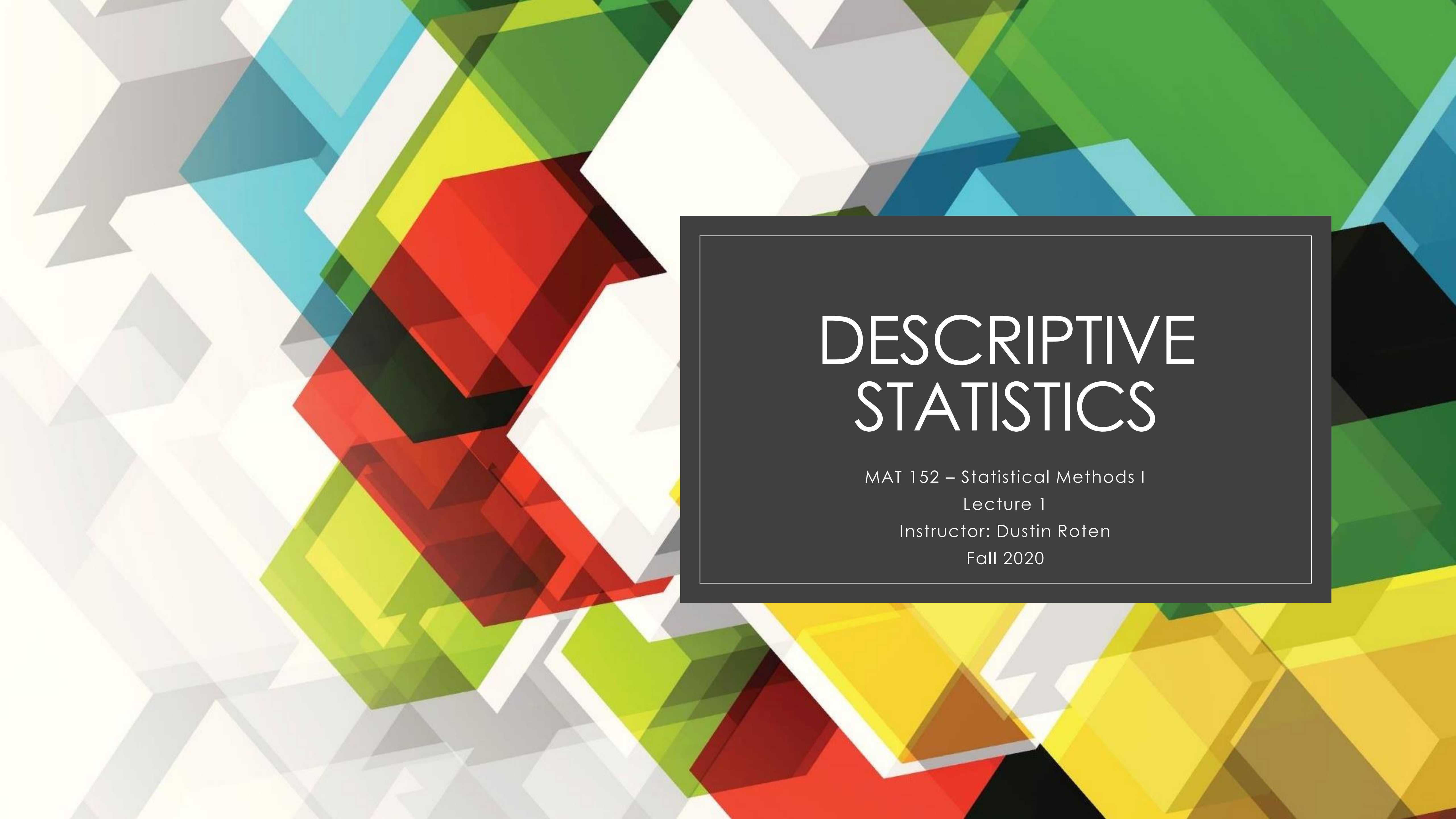


DESCRIPTIVE STATISTICS

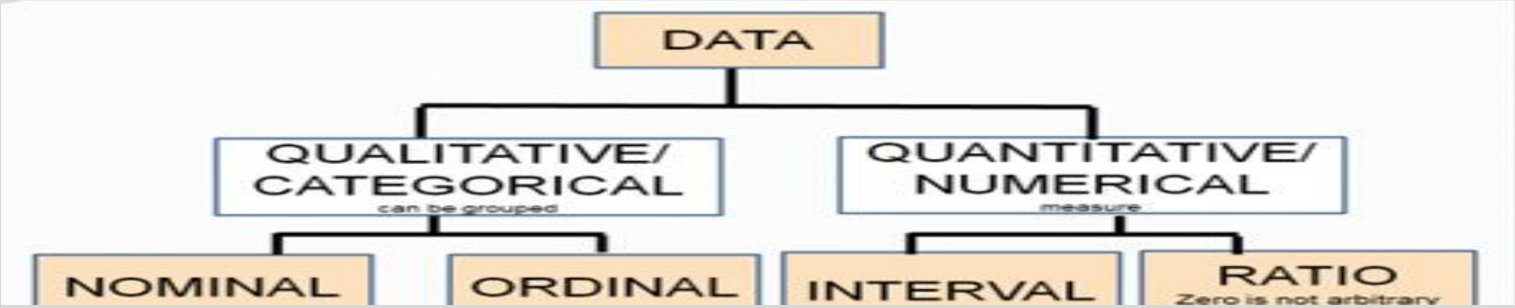
MAT 152 – Statistical Methods I

Lecture 1

Instructor: Dustin Roten

Fall 2020

A few more terms about data…



Categories, colors, type, etc.

Make/Model of car

CANNOT be ordered

Categories strength, satisfaction

Your satisfaction with service "X"?

CAN be ordered

Values with an arbitrary range

Years, Temperature

Calculations can be performed

Values that have a 0 point (not usually negative)

Height, weight, exam scores

Calculations can be performed

Suppose a researcher wishes to gain insight into a certain **parameter** of a **population**. The researcher uses a **random sampling** method to construct a **sample** of the **population**. **Statistics** about the **sample** can then be determined.

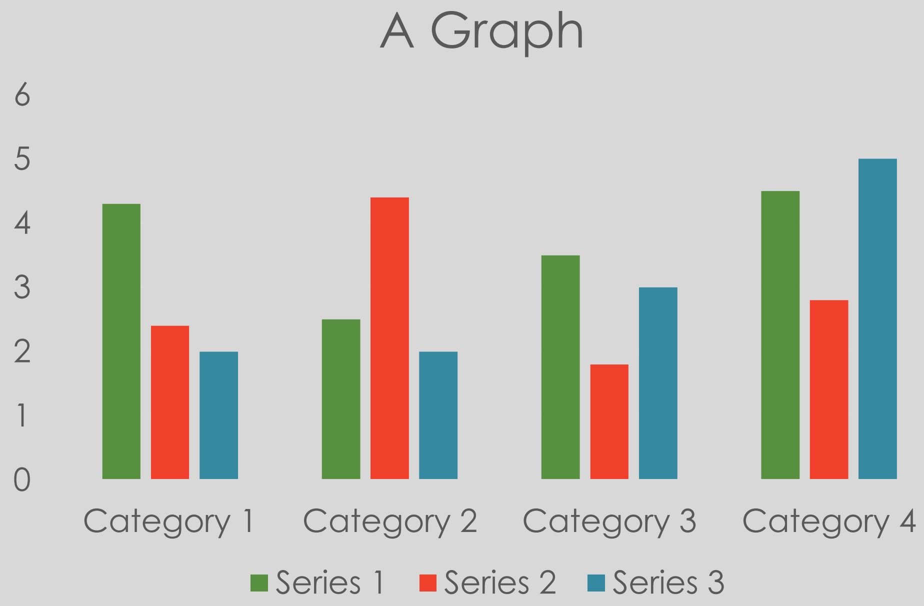
So, what can be done with these data?

We can **visualize** these data with **descriptive statistics**.

Descriptive Statistics

- **Descriptive Statistics** deals with numerical and graphical ways of describing data.

Data can be displayed in graphs and/or tables



A Table

Category	Total
Math Students	15
Biology Students	7
English Students	18
History Students	23

A First Look at Your Data: Stemplots

For small datasets, a **stem-and-leaf graph** (or **stemplot**) can reveal overall patterns and **outliers**.

An **outlier** is an observation of data that does not fit the rest of the data.

(May require further investigation)

To create a stemplot, divide each observation (value) into a stem and a leaf.

The leaf consists of the **final significant digit**

A quick example:

Consider the following scores from a final exam in a statistics course (ratio scale).

33, 42, 55, 61, 69, 72, 81, 88, 88, 89, 94, 97, 98 100

A First Look at Your Data: Stemplots

Scores: 33, 42, 55, 61, 69, 72, 81, 88, 88, 89, 94, 97, 98 100

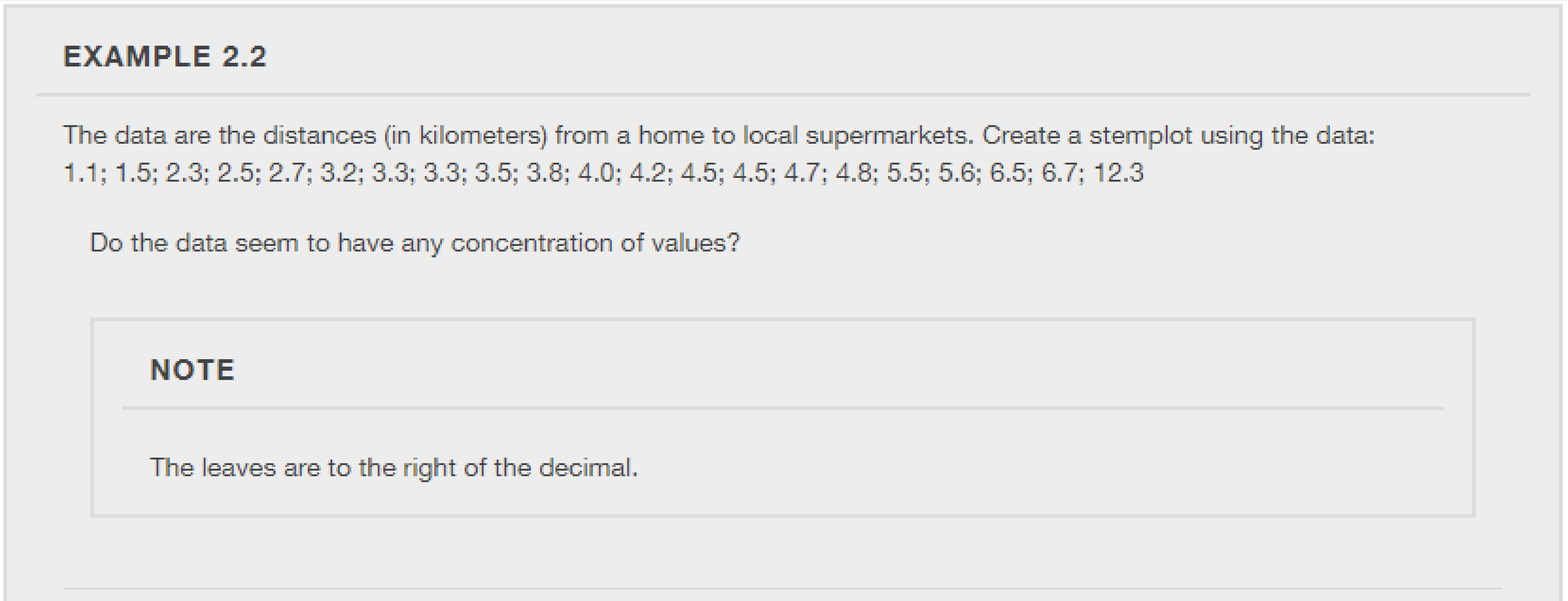
The first number is considered the **stem** (in blue above).
The last number is considered the **leaf** (in black above).

Notice that the value of 100 is broken down to 10 – 0

Stemplots are a quick way to look for any overall patterns.
Here, it's clear that most students scored in the 80's and 90's.

Stem	Leaf
3	3
4	2
5	5
6	1 9
7	2
8	1 8 8 9
9	4 7 8
10	0

Example (From textbook)



EXAMPLE 2.2

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:
1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Do the data seem to have any concentration of values?

NOTE

The leaves are to the right of the decimal.

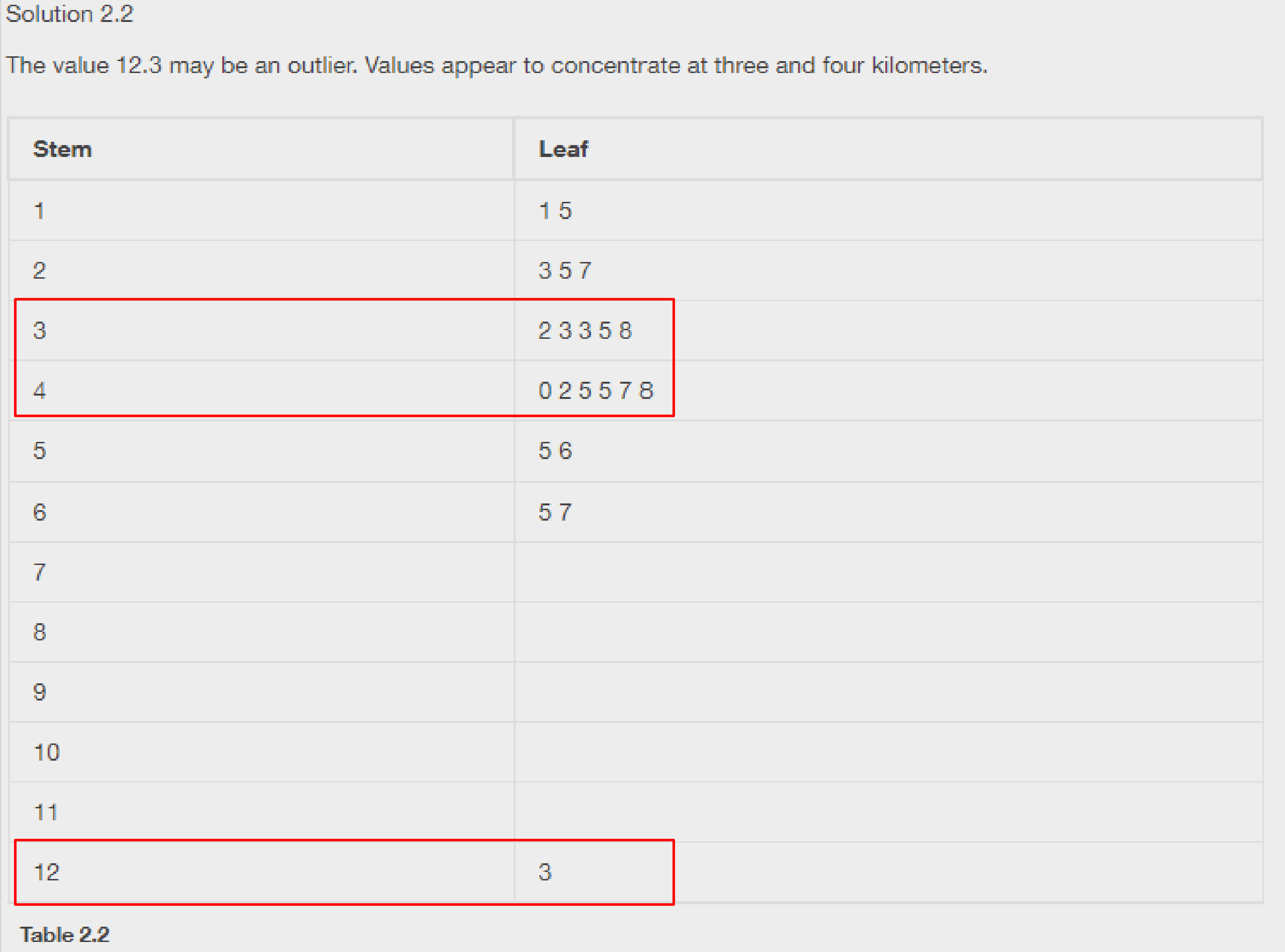
Example

- Stems are to the left of the decimal
- Leaves are the **final significant digit** (to the right of the decimal).

Overall Observations:

1.) Values tend to concentrate around 3 and 4 km.

2.) 12.3 MAY be an outlier since this value is much higher than the other values.



Solution 2.2

The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

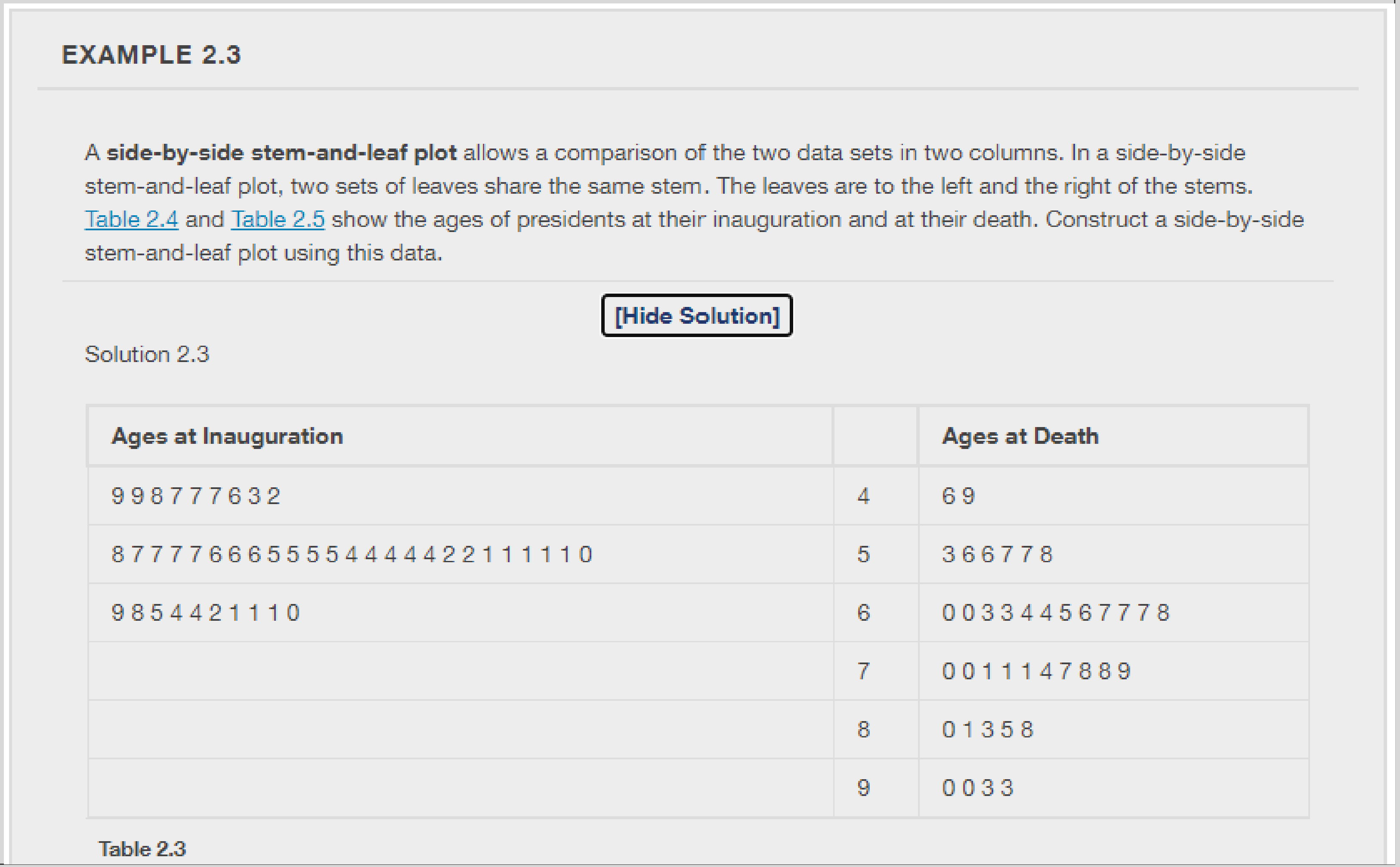
Table 2.2

Stemplots

Side-by-side stemplots can be used to compare two different sets of data.

Shifts in the data can be examined before and after an event or experiment.

Example: Pre-tests and Post-tests



EXAMPLE 2.3

A **side-by-side stem-and-leaf plot** allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. Table 2.4 and Table 2.5 show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Solution 2.3

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 8 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

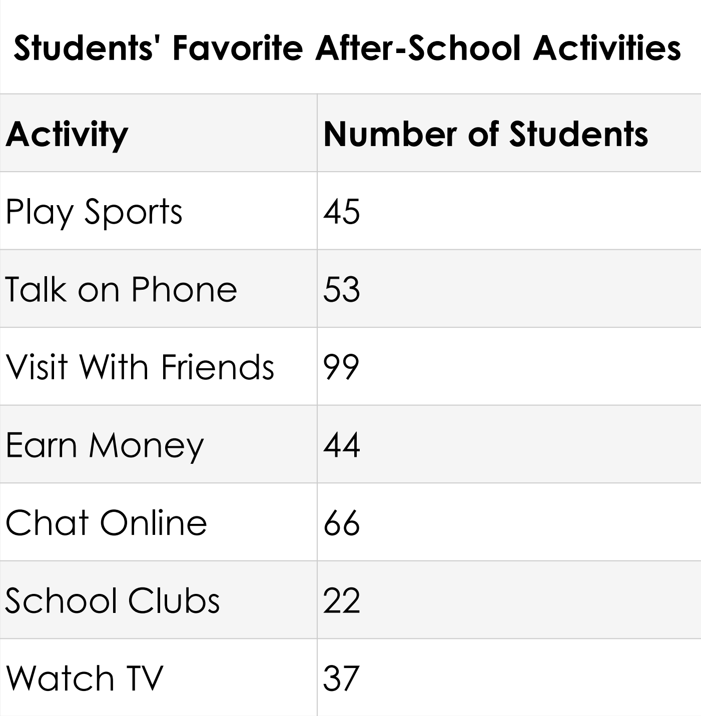
Table 2.3

Other Ways to Display Data: Bar Graphs

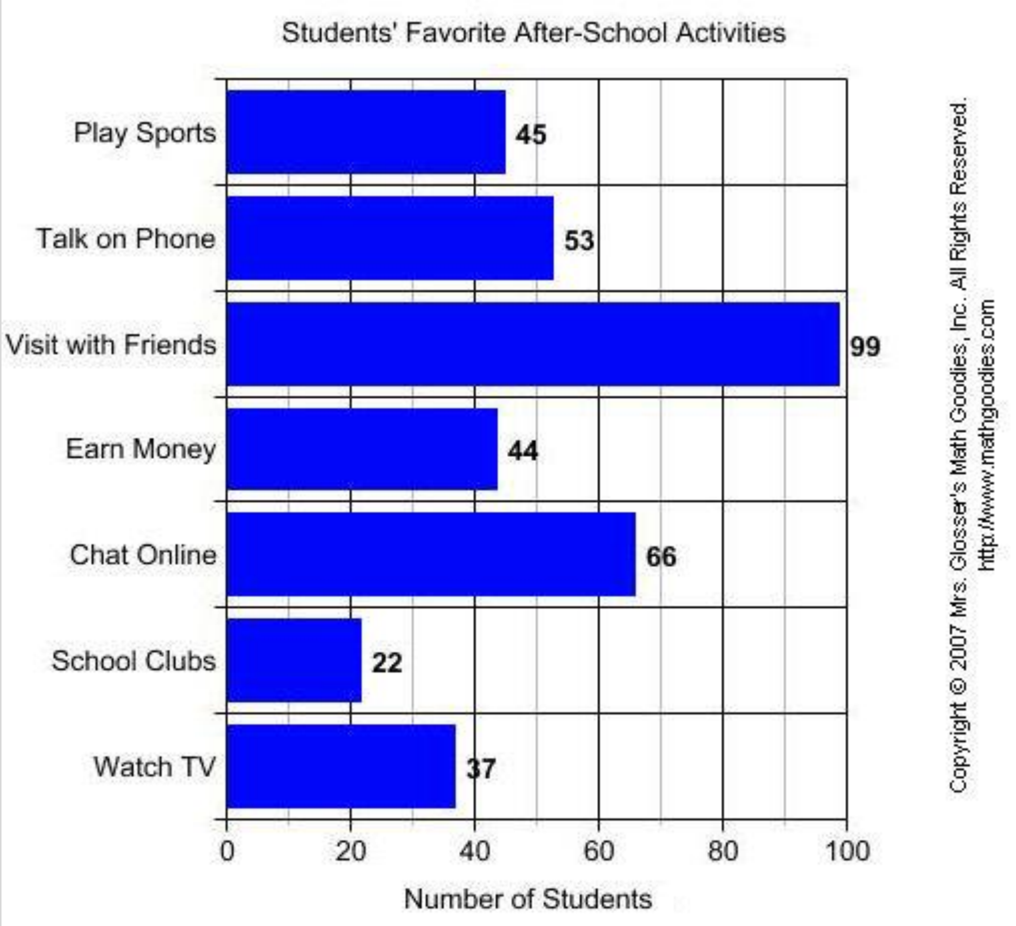
Bar graphs consist of bars that are separated from each other.

Great for nominal and ordinal data

Nominal Data

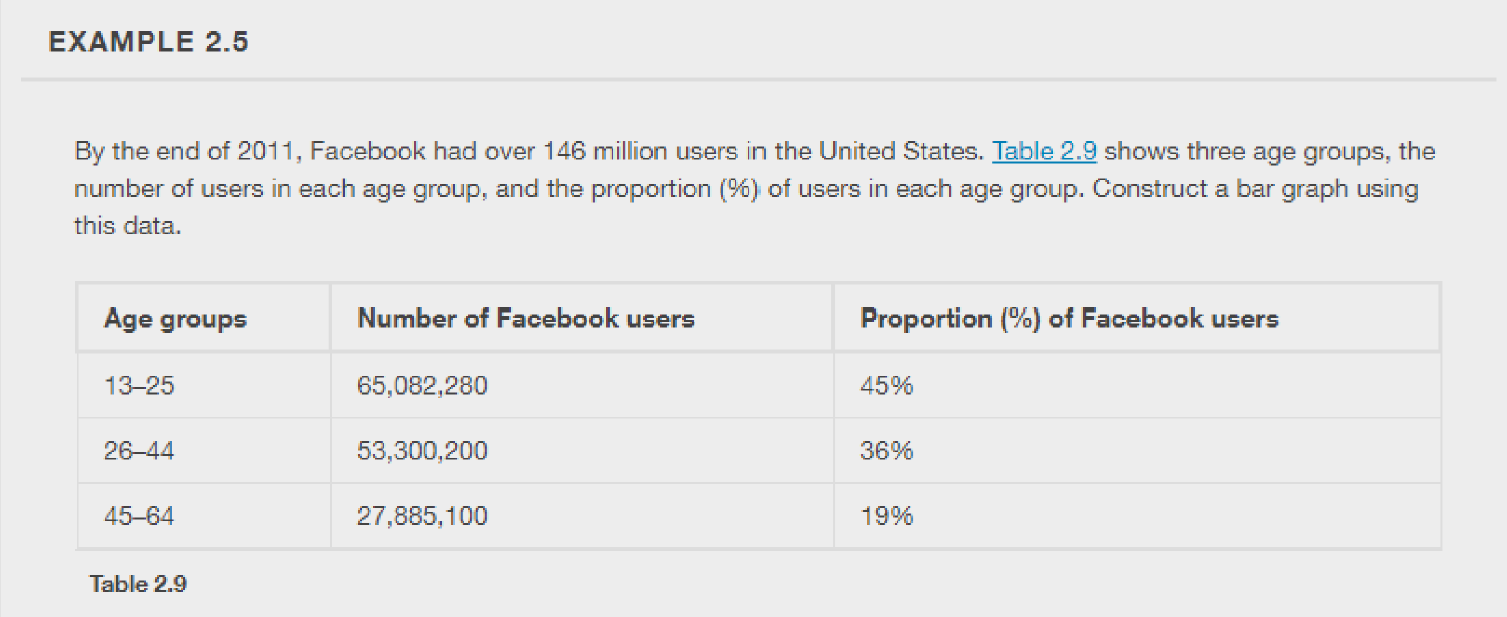


Students' Favorite After-School Activities	
Activity	**Number of Students**
Play Sports	45
Talk on Phone	53
Visit With Friends	99
Earn Money	44
Chat Online	66
School Clubs	22
Watch TV	37



Other Ways to Display Data: Bar Graphs

The **bins** of a bar plot can also represent **ranges** in quantitative data

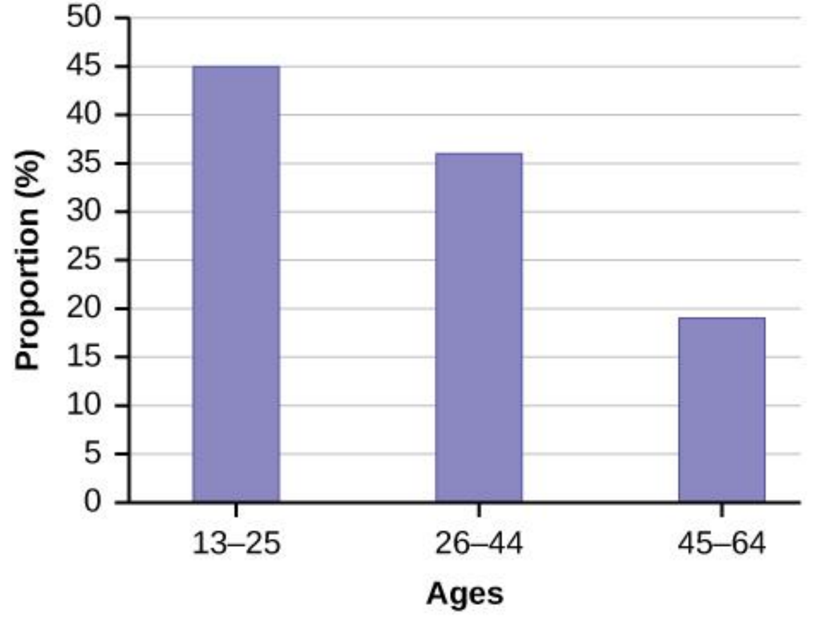


EXAMPLE 2.5

By the end of 2011, Facebook had over 146 million users in the United States. Table 2.9 shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Table 2.9



A First Look at Your Data: Frequency

- Understanding the frequency of values in your dataset provides insight into its distribution.
- **Frequency** is the number of times a value of the data occurs. ($f \equiv \text{frequency}$)
- **Relative Frequency** is the ratio of the number of times a value occurs to the total number of outcomes.
  - $n \equiv \text{total number of outcomes (number of data values)}$
  - $RF = \frac{f}{n}$
- **Cumulative Relative Frequency** is the accumulation of the previous relative frequencies.

A brief example:

Suppose a coin was tossed 10 times with the following outcomes:

Heads, Heads, Heads, Tails, Heads, Heads, Tails, Tails, Heads, Heads

Data Value	Frequency	Relative Frequency	Cumulative Relative Frequency
Heads	7	$\frac{7}{10} = 0.7$ or 70%	0.7
Tails	3	$\frac{3}{10} = 0.3$ or 30%	0.7 + 0.3 = 1.0

Example

Twenty students were asked how many hours they worked per day. Their responses are as follows:

5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3

Data Value	Frequency	Relative Frequency	Cumulative Relative Frequency
2	3	$\frac{3}{20} = 0.15$ or 15%	0.15
3	5	$\frac{5}{20} = 0.25$ or 25%	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20} = 0.15$ or 15%	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20} = 0.30$ or 30%	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20} = 0.10$ or 10%	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20} = 0.05$ or 5%	0.95 + 0.05 = 1.00
	Total = 20	Total = 1 or 100%	

Example (cont.)

What fraction of students surveyed work 4 OR 6 hours?

$\frac{3}{20} + \frac{2}{20} = \frac{5}{20} = \frac{1}{4} = 0.25$ or 25%

What fraction of students surveyed work at most 5 hours?

The cumulative relative frequency has the summed value for all previous data, so $0.85 = \frac{17}{20}$.

What fraction of students surveyed work from 3 to 5 hours, inclusive?

$\frac{5}{20} + \frac{3}{20} + \frac{6}{20} = \frac{14}{20} = 0.70$ or 70%

Data Value	Frequency	Relative Frequency	Cumulative Relative Frequency
2	3	$\frac{3}{20} = 0.15$ or 15%	0.15
3	5	$\frac{5}{20} = 0.25$ or 25%	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20} = 0.15$ or 15%	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20} = 0.30$ or 30%	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20} = 0.10$ or 10%	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20} = 0.05$ or 5%	0.95 + 0.05 = 1.00
	Total = 20	Total = 1 or 100%	

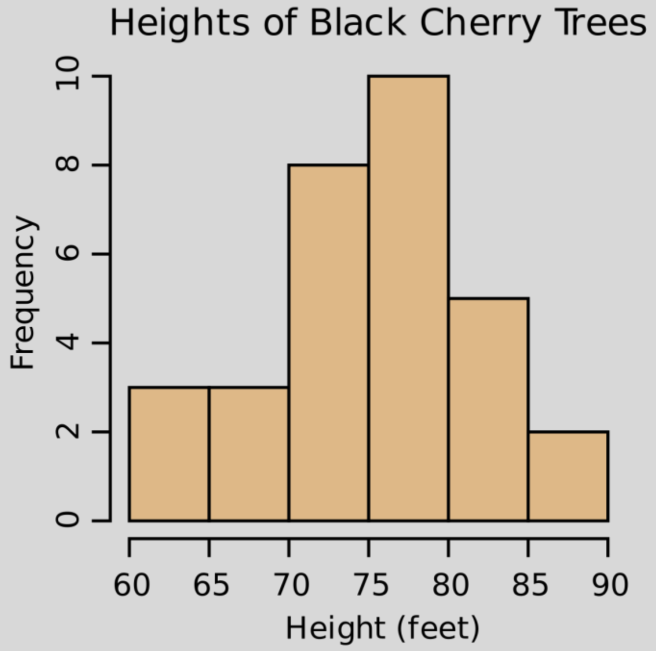
Histograms

- A **histogram** consists of contiguous boxes and has both horizontal and vertical axes.
- The **horizontal axis** is labeled with the data, the **vertical axis** contains the **frequency** or **relative frequency**.

To construct a histogram:

1.) The number of **bars** (or **intervals**) must be decided.

2.) Intervals should be picked such that data do not fall on boundaries



Picking Interval Sizes

Consider the following data:

1.1, 2.3, 2.8, 2.8, 4.2, 6.1, 6.8, 8.4, 9.8, 10.4

1.) Decide how many **bars** or **Intervals** represent the data. (usually 5 to 15 bars)

For a simple example, we will use 3.

2.) Choose a starting point for the first interval. It should be less than the smallest data value.

A lower value carried out to one more decimal place than the value with the most decimal places.

When the starting point is carried out to one additional decimal place, no data will fall on a boundary.

Example from the data:

1.1 is the smallest value and all values have the same number of decimal places.

Start at: $1.1 - 0.05 = 1.05$

End at: $10.4 + 0.05 = 10.45$

Picking Interval Sizes (cont.)

1.1, 2.3, 2.8, 2.8, 4.2, 6.1, 6.8, 8.4, 9.8, 10.4

Start at: $1.1 - 0.05 = 1.05$

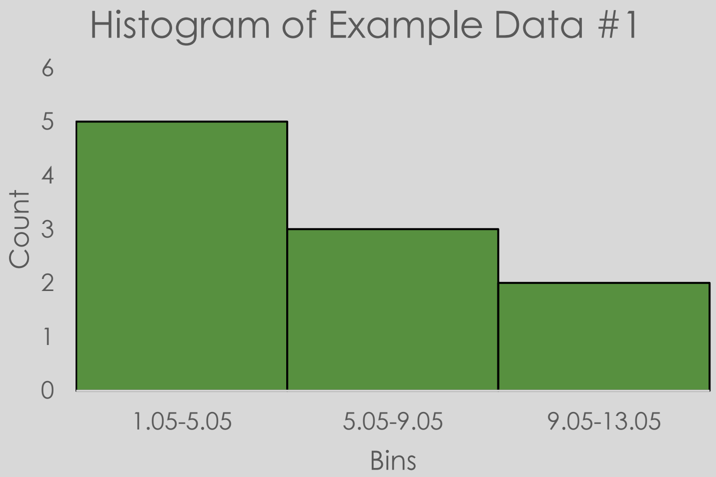
End at: $10.4 + 0.05 = 10.45$

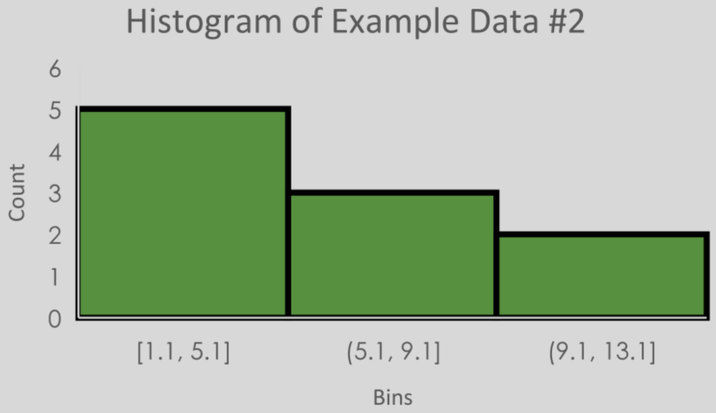
Find the width of each bar:

$$\frac{\text{End} - \text{Start}}{\text{Bars}}$$

$$\frac{10.45 - 1.05}{3} = 3.1\overline{3}$$

Rounding is another way to prevent values from falling on a boundary. Bar width: 4.





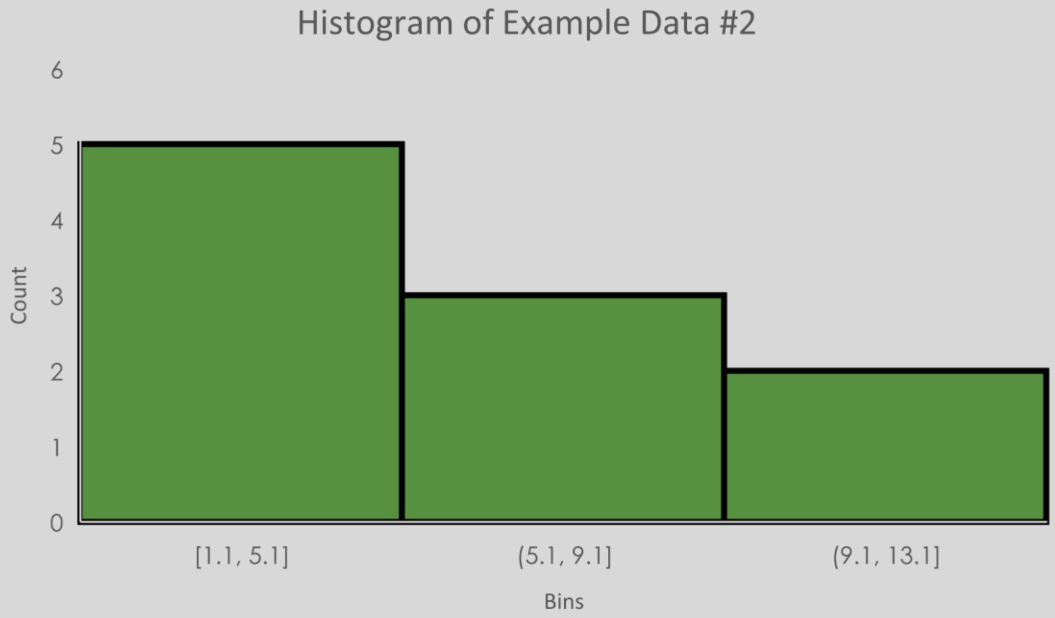
A Note on Notation

The graph presented here is a standard histogram generated by Microsoft Excel. The bin widths are labeled with a mix of parentheses and square brackets.

Bin #1 is labeled [1.1, 5.1]. This means that all values from 1.1 to 5.1 are included in this bin.

Bin #2 is labeled (5.1, 9.1]. This means that 5.1 is NOT included in Bin #2 but any number GREATER than 5.1 (up to 9.1) is included. The value 5.1 belongs in Bin #2

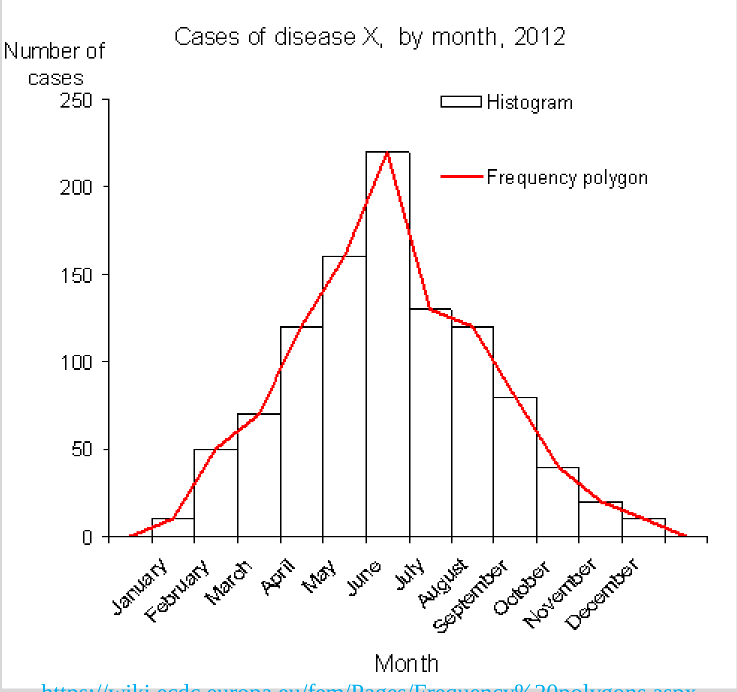
Bin #3 is labeled (9.1, 13.1]. Thus, a data value of 9.1 belongs in Bin #2 but 9.1000001 would belong in Bin #3



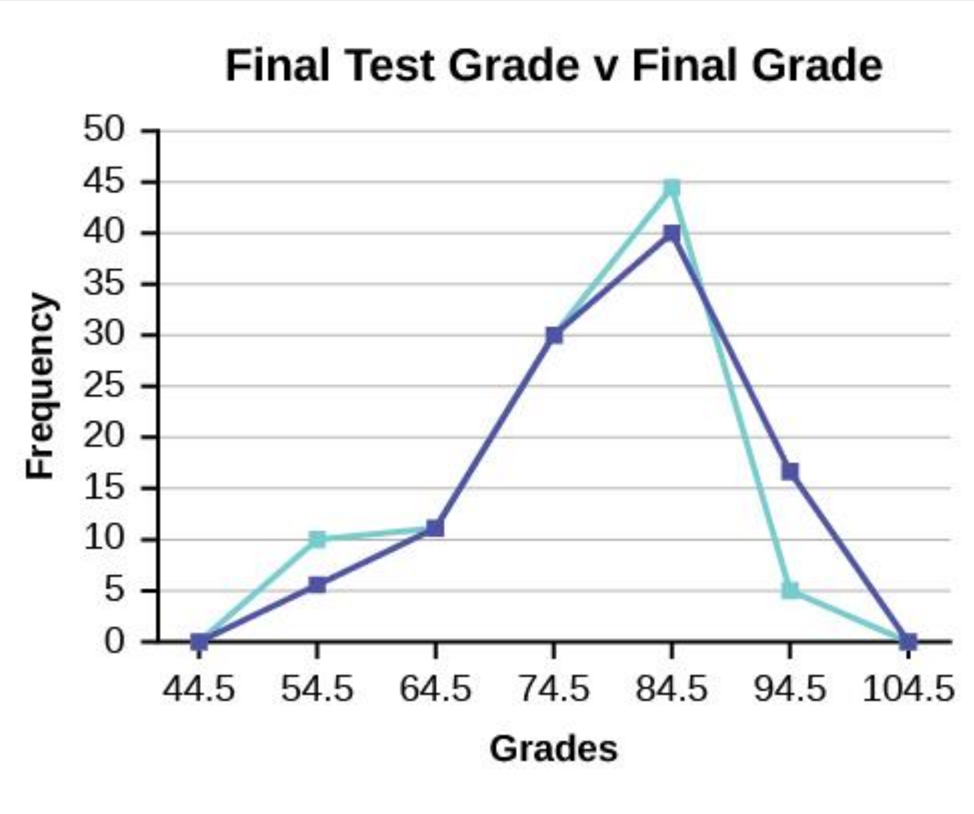
Frequency Polygons

Just like a histogram, intervals must be decided, points are placed on the center bin values, and the points are connected with straight lines.

Frequency polygons are useful for comparing distributions



https://wiki.ecdc.europa.eu/fem/Pages/Frequency%20polygons.aspx



Time Series

A graph that pairs a data value with a specific time is called a **time series graph**.

To construct a time series graph, we must look at both pieces of the **paired data set**. (The value and its associated timestamp)

On a standard plot, the horizontal axis is used to plot the time and date increments.

The vertical axis is used to plot the values of the variable.

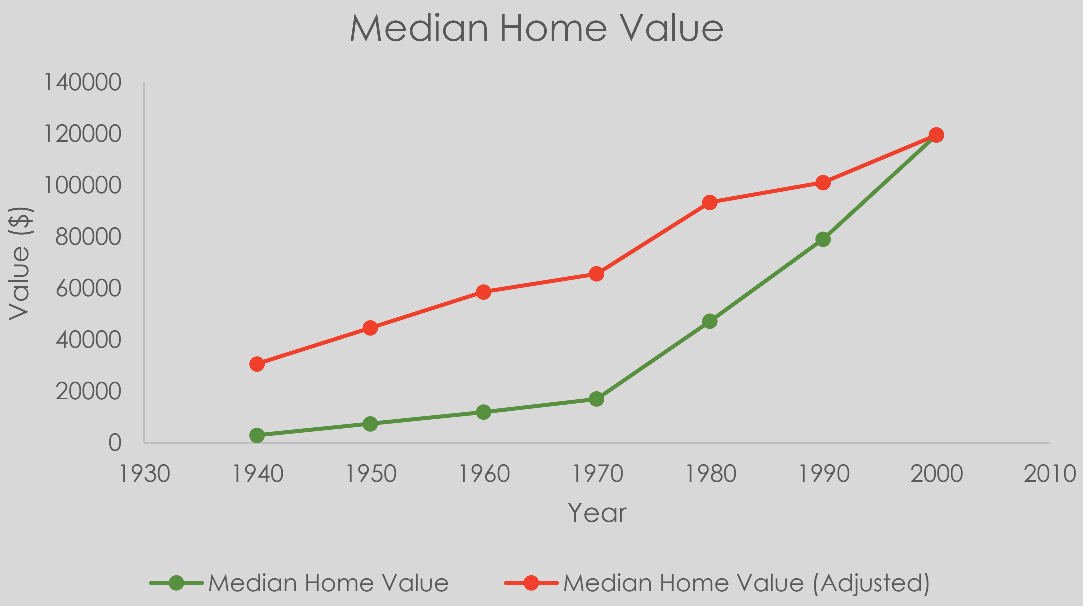
The points are connected by straight lines.

Time Series

Below is a time series of housing prices over the last 50 years

These data were obtained from a CNBC news article entitled “Here’s how much housing prices have skyrocketed over the last 50 years”.

These data are presented with and without adjustments for inflation.



Review

Four different types of data:

Qualitative

Nominal – characteristics like color, model, etc.

Ordinal – categories that can be ranked such as customer satisfaction, difficulty, etc.

Quantitative

Interval – numerical data without an “origin”; values can be positive AND negative. (Temperature)

Ratio – numerical data in which there is an “origin”; values can’t be negative. (Grades, height, etc.)

Ways to visualize data:

Stemplots, bar graphs, frequency tables, histograms, frequency polygons, time series