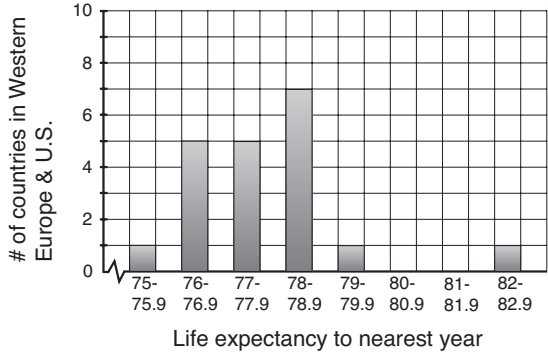


Vocabulary: Data About Us

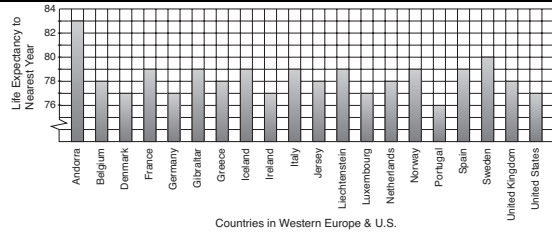
Concept	Example																		
<p>Two Types of Data</p> <p>Numerical data: is data about some attribute that <i>must</i> be organized by numerical order to show how the data varies. For example:</p> <ul style="list-style-type: none"> • Number of pets • Measure of a piece of lumber <p>A bar graph that shows numerical data has the values of the data in a fixed order on the horizontal axis. These may be individual numbers or intervals.</p> <p>Summarizing Numerical Data: The maximum, minimum and center values of the attribute make sense. For numerical data, the dot plot or bar graph always shows how frequently particular numerical values of the variable of interest occur, for example, how many times a value of “3 pets” was recorded, or how many times a value of 8.1 feet was recorded. We can compute with the values of the variable, say “number of pets,” and say what the mean or median number of pets might be, as well as the mode.</p> <p>Categorical data: is data that has been collected and recorded about some non-numerical attribute. For example:</p> <ul style="list-style-type: none"> • <i>color</i> is an attribute or variable (for which the categories or values could be red, green, blue etc.); • <i>gender</i> is an attribute (for which the categories are male and female); • <i>country</i> is an attribute (for which the categories might be U. S., Germany, etc.); • <i>yes/no</i> are two categories of response • <i>birth month</i> is an attribute (for which the categories are January, February etc.) 	<p><i>Example 1.</i> Below is the same data as in example 2, organized as a frequency bar graph of numerical data. The variable or attribute is life expectancy to nearest year, and the vertical axis indicates how frequently this life expectancy was observed.</p>  <table border="1" style="margin-left: auto; margin-right: auto;"> <caption>Data for Example 1: Frequency Bar Graph</caption> <thead> <tr> <th>Life expectancy interval (years)</th> <th>Number of countries</th> </tr> </thead> <tbody> <tr><td>75-75.9</td><td>1</td></tr> <tr><td>76-76.9</td><td>5</td></tr> <tr><td>77-77.9</td><td>5</td></tr> <tr><td>78-78.9</td><td>7</td></tr> <tr><td>79-79.9</td><td>1</td></tr> <tr><td>80-80.9</td><td>0</td></tr> <tr><td>81-81.9</td><td>0</td></tr> <tr><td>82-82.9</td><td>1</td></tr> </tbody> </table> <p>It makes sense to say that the range for this variable, life expectancy, is quite small, only 8 years or 83 – 75 years, and even smaller if the outlier, 83 years (Andorra) is not included in the calculation. There is a gap in the graph from 80 – 82 years, making Andorra’s 83 years look even more unusual. Most of the data clusters between 76 and 79 years.</p> <p><i>Example 2.</i> Below is a value bar graph of categorical data. The categories are countries. The values are life expectancies. Measures of variability of the variable of interest (country, in this case), like range or maximum or minimum, make no sense for this kind of graph. There is no maximum country, for example, though there is a maximum life expectancy. Notice that the scale on the y-axis MIGHT be misleading. It makes it look like Andorrans live twice as long as Belgians, when in fact the difference is 5 years. Life expectancy in the United States is represented by the bar furthest to the right.</p>	Life expectancy interval (years)	Number of countries	75-75.9	1	76-76.9	5	77-77.9	5	78-78.9	7	79-79.9	1	80-80.9	0	81-81.9	0	82-82.9	1
Life expectancy interval (years)	Number of countries																		
75-75.9	1																		
76-76.9	5																		
77-77.9	5																		
78-78.9	7																		
79-79.9	1																		
80-80.9	0																		
81-81.9	0																		
82-82.9	1																		

A **bar graph** that shows categorical data will have category names (red, green, blue etc) along the horizontal or independent axis. These names can be rearranged and sense can still be made of the graph.

Summarizing categorical data: Because the names can be arranged in a different order we do not talk about the “maximum color,” for example, or the “mean color.” A bar graph may show **how frequently** a particular category occurs. For example, suppose that “red” was counted 22 times, and “green” was counted 100 times and that “green” was counted more often than any other category. Therefore, we can summarize categorical data by using the **mode**, but not by using the median or mean.

The bar graph may show information about **some value of the variable**, such as population in U.S., population in Germany etc., in which case we can say which country has the largest population, or note the gap between largest and smallest countries.

Note: *Birth month* could be displayed by numbering the months, but “1” is not a numerical value for January etc.



Using standard graphical representations

- Line plot:** Each case is represented as an "X" positioned over a labeled number line. The number line represents various categories or values of the attribute which is being studied, and we can see at a glance how the data is distributed.
- Frequency bar graph:** - A bar's height is not the value of an individual case but rather the number (frequency) of cases that all have that value.
- Stem-and-leaf plot:** A plot that permits students to group data in intervals. Stem plots are often introduced as a way to group data that has few repeated values and is spread out. In such situations the use of line plots provides little information.
- Scatterplot:** The relationship between two different variables is explored by plotting data values on a Cartesian coordinate system.

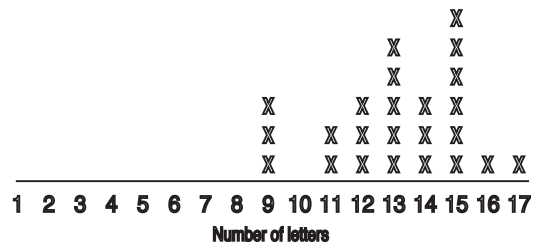
Example 3:

Line plots are visual representations of tally marks in a table.

The graph below is a **line plot made with numerical data**. The tally of the data was:

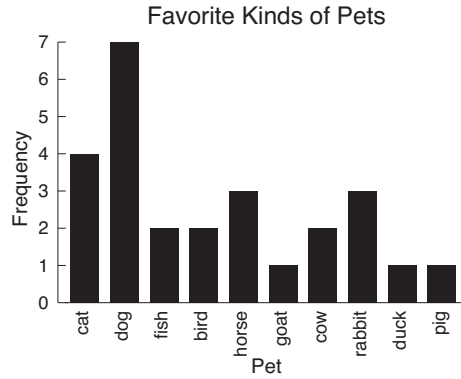
Number of Letters in name	Frequency
9	
10	
11	
12	
13	
14	
15	
16	
17	

Name Lengths of Ms. Jeckle's Students



Example 4:

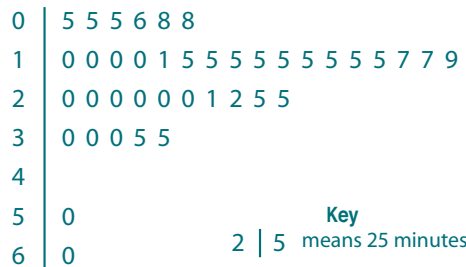
Frequency bar graphs are a variation on line plots. They can be made with **categorical data or with numerical data**. If they are made with numerical data the shape of the graph is significant. If they are made with categorical data the order is generally not important, so the shape is not important. The graph below is a frequency bar graph, where the categories on the horizontal axis are types of pets. The height of the bar indicates the frequency with which each pet occurs.



Example 5:

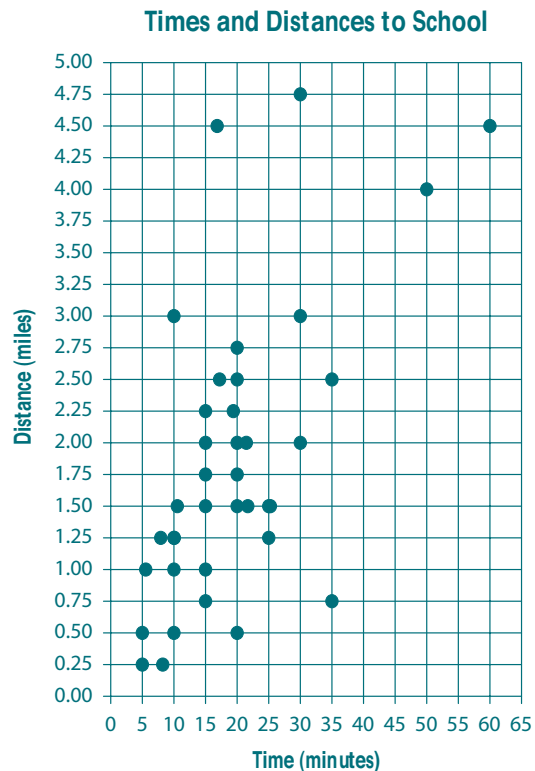
The graph below is a **stem and leaf plot**. The original list of data was 5, 5, 5, 6, 8, 8, 10, 10, 10, 10, 11, 15, 15, 15, 15, 15, 15, 15, 15, 17, 17, 19, 20, 20, 20, 20, 20, 20, 21, 22, 25, 25, 30, 30, 30, 35, 35, 50, 60. We could use every single value on a number line and made a line plot or bar graph. If we did this we would notice that 15 is the mode, and 55 (Max 60 – Min 5) is the range. But the distribution would have so many values along the horizontal axis that the shape and clusters of data would go unnoticed. By creating intervals to group the data the shape, with clusters and gaps and outliers, becomes apparent.

Travel Times to School (minutes)



Example 6:

The graph below is a **scatterplot**. The two variables are Time and Distance, and because there is an overall pattern we deduce there is a relationship. Generally we look for a pattern, either a linear pattern or a curve, or perhaps a cyclical pattern. The point of identifying a pattern is to permit predictions.



Using measures of center

Every measure of center is in some way an attempt to say what is “typical.” For numerical data there are three measures of “typical.”

Three measures of central tendency (center):

- **Mode** is the data value or category occurring with the highest frequency. It sometimes has more than one value and is unstable because a change in one or a few data values can lead to a very large change in the mode value. A distribution may be unimodal, bimodal,

Example 7:

Suppose the original list of data was 5, 5, 5, 6, 8, 8, 10, 10, 10, 10, 11, 15, 15, 15, 15, 15, 15, 15, 15, 17, 17, 19, 20, 20, 20, 20, 20, 20, 21, 22, 25, 25, 30, 30, 30, 35, 35, 50, 60 (See the example about travel times to school above.) Because this is numerical data we can find the

- **Mode** = 15, because 15 minutes is the most frequently occurring time. On a line plot or bar graph this would be the highest column.

<p>or multimodal.</p> <ul style="list-style-type: none"> ○ Median is the numerical value that marks the middle of a distribution. It is not influenced by extreme data values so is a good measure to use when working with distributions that are skewed. Graphically, the median marks the location that divides a distribution into two equal parts. ○ Mean is the numerical value that marks the balance point of a distribution; it is influenced by all values of the distribution, including extremes and outliers. It is a good measure to use when working with distributions that are roughly symmetric. <p>Note: Students must learn to choose the measure of center that is appropriate to the situation and distribution. The goal is to choose the measure that gives the most useful information about what is “typical.”.</p>	<ul style="list-style-type: none"> ● Median = 16, because in this list of 40 observations, the 20th is “15,” and the 21st is “17.” We want the exact middle of the ordered list, so this is “between” 15 and 17, even though no such data value exists in the list. On a graph this would be in the middle. ● Mean = ? There are two strategies for finding the mean. One is to “balance” out the list or graphical distribution. Thus, in the above list 5, 5, 5, 6, 8, 8, 10, 10, 10, 10, 11, 15, 15, 15, 15, 15, 15, 15, 15, 17, 17, 19, 20, 20, 20, 20, 20, 20, 21, 22, 25, 25, 30, 30, 30, 35, 35, 50, 60 we might balance a “5” and a “60” by replacing them with a “30” and a “35.” The total time would be the same. Continuing to balance we might balance another “5” and a “50” with a “25” and a “30.” At this point the list has become: 30, 25, 5, 6, 8, 8, 10, 10, 10, 11, 15, 15, 15, 15, 15, 15, 15, 15, 17, 17, 19, 20, 30, 30, 30, 35, 35, 30, 35. If we keep balancing this list we will eventually arrive at a point where we have a list of identical or nearly identical pieces of data. Graphically the mean is the balance point of the distribution. OR, we can use the usual strategy of summing the data and dividing the total by the number of pieces of data. Either way the mean = 18.975.
<p>Measures or descriptors of variability: In this unit students look at</p> <ul style="list-style-type: none"> ● the range to determine if the gap from minimum to maximum is wide, ● where data cluster to determine if most data lie within a narrow range of values ● holes in a distribution to determine if some values occur much less frequently than others ● outliers to determine if these are true 	<p>Example 8: Specifying the smallest and the largest values gives the range. In Ms. Jeckle’s class names lengths range from 9 letters to 17 letters (see above). The range is also commonly given by subtracting the smallest value from the largest value. For Ms. Jeckle’s data, the range interval is 8 letters.</p>

<p>representatives of wide variation in the data or perhaps errors</p> <p>Note: the only <i>measure</i> of variability in the above list is <i>range</i>. In future units students learn to measure variability by using the <i>interquartile range</i>.</p>	
<p>Covariation: is a way of characterizing a kind of relationship between two variables. It means that information about values from one variable helps us understand and/or explain or predict values of the other variable. If the two variables in question are plotted on a scatterplot and a visual pattern is observed, then this pattern indicates a relationship. The visual pattern can be extended, cautiously, to predict data not directly available.</p>	<p>Example 9: In the scatterplot on distances and times to school we can see a pattern. In general, as time increases distance increases. In fact it looks like a line with a slope of 5, or a rate of 5 miles per 1 hour, would roughly summarize this pattern. While the line $D = 5t$ (where distance is measured in miles and time in hours) would not fit every piece of observed data it would pass close to many pieces of data. Drawing this line would permit prediction of other pairs of data that would lie on the line, or fit the same overall pattern. We have to be cautious about predicting far from the range of the data; it may be that this relationship only works for small distances.</p>