RELATIONSHIPS IN DATA

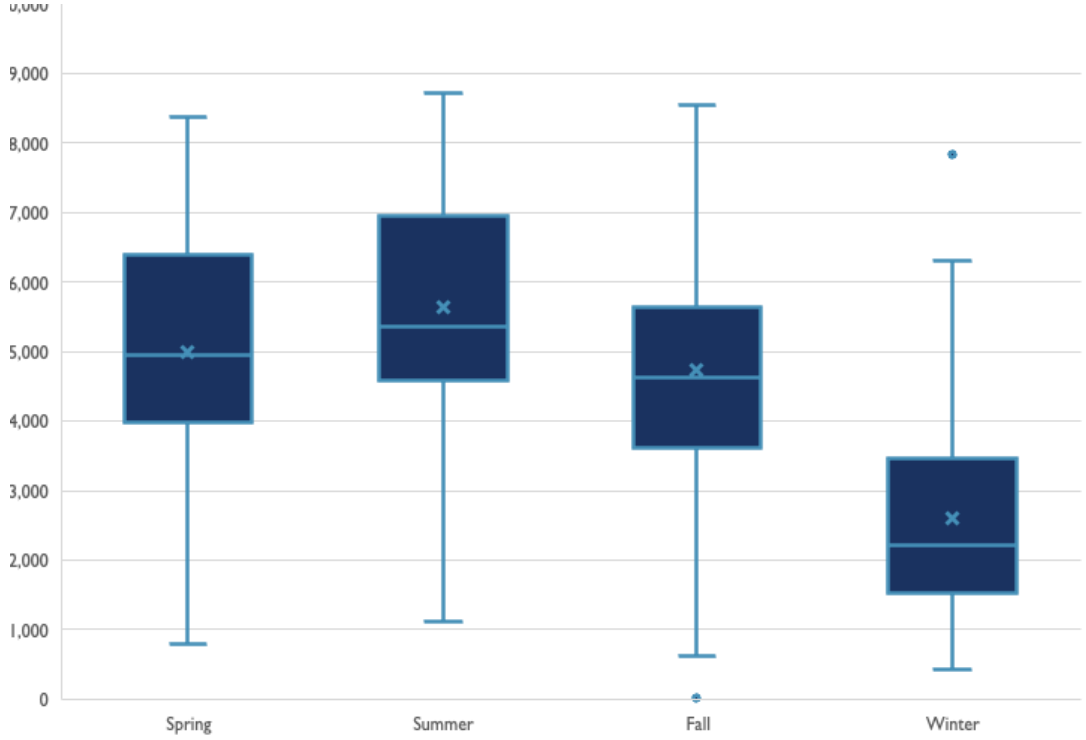
ST101 – DR. ARIC LABARR

EXPLORING DATA RELATIONSHIPS

- Exploring data reveals potential insights and valuable uses of that information.
- Visuals help explore data.
  - Distributions, bar charts, stacked bar charts, **boxplots**, **scatterplots**, etc.
- Are visuals enough?

BOXPLOTS

RELATIONSHIPS IN DATA



EXAMPLE – BIKE RENTAL DATA

5 NUMBER SUMMARY

- Boxplots are visual representations of 5 (and sometimes more!) summary measures about a set of data.
  - Minimum value
  - 1st quartile
  - Median
  - 3rd quartile
  - Maximum value

5 NUMBER SUMMARY

- Boxplots are visual representations of 5 (and sometimes more!) summary measures about a set of data.
  - Minimum value
  - 1st quartile
  - Median
  - 3rd quartile
  - Maximum value

MINIMUM VALUE

- The **minimum value** of a variable is numerically lowest value that the variable takes.
- Here are the lowest temperatures from our bike rental data (°F):

22.6°, 25.8°, 25.8°, 26.7°, …

MINIMUM VALUE

- The **minimum value** of a variable is numerically lowest value that the variable takes.
- Here are the lowest temperatures from our bike rental data (°F):

22.6°, 25.8°, 25.8°, 26.7°, …

Minimum value

QUARTILES AND PERCENTILES

- A percentile provides information about how the data are spread over the interval from the smallest values to the largest value.
- The ***p*th percentile** of a data set is a value such that at least *p* percent of the items take on this value or less and at least (100 - *p*) percent of the items take on this value or more.
- For example, the student's test score was in the 93rd percentile.

QUARTILES AND PERCENTILES

- Quartiles are specific percentiles that are commonly used.
- The **first quartile, Q_1**, is the **25^{th} percentile**.
- The **second quartile** is the **50^{th} percentile**, which we already defined – the **median**.
- The **third quartile, Q_3**, is the **75^{th} percentile**.

QUARTILE CALCULATION

- Quartiles are calculated in similar ways as the median with all data ordered from smallest to largest.
- Typically, let computers do this calculation.
- In our bike rental data, these are the quartiles:
  - Q_1 = 46.08°F
  - Median = 59.76°F
  - Q_3 = 73.08°F

INTERQUARTILE RANGE (IQR)

- The **interquartile range** (**IQR**) of a data set if the difference between the third and first quartiles:

$$IQR = Q_3 - Q_1$$

- This is the **middle 50%** of the data set and is not bothered by extreme observations in the tails of the data set.

INTERQUARTILE RANGE (IQR)

- In our bike rental data, these are the quartiles:
  - Q_1 = 46.08°F
  - Median = 59.76°F
  - Q_3 = 73.08°F

$$IQR = 73.08 - 46.08 = 27°\text{F}$$

MAXIMUM VALUE

- The **maximum value** of a variable is numerically highest value that the variable takes.
- Here are the highest temperatures from our bike rental data (°F):

… , 88.5°, 89.4°, 89.4°, 90.5°

MAXIMUM VALUE

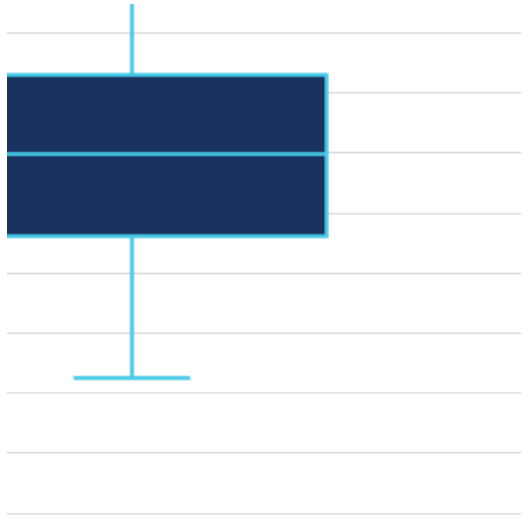
- The **maximum value** of a variable is numerically highest value that the variable takes.
- Here are the highest temperatures from our bike rental data (°F):

… , 88.5°, 89.4°, 89.4°, 90.5°

Maximum value

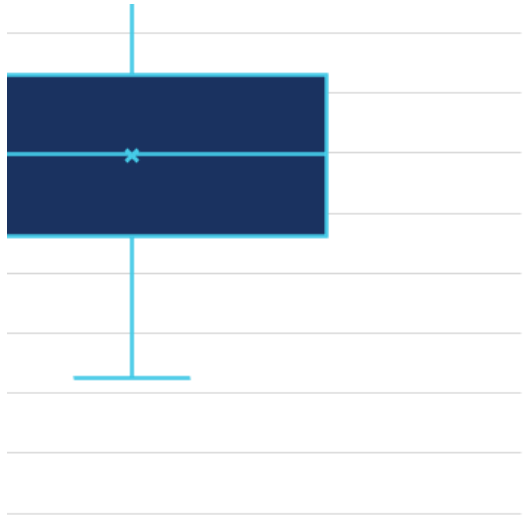
BOXPLOT

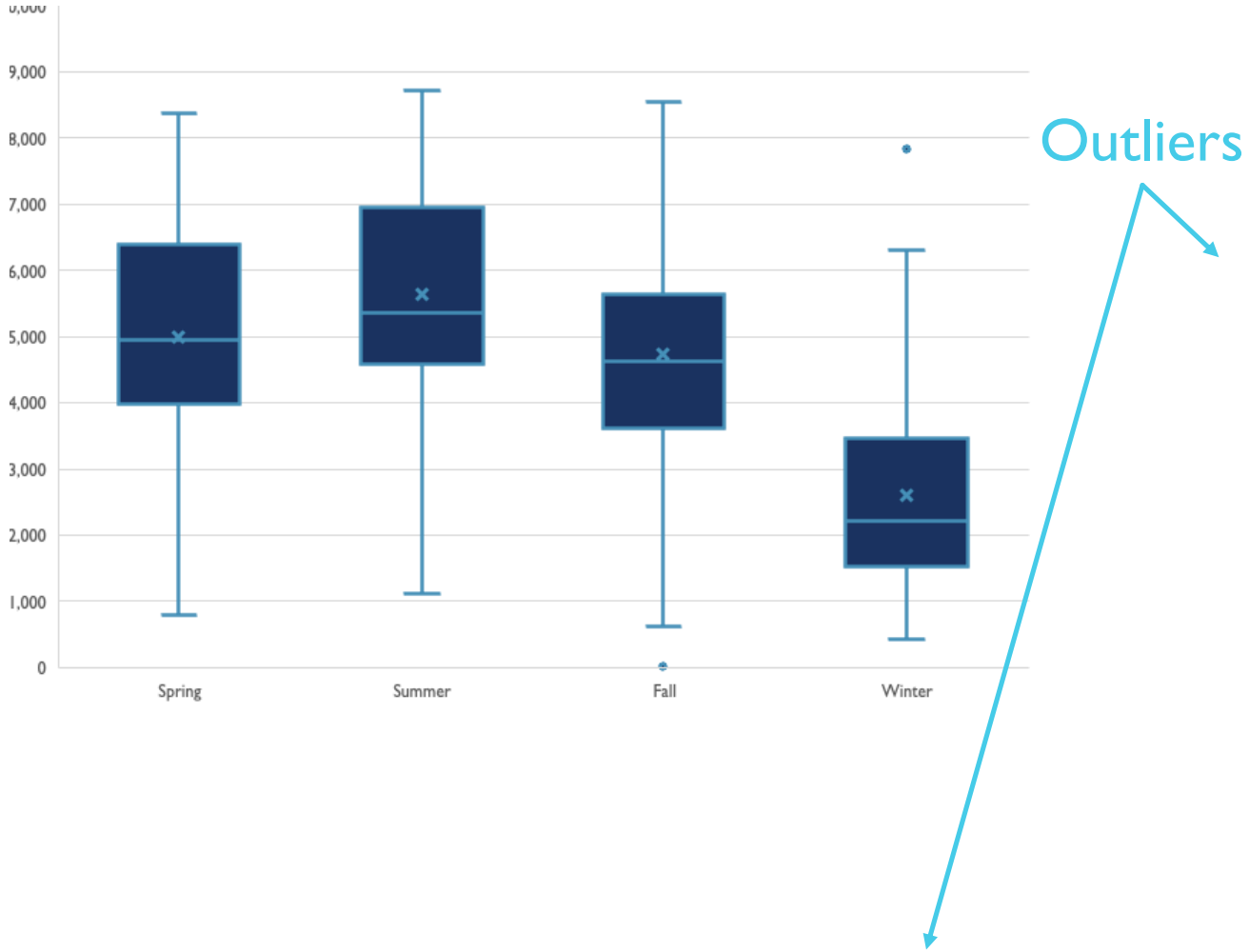
- Boxplots are visual representations of 5 (and sometimes more!) summary measures about a set of data.
  - Maximum value = 90.50°F
  - 3rd quartile = 73.08°F
  - Median = 59.76°F
  - 1st quartile = 46.08°F
  - Minimum value = 22.60°F



BOXPLOT

- Boxplots are visual representations of 5 (**and sometimes more!**) summary measures about a set of data.
  - Maximum value = 90.50°F
  - 3rd quartile = 73.08°F
  - Median = 59.76°F
  - Mean = 59.51°F
  - 1st quartile = 46.08°F
  - Minimum value = 22.60°F





EXAMPLE – BIKE RENTAL DATA

OUTLIERS ON A BOXPLOT

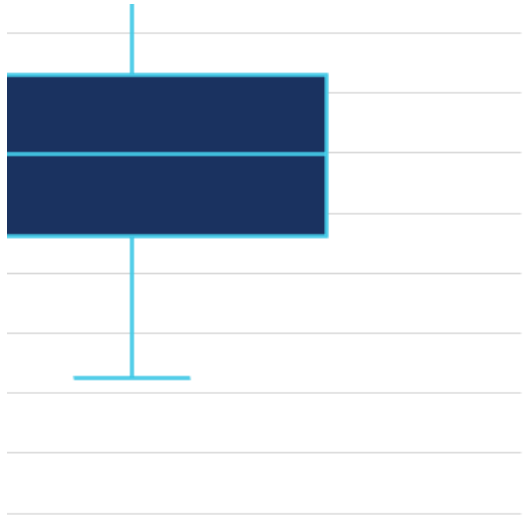
- The minimum and maximum value lines are now the maximum and minimum values **within an outlier boundary** of the IQR.
- Anything outside of this boundary (low or high) is considered an outlier.

1.5 IQR RULE

- Anything outside of this boundary (low or high) is considered an outlier.
- The boundary is between $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$.

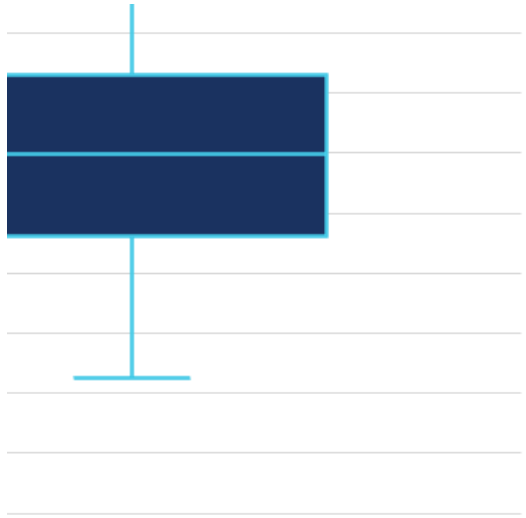
1.5 IQR RULE

- Bike Data:
  - Maximum value = 90.50°F
  - 3rd quartile = 73.08°F
  - Median = 59.76°F
  - 1st quartile = 46.08°F
  - Minimum value = 22.60°F
- IQR = 27



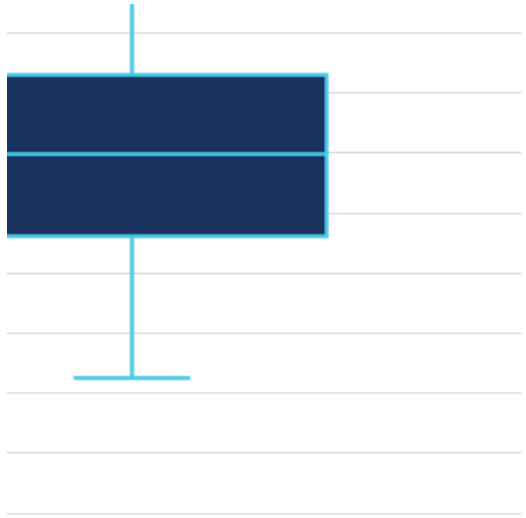
1.5 IQR RULE

- Bike Data:
  - Maximum value = 90.50°F
  - 3rd quartile = 73.08°F
  - Median = 59.76°F
  - 1st quartile = 46.08°F $-1.5 \times 27 = 5.58$
  - Minimum value = 22.60°F
- IQR = 27



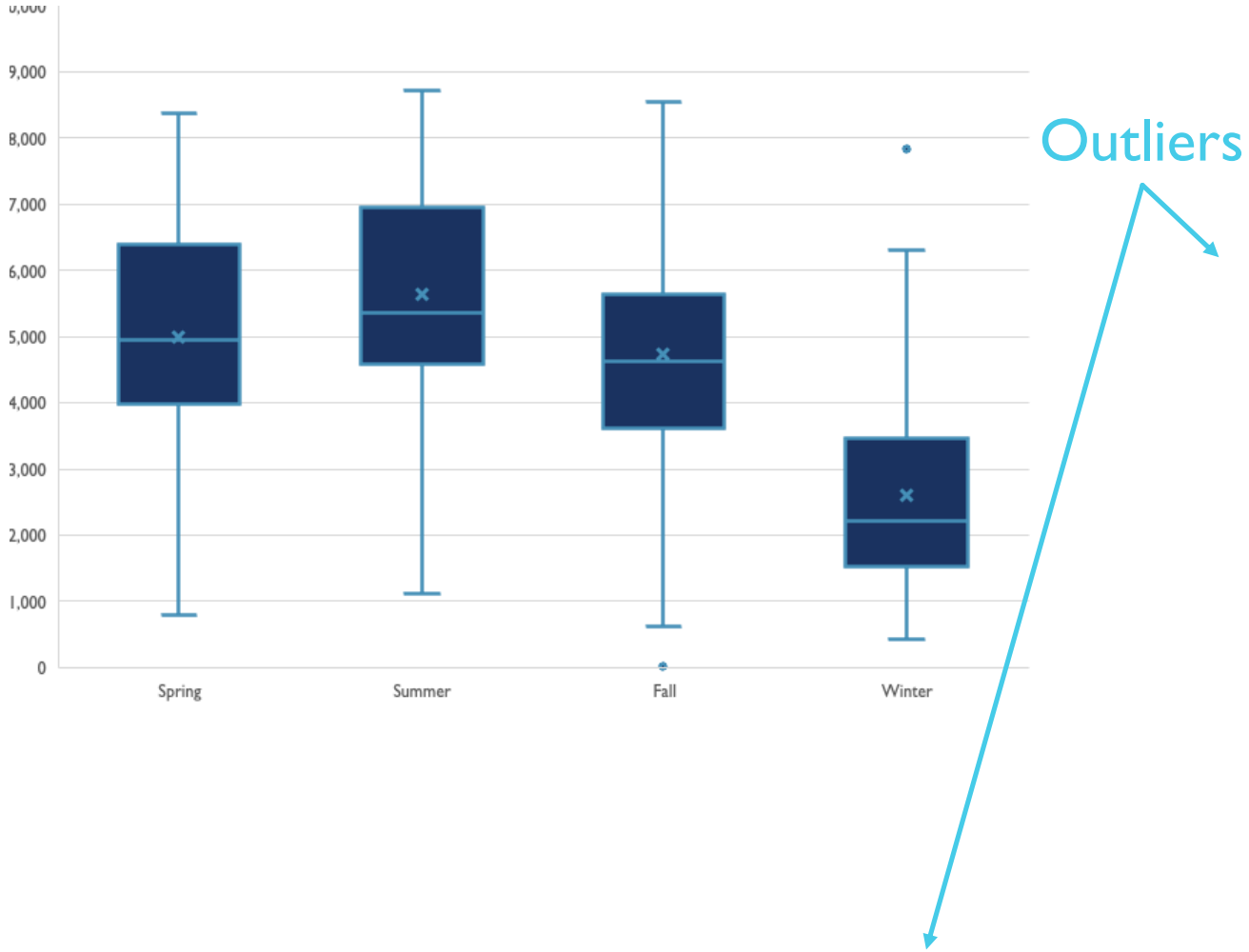
1.5 IQR RULE

- Bike Data:
  - Maximum value = 90.50°F
  - 3rd quartile = 73.08°F$+1.5 \times 27 = 113.58$
  - Median = 59.76°F
  - 1st quartile = 46.08°F$-1.5 \times 27 = 5.58$
  - Minimum value = 22.60°F
- IQR = 27



1.5 IQR RULE

- Anything outside of this boundary (low or high) is considered an outlier.
- The boundary is between $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$.
- In our bike temperature data, the boundary is **5.58** and **113.58**.
- Any temperature outside of this boundary is an outlier.



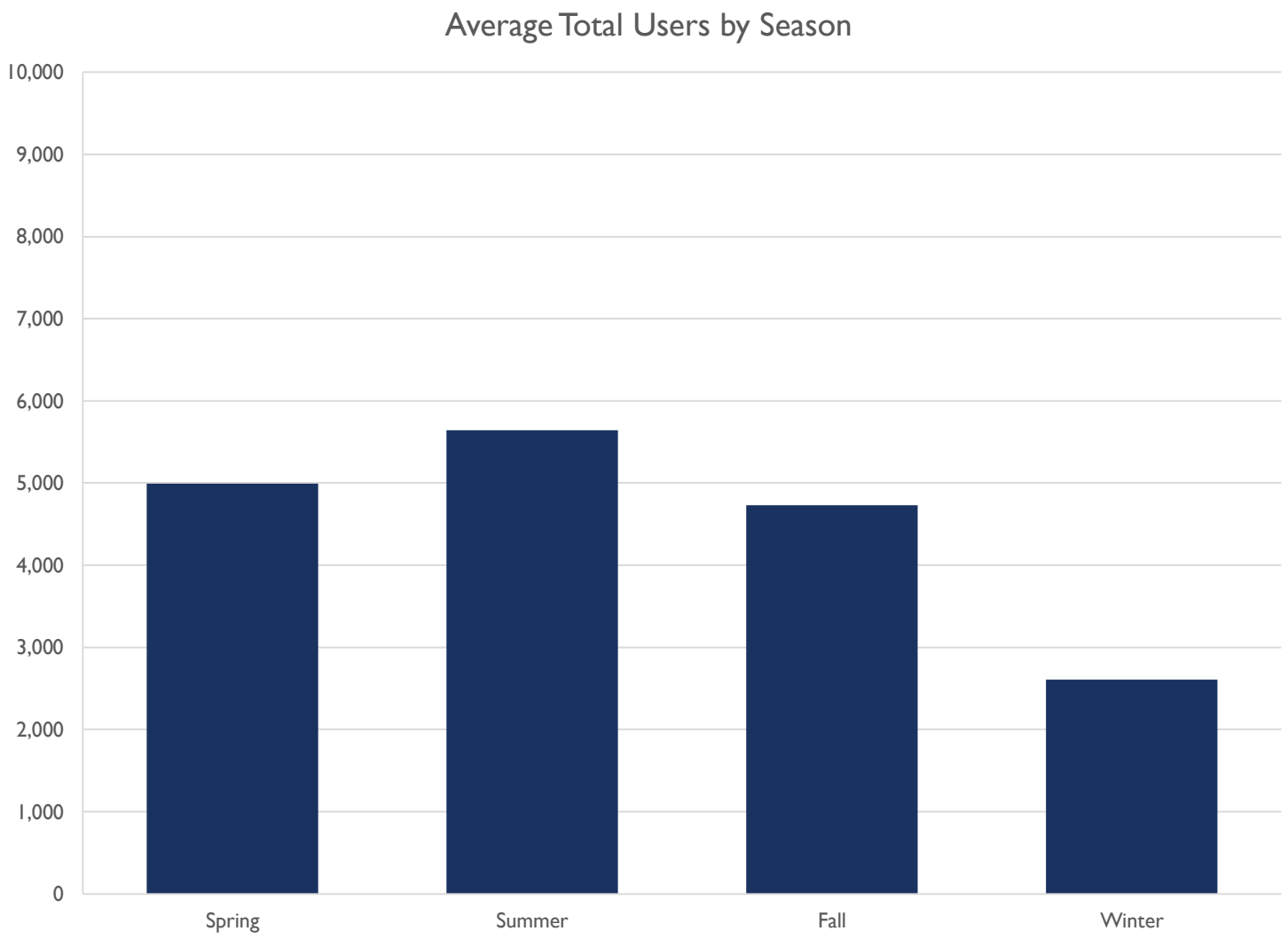
EXAMPLE – BIKE RENTAL DATA

SUMMARY

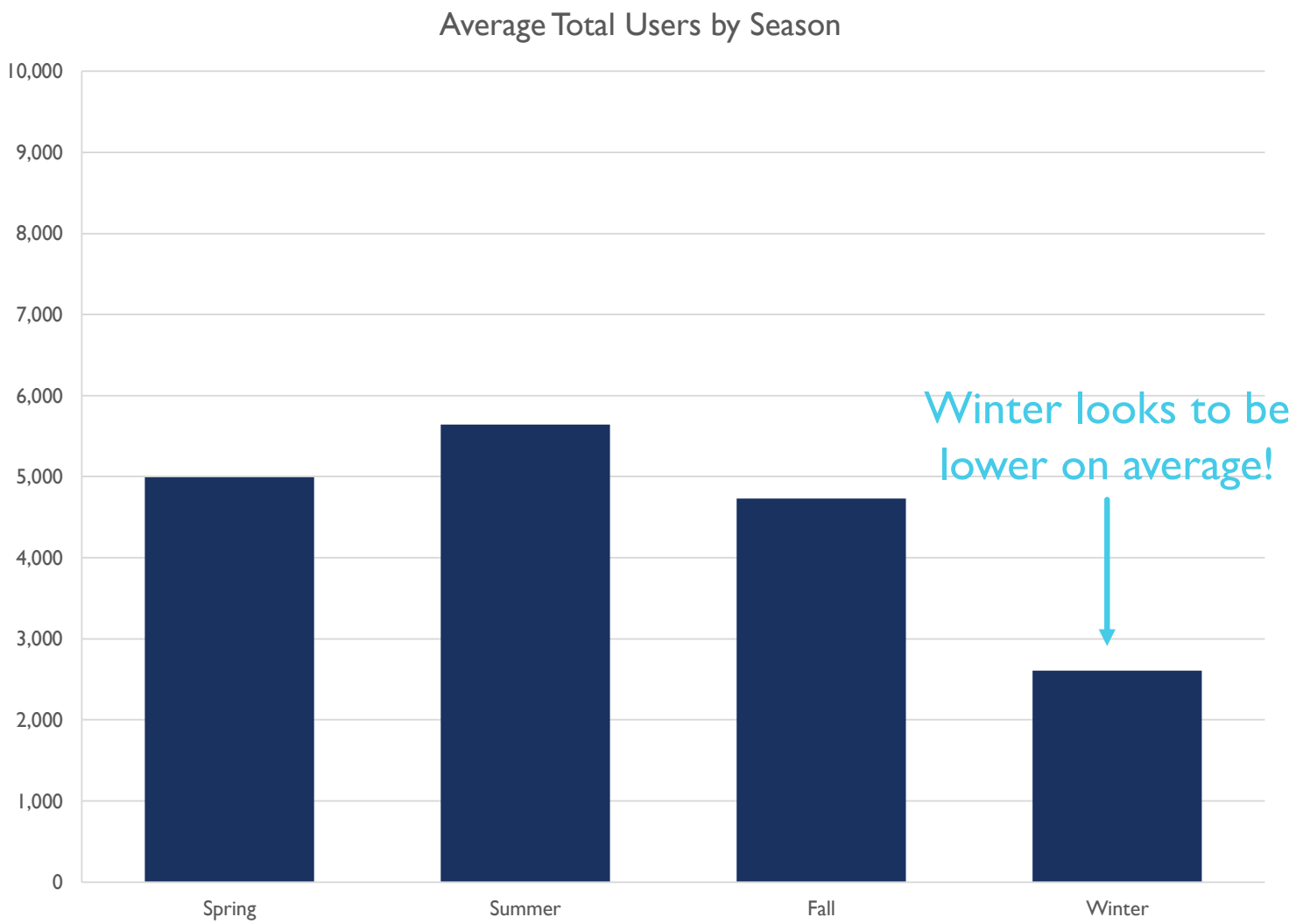
- Boxplots are visual representations of 5 (and sometimes more!) summary measures about a set of data.
  - Minimum value – smallest value in a variable
  - 1st quartile – 25th percentile of a variable
  - Median – 50th percentile of a variable
  - Mean – average value of the variable
  - 3rd quartile – 75th percentile of a variable
  - Maximum value – largest value of a variable
  - Outliers – anything outside of the 1.5 IQR Rule boundary on your variable

IDEA OF ANALYSIS OF VARIANCE (ANOVA)

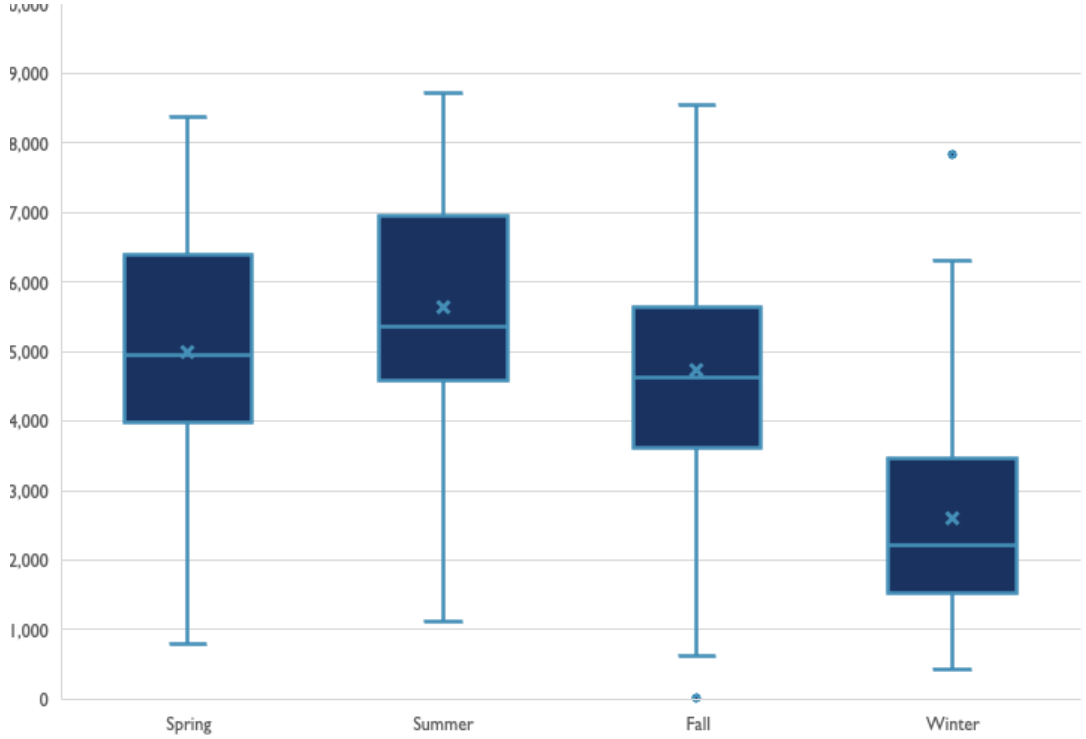
RELATIONSHIPS IN DATA



DOES WINTER HAVE LOWER AVERAGE TOTAL USERS?



DOES WINTER HAVE LOWER AVERAGE TOTAL USERS?



DOES WINTER HAVE LOWER AVERAGE TOTAL USERS?

Look at how spread out the values of users in winter are!

COMPARING AVERAGES

- When comparing averages between two groups of data, we must think about how spread out the data is.
- Compare the 5 number summary between Summer and Winter.

COMPARING AVERAGES

Summer 5ish Number Summary

- Maximum: 8714
- 3rd Quartile: 6954
- Mean: 5644.3
- Median: 5354
- 1st Quartile: 4580
- Minimum: 1115

Winter 5ish Number Summary

- Maximum: 7836
- 3rd Quartile: 3472
- Mean: 2604.1
- Median: 2209
- 1st Quartile: 1534
- Minimum: 431

COMPARING AVERAGES

Summer 5ish Number Summary

- Maximum: 8714
- 3rd Quartile: 6954
- Mean: 5644.3
- Median: 5354
- 1st Quartile: 4580
- Minimum: 1115

Winter 5ish Number Summary

- Maximum: 7836
- 3rd Quartile: 3472
- Mean: 2604.1
- Median: 2209
- 1st Quartile: 1534
- Minimum: 431

COMPARING AVERAGES

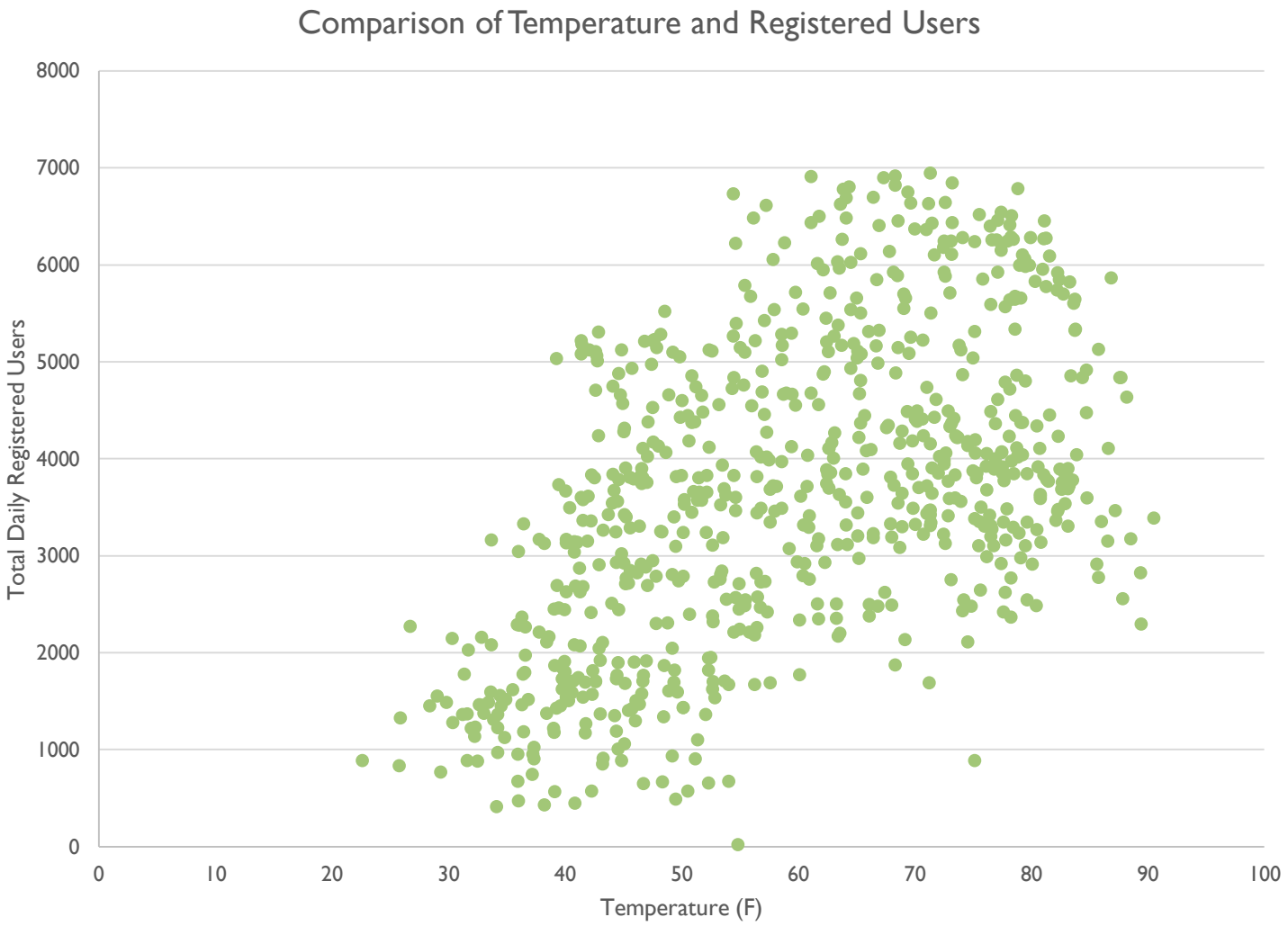
- When comparing averages between two groups of data, we must think about how spread out the data is.
- Compare the 5 number summary between Summer and Winter.
- When comparing many averages *statistically* we call this an **Analysis of Variance (ANOVA)**.
  - Need to account for the spread in the data when comparing means!

SUMMARY

- When comparing many groups' averages *statistically* we call this an analysis of variance (ANOVA).
- Need to account for the spread in the data when comparing means.

INTERPRETING SCATTERPLOTS

RELATIONSHIPS IN DATA



EXAMPLE – BIKE RENTAL DATA

SCATTERPLOT

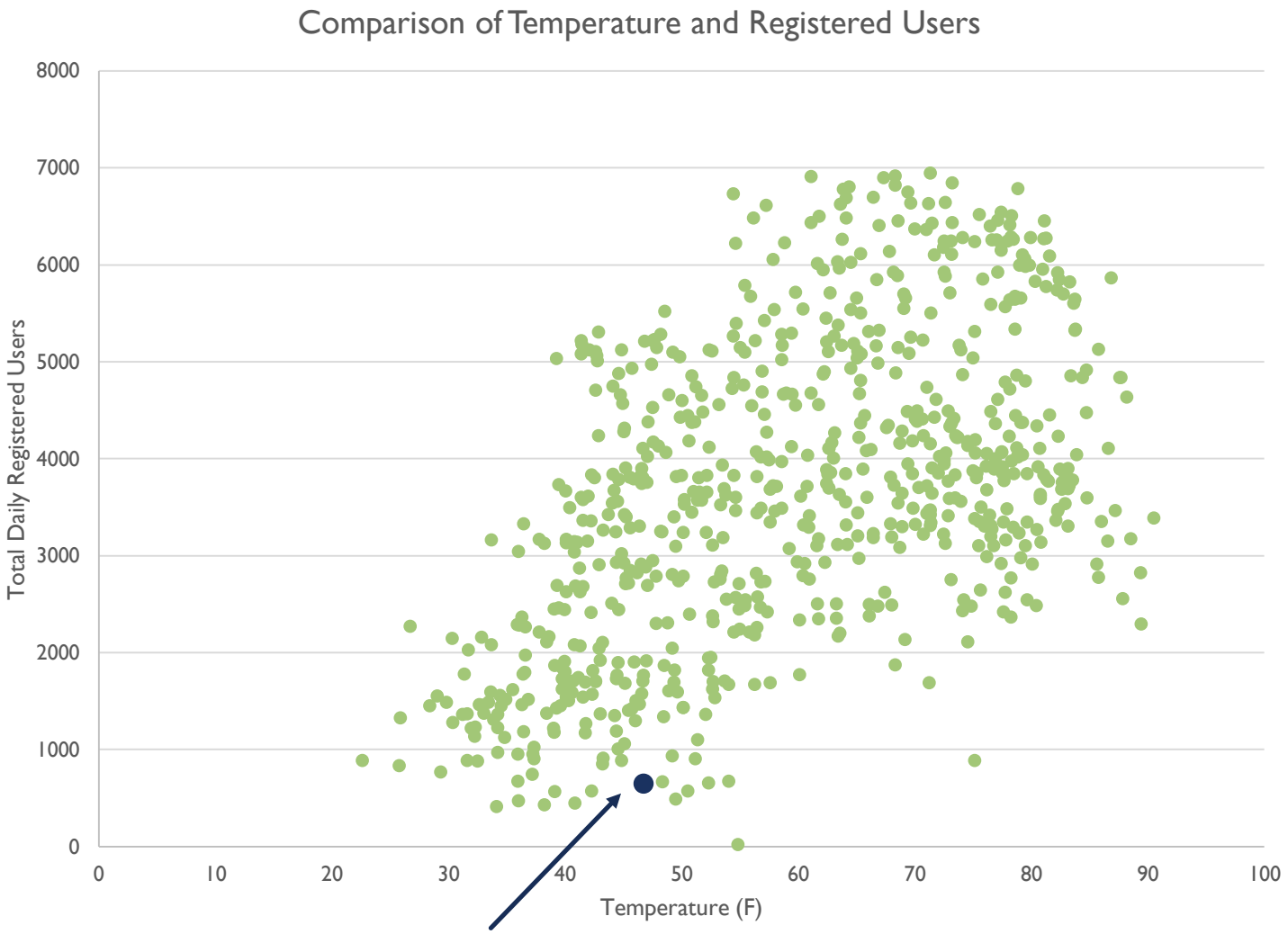
- Scatterplots are visual representations of comparing two different quantitative variables.
- For each observation in your data, you are looking at the value of **two** quantitative variables which are plotted one on each axis in the plot.

EXAMPLE – BIKE RENTAL DATA

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

⋮

EXAMPLE – BIKE RENTAL DATA



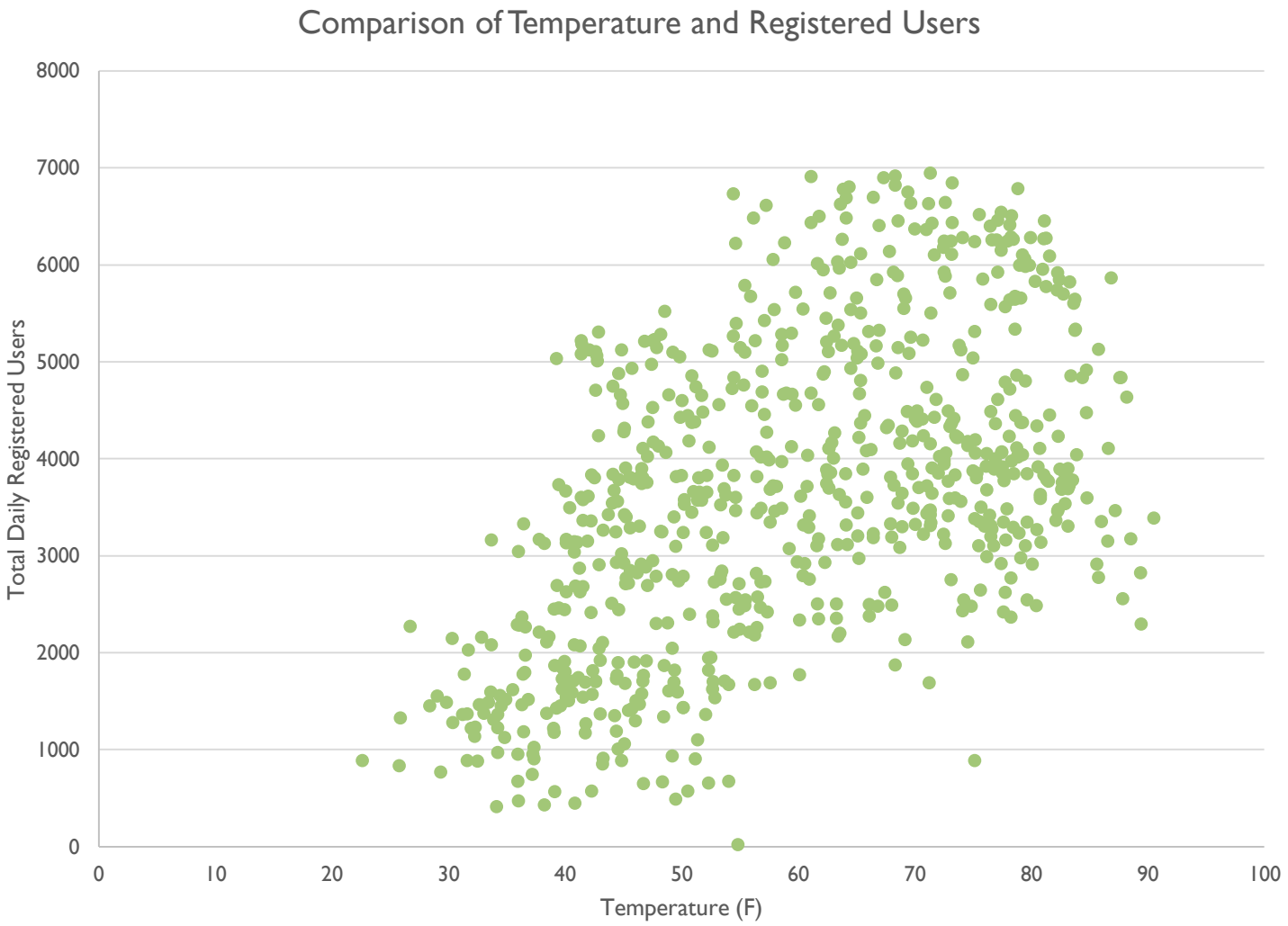
Temperature = 46.7, Registered Users = 654

VISUALIZING RELATIONSHIPS

- Viewing the relationship between two quantitative variables on a scatterplot is very beneficial.
- **Linear relationship** – relationship between variables that exhibits a fairly straight / linear pattern
- **Nonlinear relationship** – relationship between variables that exhibits a pattern that is nonlinear in nature

VISUALIZING RELATIONSHIPS

- Viewing the relationship between two quantitative variables on a scatterplot is very beneficial.
- **Positive relationship** – as one variable increases (or decreases) the other has a *tendency* to do the same
- **Negative relationship** – as one variable increases (or decreases) the other has a *tendency* to do the opposite



POSITIVE AND RELATIVELY LINEAR RELATIONSHIP

SUMMARY

- Scatterplots are visual representations of comparing two different quantitative variables, which are plotted one on each axis in the plot.
- Positive relationship – as one variable increases (or decreases) the other has a *tendency* to do the same
- Negative relationship – as one variable increases (or decreases) the other has a *tendency* to do the opposite

CORRELATION

RELATIONSHIPS IN DATA

CORRELATION

- Correlation is a popular term that is thrown around by people who may not understand the implications (or lack there of) of what they are saying.
- The **Pearson correlation coefficient**, r, is a measure of strength of the *linear* relationship between two variables.

CORRELATION COEFFICIENT

- The Pearson correlation coefficient is **unit less** – no units when describing it.

$$-1 \leq r \leq 1$$

CORRELATION COEFFICIENT

- The Pearson correlation coefficient is **unit less** – no units when describing it.

$$-1 \leq r \leq 1$$

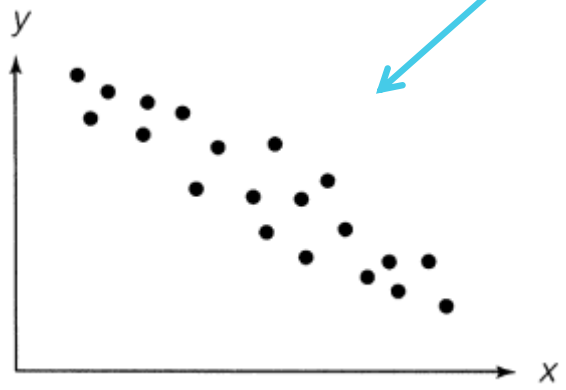
Bounded between -1 and 1

CORRELATION COEFFICIENT

- The Pearson correlation coefficient is **unit less** – no units when describing it.

$$-1 \leq r \leq 1$$

Negative values imply a negative *linear* relationship



CORRELATION COEFFICIENT

- The Pearson correlation coefficient is **unit less** – no units when describing it.

$$-1 \leq r \leq 1$$

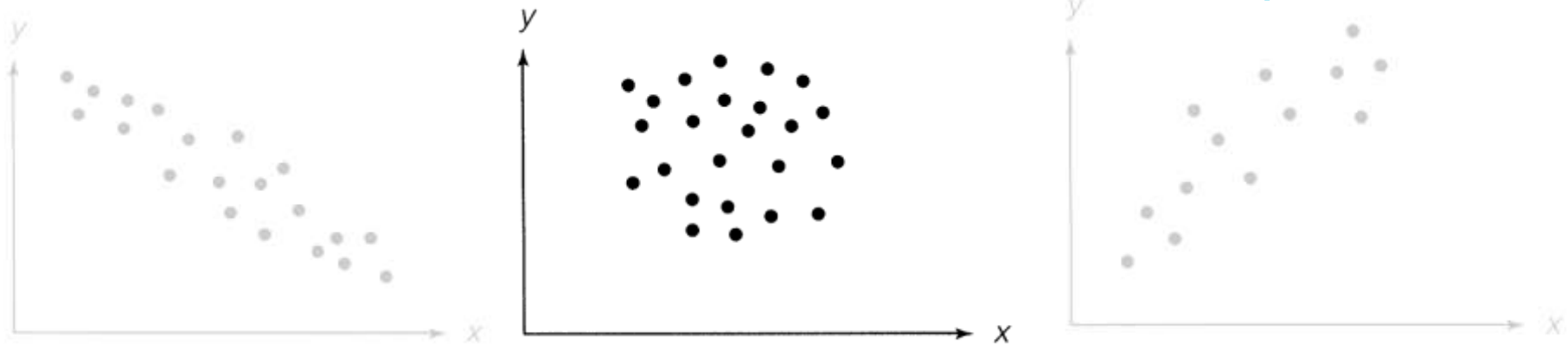
Positive values imply a positive *linear* relationship

CORRELATION COEFFICIENT

- The Pearson correlation coefficient is **unit less** – no units when describing it.

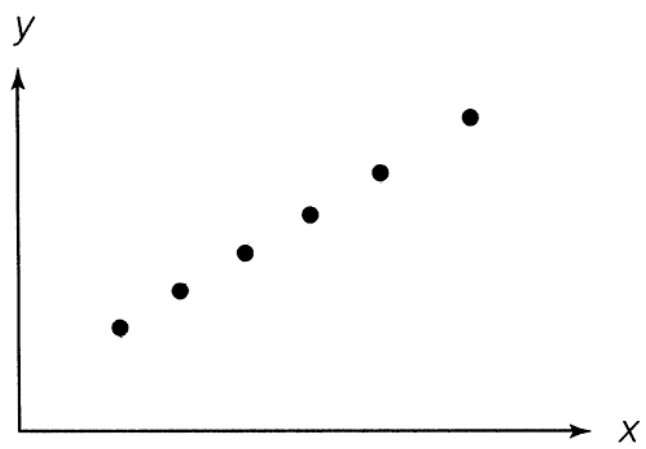
$$-1 \leq r \leq 1$$

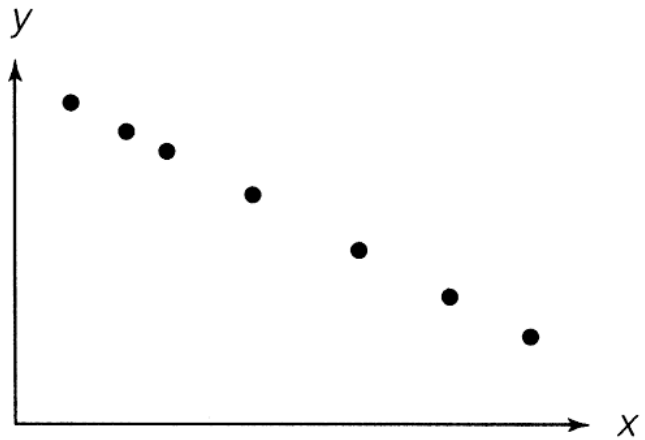
Values near 0 implies no real *linear* relationship

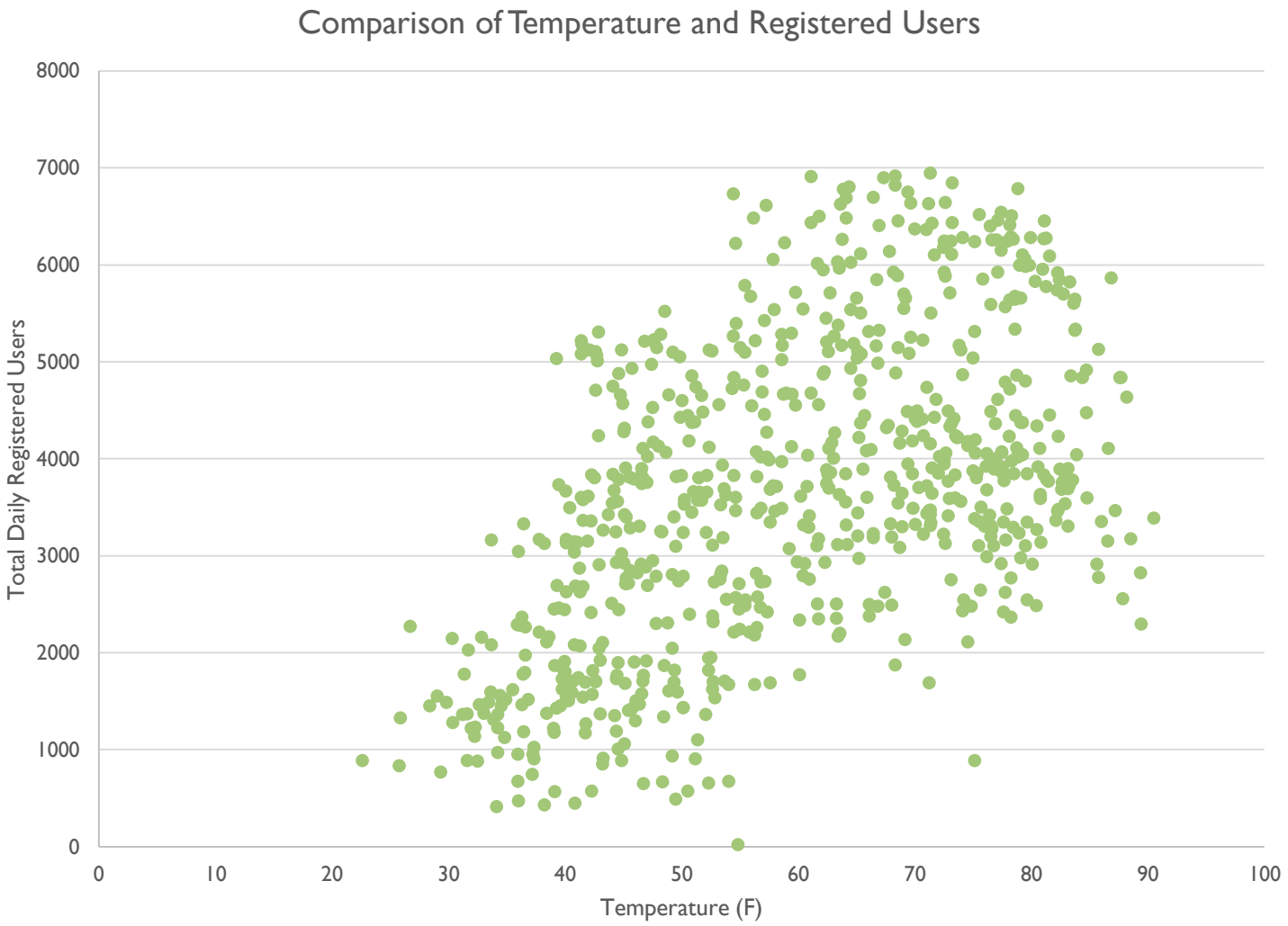


CORRELATION COEFFICIENT

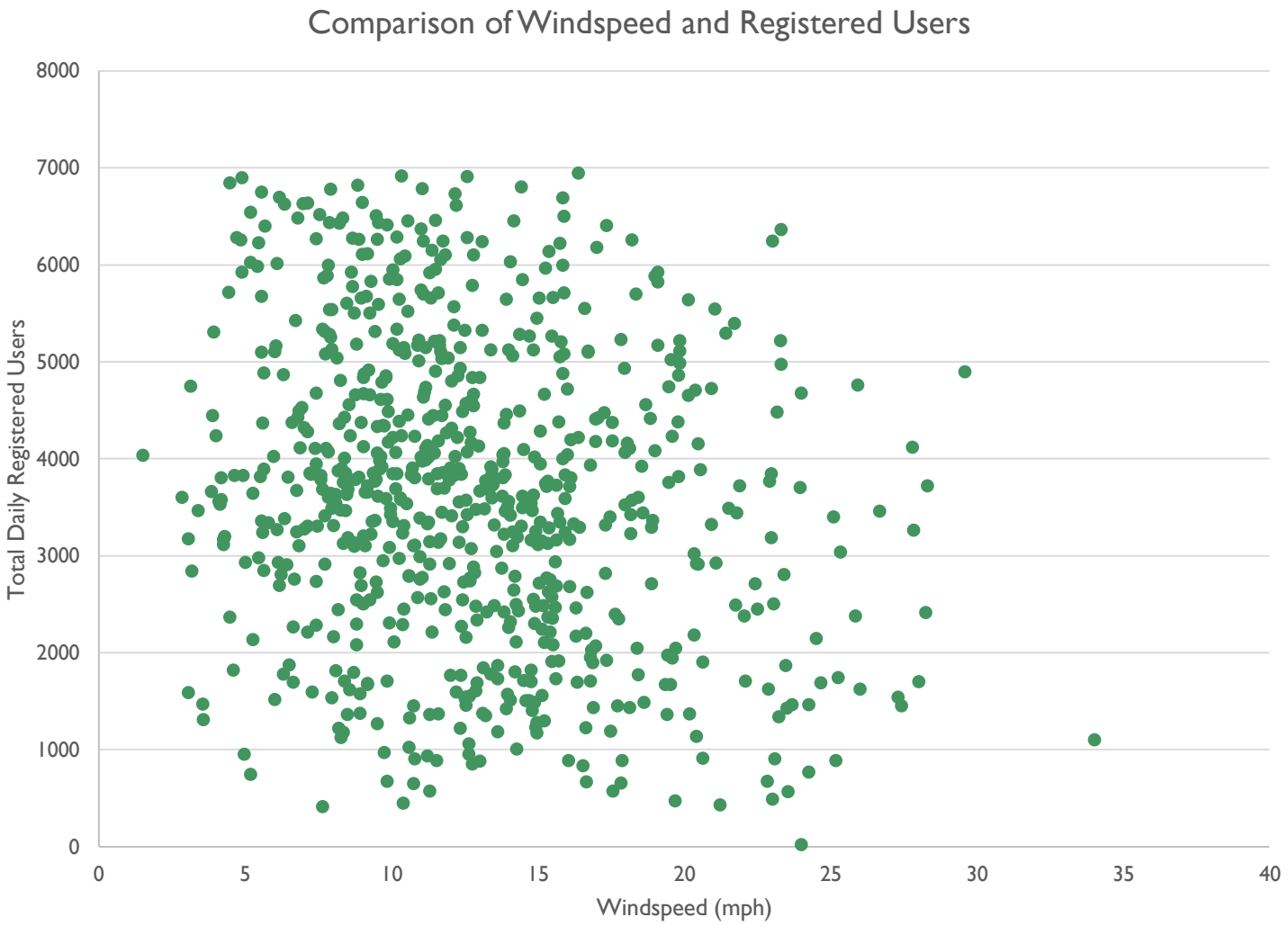
- Values of 1 or -1 imply a **perfect** linear relationship between y and x.







CORRELATION OF 0.54



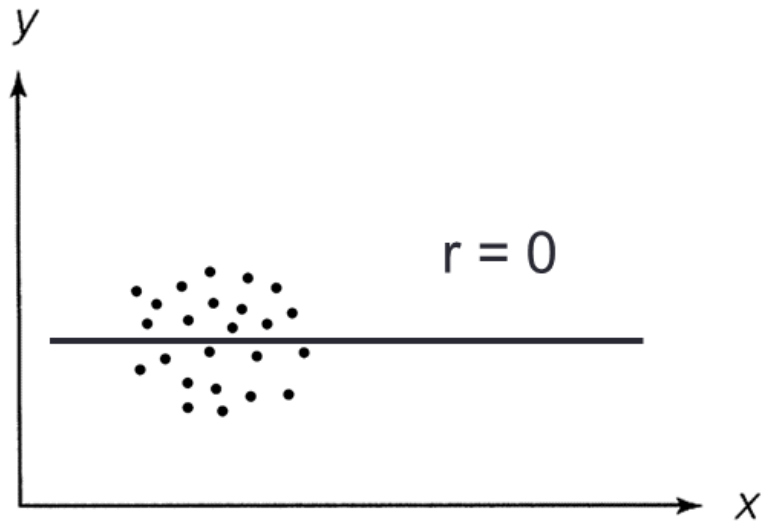
CORRELATION OF -0.22

POTENTIAL ISSUES WITH CORRELATION

- Two of the biggest problems with correlation are the following:
  1. Outliers
  2. Causation

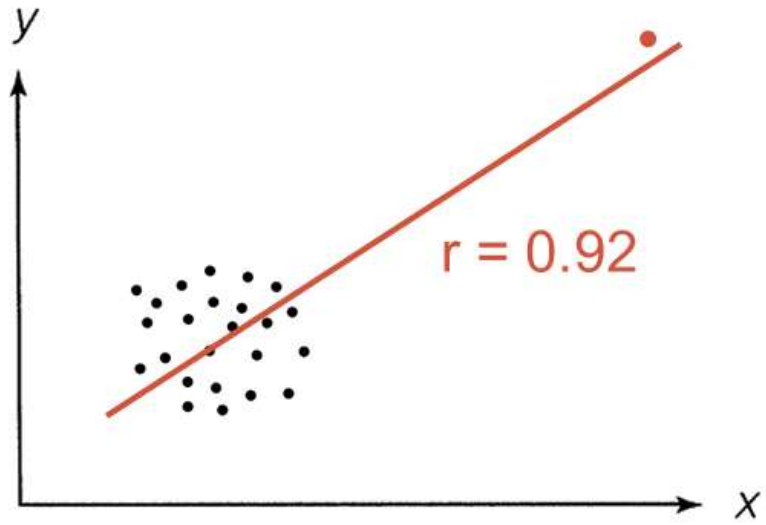
OUTLIERS IN CORRELATION

- Outliers can lead to false conclusions about correlation if you don't visualize the data to help us see what might be going on.
- Outliers can make relationships that aren't really there.



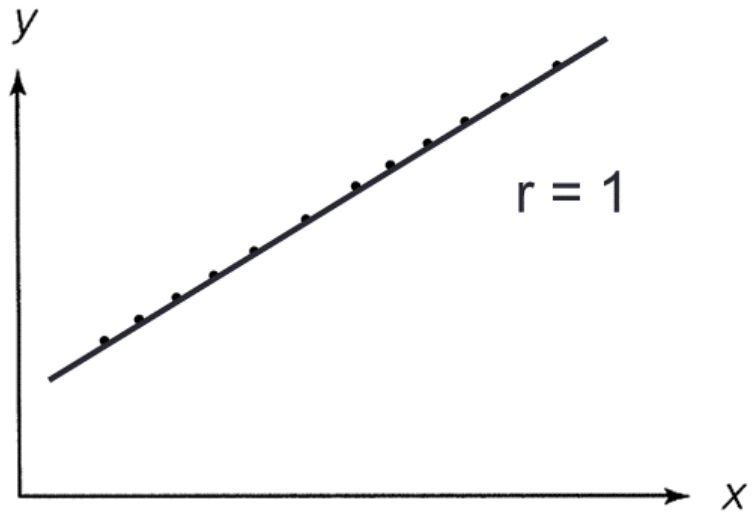
OUTLIERS IN CORRELATION

- Outliers can lead to false conclusions about correlation if you don't visualize the data to help us see what might be going on.
- Outliers can make relationships that aren't really there.



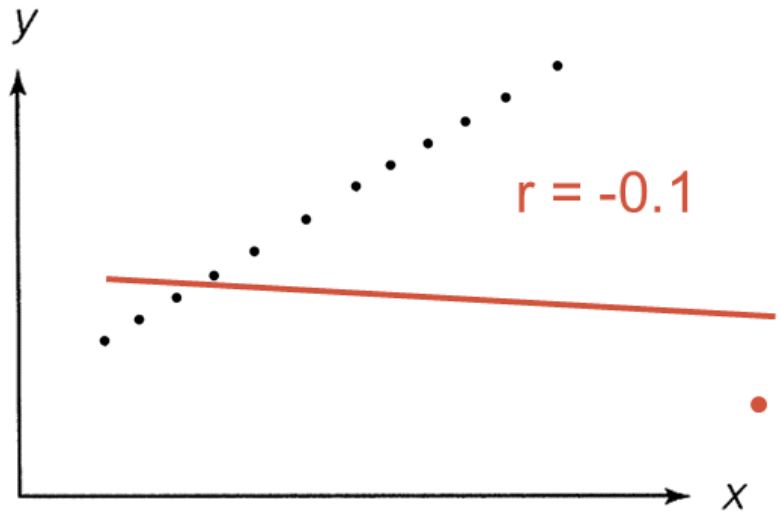
OUTLIERS IN CORRELATION

- Outliers can lead to false conclusions about correlation if you don't visualize the data to help us see what might be going on.
- Outliers can hide relationships that are really there.



OUTLIERS IN CORRELATION

- Outliers can lead to false conclusions about correlation if you don't visualize the data to help us see what might be going on.
- Outliers can hide relationships that are really there.



SUMMARY

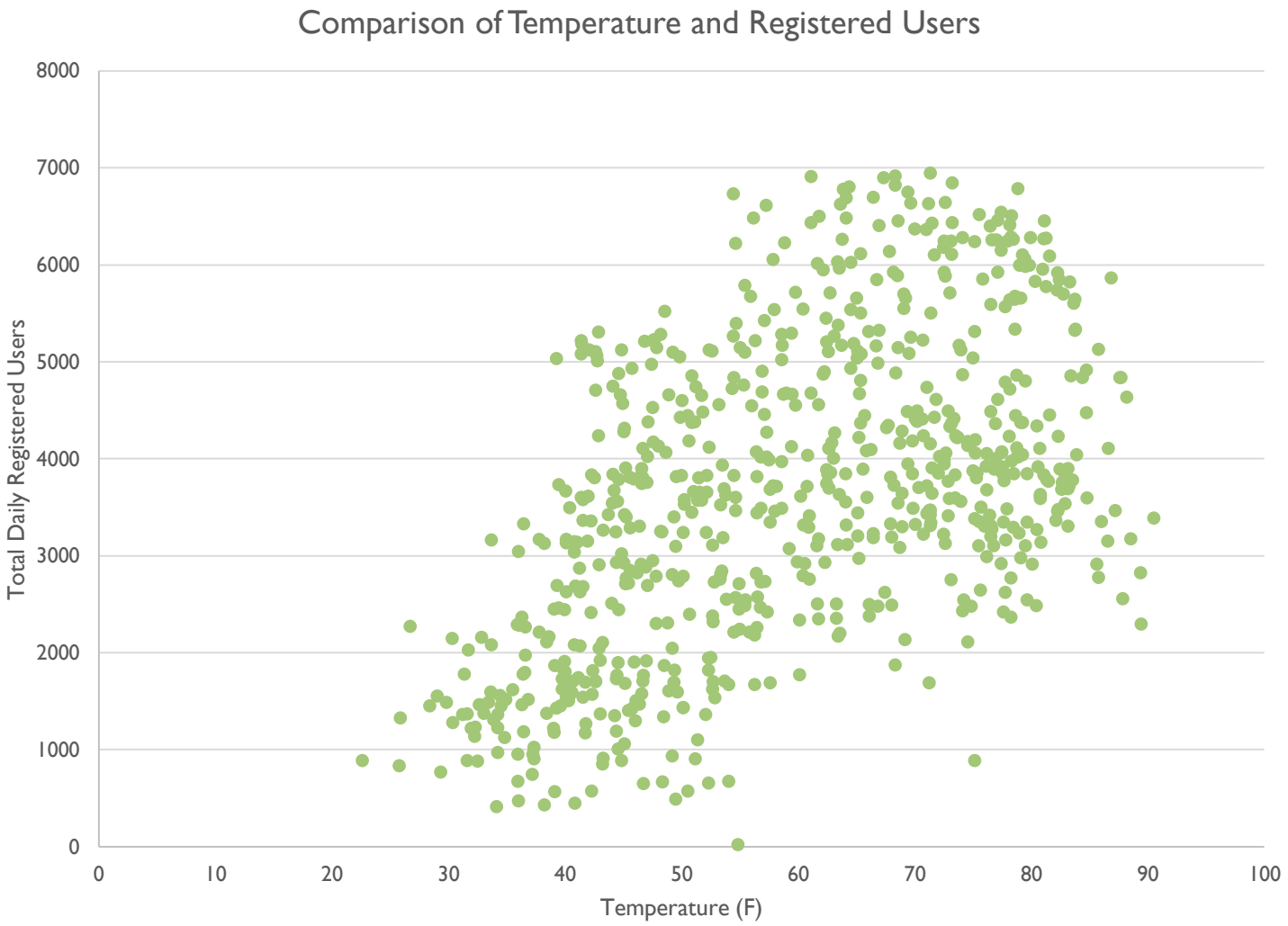
- The Pearson correlation coefficient, r, is a measure of strength of the *linear* relationship between two variables.
  - Negative values imply a negative *linear* relationship.
  - Positive values imply a positive *linear* relationship.
  - Values near 0 implies no real *linear* relationship.
- Outliers can make/hide relationships that are/aren't really there.

CORRELATION AND CAUSATION

RELATIONSHIPS IN DATA

POTENTIAL ISSUES WITH CORRELATION

- Two of the biggest problems with correlation are the following:
  1. Outliers
  2. Causation



CORRELATION OF 0.54

CORRELATION VS. CAUSATION

- Confusing correlation and causation is a common phenomena.
- All correlation implies is a linear trend may exist between two variables of interest.
- Many famous examples of correlations that are **not** causations.

CLASSIC EXAMPLE

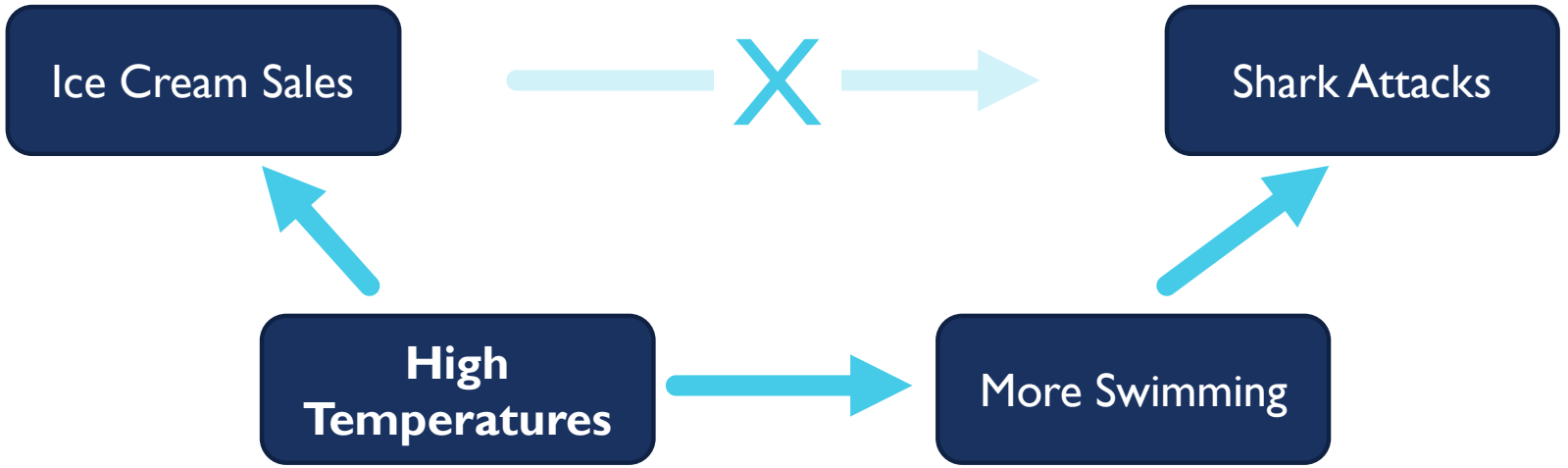
- Ice cream sales are positively correlated with shark attacks.
- Is it because we taste better with more ice cream?

CLASSIC EXAMPLE

- Ice cream sales are positively correlated with shark attacks.
- Is it because we taste better with more ice cream?
- What else may be causing this relationship?

CLASSIC EXAMPLE

- Ice cream sales are positively correlated with shark attacks.
- Is it because we taste better with more ice cream?
- What else may be causing this relationship?



CORRELATION VS. CAUSATION

- In some examples, there is an underlying factor that is related to both of the correlated variables.
- Not always the case:
  - Divorce rate in Maine and US consumption of margarine per person.
  - US consumption of mozzarella cheese (per person) and awarded PhD's in civil engineering.
  - Decrease in number of pirates and increase in global warming.
  - Many, many more…

SUMMARY

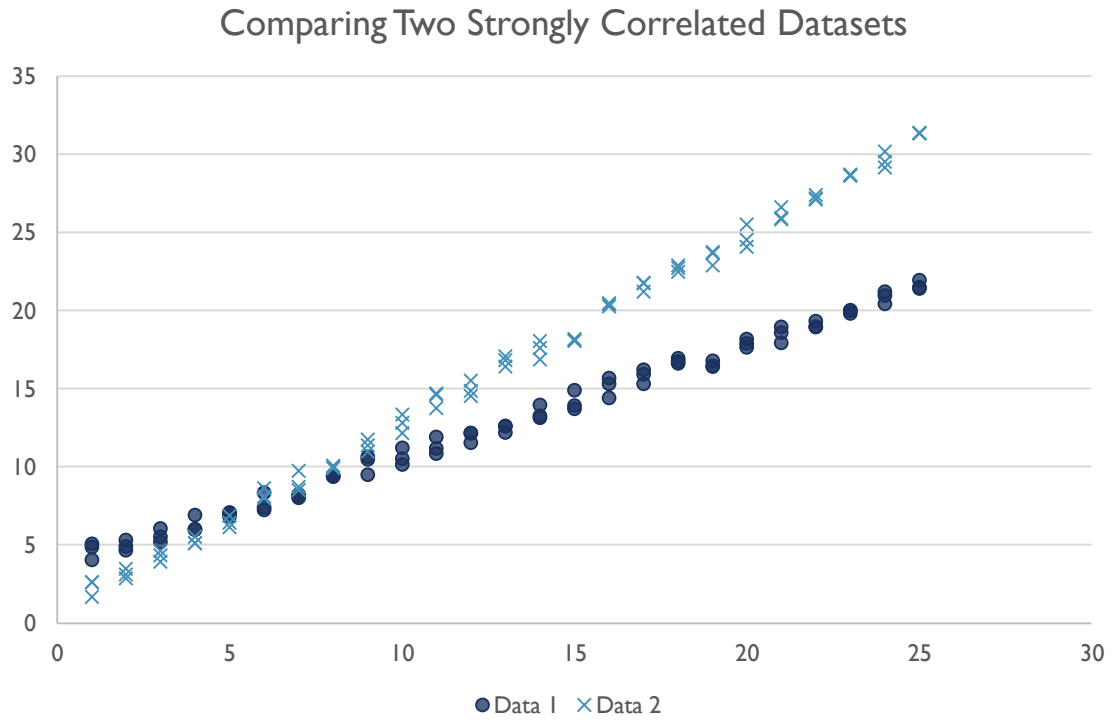
- Correlation does not imply causation.
- In some examples, there is an underlying factor that is related to both of the correlated variables, but not always the case.

IDEA OF REGRESSION

RELATIONSHIPS IN DATA

CORRELATION IS NOT EVERYTHING

- Correlation is a measure of strength of a linear relationship but does not say what the linear relationship is.
- Plot has two sets of data with exact same correlation of 0.99.
- However, the relationship is different between the two.



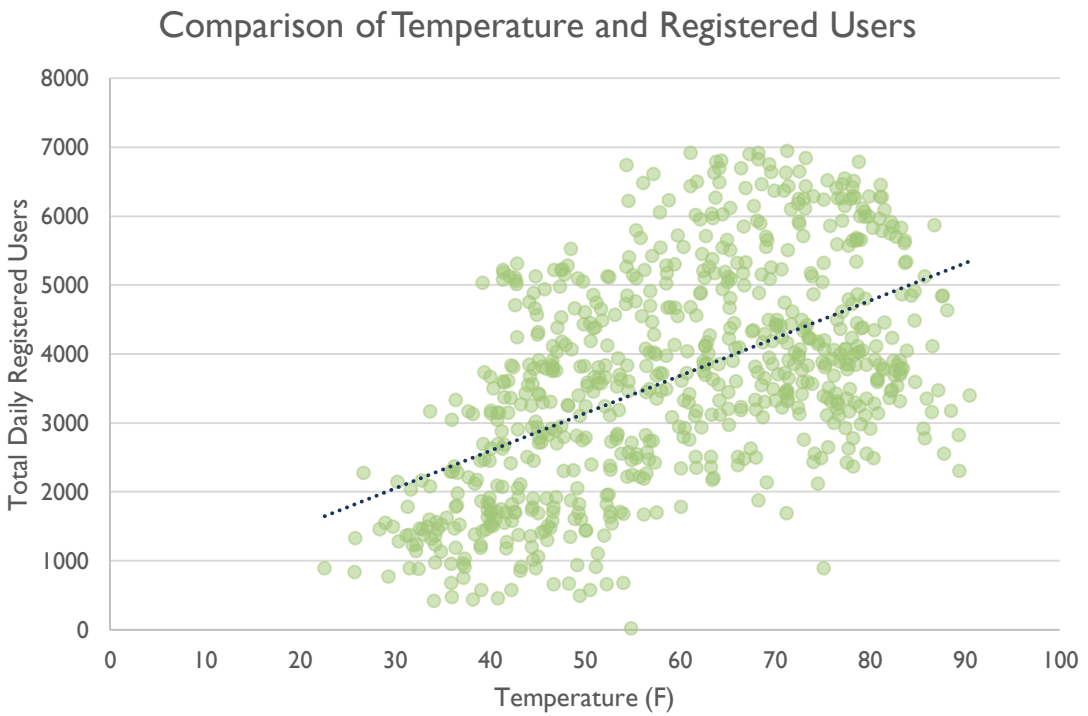
REGRESSION MODELING

- Many people across industries devote research funding to discover how variables are related (modeling).
- The simplest graphical technique to relate two quantitative variables is through a straight-line relationship – called the **simple linear regression** (**SLR**) **model**.
- Most models are more extensive and complicated than SLR models, but SLR models form a good foundation.

BIKE DATA EXAMPLE

- What if you wanted to predict the number of registered users based on the temperature outside?
- What is the best guess line for the following?

$$Predicted\ Users = \beta_0 + \beta_1 \times Temp$$



SIMPLE LINEAR REGRESSION MODEL

- Simple Linear Regression:

$$Predicted\ Users = \beta_0 + \beta_1 \times Temp$$

Intercept

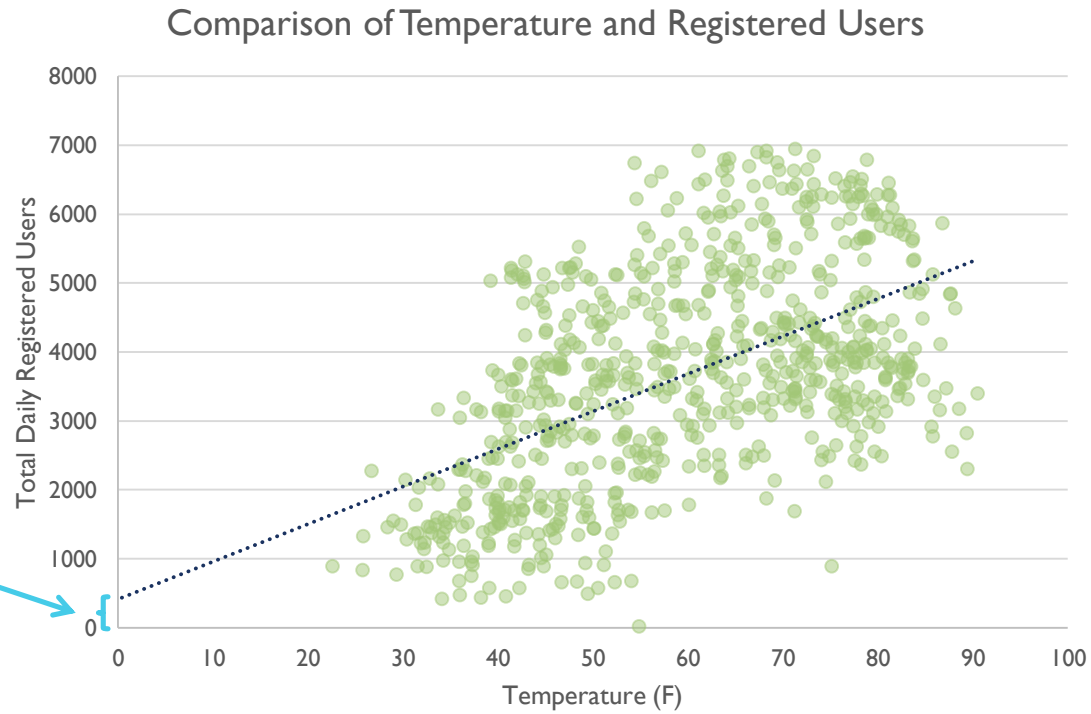
Slope

BIKE DATA EXAMPLE

- Simple Linear Regression:

$$Predicted\ Users = \beta_0 + \beta_1 \times Temp$$

Intercept

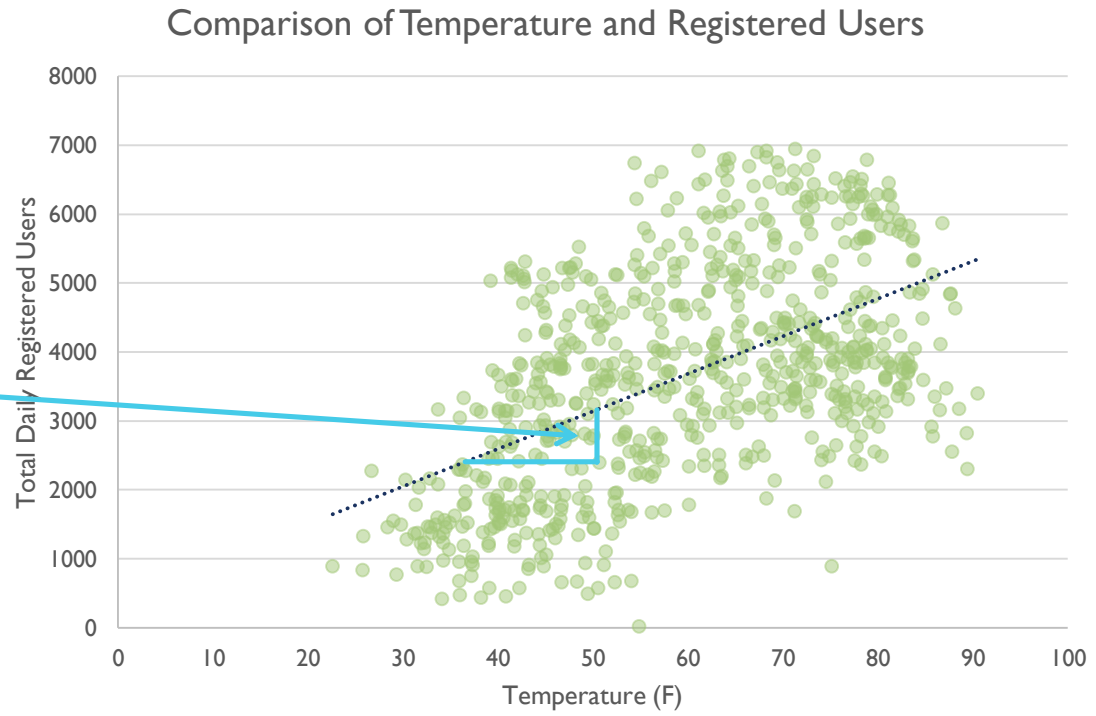


BIKE DATA EXAMPLE

- Simple Linear Regression:

$$Predicted\ Users = \beta_0 + \beta_1 \times Temp$$

Slope



SIMPLE LINEAR REGRESSION MODEL

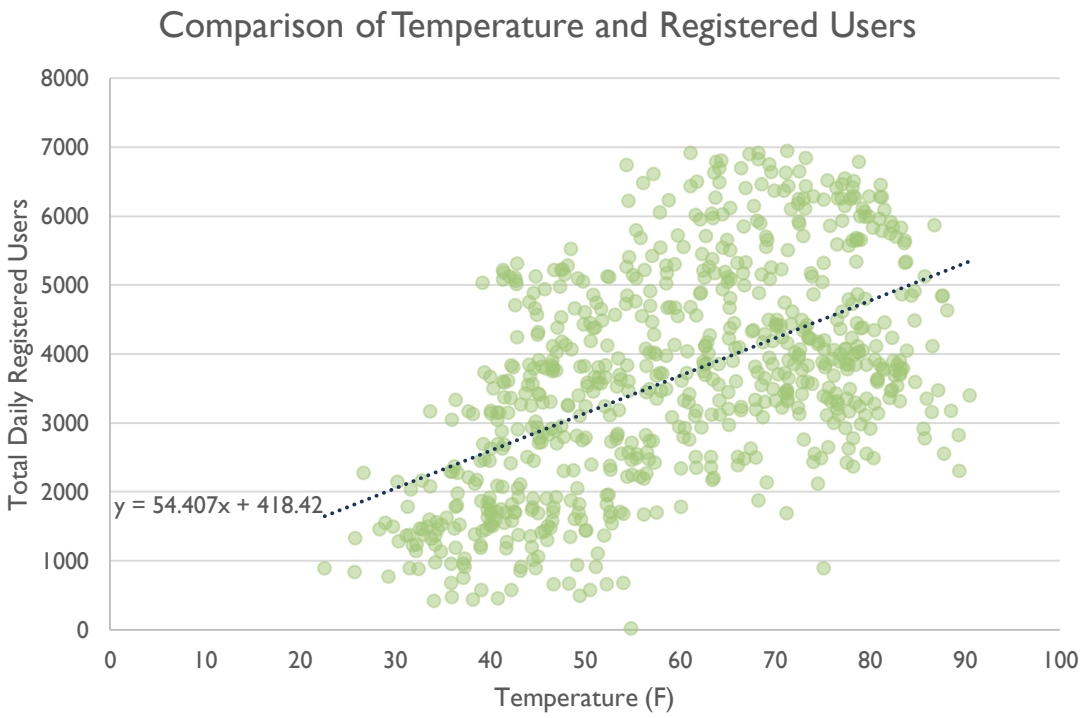
- The intercept is the value of the average of the registered users when the temperature equals zero.

- The slope is the **average** increase in the registered users with a one degree (F) increase in the temperature.

BIKE DATA EXAMPLE

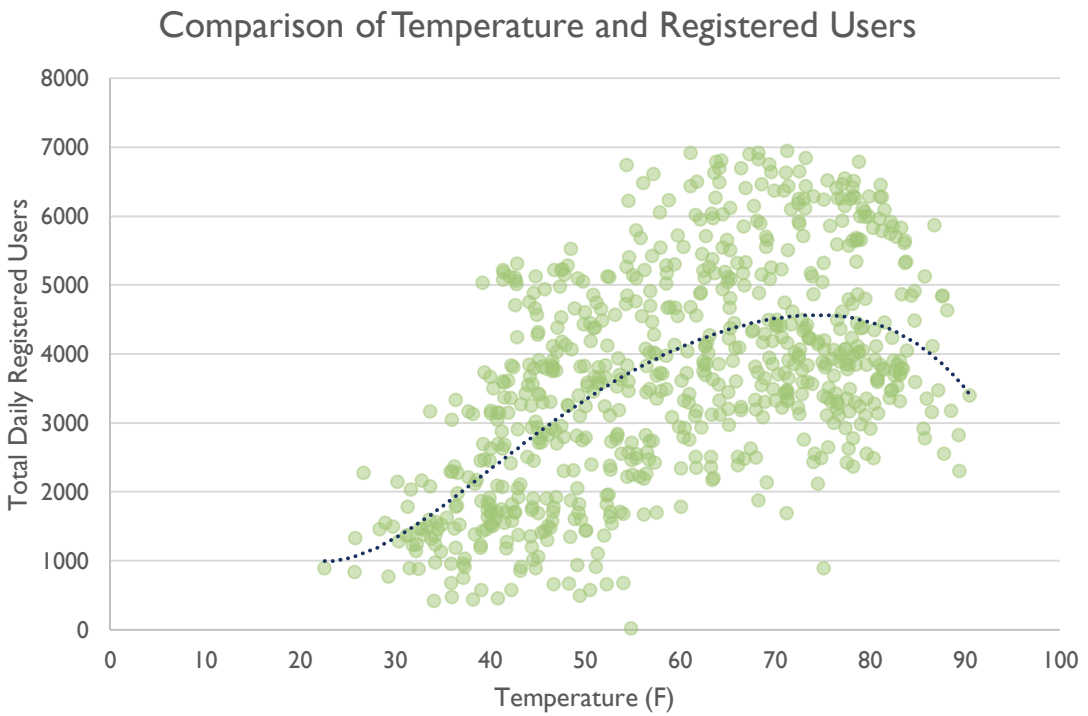
- What if you wanted to predict the number of registered users based on the temperature outside?
- What is the best guess line for the following?

$$Predicted\ Users = 418.42 + 54.4 \times Temp$$



MORE COMPLICATED MODELING

- Models can be more complicated than just straight-line relationships.
- Beyond the scope of this course.



SUMMARY

- Correlation is a measure of strength of a linear relationship but does not say what the linear relationship is.
- The simplest graphical technique to relate two quantitative variables is through a straight-line relationship – called the **simple linear regression (SLR) model**.
  - The intercept is the value of the average of the y-axis variable when the x-axis variable equals zero.
  - The slope is the **average** increase in the y-axis variable with a one-unit increase in the x-axis variable.