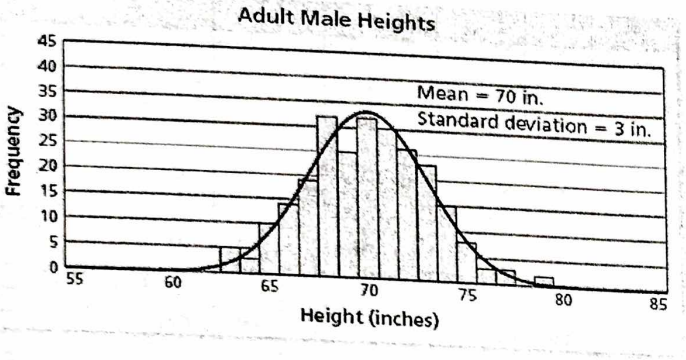
Lesson #80
Introduction – Shapes of Distributions

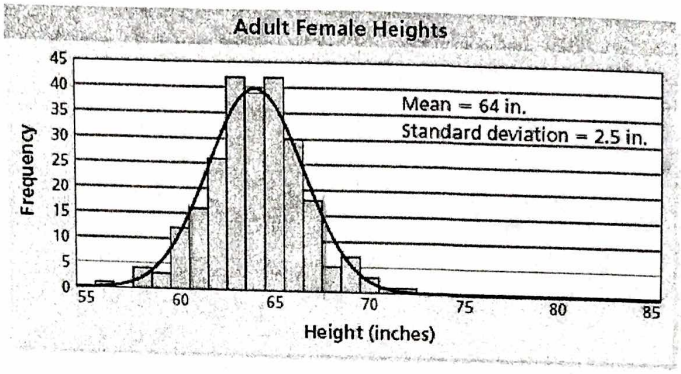
Success Criteria: I can compare distributions and determine their shape.

These histograms show the distributions of the heights (in inches) of 250 adult American males and 250 adult American females.

Each graph is also roughly outlined with a curve to help show the SHAPE of the data sets.

Note that the **mean** and **standard deviation** are given for each data set as well.





(A) Which set has the smaller mean? Explain what this means in real-life context.

Adult Female Heights

On average, most females are shorter than men.

(B) Which set has the smaller standard deviation? Explain what this means in real-life context.

Adult Female Heights

Women's heights are less dispersed.

(C) Both distributions of data are roughly symmetric. This roughly results in a bell-shape as you can see by the bell-shaped curves that are drawn around the data sets. Describe the characteristics of a symmetric (or bell-shaped) data distribution.

More values in the middle.

The sides slowly decrease from the middle then quickly decrease

Success Criteria: I can determine how much of a given data set is within one standard deviation of the mean.

(D) For the male data, roughly calculate the percentage of men whose height lies within 1 standard deviation of the mean (in other words, those whose heights fall anywhere from 67 to 73 inches).
Follow these steps:
How many men have a height of:

67? 19

68? 32

69? 25

70? 32

71? 30

72? 26

73? 23

How many total men have a height between 67 to 73 inches? 187

Divide this number by 250, then multiply by 100, to find the percentage of men whose height lies within one standard deviation of the mean:

74.8%

(E) Follow similar steps for the female data. Roughly calculate the percentage of women whose height lies within 1 standard deviation of the mean (in other words, those whose heights fall anywhere from 61 to 67 inches).

61 16

62 26

63 42

64 39

65 42

66 30

67 18

85.2%

Lesson #80b
Shapes of Distributions

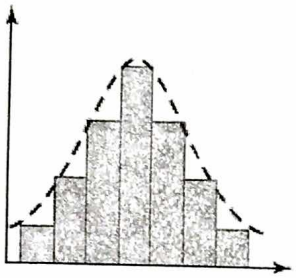
Success Criteria: I can determine if data is skewed left, skewed right or symmetric. I can determine which measure of central tendency will best describe the data. I can determine if the interquartile range or the standard deviation will best show the spread of the data.

HISTOGRAM:

Displays the frequency of data values in same size intervals using bars.

Using a histogram is another useful way to see the SHAPE of distribution in the data.

Various SHAPES of histograms:

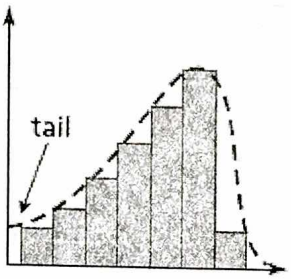


Symmetric

(1) The data are **evenly** distributed on each side of the highest bar.

(2) Both the **mean** and the **median** will be near the center of the data, but we typically just use the **mean** to measure the **center** (measure of central tendency) for this type of data distribution.

(3) The **standard deviation** will best describe (or help show) the **spread** of the data.

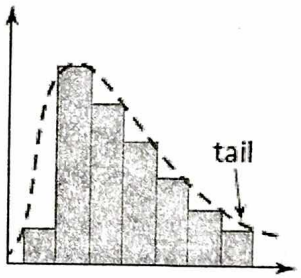


Skewed left

(1) The tail of the graph extends to the **left**, while most of the data are on the **right**.

(2) The best measure of the **center** (or measure of central tendency) will be the **median**, since the mean will be more off center than the median.

(3) The **interquartile range (the box)** of the **five number summary** will best describe (or help show) the **spread** of the data.



Skewed right

(1) The tail of the graph extends to the **right**, while most of the data are on the **left**.

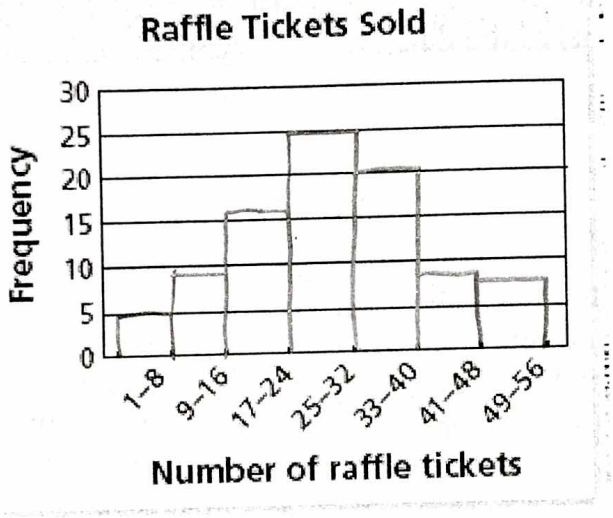
(2) The best measure of the **center** (or measure of central tendency) will be the **median**, since the mean will be more off center than the median.

(3) The **interquartile range (the box)** of the **five number summary** will best describe (or help show) the **spread** of the data.

Example 1: The frequency table below shows the number of raffle tickets sold for 90 students at a school.

Number of Tickets Sold	Frequency
1–8	5
9–16	9
17–24	16
25–32	25
33–40	20
41–48	8
49–56	7

(A) Display the data in a histogram.



(B) Describe the shape (skewed left, symmetric, or skewed right) of the data distribution.

Symmetric - data is evenly spread

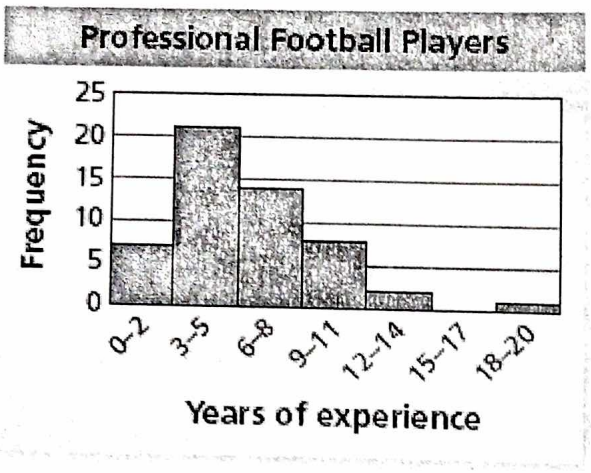
(C) Which measure of central tendency, the mean **OR** median, would best represent the data? (In other words, which could be used to measure or show the "center" of the data values)

Mean & median will both be near the center but we will use mean

(D) Which measure of dispersion (or spread), the standard deviation **OR** the interquartile range (the box) of the five number summary, would best represent the data?

Standard deviation

Example 2: Use the histogram shown below.



(A) Describe the shape (skewed left, symmetric, or skewed right) of the data distribution.

Skewed right

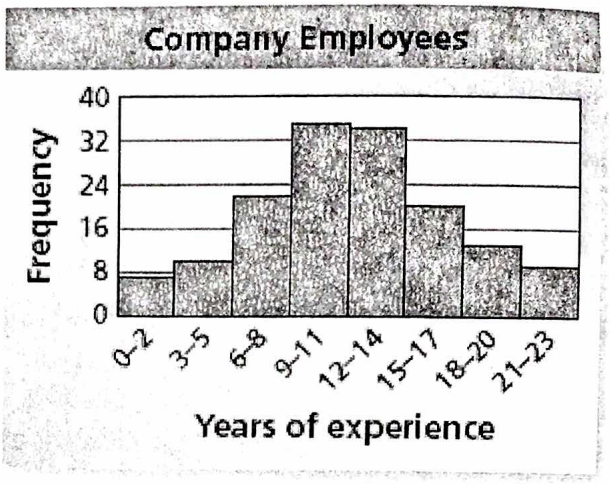
(B) Which measure of central tendency, the mean **OR** the median, could represent the data? (In other words, which could be used to measure or show the "center" of the data values)

Median

(C) Which measure of dispersion (or spread), the standard deviation **OR** the interquartile range (the box) of the five number summary, would best represent the data?

interquartile range

Example 3: Use the histogram shown below.



(A) Describe the shape (skewed left, symmetric, or skewed right) of the data distribution.

Symmetric

(B) Which measure of central tendency, the mean **OR** the median, could represent the data? (In other words, which could be used to measure or show the "center" of the data values)

mean

(C) Which measure of dispersion (or spread), the standard deviation **OR** the interquartile range (the box) of the five number summary, would best represent the data?

standard deviation

Practice: Worksheet 12.4 Practice A #2, 3

Lesson #81
Scatter Plots and Lines of Fit

Success Criteria: I can create a scatter plot given a table. I can use a scatter plot to predict future data and answer questions about that data.

Problem #1:

You have been working on a science project for 8 months. Each month, you have measured the length of a baby alligator.

The table shows your measurements.

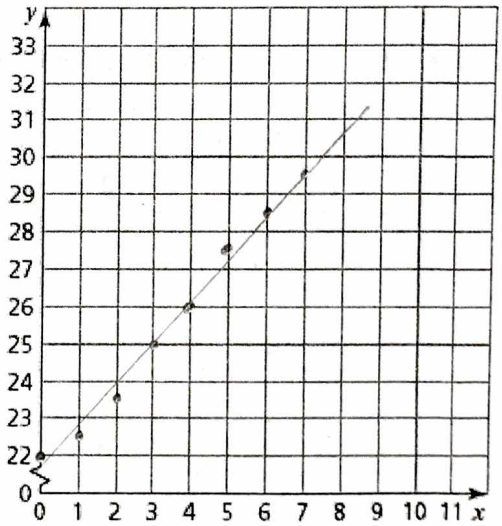
September ↓ April ↓

Month, x	0	1	2	3	4	5	6	7
Length (in.), y	22.0	22.5	23.5	25.0	26.0	27.5	28.5	29.5

Use the following steps to predict the baby alligator's length next September.

a. Graph the data in the table.

b. Draw the straight line that you think best approximates the points.



c. Write an equation of the line you drew.

$y = 1.1x + 21.6$

d. Use the equation to predict the baby alligator's length next September.

$y = 1.1(12) + 21.6$

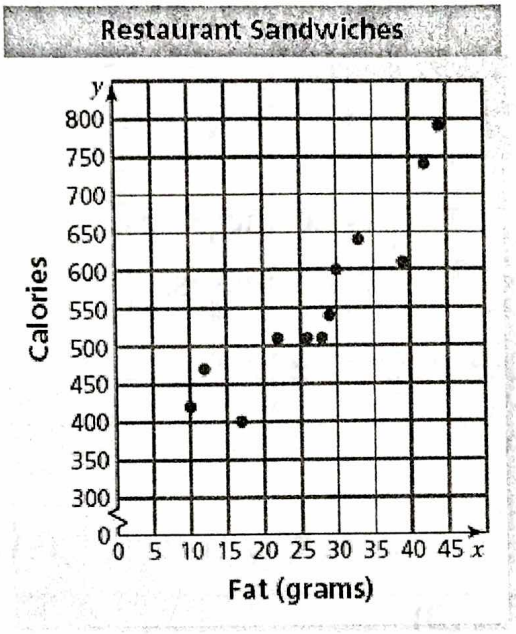
$y = 34.8$ in

Scatter Plot:

A **scatter plot** is a graph that shows the relationship between two data sets.

We graph the two sets of data as ordered pairs in a coordinate plane.

Example 1: The scatter plot below shows the amounts of fat (in grams) and the number of calories in 12 restaurant sandwiches.



(A) How many calories are in the sandwich that contains 17 grams of fat?

400 calories

(B) How many grams of fat are in the sandwich that contains 600 calories?

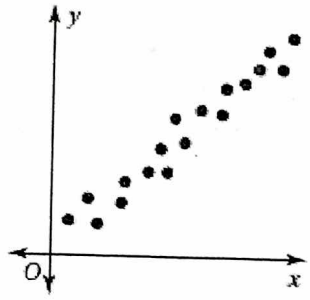
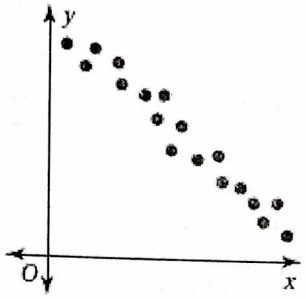
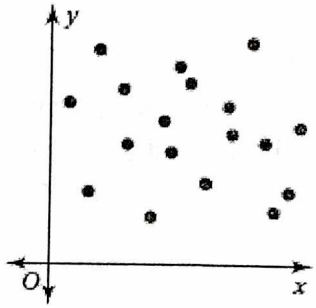
30 grams

(C) As the number of grams of fat increases, what tends to happen to the number of calories?

Calories increase

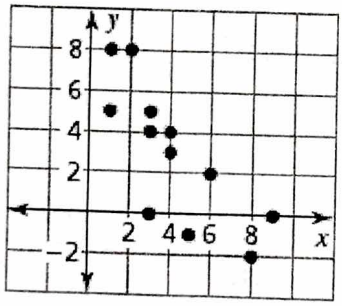
Success Criteria: I can determine if a scatter plot with have a positive, negative or no relationship..

Relationships between two data sets:

Positive Relationship	*Negative Relationship*	*No Relationship*
		
As x increases, y increases.	As x increases, y decreases.	The data points show no pattern.

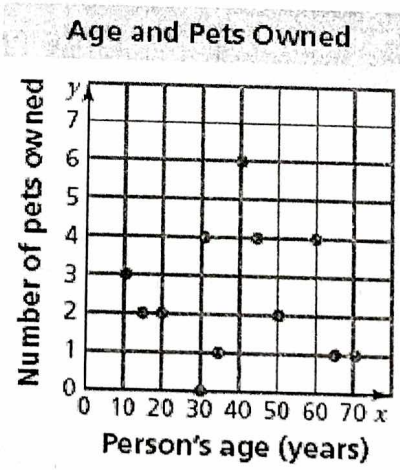
Example 2: Tell whether the data show a POSITIVE, a NEGATIVE, or NO relationship.

(A)



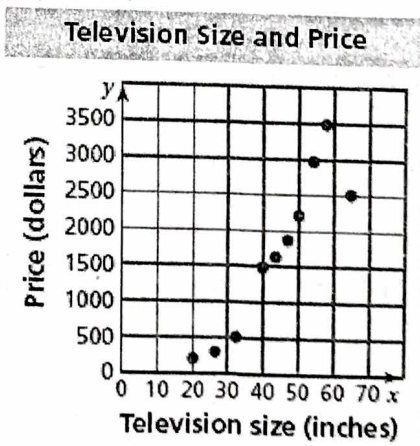
Negative correlation

(B)



No correlation

(C)



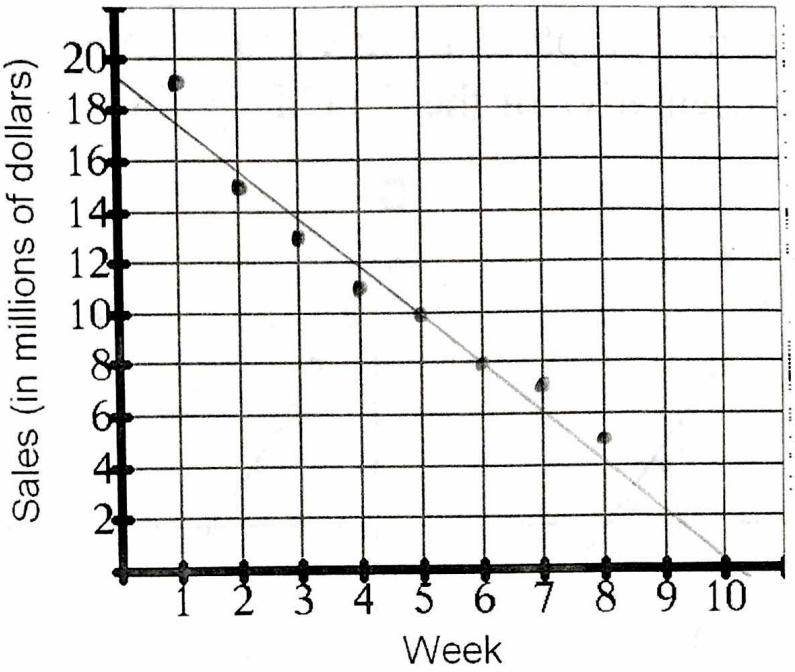
Positive correlation

Success Criteria: I can use my calculator to find an equation of best fit.

Example 3: The table below shows the weekly sales (in millions of dollars) of a DVD and the number of weeks since its release.

(A) Make a scatter plot of the data and then draw a "line of fit".

Week, x	Sales (millions), y
1	\$19
2	\$15
3	\$13
4	\$11
5	\$10
6	\$8
7	\$7
8	\$5



(B) Use your calculator to type in the x values in L1 and the y values in L2. Go to STAT then arrow over to CALC. Find the "linreg" category. Make sure your x values are L1 and your y values are L2. Write your equation below.

Equation: $y = 19.25 - 1.83x$

(C) Using your equation from part B, graph the line using your calculator.

(D) Interpret and explain the meaning of the slope of your equation.

Each week the sales go down by \$1.83 million

(E) Use the equation for the line of fit to predict the sales in week 9.

$y = -1.83(9) + 19.25$

\$2.78 mil

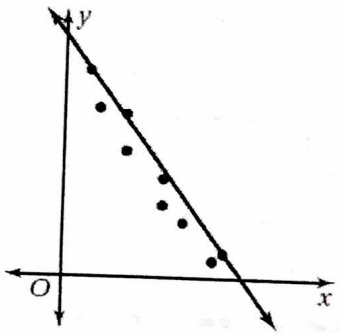
Lesson #81b
Scatter Plots and Lines of Fit

Success Criteria: I can create a scatter plot given a table. I can use a scatter plot to predict future data and answer questions about that data. I can determine if a scatter plot with have a positive, negative or no relationship. I can use my calculator to find an equation of best fit.

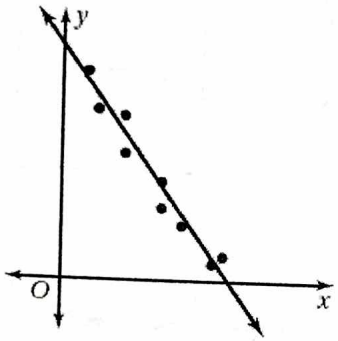
Problem #1:

In which graph is the line shown most representative of the data? Explain.

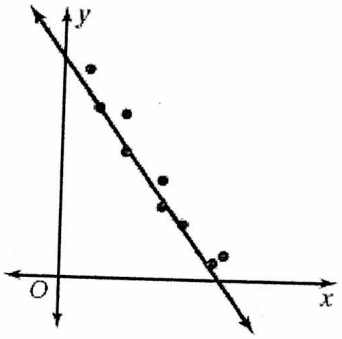
A



B



C



Problem #2

The table shows the weights of *x* bananas.

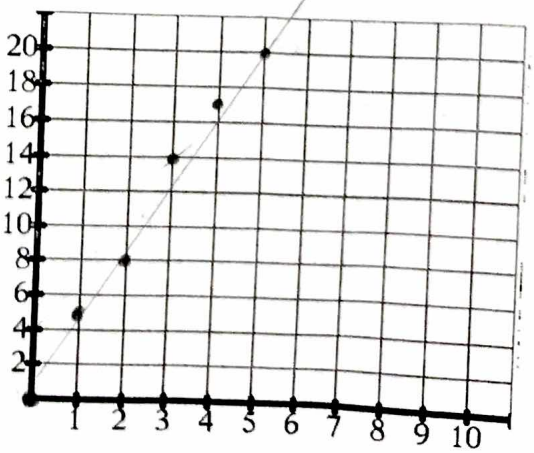
Number of Bananas	0	1	2	3	4	5
Weight (ounces)	0	5	8	14	17	20

1. Graph the data in the table.
2. Draw the straight line that you think best approximates the points.
3. Write an equation of the line you drew.

   $y = 4x + 0.5$
4. Use the equation to predict the weight of 10 bananas.

   $y = 4(10) + 0.5$

   $y = 40.5$



Problem 3: The table below shows the number of people who attended a neighborhood festival over an 8-year period.

Year, x	1	2	3	4	5	6	7	8
Attendance, y	420	500	650	900	1100	1500	1750	2400

(A) What kind of relationship do the data have (positive, negative, or none)?

Positive

(B) Find the equation for the line of fit on a graphing calculator.

Line of fit equation: $y = 272.14x - 72.14$

(C) Interpret and explain the meaning of the slope.

Each year about 272 more people attend

(D) Use part (B) to predict the number of people who will attend the festival in year 14.

$y = 272.14(14) - 72.14$

$y = 3738$ people

Lesson #82
Analyzing Lines of Fit

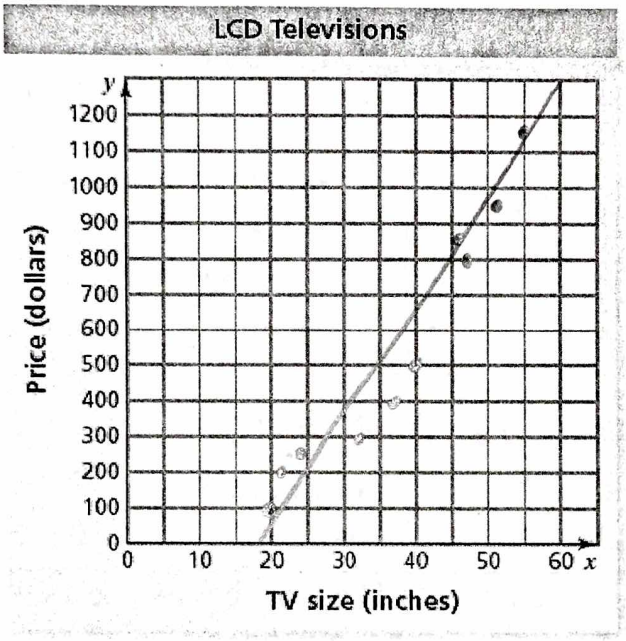
Success Criteria: I can determine if a line is a good fit for a set of data.

Problem #1:

You researched and gathered data on the price of liquid crystal display (LCD) televisions. The table below shows the size and price of several LCD televisions.

x (TV size in inches)	20	22	24	32	37	40	46	47	52	55
y (price in dollars)	100	200	250	300	400	500	850	800	950	1150

(A) Make a scatter plot of the data.



(B) Describe the relationship between the TV size and the price. Is this true in real life?

Positive correlation

(C) The line y = 30x – 500 is drawn on the graph. Is it a good fit for the data? Explain.

Yes - even amount above & below

(D) About how much do you expect a 60-inch TV to cost?

$y = 30(60) - 500$

$y = 1800 - 500$ $1300

(E) The largest HD screen in the world is at Dallas Cowboy stadium. It is approximately a 175 foot TV. How much would this line predict that TV would cost?

$y = 30(175) - 500$ $4750

$y = 5250 - 500$

(F) Find how much it actually cost. What do you make of that?

$40 million

Success Criteria: I can determine if a graph has positive, negative or no correlation based on the correlation coefficient.

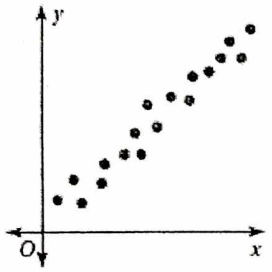
The **correlation coefficient** value r tells us two things:

(1) The closer the value of r is to -1 or to 1 the more closely the linear regression equation will fit the data set.

Generally, an r value around -0.9 or 0.9 (or better) will mean it's a **good fit.

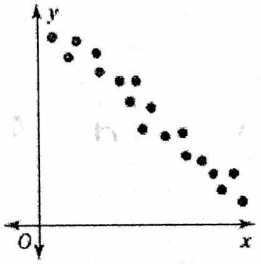
(2) A positive value of r means there is a positive correlation between the variables, while a negative value of r means there is a negative correlation between the variables. However, if the value of r is close to 0, there is relatively no correlation.

Positive Correlation



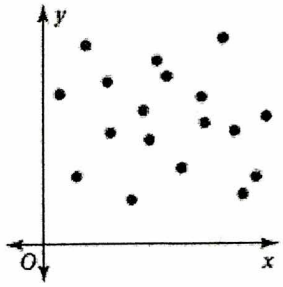
As x increases, y increases.

Negative Correlation



As x increases, y decreases.

No Correlation



The data points show no relationship.

Example 1: (Do on a graphing calculator) The table below shows the worldwide movie ticket sales (in billions of dollars) from the year 2000 to 2010, where $x = 0$ represents the year 2000.

Year, x	0	1	2	3	4	5	6	7	8	9	10
Ticket Sales, y	16	17	20	20	25	23	26	26	28	29	32

(A) Enter the data into the lists.

(B) Find the equation for the line of fit: $y =$ 1.5x+16.32

(C) Find the correlation coefficient: $r =$ 0.98

(D) Based on the value of r, will the line be a good fit (strong fit) for the data or not?

Yes

(E) Based on the value of r, is there a positive correlation, a negative correlation, or relatively no correlation between the year and ticket sales?

Positive correlation

Success Criteria: I can determine if there is correlation and/or causation in a situation.

We know that a correlation between two variables means that when one variable changes (increases or decreases), the other variable changes (increases or decreases).

This can happen **without** one variable directly causing the other variable to change.

In other words, there can be a correlation with or without **causation**.

For example: In one school, as the average test scores in history increased, the average test scores in science also increased.

Although there is a correlation here between the history scores and the science scores, there is no reason why the change in history scores would directly affect the science scores.

What might have caused both scores to go up?

Students have been studying more.

Example 2: For each situation described, tell whether there is likely to be a correlation between the variables. If there is a correlation, tell whether there is also causation.

(A) The time spent exercising and the number of calories burned. Correlation? Yes Causation? Yes	**(B)** The number of gallons of milk a family consumes per month and the number of people in the family. Correlation? Yes Causation? No
(C) The number of hours a person works per week and the number of pickles they eat per week. Correlation? No Causation? No	**(D)** Years of education and yearly earnings. Correlation? Yes Causation? No & yes

Lesson #83
Two-Way Tables

Success Criteria: I can complete a two way frequency table and find the marginal frequencies. I can find information and percentages given a two way frequency table.

Problem #1:

The table below (called a two-way table) shows the number of soccer T-shirts a sports shop has left in stock at the **END** of the season.

		T-Shirt Size					
		S	M	L	XL	XXL	Total
Color	Blue/White	5	4	1	0	2	12
	Blue/Gold	3	6	5	2	0	16
	Red/White	4	2	4	1	3	14
	Black/White	3	4	1	2	1	11
	Black/Gold	5	2	3	0	2	12
	Total	20	18	14	5	8	65

(A) Find the totals for each row and then for each column. Record them in the table.

(B) Are there any black and gold (Black/Gold) size XL T-shirts left in stock? If so, how many?

No

(C) The number of soccer T-shirts the shop ordered at the **BEGINNING** of the season are shown in the two-way table below. Complete this table by filling in the totals.

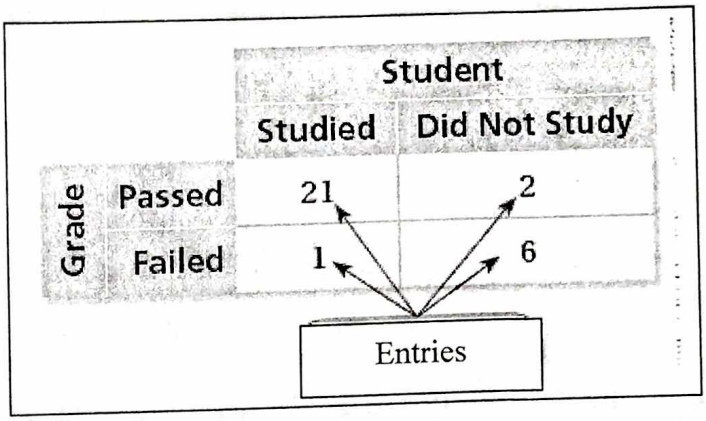
		T-Shirt Size					
		S	M	L	XL	XXL	Total
Color	Blue/White	5	6	7	6	5	29
	Blue/Gold	5	6	7	6	5	29
	Red/White	5	6	7	6	5	29
	Black/White	5	6	7	6	5	29
	Black/Gold	5	6	7	6	5	29
	Total	25	30	35	30	25	145

(D) Based on the two tables, how should the store manager alter the number of T-shirts he/she orders for next season? Explain your reasoning.

- Buy more XL + XXL shirts
- Less S/M/L shirts

A **two-way table** displays two categories of data collected from the same source.

EXAMPLE: Suppose you randomly surveyed students in your school about their grades on the last test and whether they studied for the test. You could record your results in a two-way table as shown.



Example 1:

A) How many students took the test?

30

B) How many students did not study and then failed the test?

6

C) How many students passed the test? This is called a **marginal frequency**.

23

The SUMS of the rows/columns are called **marginal frequencies. (In the MARGINS)**

The marginal frequencies for the table data above are completed and explained below.

		Student		
		Studied	Did Not Study	Total
Grade	Passed	21	2	23
	Failed	1	6	7
	Total	22	8	30

23 students passed.

7 students failed.

30 students were surveyed.

22 students studied.

8 students did not study.

Example 2: Kelsey randomly surveyed students between the ages of 12 and 17 about whether they ride the bus to school. The two-way table shows the results.
(** Note that the marginal frequencies are totaled and recorded as well)

		Age			
		12–13	14–15	16–17	Total
Student	Rides Bus	24	12	14	50
	Does Not Ride Bus	16	13	21	50
	Total	40	25	35	100

(A) How many students of age 14 to 15 do not ride the bus?

13

(B) How many students of age 16 to 17 ride the bus?

14

(C) Another way to use a two-way table to display and analyze such data is to calculate and record the **PERCENTAGES**.

For example, 14 of the 35 students surveyed who were of age 16-17 said they ride the bus.

Calculating this percentage, we get $\frac{14}{35} = 0.4$ or 40%.

This means that 40% of students surveyed who were of age 16-17 ride the bus.

(Note where this and other such percentages are recorded in the two-way table below)

		Age		
		12–13	14–15	16–17
Student	Rides Bus	60%	48%	40%
	Does Not Ride Bus	40%	52%	60%

(D) Explain what the 52% represents.

52% of 14-15 year olds do not ride the bus

(E) Does this table show a relationship between age and whether students ride the bus? Explain.

Yes - as they get older less ride the bus

Example 3: Jim randomly surveyed students in his school about whether they buy a school lunch or pack a lunch. His results are shown below.

Grade 6 students
11 pack lunch, 9 buy school lunch
Grade 7 students
23 pack lunch, 27 buy school lunch
Grade 8 students
16 pack lunch, 14 buy school lunch

(A) Record the data in the two-way table below. Total and record the marginal frequencies.

Lunch Type	Grade 6	Grade 7	Grade 8	
Pack	11	13	16	50
Buy	9	27	14	50
	20	50	30	100

(B) How many students in grade 7 buy a lunch?

27

(C) What percentage of students in grade 7 buys a lunch?

$\frac{27}{50} = 54\%$

(D) What percentage of all the students pack a lunch?

$\frac{50}{100} = 50\%$

Lesson #84
Misleading Data Displays

Success Criteria: I can look at different data representations and determine if they are misleading. I can identify the reasons why someone might want to represent data in a different way.

Example 1: One of the line graphs below is misleading in showing how box office sales changed over a seven year period. Decide which one it is and explain your reasoning.

Table #1

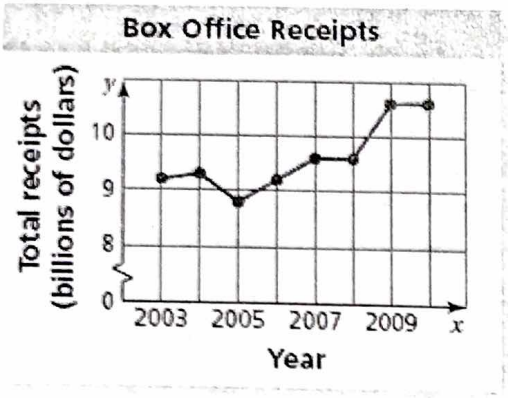
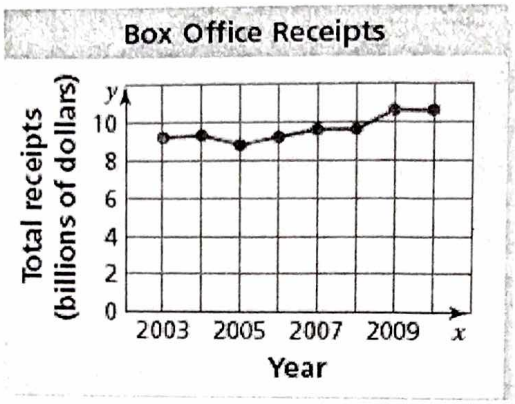
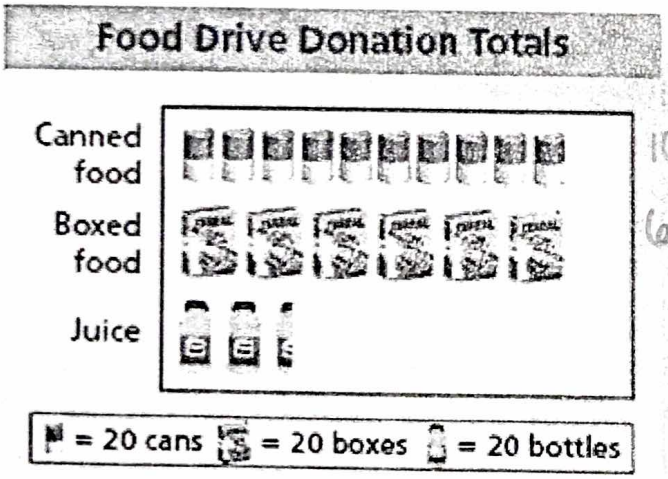


Table #2



Misleading - doesn't show whole y-axis
Looks like more changes

Example 2: Looking at the pictograph below, a volunteer concludes that the number of cans of food donated was about the same as the number boxes of food donated.



$10 \cdot 20 = 200$

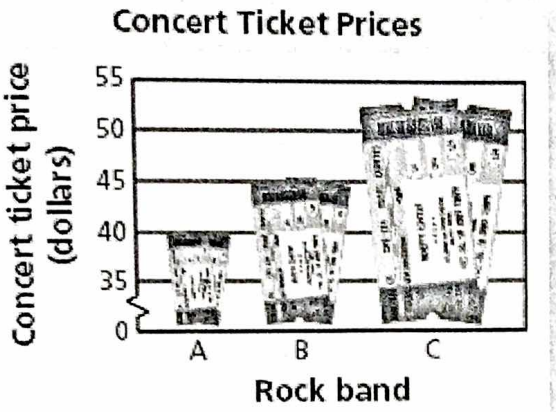
$6 \cdot 20 = 120$

Is this conclusion accurate? Explain.

No the boxes are bigger so they look like they take up more space

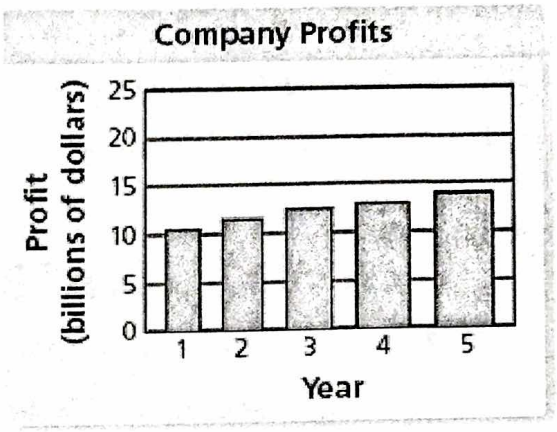
Example 3: (Group Problem) Explain why each data display is misleading.

(A)



Ticket size is much bigger for band C but only $12 difference

(B)



Profit didn't increase that much - The bars get wider to look bigger