EXPLORING DATA

ST101 – DR. ARIC LABARR

EXPLORATION

- Why do we organize and explore our data?
- A lot of insights can be drawn from just organizing, exploring, and looking at your data.
- Different types of data need to be summarized differently.

TYPES OF VARIABLES

- There are two main types of variables:
  - Qualitative – data with a measurement scale inherently categorical.
  - Quantitative – data that are numeric and define a value or quantity.

EXPLORING DIFFERENT TYPES OF VARIABLES

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy	34.5	59.0	108	1454
1/5/2011	Wednesday	Winter	Clear / Partly Cloudy	36.8	43.7	82	1518

⋮

EXPLORING DIFFERENT TYPES OF VARIABLES

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy	34.5	59.0	108	1454
1/5/2011	Wednesday	Winter	Clear / Partly Cloudy	36.8	43.7	82	1518

Qualitative – explore within a category or across categories.

EXPLORING QUALITATIVE DATA

- Qualitative – explore within a category or across categories.
- Example questions:
  - What do Saturdays look like?
  - What do clear days look like in comparison to rainy days?
  - Is the winter different than the summer?

EXPLORING DIFFERENT TYPES OF VARIABLES

Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy	34.5	59.0	108	1454
1/5/2011	Wednesday	Winter	Clear / Partly Cloudy	36.8	43.7	82	1518

⋮

Quantitative – explore center, spread, and "look" of variables.

EXPLORING QUANTITATIVE DATA

- Quantitative – explore center, spread, and "look" of variables.
- Example questions:
  - What is the typical temperature in my data?
  - Is the number of users trending up or down?
  - What is the range of humidity values? Are they narrow or very spread out?

SUMMARY

- A lot of insights can be drawn from just organizing, exploring, and looking at your data.
- Different types of data need to be summarized differently.
- Qualitative – explore within a category or across categories.
- Quantitative – explore center, spread, and "look" of variables.

DISPLAYING QUALITATIVE DATA

EXPLORING DATA

EXPLORING QUALITATIVE DATA

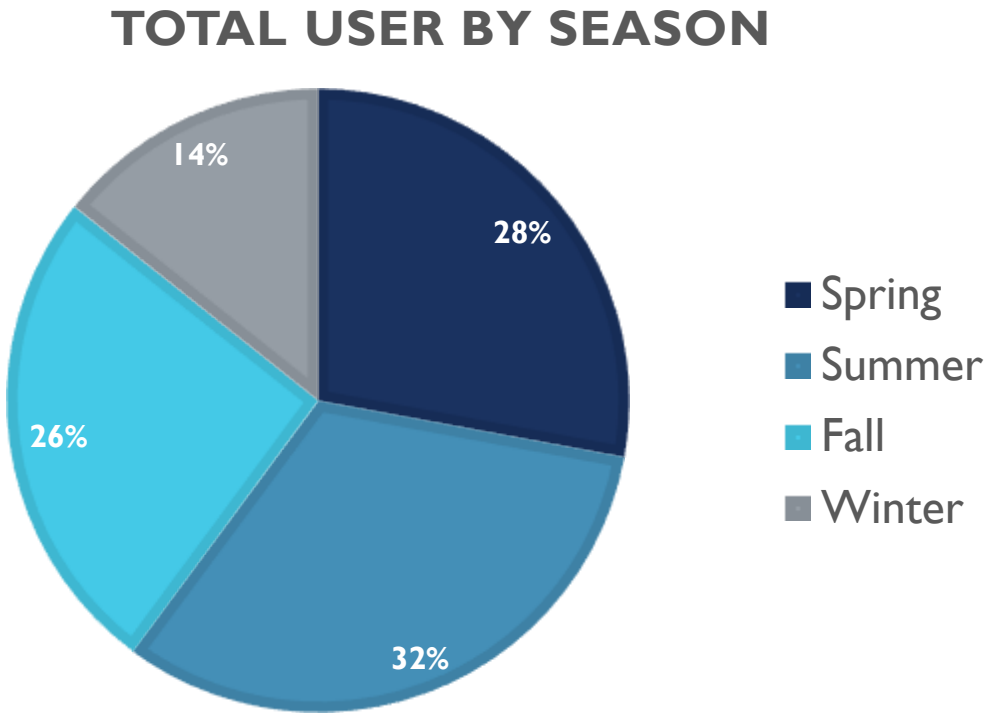
- Qualitative – explore within a category or across categories.
- Example questions:
  - What do Saturdays look like?
  - What do clear days look like in comparison to rainy days?
  - Is the winter different than the summer?

EXPLORING QUALITATIVE DATA

- Qualitative – explore within a category or across categories.
- Different graphs are used for different tasks:
  - Pie chart – comparison across all categories (distribution of categories).
  - Bar chart – comparison across specific categories
    - Regular
    - Side-by-side
    - Stacked

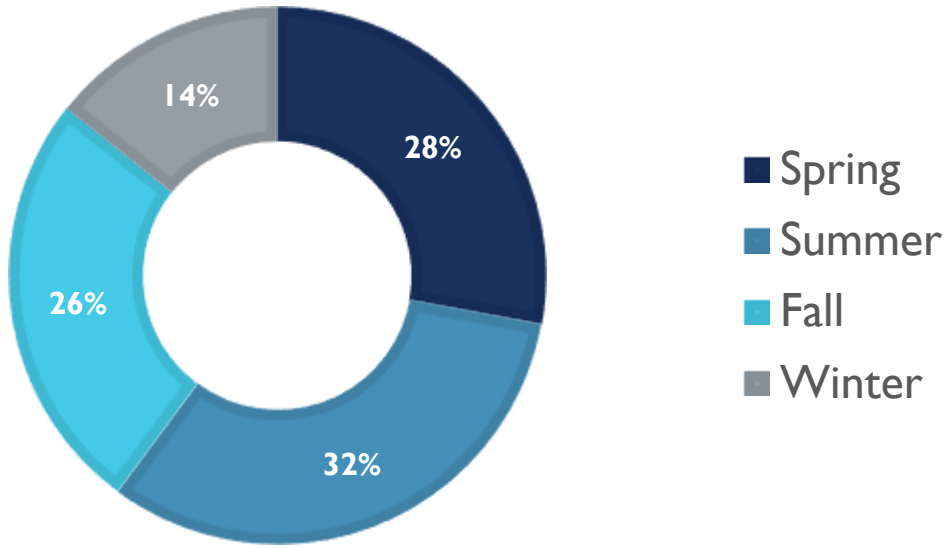
PIE CHART

- **Pie chart** – graph in which a circle is divided into sections that each represent a proportion of the whole.
- Best used when looking to show entire distribution (or set of categories) for a qualitative variable.



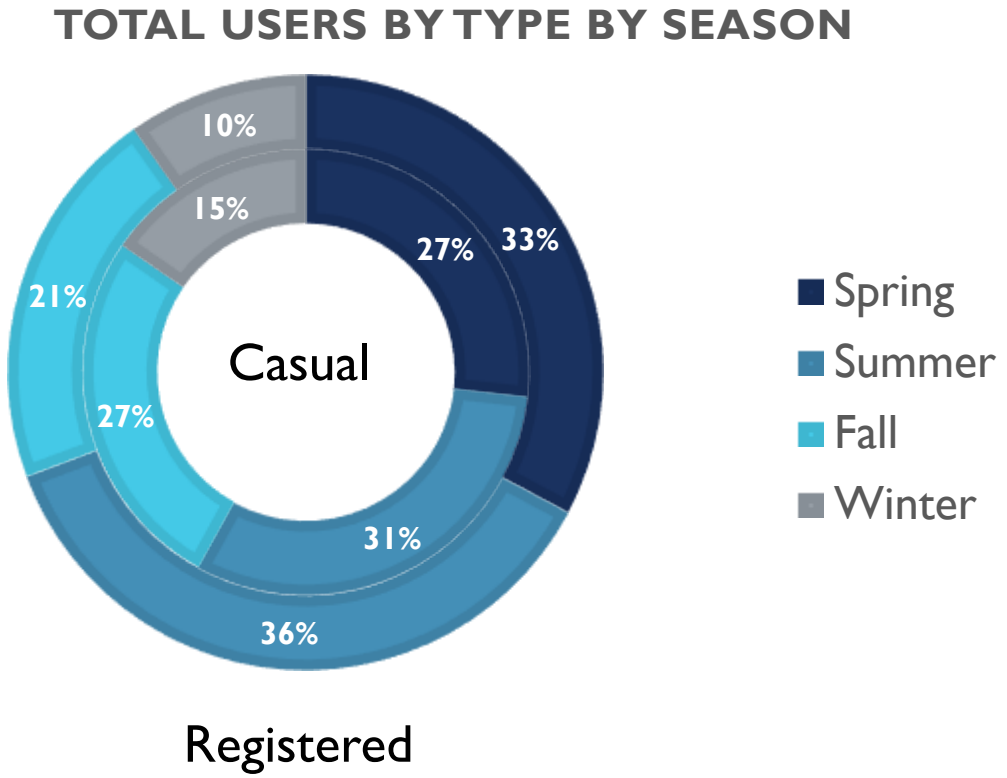
PIE CHART

- **Pie chart** – graph in which a circle is divided into sections that each represent a proportion of the whole.
- Best used when looking to show entire distribution (or set of categories) for a qualitative variable.
- **Donut chart** is basically the same thing with the center missing.



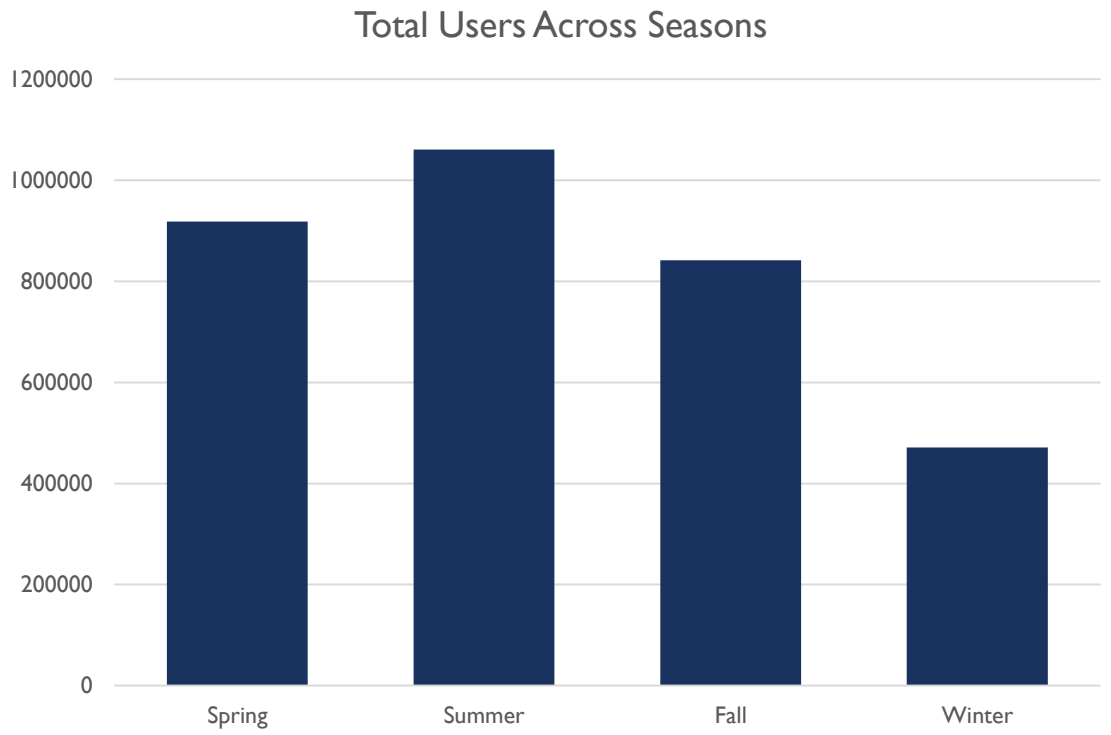
PIE CHART

- Donut charts allow us to compare different groups' distributions across all categories.



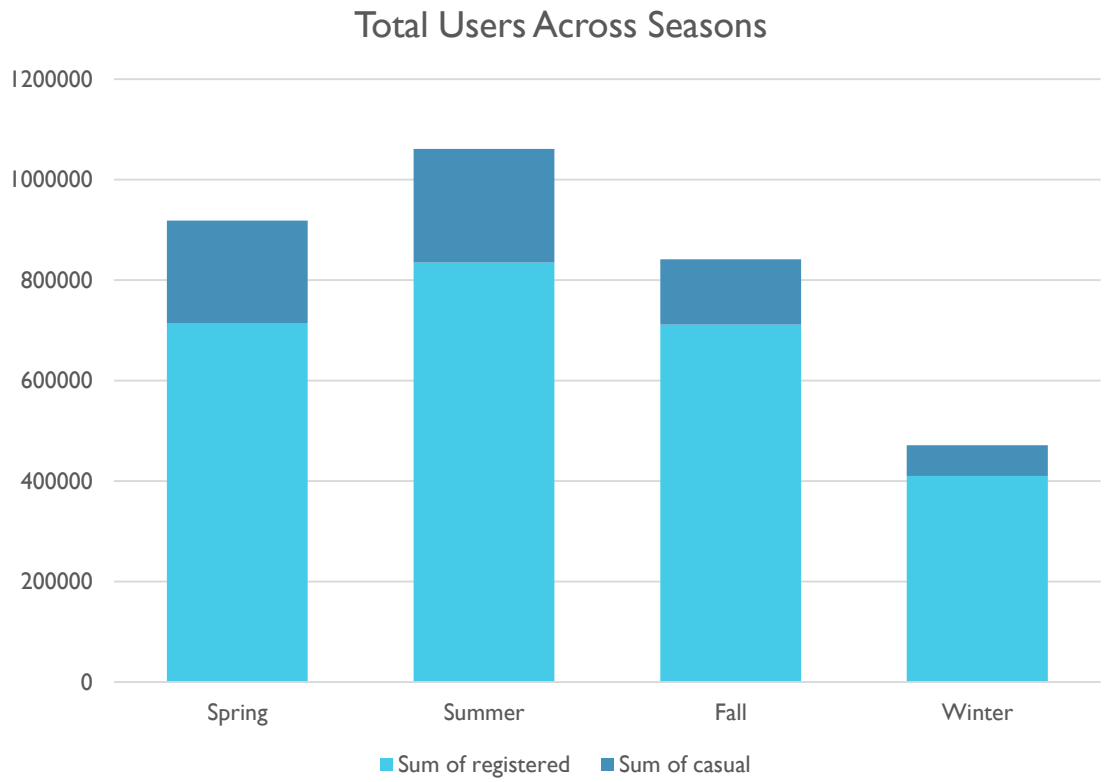
BAR CHART

- **Bar chart** – numerical values of variables are represented by the height or length of lines or rectangles of equal width.
- Best used when looking to compare specific categories to each other.



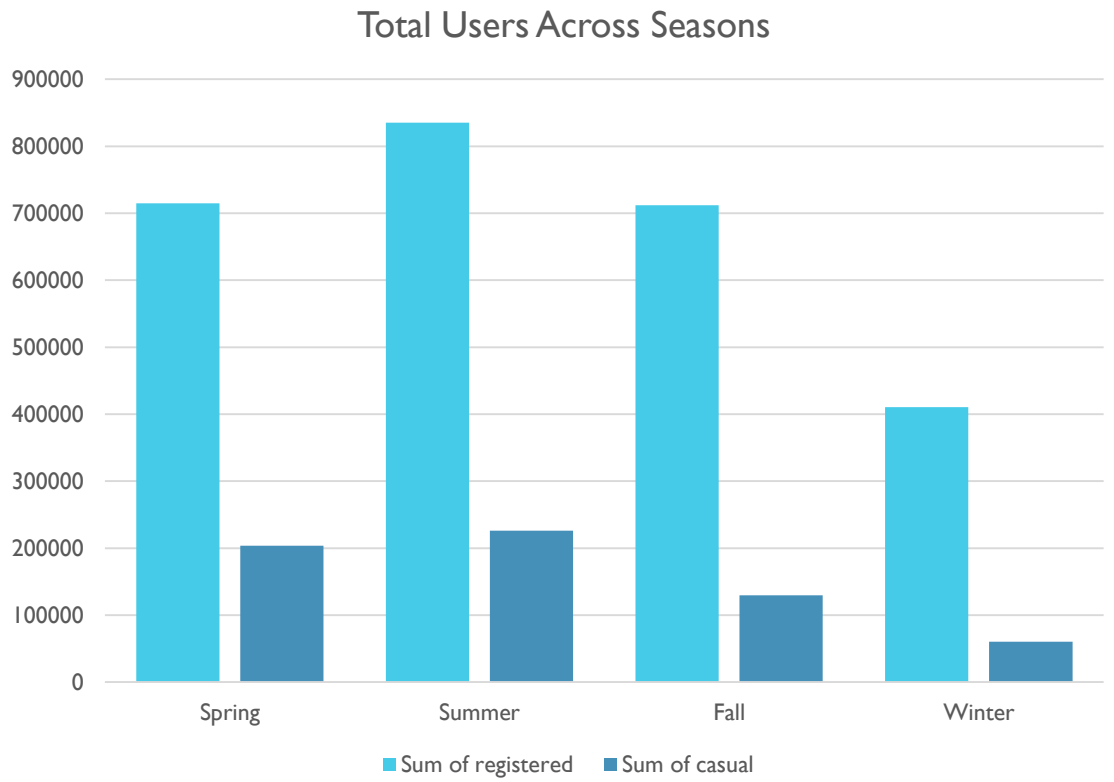
BAR CHART

- **Bar chart** – numerical values of variables are represented by the height or length of lines or rectangles of equal width.
- Best used when looking to compare specific categories to each other.
- **Stacked bar charts** break down the first categories into subcategories.



BAR CHART

- **Bar chart** – numerical values of variables are represented by the height or length of lines or rectangles of equal width.
- Best used when looking to compare specific categories to each other.
- **Side-by-side bar charts** look at these comparisons across multiple categories.



SUMMARY

- Qualitative – explore within a category or across categories.
- Pie chart – graph in which a circle is divided into sections that each represent a proportion of the whole.
- Bar chart – numerical values of variables are represented by the height or length of lines or rectangles of equal width.

DISPLAYING QUANTITATIVE DATA

EXPLORING DATA

EXPLORING QUANTITATIVE DATA

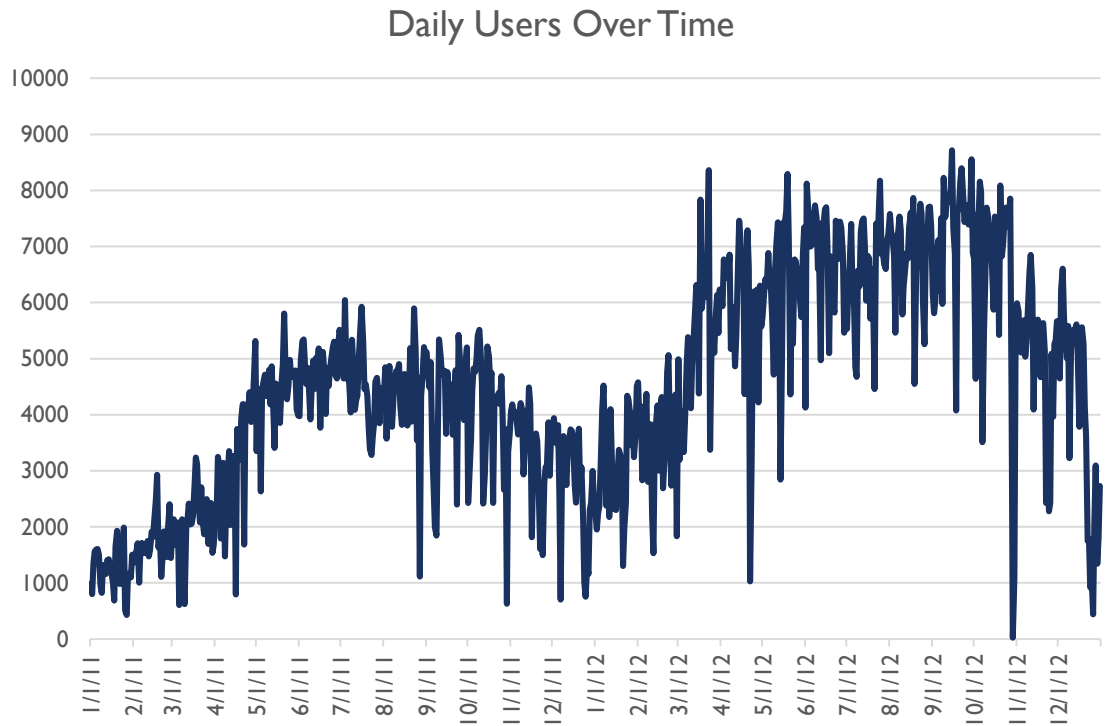
- Quantitative – explore center, spread, and "look" of variables.
- Example questions:
  - What is the typical temperature in my data?
  - Is the number of users trending up or down?
  - What is the range of humidity values? Are they narrow or very spread out?

EXPLORING QUANTITATIVE DATA

- Quantitative – explore center, spread, and "look" of variables.
- Different graphs are used for different tasks:
  - Line graph – look at how a variable changes over time.
  - Scatterplot – comparison between different quantitative variables.

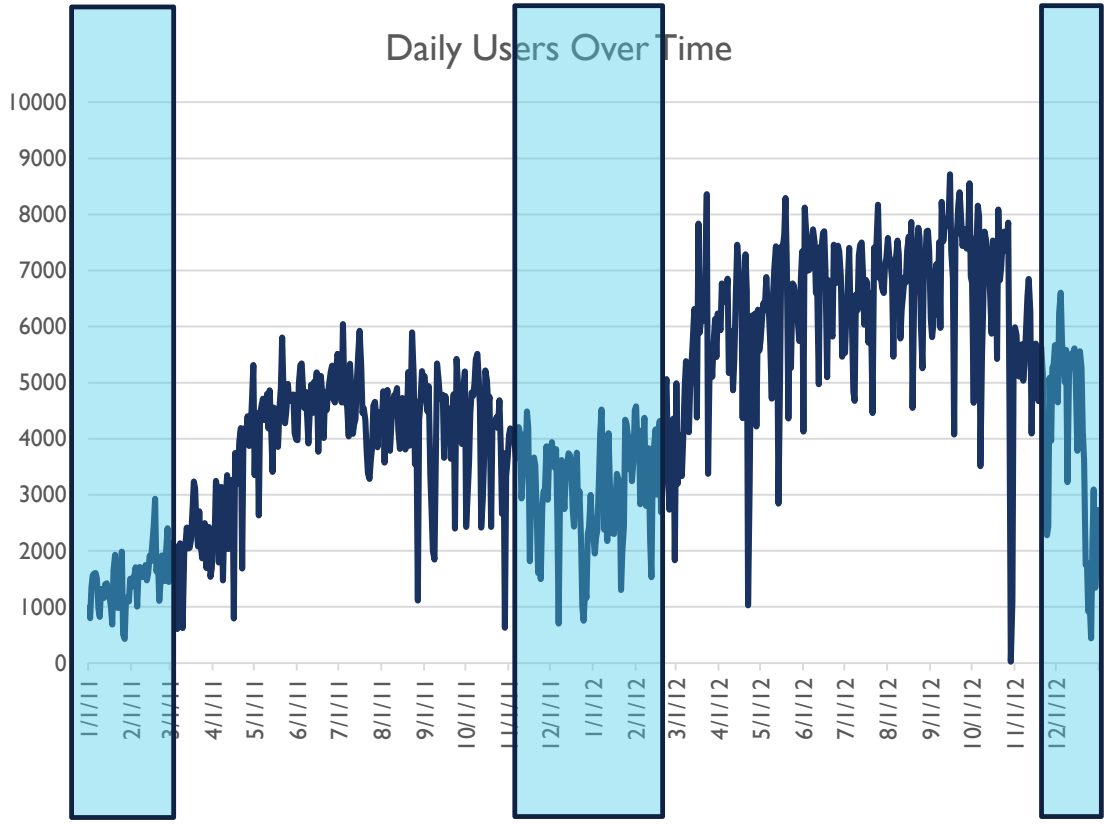
LINE GRAPH

- **Line graph** – uses lines to connect individual data points over time.
- Best when wanting to see how things change across time.



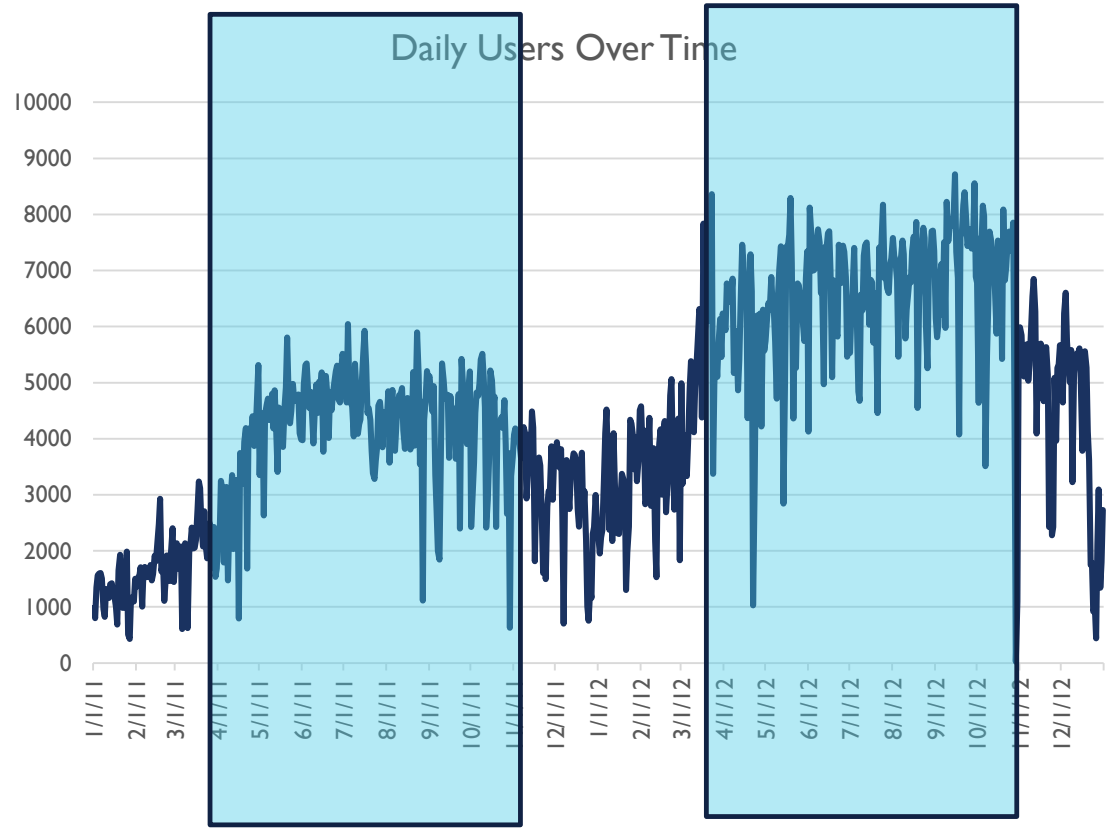
LINE GRAPH

- **Line graph** – uses lines to connect individual data points over time.
- Best when wanting to see how things change across time.



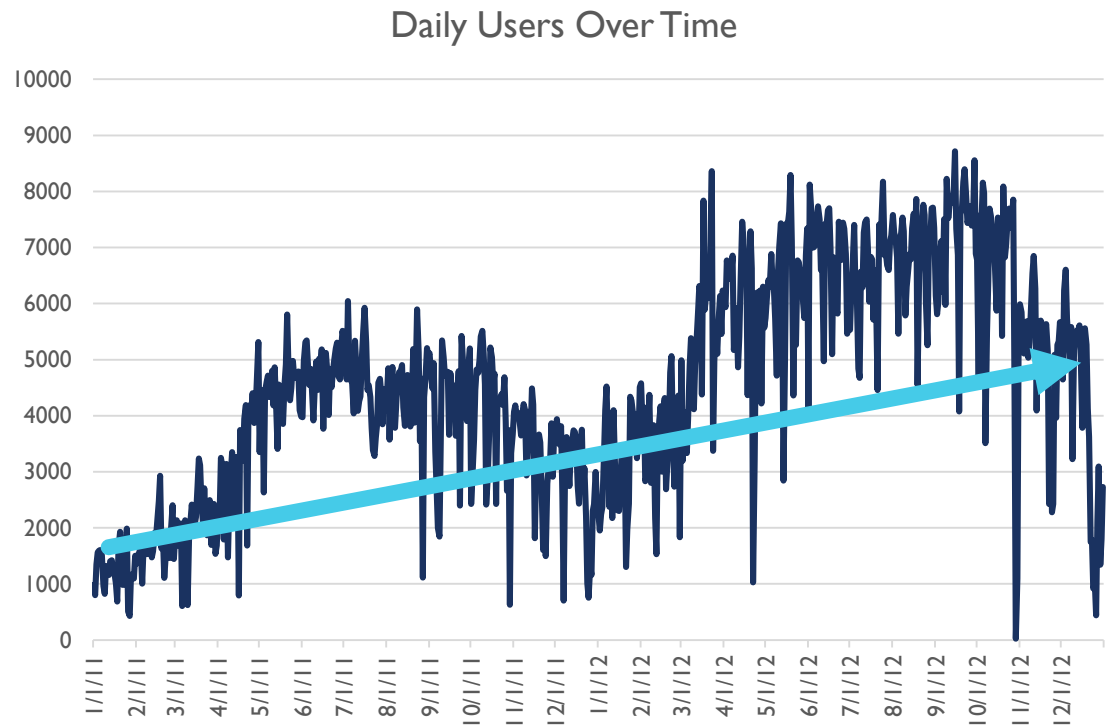
LINE GRAPH

- **Line graph** – uses lines to connect individual data points over time.
- Best when wanting to see how things change across time.



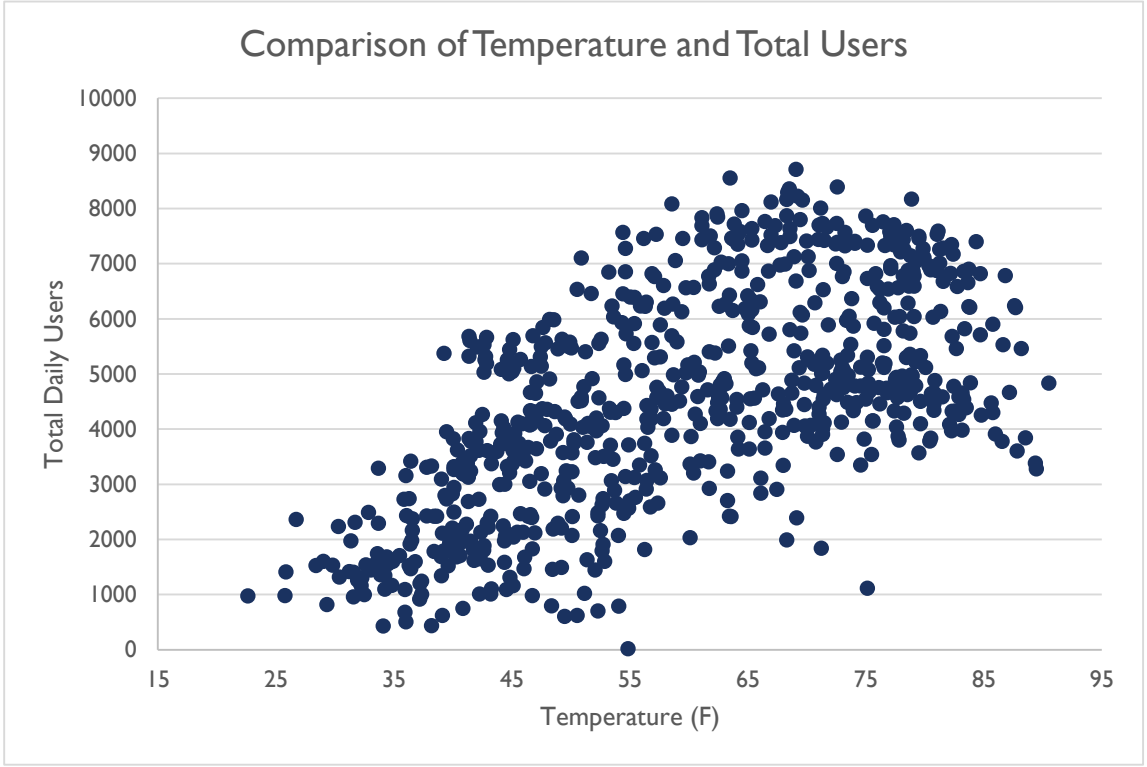
LINE GRAPH

- **Line graph** – uses lines to connect individual data points over time.
- Best when wanting to see how things change across time.



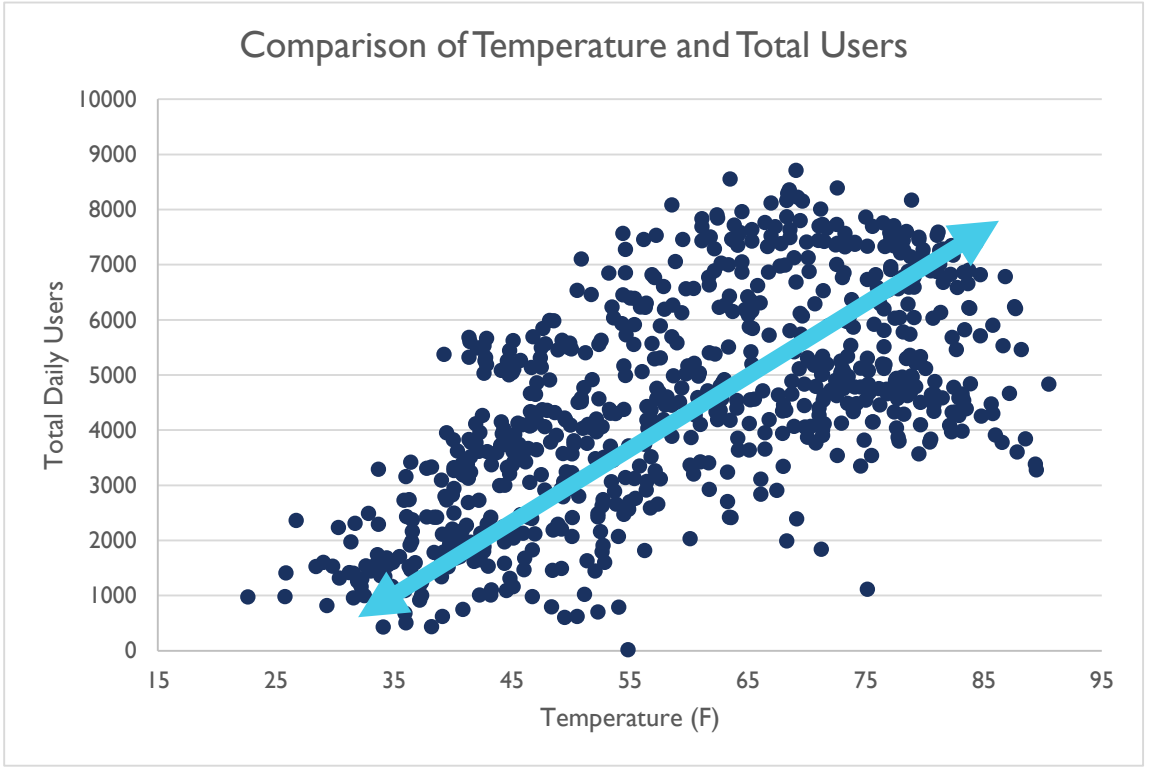
SCATTERPLOT

- **Scatterplot** – the values of two variables are plotted along two axes, the pattern of the resulting points revealing any relationship present.
- Best used when trying to explore relationships between two quantitative variables.



SCATTERPLOT

- **Scatterplot** – the values of two variables are plotted along two axes, the pattern of the resulting points revealing any relationship present.
- Best used when trying to explore relationships between two quantitative variables.



SUMMARY

- Quantitative – explore center, spread, and "look" of variables.
- Line graph – uses lines to connect individual data points over time.
- Scatterplot – the values of two variables are plotted along two axes, the pattern of the resulting points revealing any relationship present.

DESCRIBING CENTER

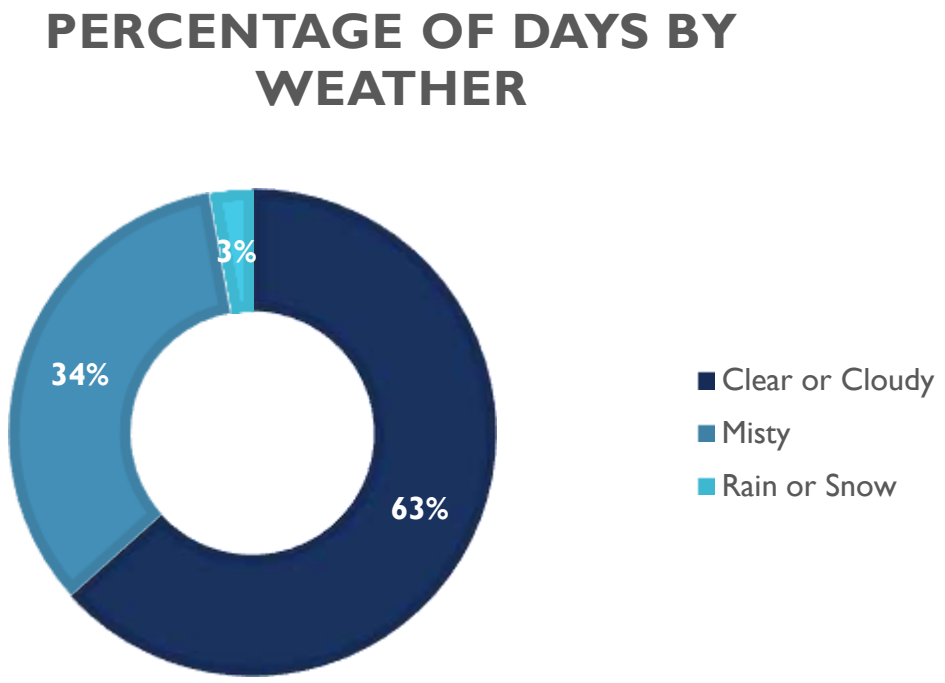
EXPLORING DATA

"TYPICAL" VALUE

- When exploring data, a good summary of a variable might be what a "typical" value of that variable would look like.
- What is "typical" really mean?
  - Qualitative variable – most common category.
  - Quantitative variable – focus on the **center** of the values of the variable.

MODE

- **Mode** – the mode of a variable is the most common value.
  - Typically reported with qualitative variables more than quantitative variables.
- In our data, the "typical" weather day (according to mode) is clear or cloudy.



MEAN

- **Mean** – the mean of a variable is the sum of all the values divided by the number of values.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Number of the observations in the sample

MEAN

- **Mean** – the mean of a variable is the sum of all the values divided by the number of values.

Sum of the values of the *n* observations

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

MEAN

- **Mean** – the mean of a variable is the sum of all the values divided by the number of values.

$$\bar{x} = \frac{46.7 + 48.4 + 34.2 + \cdots}{731}$$

MEAN

- **Mean** – the mean of a variable is the sum of all the values divided by the number of values.

$$59.51 = \frac{46.7 + 48.4 + 34.2 + \cdots}{731}$$

- In our data, the "typical" weather day (according to mean) is 59.51°F.

MEDIAN

- **Median** – value in the middle when the data items are arranged in ascending order.

- For an **odd number** of observations:

Original Data

26	30	27	22	24	29	13

MEDIAN

- **Median** – value in the middle when the data items are arranged in ascending order.

- For an **odd number** of observations:

Original Data

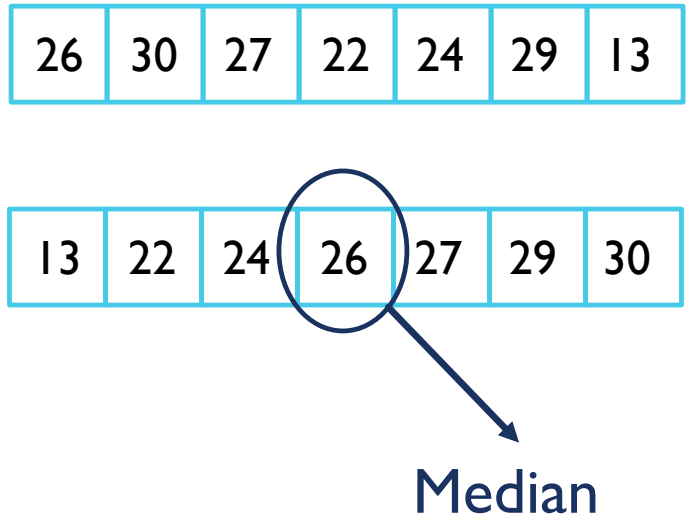
26	30	27	22	24	29	13

13	22	24	26	27	29	30

Ascending Order

MEDIAN

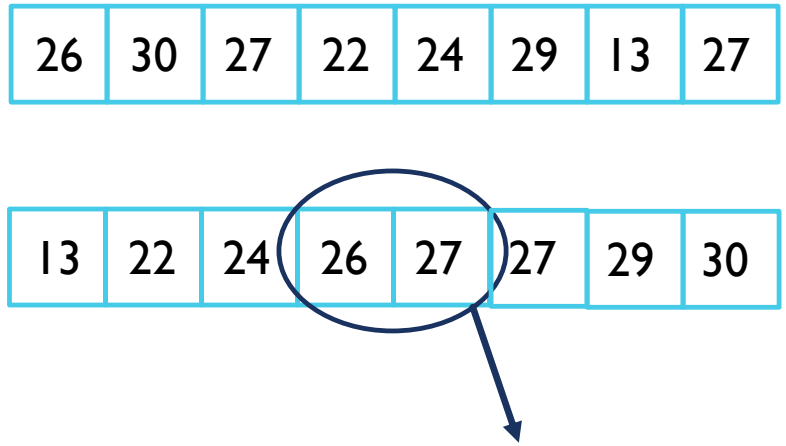
- **Median** – value in the middle when the data items are arranged in ascending order.
- For an **odd number** of observations:



MEDIAN

- **Median** – value in the middle when the data items are arranged in ascending order.

- For an **even number** of observations:



$$\text{Median} = \frac{26+27}{2} = 26.5$$

MEDIAN

- **Median** – value in the middle when the data items are arranged in ascending order.

- In our data, the "typical" weather day (according to median) is 59.76°F.

MEAN VS. MEDIAN

- Whenever a data set has extreme values, the median is the preferred measure of center.
- Mean is bothered by extreme values, while median is not.

26	30	27	22	24	29	13

13	22	24	26	27	29	30

Mean = 24.42

Median = 26

MEAN VS. MEDIAN

- Whenever a data set has extreme values, the median is the preferred measure of center.
- Mean is bothered by extreme values, while median is not.

26	300	27	22	24	29	13

13	22	24	26	27	29	300

Mean = 63

Median = 26

SUMMARY

- When exploring data, a good summary of a variable might be what a "typical" (or center) value of that variable would look like.
- Mode – the mode of a variable is the most common value.
- Mean – the mean of a variable is the sum of all the values divided by the number of values.
- Median – value in the middle when the data items are arranged in ascending order.

DESCRIBING SPREAD

EXPLORING DATA

MEASURES OF VARIABILITY

- Center can only get you so far with describing a variable's "typical" value.
- Typically, we also consider how spread out a data set is as well.
- This is called **variability** or dispersion.

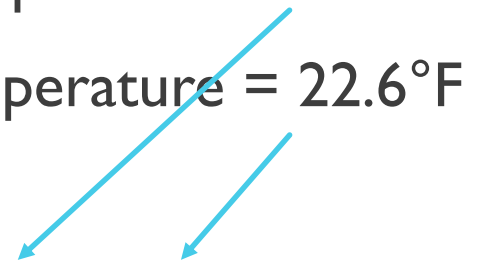
RANGE

- **Range** – difference between the largest and smallest values.
  - Highest temperature = 90.5°F
  - Lowest temperature = 22.6°F

RANGE

- **Range** – difference between the largest and smallest values.

- Highest temperature = 90.5°F
- Lowest temperature = 22.6°F

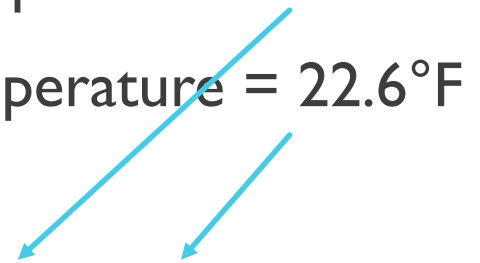


Range = 90.5 – 22.6 = 67.9°F

RANGE

- **Range** – difference between the largest and smallest values.
- Very sensitive to observations with extreme values as it only focuses on the largest and smallest values in the data.

- Highest temperature = 90.5°F
- Lowest temperature = 22.6°F



Range = 90.5 – 22.6 = 67.9°F

RANGE

- **Range** – difference between the largest and smallest values.

- In our data, the "spread" of temperature (according to range) is 67.9°F.

- Highest temperature = 90.5°F
- Lowest temperature = 22.6°F

Range = 90.5 – 22.6 = 67.9°F

VARIANCE

- **Variance** – measure of dispersion around the mean of the data set.
- Average of the squared distances between each data value and mean.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

VARIANCE

- **Variance** – measure of dispersion around the mean of the data set.
- Average of the squared distances between each data value and mean.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Temperature (°F)	$(x_i - \bar{x})$ $\bar{x} = 59.5$	$(x_i - \bar{x})^2$
46.7	-12.8	163.8
48.4	-11.1	123.2
34.2	-25.3	640.1
34.5	-25.0	625.0
36.8	-22.7	515.3
…	…	…
72.5	13.0	169.0

VARIANCE

- **Variance** – measure of dispersion around the mean of the data set.
- Average of the squared distances between each data value and mean.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Temperature (°F)	$(x_i - \bar{x})$ $\bar{x} = 59.5$	$(x_i - \bar{x})^2$
46.7	-12.8	163.8
48.4	-11.1	123.2
34.2	-25.3	640.1
34.5	-25.0	625.0
36.8	-22.7	515.3
…	…	…
72.5	13.0	169.0

Add together and divide by 730 (= 731 – 1)

VARIANCE

- **Variance** – measure of dispersion around the mean of the data set.

- In our data, the "spread" of temperature (according to variance) is 239.8°F squared.

Temperature (°F)	$(x_i - \bar{x})$ $\bar{x} = 59.5$	$(x_i - \bar{x})^2$
46.7	-12.8	163.8
48.4	-11.1	123.2
34.2	-25.3	640.1
34.5	-25.0	625.0
36.8	-22.7	515.3
…	…	…
72.5	13.0	169.0

VARIANCE

- **Variance** – measure of dispersion around the mean of the data set.

- In our data, the "spread" of temperature (according to variance) is 239.8°F squared.

  ???

Temperature (°F)	$(x_i - \bar{x})$ $\bar{x} = 59.5$	$(x_i - \bar{x})^2$
46.7	-12.8	163.8
48.4	-11.1	123.2
34.2	-25.3	640.1
34.5	-25.0	625.0
36.8	-22.7	515.3
…	…	…
72.5	13.0	169.0

STANDARD DEVIATION

- The problem with variance is that it is in terms of squared units of the data.
- To correct for this, we have the **standard deviation**, which is just the square root of the variance.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

STANDARD DEVIATION

- The problem with variance is that it is in terms of squared units of the data.
- To correct for this, we have the **standard deviation**, which is just the square root of the variance.
- In our data, the "spread" of temperature (according to standard deviation) is 15.5°F.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2 CHARACTERISTICS OF VARIANCE

- Variance (and standard deviation) possess two common characteristics:
  1. If the variance equals zero, then all of the data in the data set has the same value.
  2. All measures of spread are positive (or nonnegative if zero spread) in value.

SUMMARY

- Don’t only look at center, but also variability of a variable.
- Range – difference between the largest and smallest values.
- Variance – measure of dispersion around the mean of the data set.
- Standard deviation – the square root of the variance (helps with units of variance).