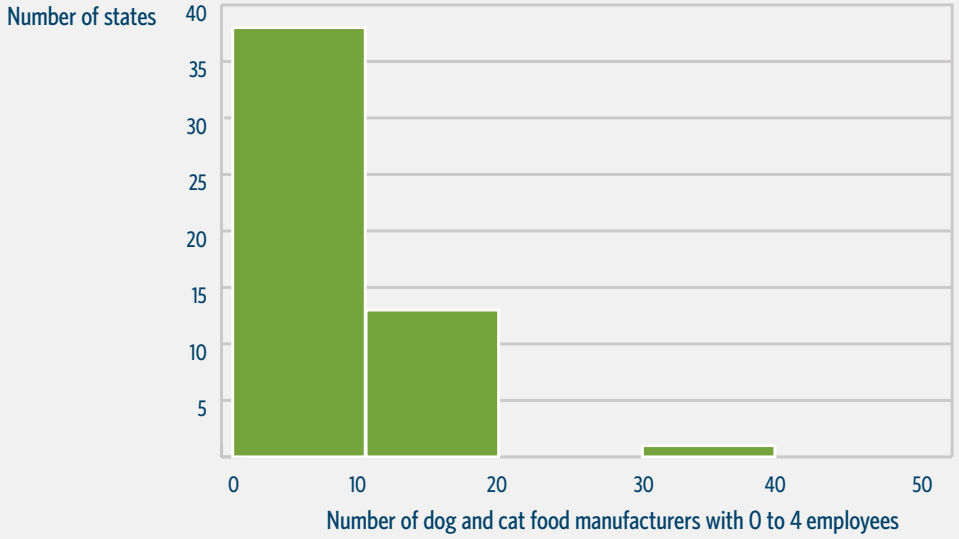


Graph B: Number of Dog and Cat Food Manufacturers With 0 to 4 Employees in Each State (Intervals of 2)



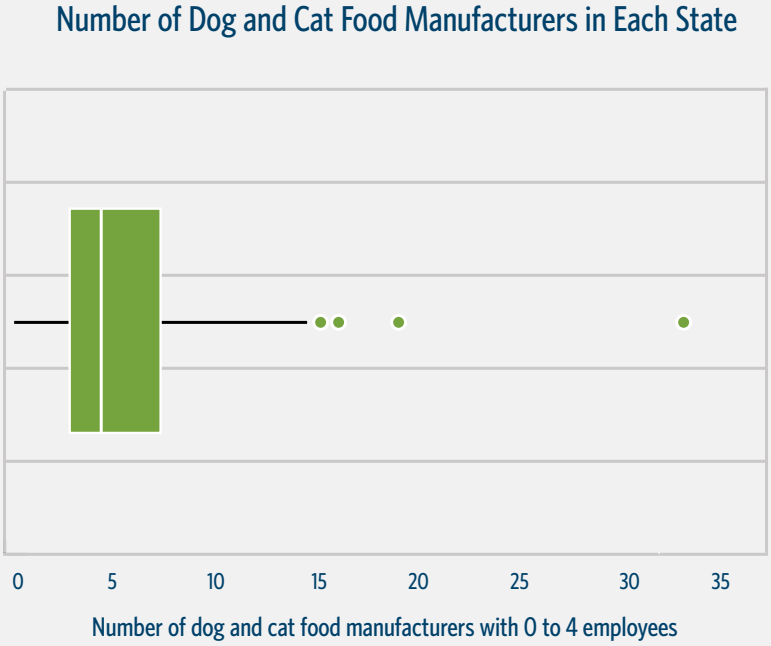
Item 4: Histograms of Data With Varying Intervals (Continued)

Graph C: Number of Dog and Cat Food Manufacturers With 0 to 4 Employees in Each State (Intervals of 10)



Source: U.S. Census Bureau 2012 Economic Census

Item 5: Box Plot of Data



Source: U.S. Census Bureau 2012 Economic Census

Item 6: Dog and Cat Food Manufacturers With 0 to 4 Employees, by State, 2012

State	Number of Dog and Cat Food Manufacturers With 0 to 4 Employees
Alabama	1
Alaska	1
Arizona	2
Arkansas	0
California	13
Colorado	4
Connecticut	0
Delaware	1
Florida	4
Georgia	2
Hawaii	0
Idaho	1
Illinois	7
Indiana	4
Iowa	4
Kansas	0
Kentucky	1
Louisiana	0
Maine	0
Maryland	2
Massachusetts	1
Michigan	4
Minnesota	2
Mississippi	1
Missouri	5

State	Number of Dog and Cat Food Manufacturers With 0 to 4 Employees
Montana	0
Nebraska	3
Nevada	0
New Hampshire	3
New Jersey	3
New Mexico	1
New York	3
North Carolina	4
North Dakota	0
Ohio	3
Oklahoma	2
Oregon	4
Pennsylvania	5
Rhode Island	1
South Carolina	1
South Dakota	3
Tennessee	2
Texas	7
Utah	0
Vermont	5
Virginia	1
Washington	8
West Virginia	1
Wisconsin	5
Wyoming	0

Source: U.S. Census Bureau 2012 Economic Census