

Looking at data: scatter plots and relationships

A **scatter plot** is an informative way to display numerical data with two variables. In your previous work in Grade 8, you saw how to construct and interpret scatter plots. Recall that if the two numerical variables are denoted by x and y , the scatter plot of the data is a plot of the (x, y) data pairs.

Example 1: Elevation vs. Amount of Sunshine

The National Climate Data Center collects data on weather conditions at various locations. They classify each day as clear, partly cloudy, or cloudy. Using data taken over a number of years, they provide data on the following variables.

x represents elevation above sea level (in feet).

y represents mean number of clear days per year.

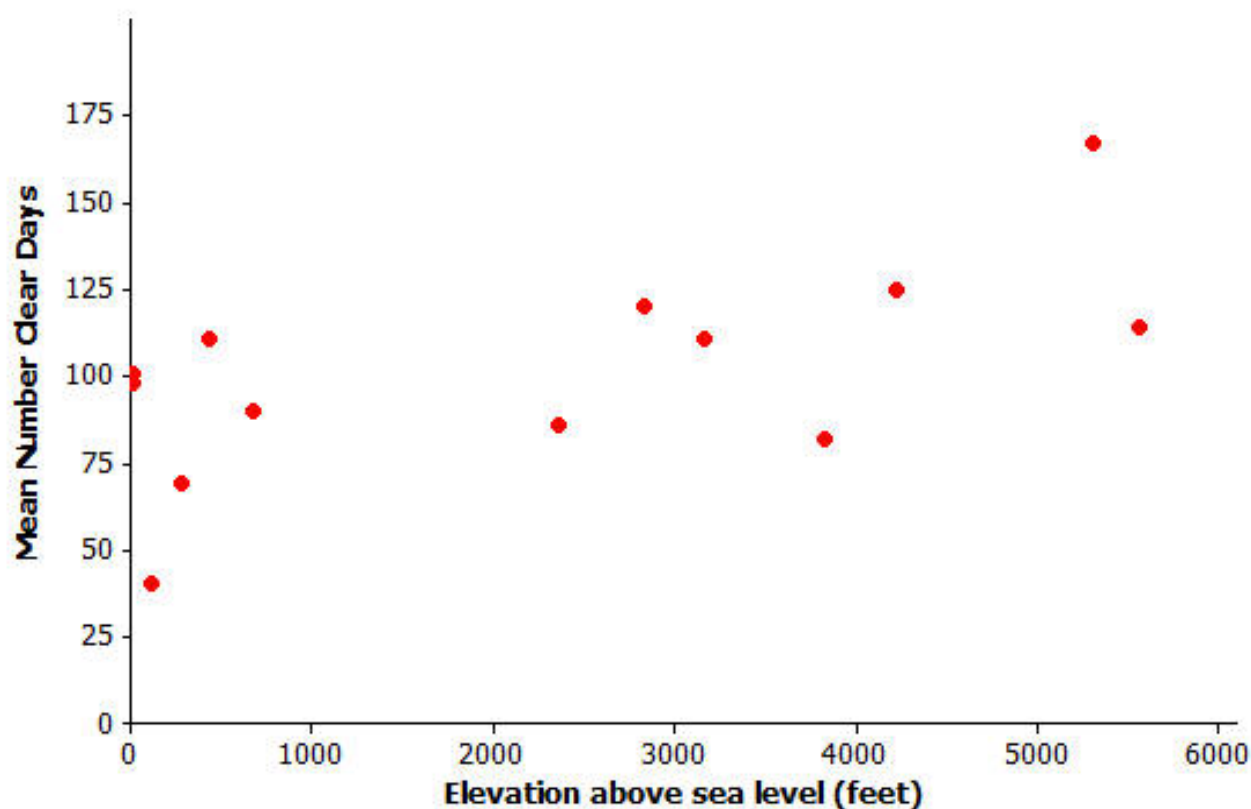
w represents mean number of partly cloudy days per year.

z represents mean number of cloudy days per year.

The table below shows data for 14 U.S. cities.

City	x (Elevation Above Sea Level in Feet)	y (Mean Number of Clear Days per Year)	w (Mean Number of Partly Cloudy Days per Year)	z (Mean Number of Cloudy Days per Year)
Albany, NY	275	69	111	185
Albuquerque, NM	5,311	167	111	87
Anchorage, AK	114	40	60	265
Boise, ID	2,838	120	90	155
Boston, MA	15	98	103	164
Helena, MT	3,828	82	104	179
Lander, WY	5,557	114	122	129
Milwaukee, WI	672	90	100	175
New Orleans, LA	4	101	118	146
Raleigh, NC	434	111	106	149
Rapid City, SD	3,162	111	115	139
Salt Lake City, UT	4,221	125	101	139
Spokane, WA	2,356	86	88	191
Tampa, FL	19	101	143	121

Here is a scatter plot of the data on elevation and mean number of clear days.



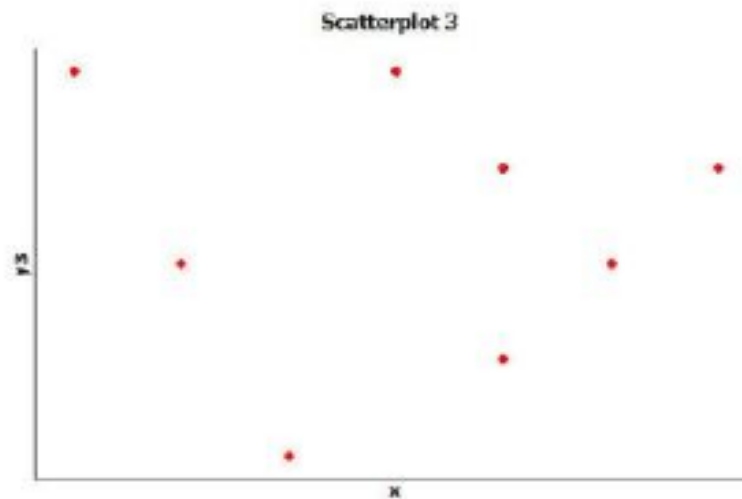
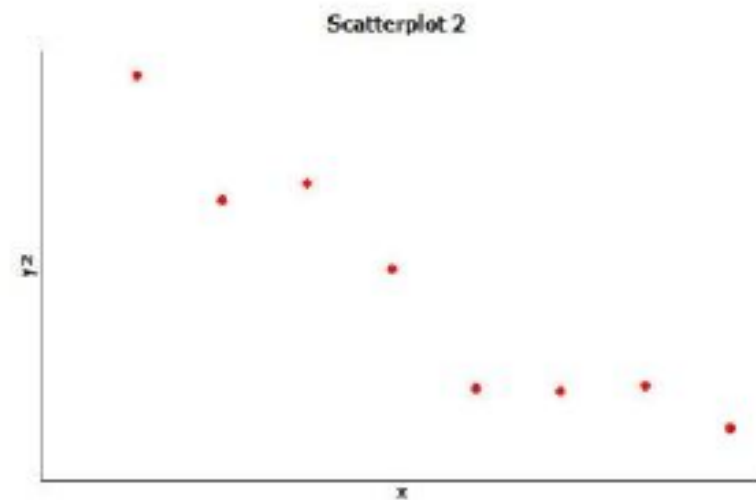
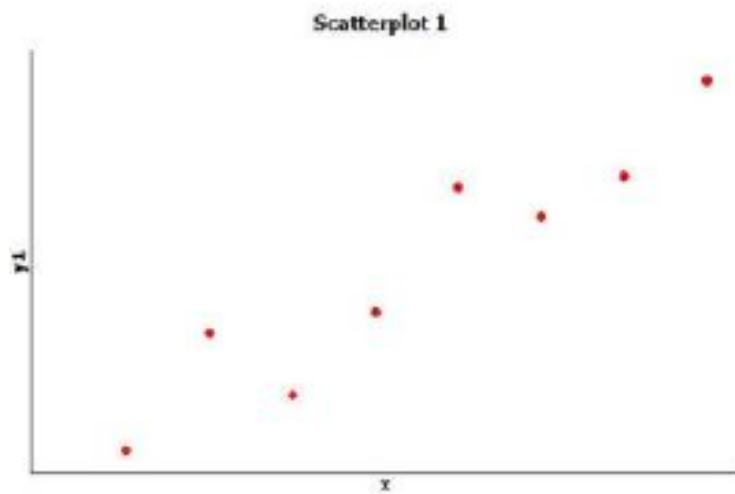
Data Source: www.ncdc.noaa.gov

1. Do you see a pattern in the scatter plot, or does it look like the data points are scattered?
2. How would you describe the relationship between elevation and mean number of clear days for these 14 cities? That is, does the mean number of clear days tend to increase as elevation increases, or does the mean number of clear days tend to decrease as elevation increases, or are they unrelated?
3. Do you think that a straight line would be a good way to describe the relationship between the mean number of clear days and elevation? Why do you think this?

Example 2: What relationship might this be?

Below are three scatter plots. Each one represents a data set with eight observations.

The scales on the x - and y -axes have been left off these plots on purpose, so you have to think carefully about the relationships.



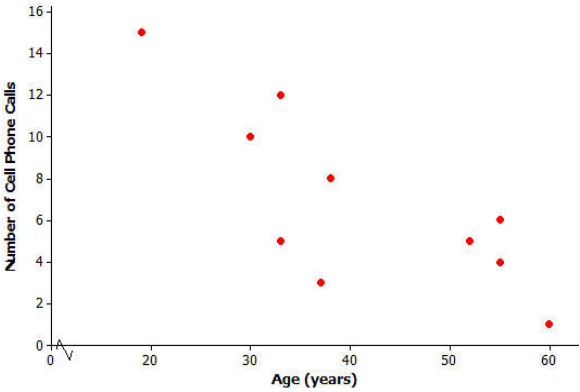
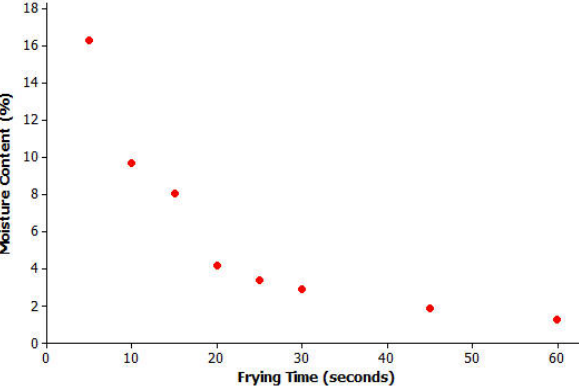
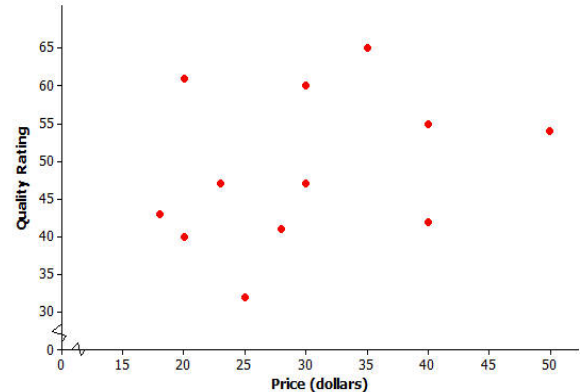
1. What are some possible data relationships that these scatter plots might illustrate? Come up with two-three such examples for each graph.

2. If one of these scatter plots represents the relationship between height and weight for eight adults, which scatter plot do you think it is and why?
3. If one of these scatter plots represents the relationship between height and SAT math score for eight high school seniors, which scatter plot do you think it is and why?
4. If one of these scatter plots represents the relationship between the weight of a car and fuel efficiency for eight cars, which scatter plot do you think it is and why?
5. Which of these three scatter plots does *not* appear to represent a linear relationship? Explain the reasoning behind your choice.

Example 3: Not Every Relationship Is Linear

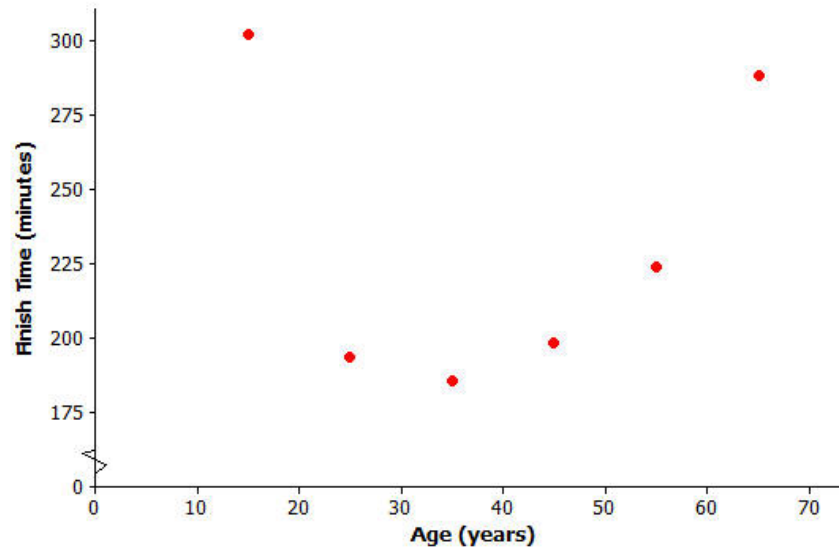
When a straight line provides a reasonable summary of the relationship between two numerical variables, we say that the two variables are *linearly related* or that there is a *linear relationship* between the two variables.

Take a look at the scatter plots below, and answer the questions that follow.

<p style="text-align: center;">Scatter Plot 1</p> 	<p style="text-align: center;">Scatter Plot 2</p>  <p>Data Source: R.G. Moreira, J. Palau, V.E. Sweat, and X. Sun, "Thermal and Physical Properties of Tortilla Chips as a Function of Frying Time," <i>Journal of Food Processing and Preservation</i>, 19 (1995): 175.</p>	<p style="text-align: center;">Scatter Plot 3</p>  <p>Data Source: www.consumerreports.org/health</p>
<ol style="list-style-type: none"> 1. Is there a relationship between the number of cell phone calls and age, or does it look like the data points are scattered? 2. If there is a relationship between the number of cell phone calls and age, does the relationship appear to be linear? 	<ol style="list-style-type: none"> 1. Is there a relationship between moisture content and frying time, or do the data points look scattered? 2. If there is a relationship between moisture content and frying time, does the relationship look linear? 	<ol style="list-style-type: none"> 1. Scatter Plot 3 shows data for the prices of bike helmets and the quality ratings of the helmets (based on a scale that estimates helmet quality). Is there a relationship between quality rating and price, or are the data points scattered? 2. If there is a relationship between quality rating and price for bike helmets, does the relationship appear to be linear?

Try It Yourself:

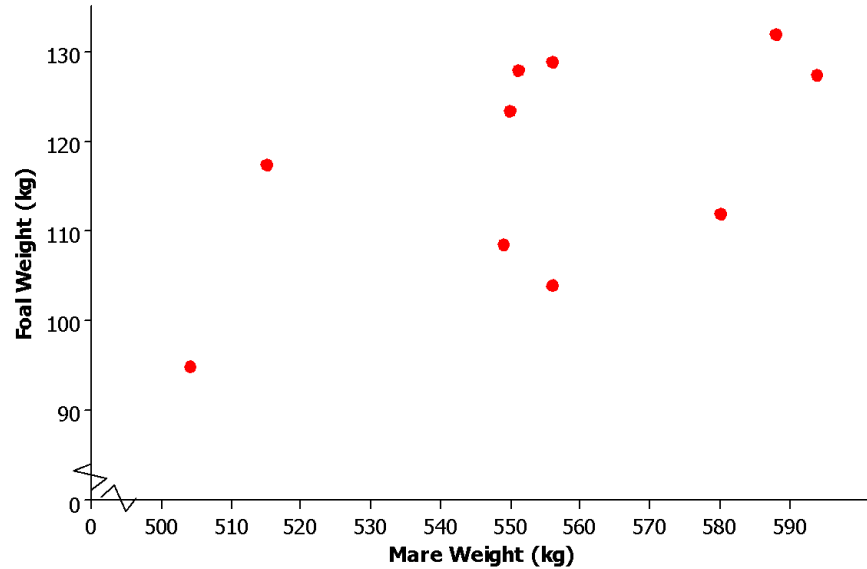
1. Use the table from the first page to construct a scatter plot that displays the data for x (elevation above sea level in feet) and z (mean number of *cloudy days per year*).
2. Based on the scatter plot you constructed in Problem 1, is there a relationship between elevation and the mean number of cloudy days per year? If so, how would you describe the relationship? Explain your reasoning.



Data Source: Sample of six women who ran the 2003 NYC marathon

Consider the scatter plot on the left for Problems 3 and 4.

3. Is there a relationship between finish time and age, or are the data points scattered?
4. Do you think there is a relationship between finish time and age? If so, does it look linear?



Consider the scatter plot on the left for Problems 5 and 6.

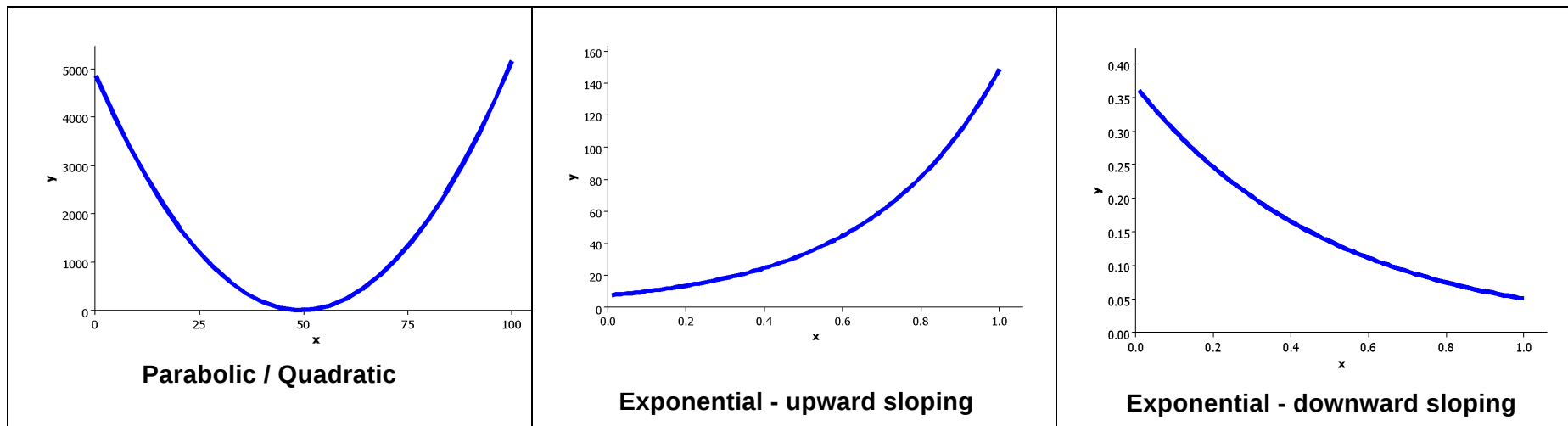
5. A mare is a female horse, and a foal is a baby horse. Is there a relationship between a foal's birth weight and a mare's weight, or are the data points scattered?
6. If there is a relationship between baby birth weight and mother's weight, does the relationship look linear?

Data Source: Elissa Z. Cameron, Kevin J. Stafford, Wayne L. Linklater, and Clare J. Veltman, "Suckling behaviour does not measure milk intake in horses, equus caballus," *Animal Behaviour*, 57 (1999): 673.

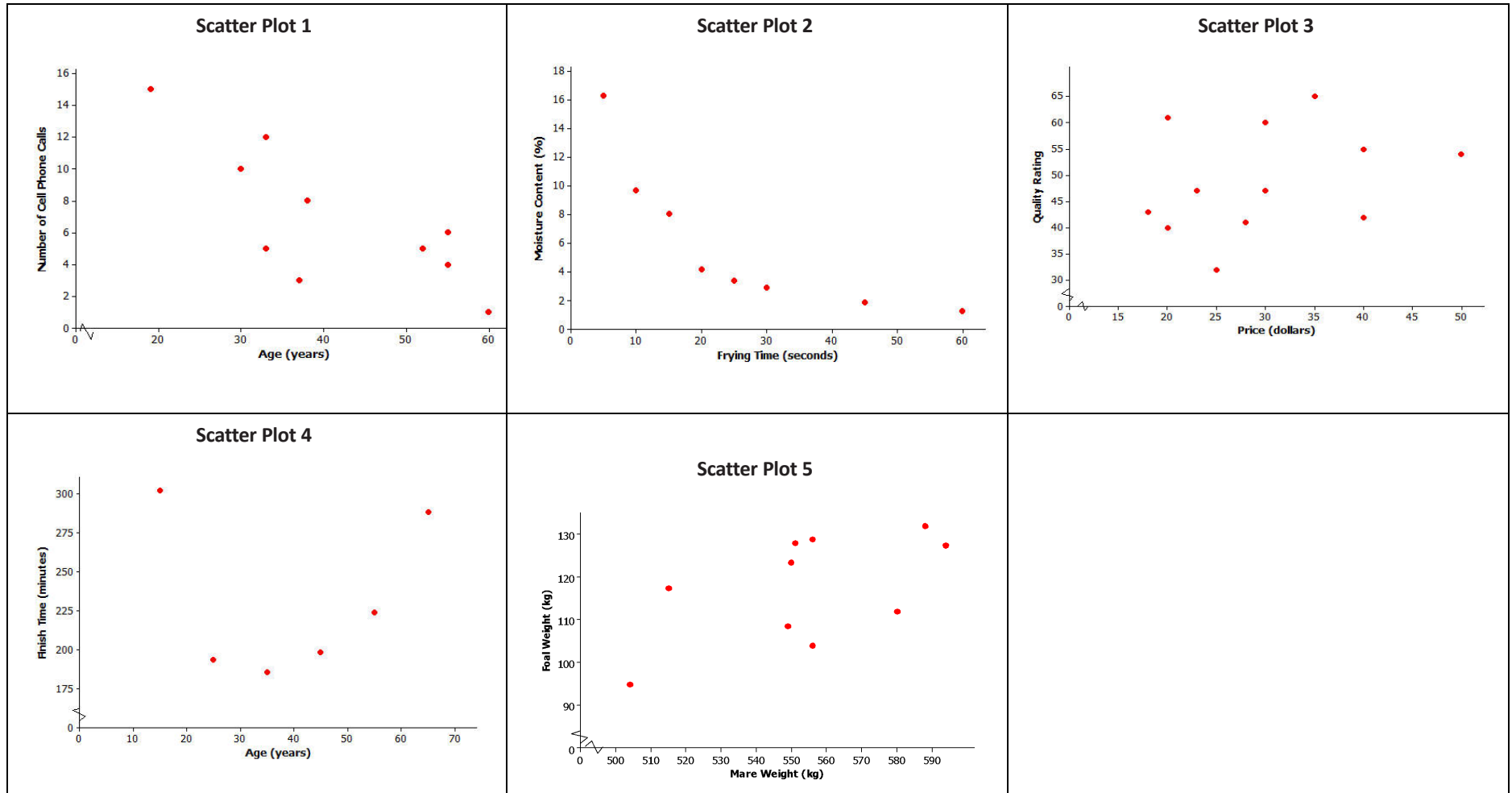
Relationships Between Two Numerical Variables

Not all relationships between two numerical variables are linear. There are many situations where the pattern in the scatter plot would best be described by a curve. Two types of functions often used in modeling nonlinear relationships are quadratic and exponential functions.

Some frequently encountered non-linear shapes of data are shown below:



Example 4: Consider the five scatter plots below

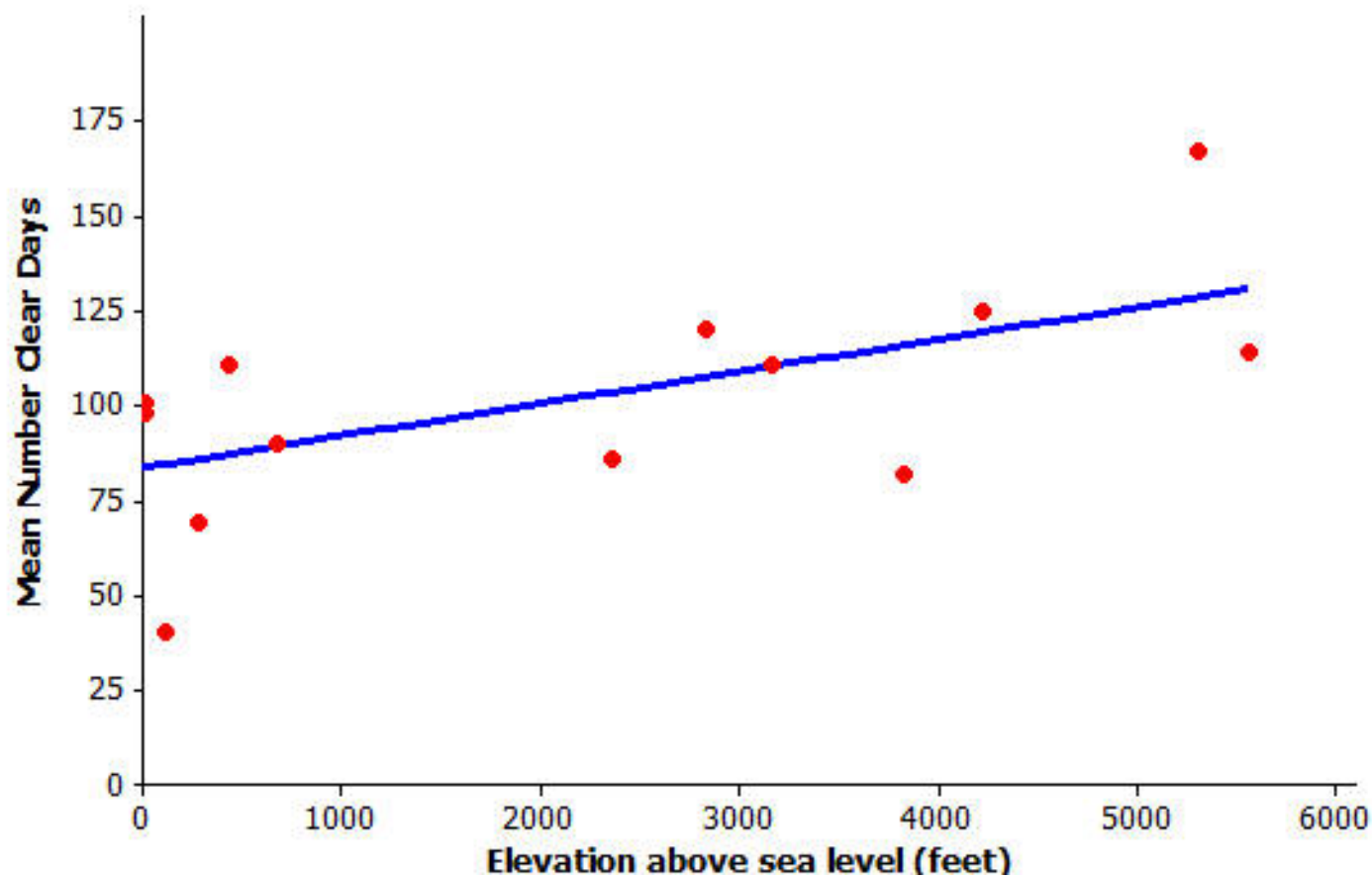


1. Which of the five scatter plots shows a pattern that could be reasonably described by a quadratic curve?
2. Which of the five scatter plots shows a pattern that could be reasonably described by an exponential curve?

Example 5: Altitude vs. Amount of Sunshine - Revisited

Let's revisit the data on elevation (in feet above sea level) and mean number of clear days per year. The scatter plot of this data is shown below. The plot also shows a straight line that can be used to model the relationship between elevation and mean number of clear days.

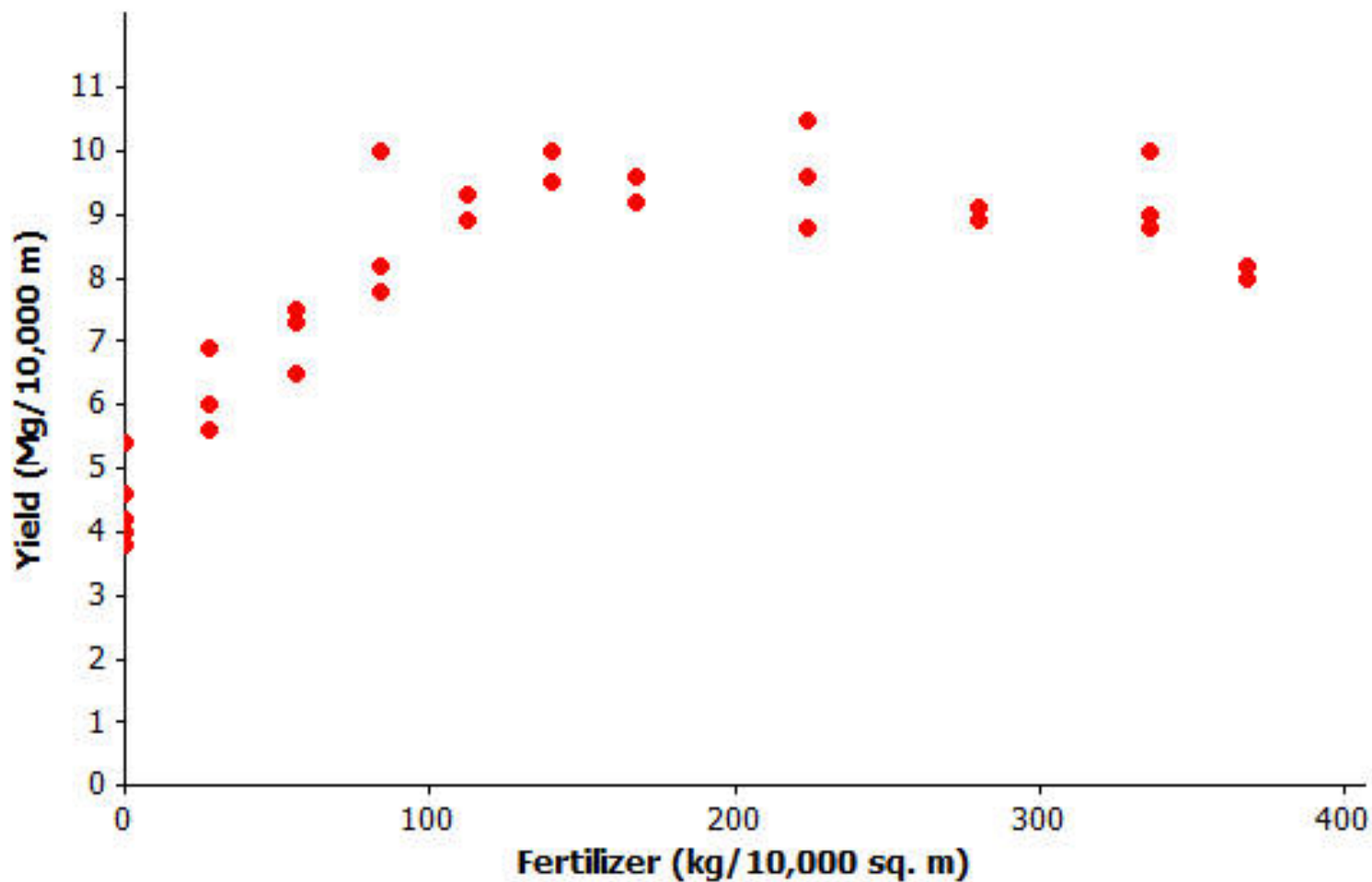
The equation of this line is $y = 83.6 + 0.008x$.



3. Assuming that the 14 cities used in this scatter plot are representative of cities across the United States, should you see more clear days per year in Los Angeles, which is near sea level, or in Denver, which is known as the mile-high city? Justify your choice with a line showing the relationship between elevation and mean number of clear days.
4. One of the cities in the data set was Albany, New York, which has an elevation of 275 ft . If you did not know the mean number of clear days for Albany, what would you predict this number to be based on the line that describes the relationship between elevation and mean number of clear days?
5. Another city in the data set was Albuquerque, New Mexico. Albuquerque has an elevation of $5,311\text{ ft}$. If you did not know the mean number of clear days for Albuquerque, what would you predict this number to be based on the line that describes the relationship between elevation and mean number of clear days? Was the prediction of the mean number of clear days based on the line closer to the actual value for Albany with 69 clear days or for Albuquerque with 167 clear days? How could you tell this from looking at the scatter plot with the line shown above?

Example 6: A Quadratic Model of Corn

Farmers sometimes use fertilizers to increase crop yield but often wonder just how much fertilizer they should use. The data shown in the scatterplot below are from a study of the effect of fertilizer on the yield of corn.



Data Source: M.E. Cerrato and A.M. Blackmer, "Comparison of Models for Describing Corn Yield Response to Nitrogen Fertilizer" *Agronomy Journal*, 82 (1990): 138.

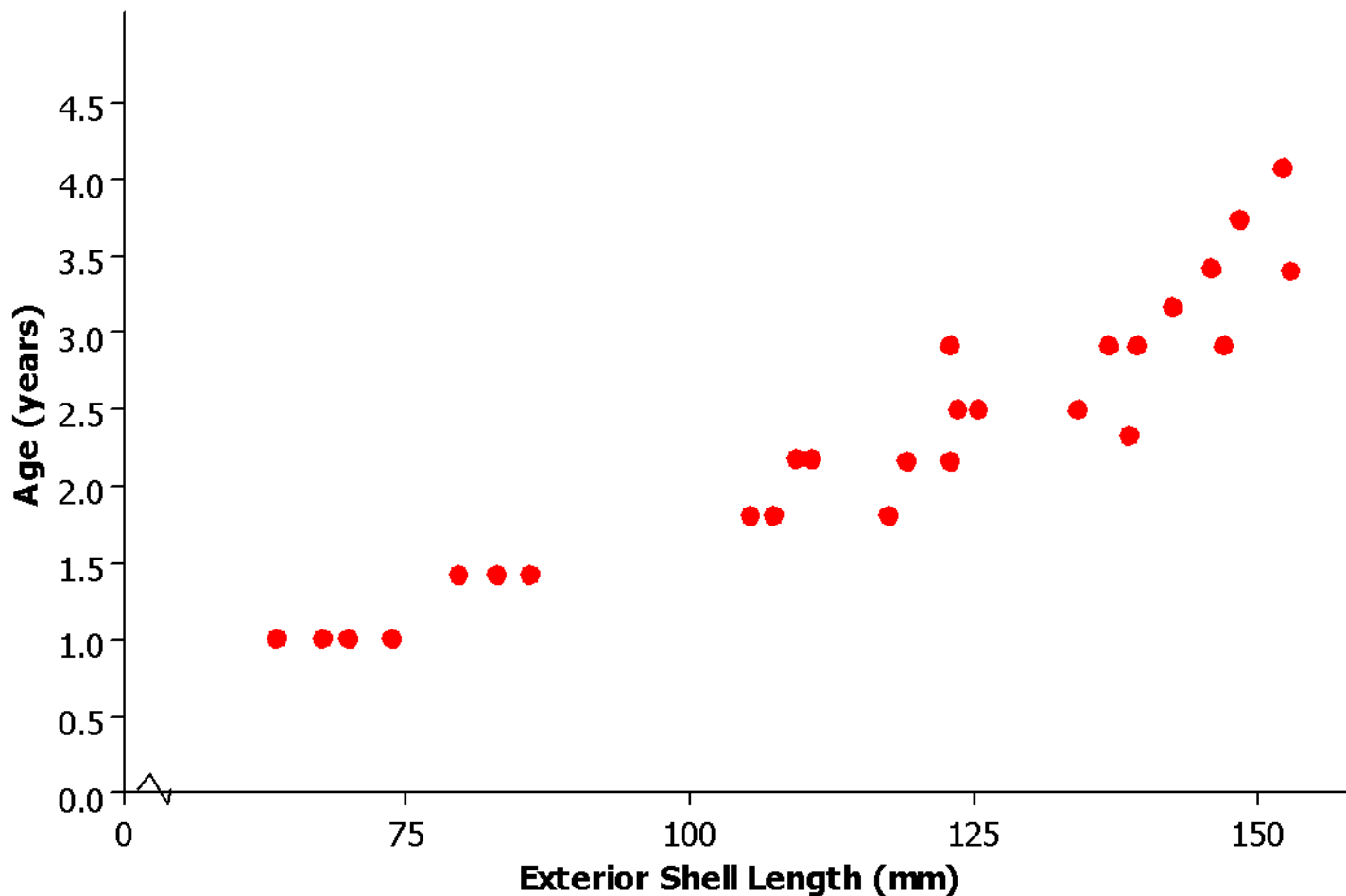
1. The researchers who conducted this study decided to use a quadratic curve to describe the relationship between yield and amount of fertilizer. Explain why they made this choice.
2. The model that the researchers used to describe the relationship was $y = 4.7 + 0.05x - 0.0001x^2$, where x represents the amount of fertilizer (kg per 10,000 $sq. m$) and y represents corn yield (Mg per 10,000 $sq. m$).
Use this quadratic model to complete the following table. Then sketch the graph of this quadratic equation on the scatter plot.

x	y
0	
100	
200	
300	
400	

3. Based on this quadratic model, how much fertilizer per 10,000 $sq. m$ would you recommend that a farmer use on his cornfields in order to maximize crop yield? Justify your choice.

Example 7: An Exponential Model of a Lobster

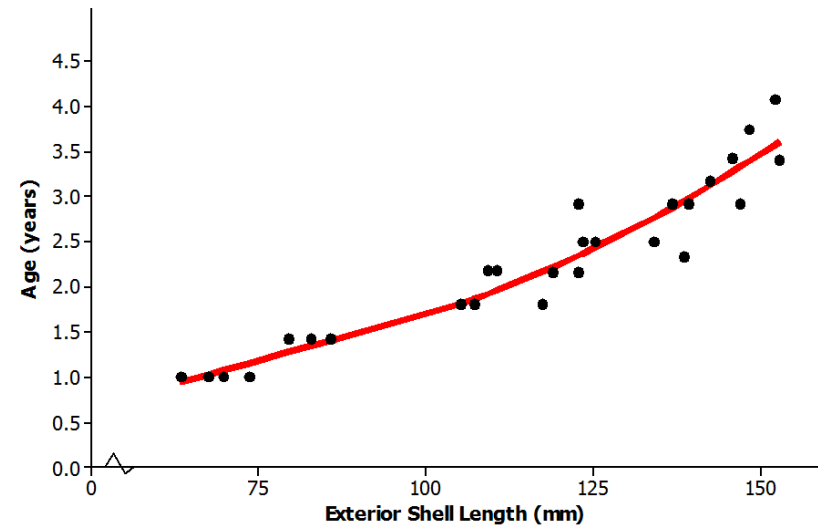
How do you tell how old a lobster is? This question is important to biologists and to those who regulate lobster trapping. To answer this question, researchers recorded data on the shell length of 27 lobsters that were raised in a laboratory and whose ages were known.



Data Source: Kerry E. Maxwell, Thomas R. Matthews, Matt R.J. Sheehy, Rodney D. Bertelsen, and Charles D. Derby, "Neurolipofuscin is a Measure of Age in *Panulirus argus*, the Caribbean Spiny Lobster, in Florida" *Biological Bulletin*, 213 (2007): 55.

1. The researchers who conducted this study decided to use an exponential curve to describe the relationship between age and exterior shell length. Explain why they made this choice.

2. The model that the researchers used to describe the relationship is $y = 10^{-0.403 + 0.0063x}$, where x represents the exterior shell length (mm), and y represents the age of the lobster (in years). The exponential curve is shown on the scatter plot on the right. Does this model provide a good description of the relationship between age and exterior shell length? Explain why or why not.



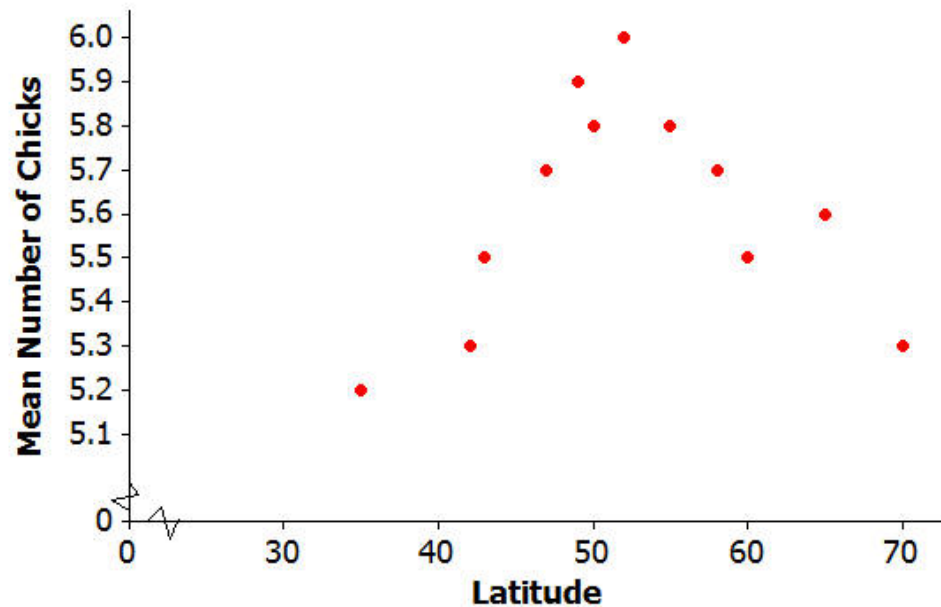
3. Based on this exponential model, what age is a lobster with an exterior shell length of 100 mm?
4. Suppose that trapping regulations require that any lobster with an exterior shell length less than 75 mm or more than 150 mm must be released. Based on the exponential model, what are the ages of lobsters with exterior shell lengths less than 75 mm? What are the ages of lobsters with exterior shell lengths greater than 150 mm? Explain how you arrived at your answer.

Try It Yourself:

Biologists conducted a study of the nesting behavior of a type of bird called a flycatcher. They examined a large number of nests and recorded the latitude for the location of the nest and the number of chicks in the nest.

1. What type of model (linear, quadratic, or exponential) would best describe the relationship between latitude and mean number of chicks?
2. One model that could be used to describe the relationship between mean number of chicks and latitude is $y = 0.175 + 0.21x - 0.002x^2$, where x represents the latitude of the location of the nest and y represents the number of chicks in the nest. Use the quadratic model to complete the following table. Then sketch a graph of the quadratic curve on the scatter plot provided at the beginning of the Problem Set.

x (degrees)	y
30	
40	
50	
60	
70	



Data Source: Juan José Sanz, "Geographic variation in breeding parameters of the pied flycatcher *Ficedula hypoleuca*" *Ibis*, 139 (1997): 107.

3. Based on this quadratic model, what is the best latitude for hatching the most flycatcher chicks? Justify your choice.

Suppose that social scientists conducted a study of senior citizens to see how the time (in minutes) required to solve a word puzzle changes with age. The scatterplot on the right displays data from this study.

Let x equal the age of the citizen and y equal the time (in minutes) required to solve a word puzzle for the seven study participants.

4. What type of model (linear, quadratic, or exponential) would you use to describe the relationship between age and time required to complete the word puzzle?
5. One model that could describe the relationship between age and time to complete the word puzzle is $y = 10^{-1.01 + 0.017x}$. This exponential curve is shown on the scatter plot on the right. Does this model do a good job of describing the relationship between age and time to complete the word puzzle? Explain why or why not.
6. Based on this exponential model, what time would you predict for a person who is 78 years old?

