

Key Concepts



Statistics: Scatter Plots and Best-Fit Lines

Objective Introduce students to two methods that are very useful in data analysis, that of scatter plots and best-fit lines.

Note to the Teacher *In this lesson, the emphasis is much more on display and interpretation of data than on mathematical techniques and concepts.*

Data Sets

First, have a discussion about data sets. Here are the main points.

- (1) Data sets are collections of data. A piece of data can be a number, an ordered pair, or any ordered “tuple” of numbers, such as ordered triples or quadruples of numbers.
- (2) Data sets are often obtained by measuring particular attributes of persons or entities.
 - We could get a data set by asking each student in a school to measure the length of his/her left arm and left leg. We could then arrange this information in an ordered pair.

(arm length, leg length)

There would be one ordered pair for each student. Taken together, they would form a collection of data consisting of ordered pairs.

- Another example would be to measure the blood pressure and caffeine consumption of a population of people, say all of the inhabitants of Springfield. We could arrange this information as an ordered pair and get a data set by forming the collection of all these ordered pairs.

(blood pressure, caffeine consumption)

- Measure, for each car in a collection of cars, the weight, the gas mileage, and the horsepower rating. For each car, we would form an ordered triple.

(weight, gas mileage, horsepower)

Taken together, they would give a data set consisting of ordered **triples** of numbers.

- (3) It is often desirable to understand properties of data sets, since they can give us information about the relationship between the various numbers that form the data set. For instance, we might find that cars with high horsepower rating or large weight tend to have low gas mileage. This is useful information in designing and buying cars. We might also find that people with high caffeine consumption tend to have higher blood pressure. This is extremely useful information for people who are deciding whether or not to consume caffeine.

Scatter Plots

Note to the Teacher *Scatter plots* are a simple way to display data sets visually, so that one can get some intuition about properties of the set. It applies only to data sets that consist of ordered pairs of numbers.

First, remind the class about plotting points. Then do an example on the chalkboard or on an overhead transparency.

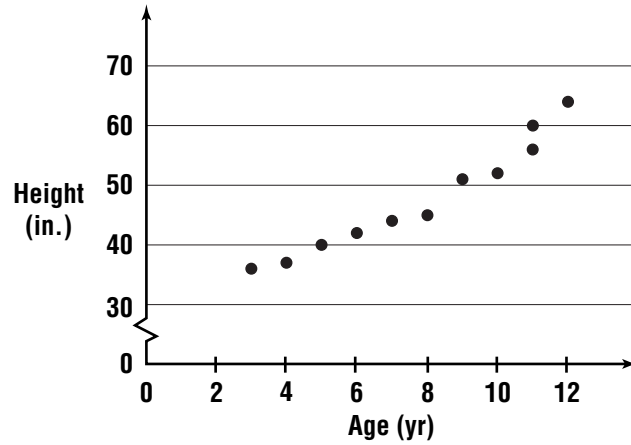
Example 1 The table at the right shows a data set that is obtained by collecting information about age and height of a group of children. Draw a scatter plot of the data and determine what relationship, if any, exists in the data.

Solution The data are given in the form of a table. This table gives the age and height of each child. We organize them into ordered pairs to get the following set of data.

$\{(3, 36), (5, 40), (8, 45), (11, 56), (12, 64), (10, 52), (4, 37), (9, 51), (6, 42), (7, 44), (11, 60)\}$

Age (yr.)	Height (in.)
3	36
5	40
8	45
11	56
12	64
10	52
4	37
9	51
6	42
7	44
11	60

To make the scatter plot, plot the points on a graph in which the axes are drawn with an appropriate scale for the numbers we are given. The graph will look like this.



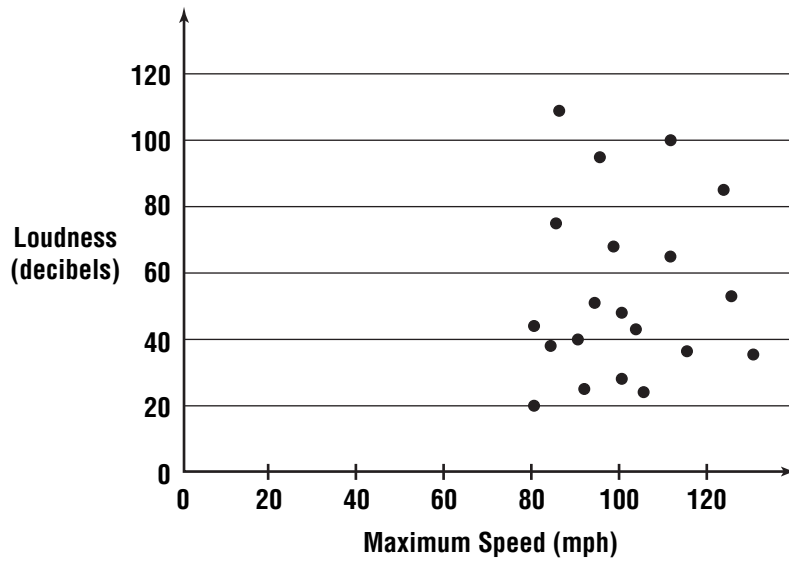
Plotting these points shows the class the mechanics of producing a scatter plot. The graph can be interpreted as saying that in general, the older the child, the taller he or she will be, because the y values grow as the x values grow.

Sometimes a scatter plot shows no relationship between the quantities in the ordered pairs. It is good to show such an example.

Example 2 For each car in a collection of cars, the maximum speed of the car (in miles per hour) and the loudness of the horn (in decibels) are measured and recorded in the following table. Draw a scatter plot of the data and determine what relationship, if any, exists in the data.

Maximum Speed (mph)	Loudness (decibels)	Maximum Speed (mph)	Loudness (decibels)
80	20	85	110
90	40	95	95
95	52	103	43
110	100	115	37
100	28	98	68
85	75	92	25
80	44	125	53
130	35	111	65
105	24	122	84
100	48	84	38

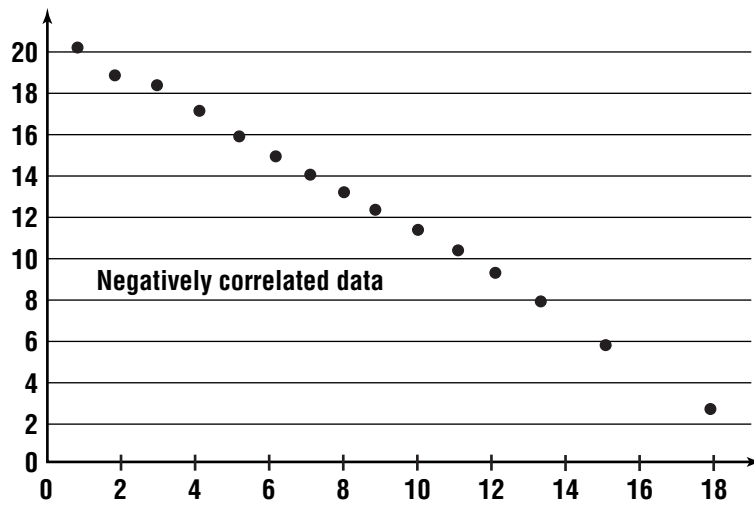
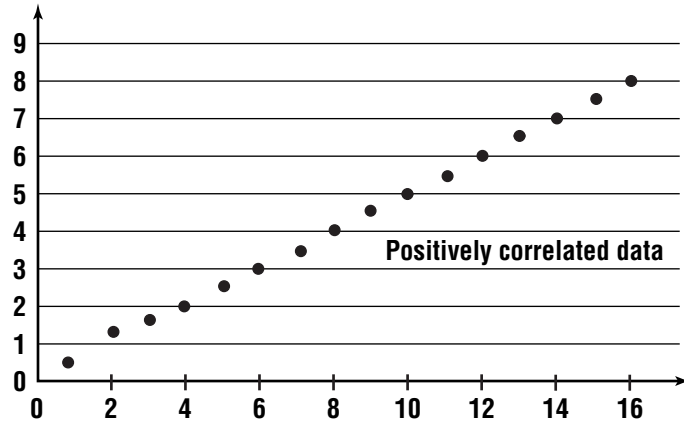
Solution Here is what the scatter plot for this data set looks like.



Notice that there is no particular pattern to these data. As we might expect, the loudness of the horn has nothing to do with the maximum speed of the car. A data set in which the plot spreads out in this way is called **uncorrelated**. The data set in Example 1, in which there is a real correlation between age and height of a child, is called **correlated**. When data are correlated, there are two possibilities, **positive** or **negative** correlation.

Key Idea	<p>In a scatter plot, the two variables are</p> <ul style="list-style-type: none">• positively correlated if the y-coordinate tends to grow as the x-coordinate grows, and• negatively correlated if the y-coordinate tends to decrease as the x-coordinate grows.
-----------------	---

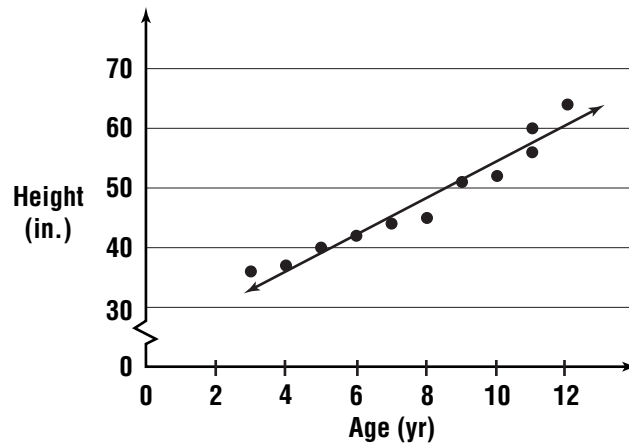
See the examples on the next page.



Best-Fit Line

Note to the Teacher *The idea of a “best-fit line” introduced here is an informal one. The line is obtained by drawing freehand a line that appears to fit the data well. There is a more formal notion of a line of best fit, found by a method called **linear regression**, which is the best such line in a formal mathematical sense.*

Using one of the well-correlated data sets from before, draw a line through the data set that appears to fit the data well.



Explain to students the value of drawing the line. It is useful to have a line that describes the data for the following three reasons.

- (1) It is much easier to keep track of or store the equation of a line than it is to keep track of all the points in a data set.
- (2) The line gives us a better understanding of the data. For instance, if we know the slope, it tells us that a certain change in the x -coordinate will produce a change in the y -coordinate, which will be the slope times that change.
- (3) Having the line allows us to predict the y values for corresponding x values.

