DATA ANALYSIS

In the first part of this unit you will process data using measures of central tendency and measures of variation. You will then examine and analyze displays of data in various types of graphs. You will also compare biased and unbiased samples of data.

Measures of Central Tendency

Measures of Variation

Box-and-Whiskers Plots

Calculate Outliers

Scatter Plots

Histograms

Sampling

Measures of Central Tendency

To analyze sets of data, researchers often try to find a number that can represent the whole set. These numbers or pieces of data are called **measures of central tendency**. The three common measures we are going to study are: **mean, mode,** and **median**.

Mean: The mean of a set of data is the sum of all the data divided by the number of pieces of data (average).

Mode: The mode of a set of data is the number that occurs most often.

Median: The median of a set of data is the number in the middle when the data are arranged in order. When there are two middle numbers, the median is the average (mean) of the two numbers.

Let's take a look at Amanda's data from her class contest and determine the mean, mode, and median.

6	**4**	**4**	**10**	**6**
5	**2**	**4**	**1**	**5**
3	**3**	**7**	**4**	**2**
0	**9**	**5**	**7**	**10**

Mean: **Add all the numbers, and then divide the sum by the number of numbers in the set.**

6 + 4 + 4 + 10 + 6 + 5 + 2 + 4+ 1 + 5 + 3 + 3 + 7 + 4 + 2 + 0 + 9 + 5 + 7 + 10 = 97

$$97 \div 20 = 4.85$$

4.85 is the mean of the set of data.

Mode: Arrange the numbers in order from the smallest to the largest and determine which number occurs most often.

0, 1, 2, 2, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 9, 10, 10

4 is the mode of the set of data because it occurs most often.

Median: Use the arranged numbers from the mode and determine what the middle number is.

0, 1, 2, 2, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 9, 10, 10

Since there are two middle numbers, we must add them together and find the mean.

$$4 + 5 = 9$$
$$9 \div 2 = 4.5$$

The median of the set of data is **4.5**.

Measures of Variation

The study of **quartiles** helps us learn about the nature and tendency of data. Quartiles divide data that is arranged in order from least to greatest into four equal parts. The ***median***, sometimes referred to as the Second Quartile, separates the data in half. The ***Lower Quartile (LQ),*** sometimes referred to as the First Quartile, is the median of the first half of the data. The ***Upper Quartile (UQ)***, sometimes referred to as the Third Quartile, is the median of the second half of the data. The ***range*** of the data is the difference between the highest and lowest data values and is found by subtracting these values. The ***Interquartile Range*** is the range of the middle half of the data and is found by subtracting the lower quartile from the upper quartile **(UQ – LQ).** Many companies analyze data to determine the promotion and implementation of their product.

Below are several examples of data arranged in order from least to greatest. The median, LQ, UQ, and interquartile range, and range have been determined for each set of data.

If there is an even number of numbers in the data, there will be two middle numbers. To find the median, calculate the average of the two middle numbers.

*Note: The first step in determining quartiles is to put the **data in order** from least to greatest.

Example 1:

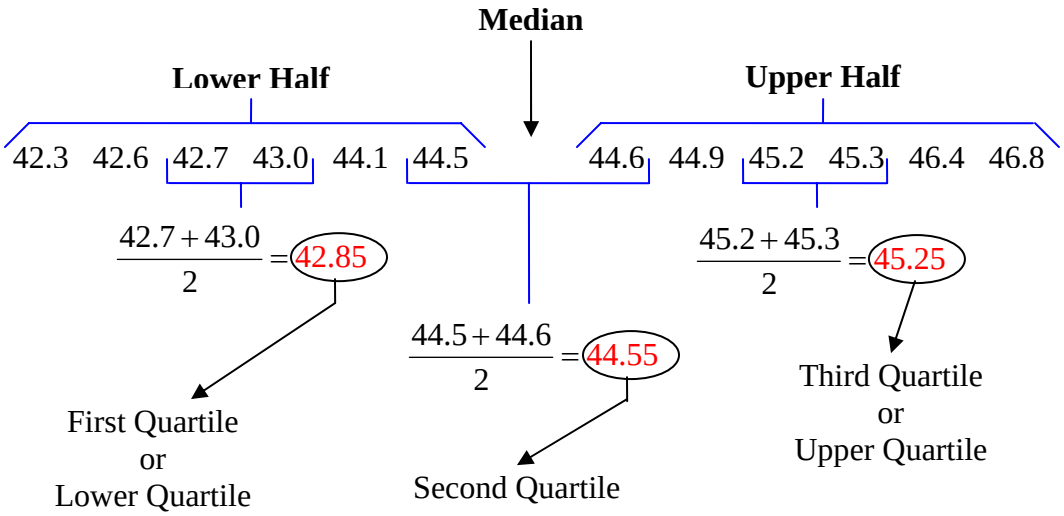
In this example there are an even number of data points in the data; thus, the **Second Quartile (median)** is the average of the two middle numbers. (44.55)

There are an even number of data points in the lower half of the data; thus, the **First Quartile** is the average of the two middle numbers in the lower half. (42.85)

There are also an even number of data points in the upper half of the data; thus, the **Third Quartile** is the average of the two middle numbers in the upper half. (45.25)

The **range** is the difference between the highest and lowest data points. (4.5)

The **interquartile range** is the difference between the Upper Quartile (Third Quartile) and the Lower Quartile (First Quartile). (2.4)



Range = Highest Value – Lowest Value
Range = 46.8 – 42.3 = 4.5
Interquartile Range = UQ – LQ
Interquartile Range = 45.25 – 42.85 = 2.4

Example 2:

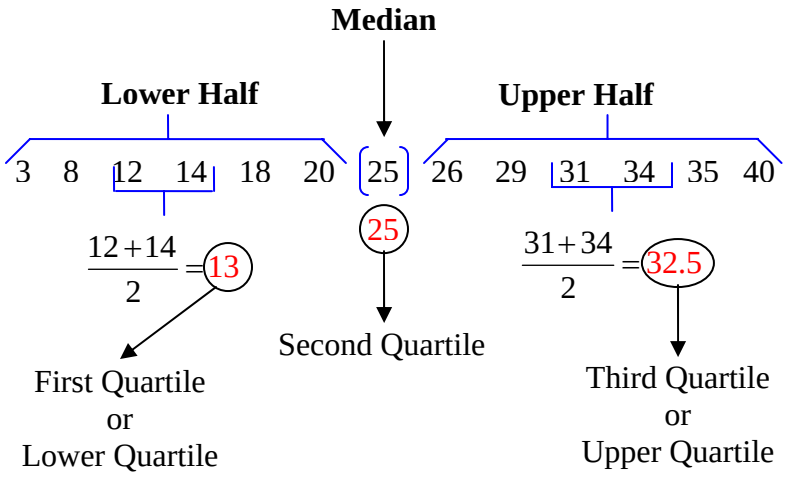
In this example there are an odd number of data points in the data; thus, the **Second Quartile (median)** is the middle number. (25)

There are an even number of data points in the lower half of the data; thus, the **First Quartile** is the average of the two middle numbers in the lower half. (13)

There are also an even number of data points in the upper half of the data; thus, the **Third Quartile** is the average of the two middle numbers in the upper half. (32.5)

The **range** is the difference between the highest and lowest data points. (37)

The **interquartile range** is the difference between the Upper Quartile (Third Quartile) and the Lower Quartile (First Quartile). (19.5)



Range = Highest Value – Lowest Value
Range = 40– 3 = 37
Interquartile Range = UQ – LQ
Interquartile Range = 32.5 – 13 = 19.5

Example 3:

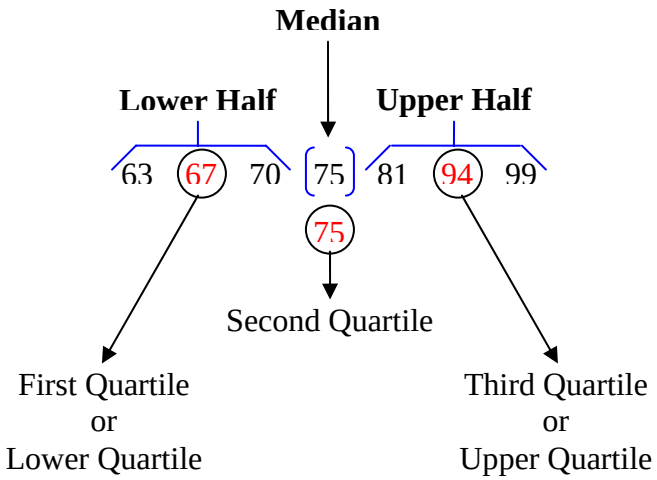
In this example there are an odd number of data points in the data; thus, the **Second Quartile (median)** is the middle number. (75)

There are an odd number of data points in the lower half of the data; thus, the **First Quartile** is the middle number in the lower half. (67)

There are also an odd number of data points in the upper half of the data; thus, the **Third Quartile** is the middle number in the upper half. (94)

The **range** is the difference between the highest and lowest data points. (36)

The **interquartile range** is the difference between the Upper Quartile (Third Quartile) and the Lower Quartile (First Quartile). (27)



Range = Highest Value – Lowest Value
Range = 99 – 63 = 36
Interquartile Range = UQ – LQ
Interquartile Range = 94 – 67 = 27

Box-and-Whiskers Plots

Box-and-whiskers plots are used to separate data into four sections. The parts will differ in length in the graph, but each part will contain one fourth of the data, with the exception of the outliers.

- **Median** - separates the entire data set in half
- LQ – **Lower Quartile** - median of the lower half of the data
- UQ – **Upper Quartile -** median of the upper half of the data
- **Interquartile Range** – difference between the UQ and LQ
- **Outliers –** data points that fall beyond the upper quartile or below the lower quartile (They are points that are more than 1.5 times the value of the interquartile range plus the UQ or minus the LQ.)
- **Lower Extreme** – smallest data point (excluding the outlier)
- **Upper Extreme** – largest data point (excluding the outlier)

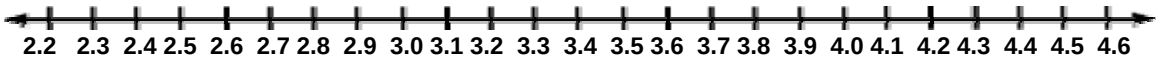
Follow the steps below to make a box-and-whiskers plot for the given data set. The data set represents the average monthly precipitation that occurred in Columbus, OH, in 2003. Also check the data set for outliers.

Average Monthly Precipitation Columbus, OH (in inches) 2003			
2.5	2.2	2.9	3.3
3.9	4.1	4.6	3.7
2.9	2.3	3.2	2.9

Step 1: Put the data in **order** from **least to greatest**.

2.2, 2.3, 2.5, 2.9, 2.9, 2.9, 3.2, 3.3, 3.7, 3.9, 4.1, 4.6

Step 2: Draw a **number line** with a **scale** that fits the data.



Step 3: Find the **median.**

$$\frac{2.9+3.2}{2}=3.05$$

Step 4: Find the quartiles (**LQ** and **UQ**).

- lower quartile (LQ) - median of the lower half

$$\frac{2.5+2.9}{2}=2.7$$

- upper quartile (UQ) - median of the upper half

$$\frac{3.7+3.9}{2}=3.8$$

Step 5: Calculate the interquartile range. (UQ – LQ)

$$3.8-2.7=1.1$$

Step 6: Check for outliers.

First, multiply the interquartile range by 1.5.

$$1.1\times 1.5=1.65$$

Second, Add 1.65 to the upper quartile. (UQ + 1.65)

$$3.8+1.65=5.45$$

Compare data points to 5.45. Are there any data points greater than 5.45? No!

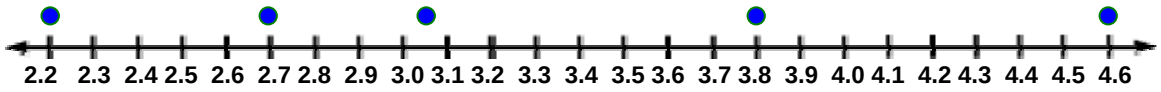
Third, Subtract 1.65 from the lower quartile. (LQ – 1.65)

$$2.7-1.65=1.05$$

Compare data points to 1.05. Are there any data points lower than 1.05? No!

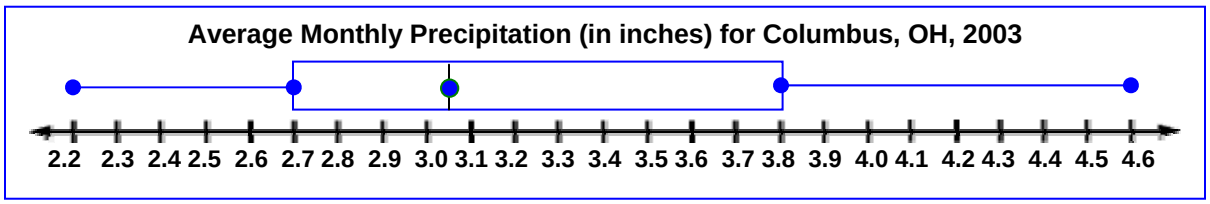
There are no outliers in this data set.

Step 7: **Plot the points** for the following data points: **lower extreme, LQ, median, UQ, and upper extreme.**



Step 8: Draw the **box-and-whiskers** graph.

- A rectangle (box) extends from the lower quartile (LQ) to the upper quartile (UQ).
- A vertical line is drawn through the median of the data set.
- Line segments (whiskers) extend from LQ to the lower extreme data point and from UQ to the upper extreme data point.
- Add a title to the graph.



- Data points 2.2, 2.3, and 2.5 are located under the left "whisker". ($\frac{1}{4}$th of the data points)
- Data points 2.9, 2.9, and 2.9 are located to the left of the median under the "box". ($\frac{1}{4}$th of the data points)
- Data points 3.2, 3.3, and 3.7 are located to the right of the median under the "box". ($\frac{1}{4}$th of the data points)
- Data points 3.9, 4.1, and 4.6 are located under the right "whisker". ($\frac{1}{4}$th of the data points)

*Note: If any outliers did exist in the data set, they would be plotted as separate points above the number line.

Calculate Outliers

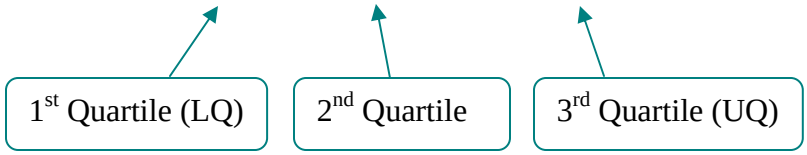
Outlier – An outlier is a data point widely separated from the main cluster of points in a sample of data.

To check for outliers follow these steps:

1) Subtract the first quartile from the third quartile. (UQ – LQ)
2) Multiply the difference by 1.5
3) Add the product found in Step 2 to the third quartile (UQ). Compare to see if any numbers in the data are greater than this sum. If so, the data point is an outlier.
4) Subtract the product found in Step 2 from the first quartile (LQ). Compare to see if any numbers in the data are less than this sum. If so, the data point is an outlier.

Example: Check for outliers.

2 6 7 8 9 12 15 15 18 24 45



Step 1: 18 – 7 = 11

Step 2: 11 × 1.5 = 16.5

Step 3: 18 + 16.5 = 34.5 *Any data points above 34.5?* Yes! Data point 45 is above 34.5

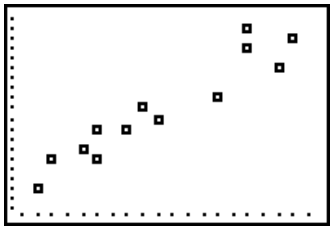
Step 4: 7 – 16.5 = −9.5 *Any data points below −9.6?* No!

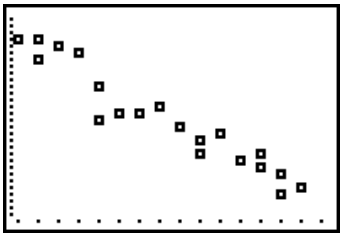
There is one outlier, **45**.

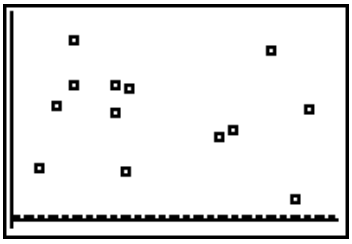
Scatter Plots

Scatter plots are an easy way of determining if there is a relationship between two variables. This relationship is called a **correlation**. A correlation is based on the slope of the line of best fit, a line that is drawn through the data and represents the overall trend of the data.

There are three possible types of correlation: a) positive, b) negative, or c) no correlation. The illustrations below show the graph of each correlation.

a) 

b) 

c) 

In graph "a", notice how the points cluster in a rise to the right. Recall from a previous lesson that this suggests a **positive slope**. Graph "b" shows points that cluster in a fall to the right, which suggests a **negative slope.** Graph "c" shows no cluster pattern and suggests the two variables have **no relationship** to each other.

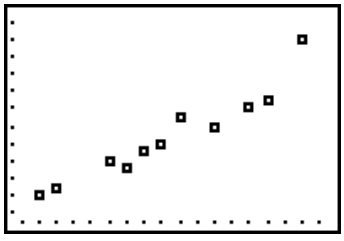
Let's take a look at a few examples and determine if each situation has a positive correlation, a negative correlation, or no correlation.

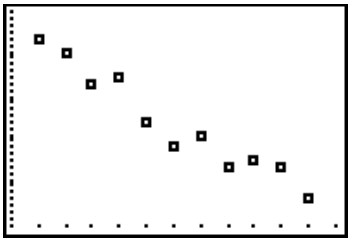
Example 1: Determine which scatter plot represents each situation.

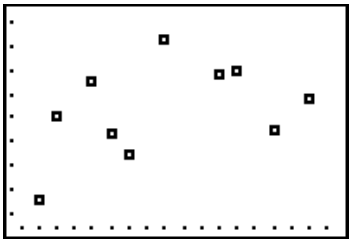
a) your height and your hourly wage

b) your height and your shoe size

c) your age and the time needed to run 100 yards

1) 

2) 

3) 

Scatter plot 1 shows a strong positive correlation. A **positive correlation** occurs when **both variables increase**. As you grow taller, your shoe size increases; therefore plot 1 represents situation "b".

Scatter plot 2 shows a strong negative correlation. A **negative correlation** occurs when **one variable increases as the other variable decreases**. In situation "c", as your age increases, the time it takes you to run 100 yards decreases. (Consider the time period from birth through young adulthood.)

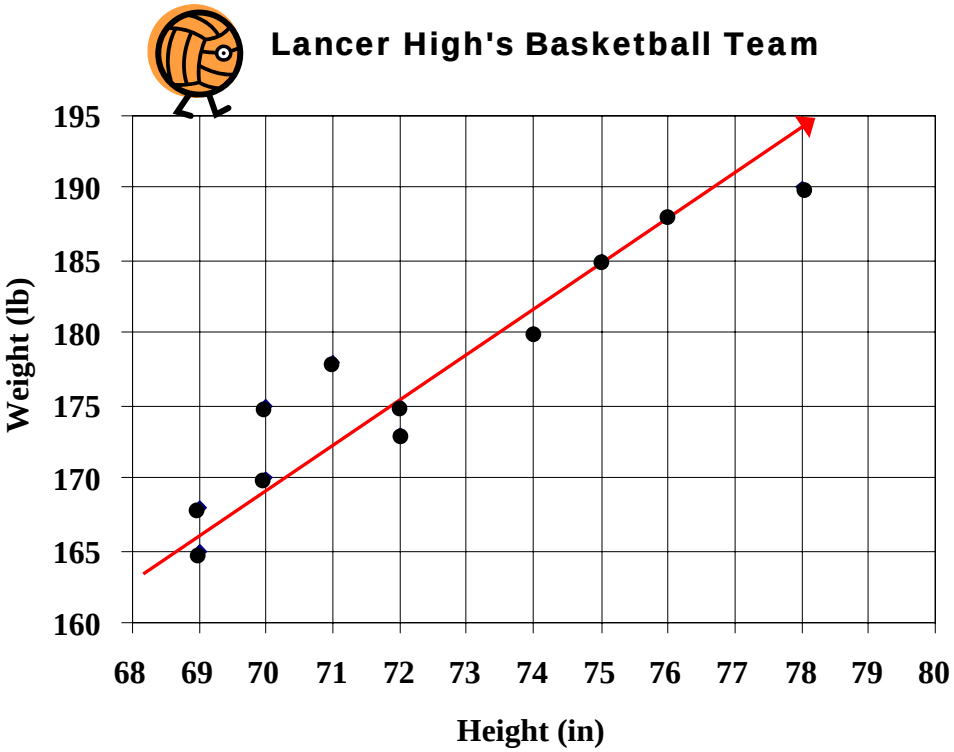
The third scatter plot shows **no correlation** because the data points are **randomly scattered**. Your height has no relationship with your hourly wage; therefore, this plot represents situation a.

Example 2: Use the data shown below to make a scatter plot of the weight and height of each member of Lancer High's basketball team.

Height (in)	Weight (lb)
70	170
69	165
72	175
74	180
75	185
70	175
69	168
72	173
71	178
78	190
76	188
69	165

Step 1: Make a scatter plot of the data pairs. The points on the scatter plot are (70, 170), (69, 165), (72, 175), (74, 180), (75, 185), (70, 175), (69, 168), (72, 173), (71, 178), (78, 190), (76, 188), and (69, 165).

Step 2: Draw the line that appears to best fit the data points. There should be about the same number of points above the line as below it. The line does not have to pass through any of the data points.



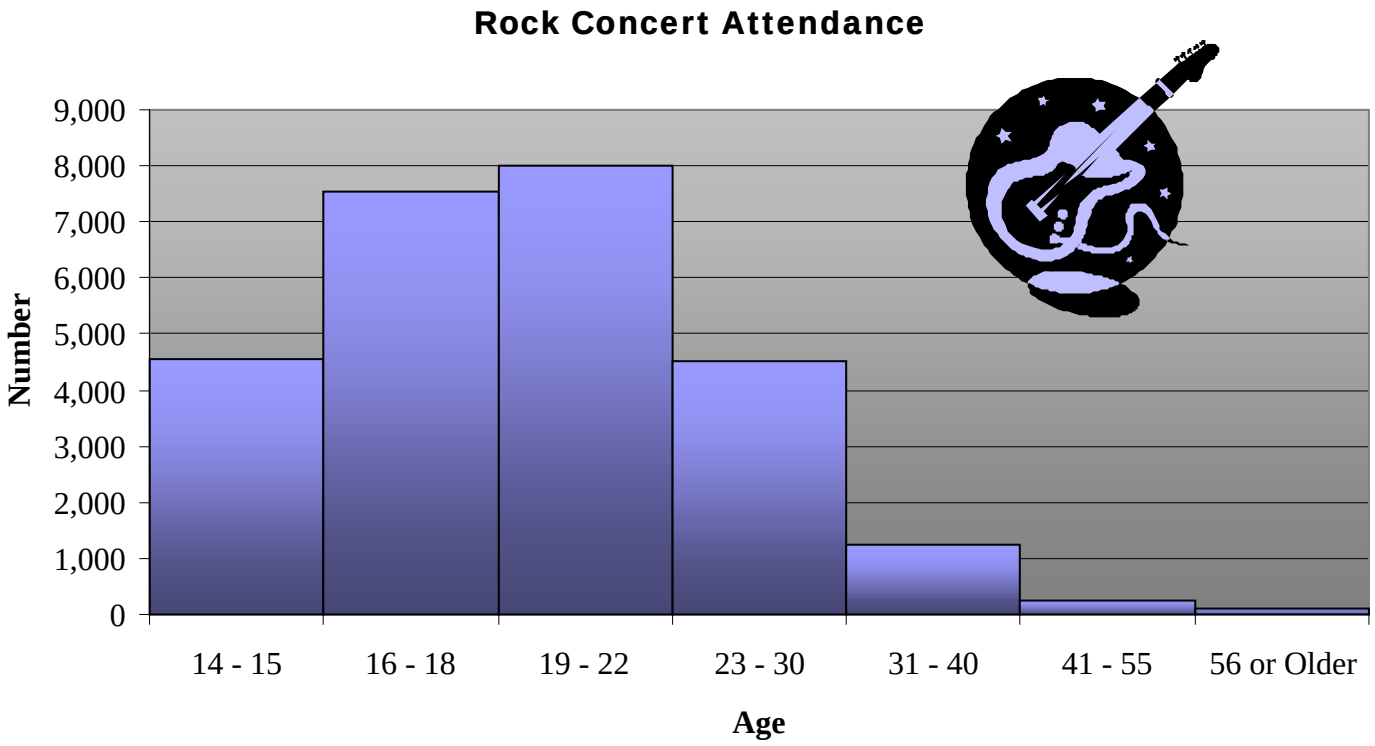
What kind of correlation exists between the data sets?

The scatter plot displays a positive correlation. The line of best fit slants upward and to the right indicating a positive correlation.

Histograms

Darin wanted to attend the rock concert being held at the local civic center. His mom agreed that he could go, but only if she went also. The data below depicts the ages of the persons attending the rock concert. It is also displayed in a histogram. Histograms are used to display ranges of data.

Ages	Frequency
14 – 15 (Young Teens)	4538
16 – 18 (Teens)	7536
19 – 22 (Young Adults)	7999
23 – 30 (Adults)	4520
31 – 40 (Middle-Aged Adults)	1236
41 – 55 (Older Adults)	233
56 or Older (Seniors)	120



What age group had the highest attendance? The 19-22 year olds

What age group had the lowest attendance? 56 or Older

In a histogram, a display of data which possesses a **symmetric** distribution is one in which the two “halves” of the data appear as mirror-images to one another.

A **skewed** (non-symmetric) distribution of data is a distribution in which the data appears at more diverse values. Distributions of data are skewed if one of the tails of a histogram (the part that stretches out from the peak) is longer than the other.

A "**skewed right**" distribution is one in which the right tail is longer.

A "**skewed left**" distribution is one in which the left tail is longer.

How would the distribution of the data of the rock concert attendance be described?

The data is “**skewed right**”. Most of the rock concert attendees were younger than 30 years old.

Sampling

When collecting data to make predictions, it is necessary to get an **unbiased** sample selection (small group) that will be representative of the population (whole group).

Suppose Rita wanted to determine the favorite after-school activity of the students in her class by surveying a sample of the entire class.

A **biased** sample would be a survey of the members of the computer club. These members have a common interest in computers so surveying them would probably reflect a lot of computer-related activities.

An **unbiased** sample would be a survey of every fifth person listed on the class roster in alphabetical order based on his/her last name.

Let's take a closer look at the types of biased and unbiased samples that are considered for surveys.

Types of biased samples:

convenience sample – A convenience sample is a sample that includes members of a population that are easily accessed.

voluntary response sample - A voluntary response sample is a sample that involves only those people that want to participate in the sampling.

Types of unbiased samples:

simple random sample - A simple random sample is a sample where each item or person in the population is as likely to be chosen as any other.

stratified random sample – A stratified random sample is a sample in which the population is divided into similar, non-overlapping groups.

systematic random sample – A systematic random sample is a sample in which the items or people are selected according to a specific time or item interval.

Examples: Identify the type of sample described.

1. A person employed by the local mall solicits shoppers to fill in a survey about new products by offering them a lottery ticket if they take the time to complete the survey.

This is a **voluntary response sample** (biased) because the participants are choosing to take the survey (most likely because they want the chance to win a prize with the lottery ticket).

2. Every person whose telephone number ends with a 48 is contacted to find out which presidential candidate he or she favors for the next election.

This is a **systematic random sample (unbiased)** because each person surveyed is selected from a list of most of the persons in the community and based upon the condition that his/her telephone number ends with a 48 (item interval).

3. The parents, grandparents, relatives, and friends who attend the school Christmas concert are surveyed and asked if they will support the next operating school tax levy.

This is a **convenience sample (biased)** because the people surveyed were the ones that were easily accessible because they attended the school's Christmas concert.

4. Persons, ages 30 to 39, are surveyed to see which car model they prefer to purchase.

This is a **stratified random sample (unbiased)** because a select age group, ages 30 to 39, were surveyed to see which type of car they prefer. Selections from other age groups may vary considerably but if the target group for sales is this age group, then the survey is unbiased.