WHAT IS DATA?

ST101 – DR. ARIC LABARR

WHAT IS/ARE DATA?

data

noun

\ˈdā - tə\

factual information used as a basis for reasoning, discussion, or calculation

WHAT IS/ARE DATA?

data
noun

\ˈdā - tə\

factual **information** used as a basis for reasoning, discussion, or calculation

- Information – measurements or values describing an object, person, place, thing, etc.
- Examples:
  - Person – height, weight, age, race, spending habits, etc.
  - Car – mileage, gas mileage, color, motor size, etc.
  - Website – # of clicks, page views, ad revenue, etc.

WHAT IS/ARE DATA?

data

noun

\ˈdā - tə\

factual information used as a basis for **reasoning**, **discussion**, or **calculation**

- Inference – using information to come to some conclusion.
- Want to use the information to draw conclusions and make better decisions in the context of our problem.
- Who, what, where, when, why, how?

DATA TABLE

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

⋮

DATA TABLE

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

Observations

⋮

DATA TABLE STRUCTURE

- Rows in a data table typically denote **observations**.
  - Observations – individuals or objects that we are collected information about.

DATA TABLE

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

⋮

DATA TABLE STRUCTURE

- Rows in a data table typically denote **observations**.
  - Observations – individuals or objects that we are collected information about.
- Columns in a data table typically denote **variables**.
  - Variables – different characteristics that describe the observations.

TYPES OF VARIABLES

- There are two main types of variables:
  - Qualitative
  - Quantitative

TYPES OF VARIABLES

- There are two main types of variables:
  - **Qualitative** – data with a measurement scale inherently categorical.
  - Quantitative

QUALITATIVE (CATEGORICAL) 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

⋮

TYPES OF VARIABLES

- There are two main types of variables:
  - **Qualitative** – data with a measurement scale inherently categorical.
    - **Nominal** – categories with no logical ordering.
    - **Ordinal** – categories with a logical ordering.
  - Quantitative

QUALITATIVE (CATEGORICAL) 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

⋮

TYPES OF VARIABLES

- There are two main types of variables:
  - Qualitative – data with a measurement scale inherently categorical.
    - Nominal – categories with no logical ordering.
    - Ordinal – categories with a logical ordering.
  - **Quantitative** – data that are numeric and define a value or quantity.

QUANTITATIVE (NUMERICAL) 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

⋮

TYPES OF VARIABLES

- There are two main types of variables:
  - Qualitative – data with a measurement scale inherently categorical.
    - Nominal – categories with no logical ordering.
    - Ordinal – categories with a logical ordering.
  - **Quantitative** – data that are numeric and define a value or quantity.
    - Not all variables that are numeric are quantitative.
    - Examples – date, SSN, ZIP code, etc.
    - Need to be able to do basic arithmetic and remain meaningful.

QUANTITATIVE (NUMERICAL) 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

⋮

SUMMARY

- Data – factual information used as a basis for reasoning, discussion, or calculation.
- Data typically structured with data tables.
  - Rows – observations.
  - Columns – variables.
- Types of variables:
  - Qualitative – categorical.
  - Quantitative – numerical.

EXPLORING RELATIONSHIPS WITH DATA

WHAT IS DATA?

DATA INTO INSIGHTS

- Data by itself is just information.
- Need to draw insights from data to make decisions.
- Insights come from **exploring your data**.

EXAMPLE – BIKE RENTAL DATA

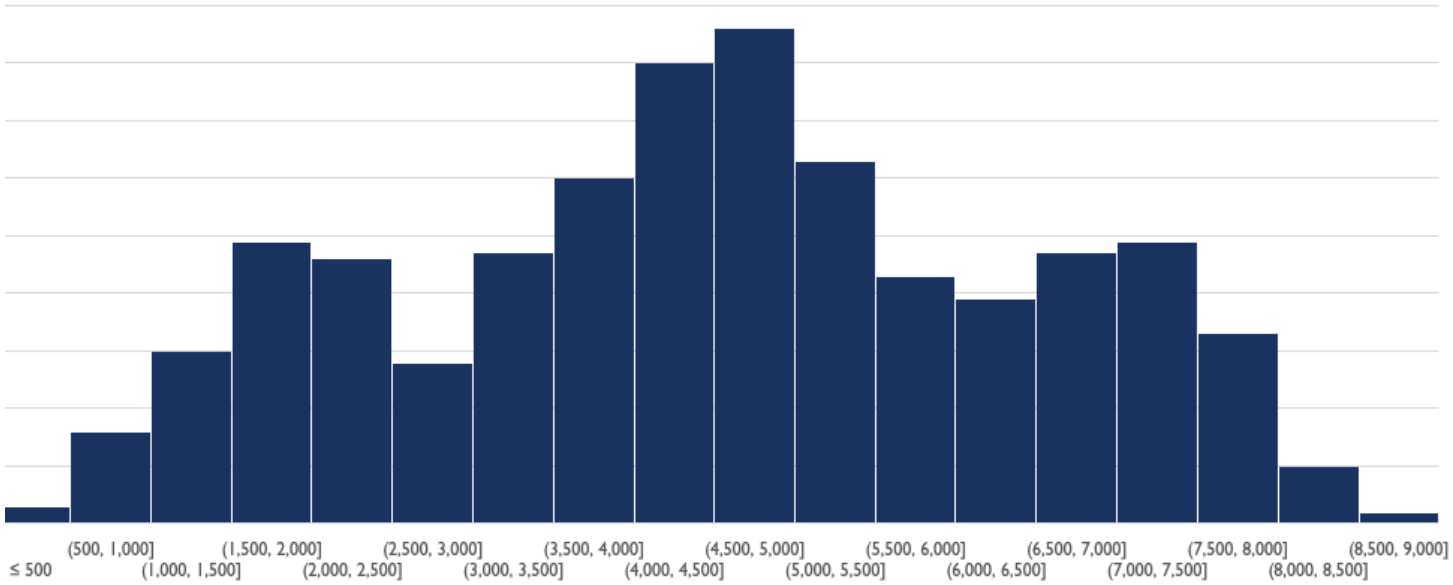
Date	Weekday	Season	Weather Type	Temperature (°F)	Humidity (%)	# Casual Users	# Registered Users
1/1/2011	Saturday	Winter	Misty	46.7	80.6	331	654
1/2/2011	Sunday	Winter	Misty	48.4	69.6	131	670
1/3/2011	Monday	Winter	Clear / Partly Cloudy	34.2	43.7	120	1229
1/4/2011	Tuesday	Winter	Clear / Partly Cloudy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

⋮

EXAMPLE – BIKE RENTAL DATA

- Historical average bike rentals is 4,000 per day.
- New employee sees low bike rental numbers over the first few days of the new year.
- Trouble?

EXAMPLE – BIKE RENTAL DATA



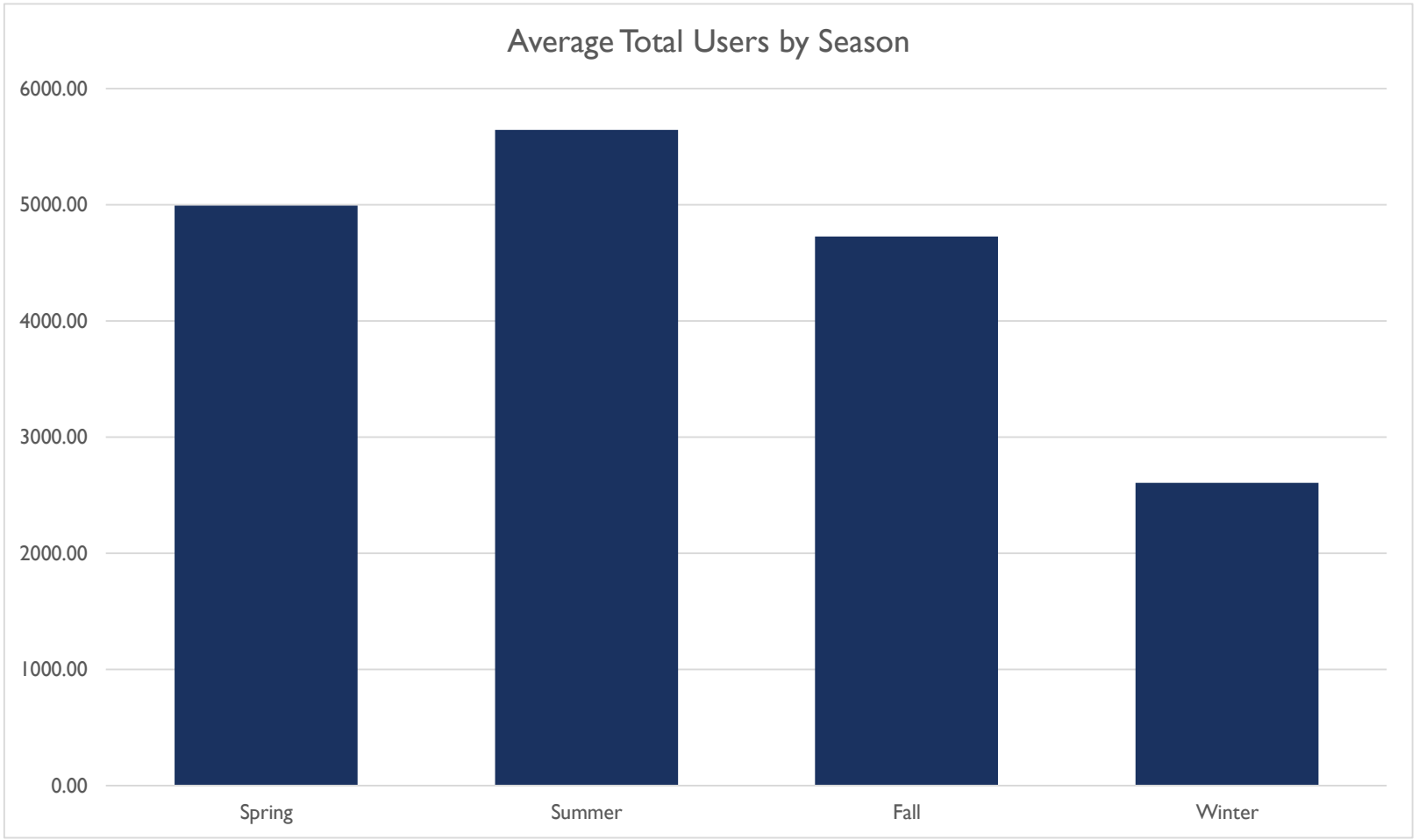
EXAMPLE – BIKE RENTAL DATA

- Historical **average** bike rentals is 4,000 per day.
- New employee sees low bike rental numbers over the first few days of the new year.
- Trouble? – Can look at the **distribution** of daily bike rentals.

EXAMPLE – BIKE RENTAL DATA

- Historical average bike rentals is 4,000 per day.
- New employee sees low bike rental numbers over the first few days of the new year.
- Maybe bike rentals drop in the winter?

EXAMPLE – BIKE RENTAL DATA



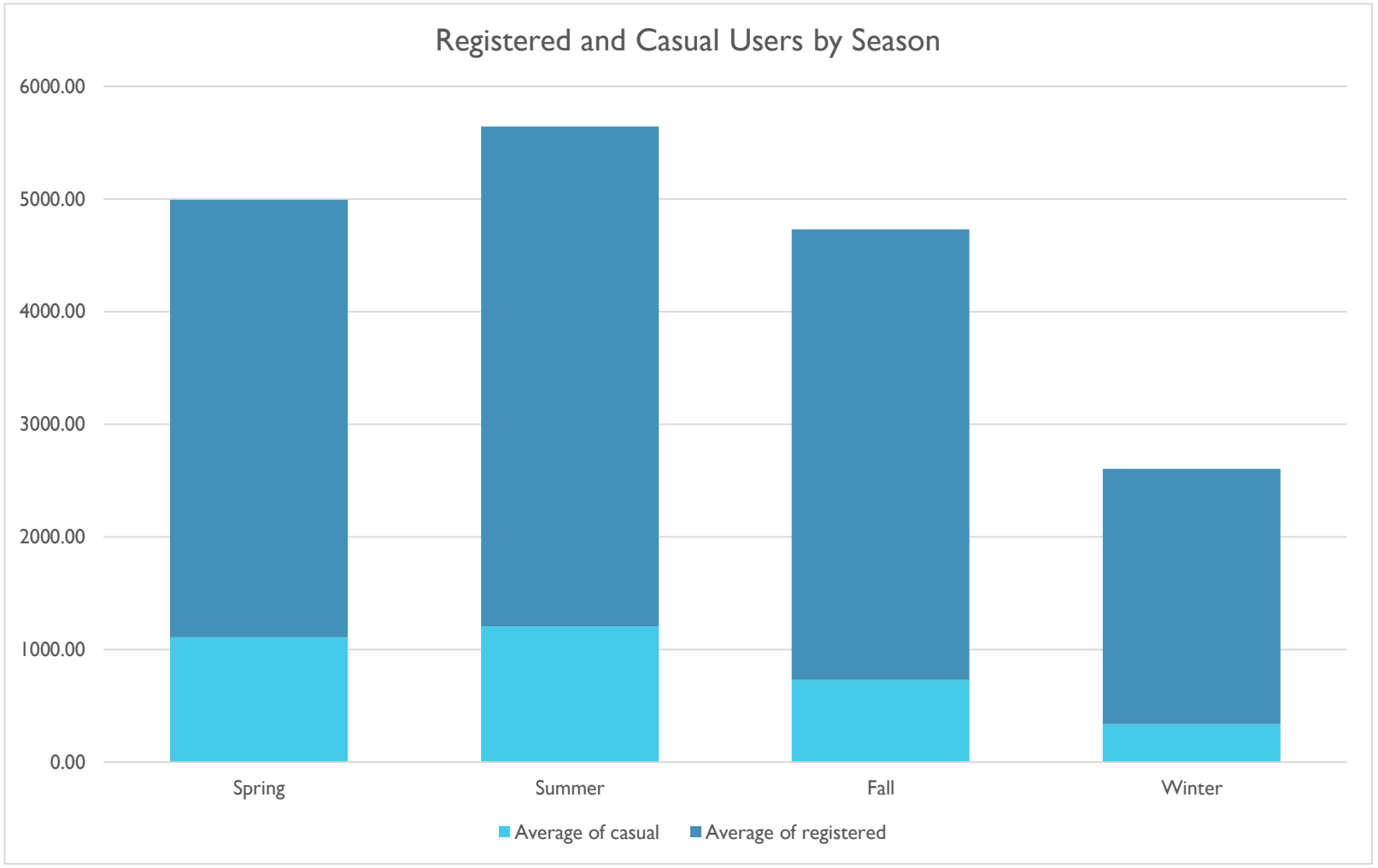
EXAMPLE – BIKE RENTAL DATA

- Historical average bike rentals is 4,000 per day.
- New employee sees low bike rental numbers over the first few days of the new year.
- Maybe bike rentals drop in the winter?
  - Can look at a **bar chart** of the data to see possible association.

EXAMPLE – BIKE RENTAL DATA

- Historical average bike rentals is 4,000 per day.
- New employee sees low bike rental numbers over the first few days of the new year.
- Very intriguing!
- Is the drop in winter the same for registered and casual users?

EXAMPLE – BIKE RENTAL DATA



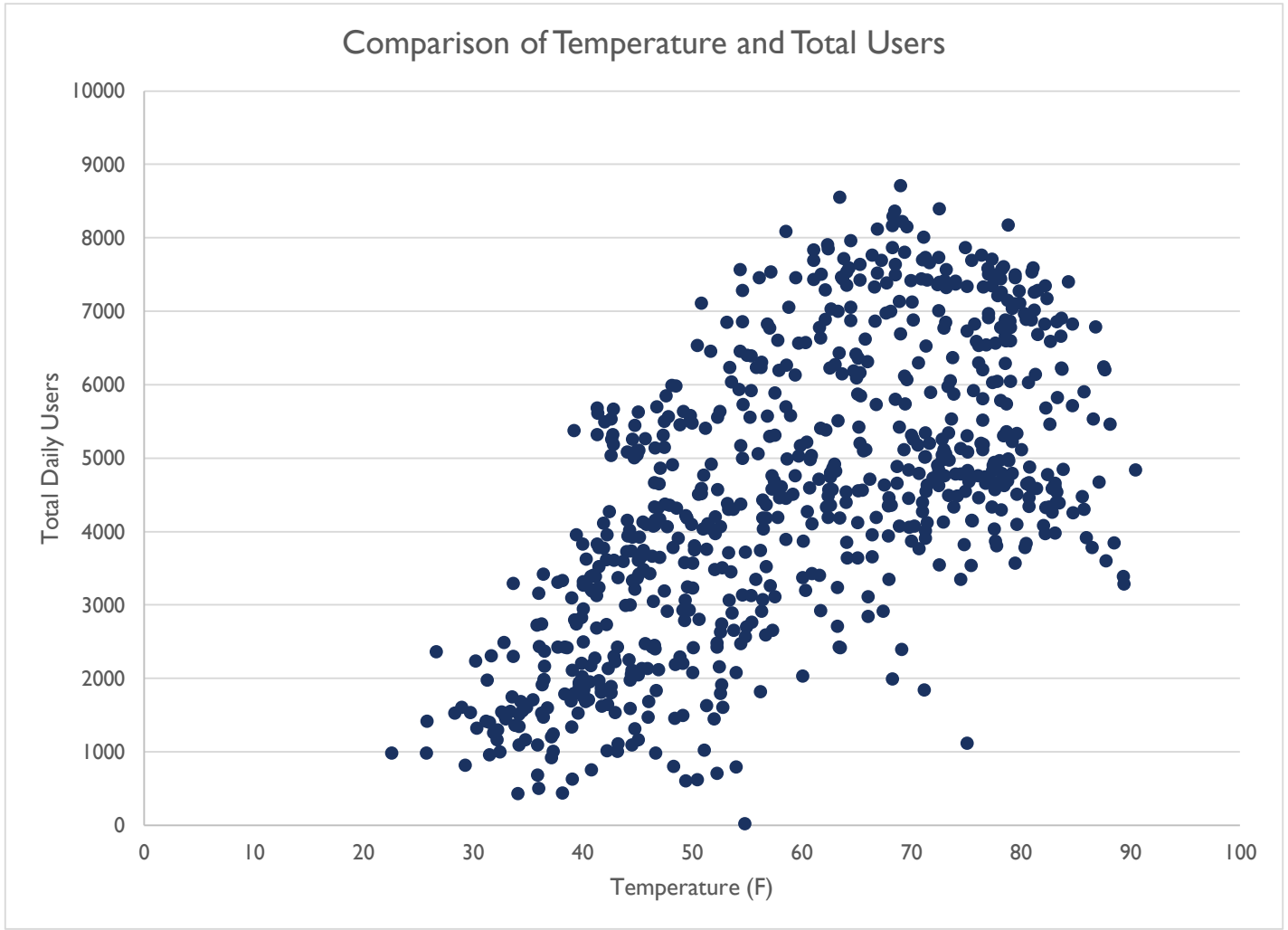
EXAMPLE – BIKE RENTAL DATA

- Historical average bike rentals is 4,000 per day.
- New employee sees low bike rental numbers over the first few days of the new year.
- Very intriguing!
- Is the drop in winter the same for registered and casual users?
  - Can look at **stacked bar chart** to see how users break down into registered and casual users.

EXAMPLE – BIKE RENTAL DATA

- Very intriguing!
- Tons of things revealed by data.
- Why do customers use bike rentals less in winter? Temperature?

EXAMPLE – BIKE RENTAL DATA



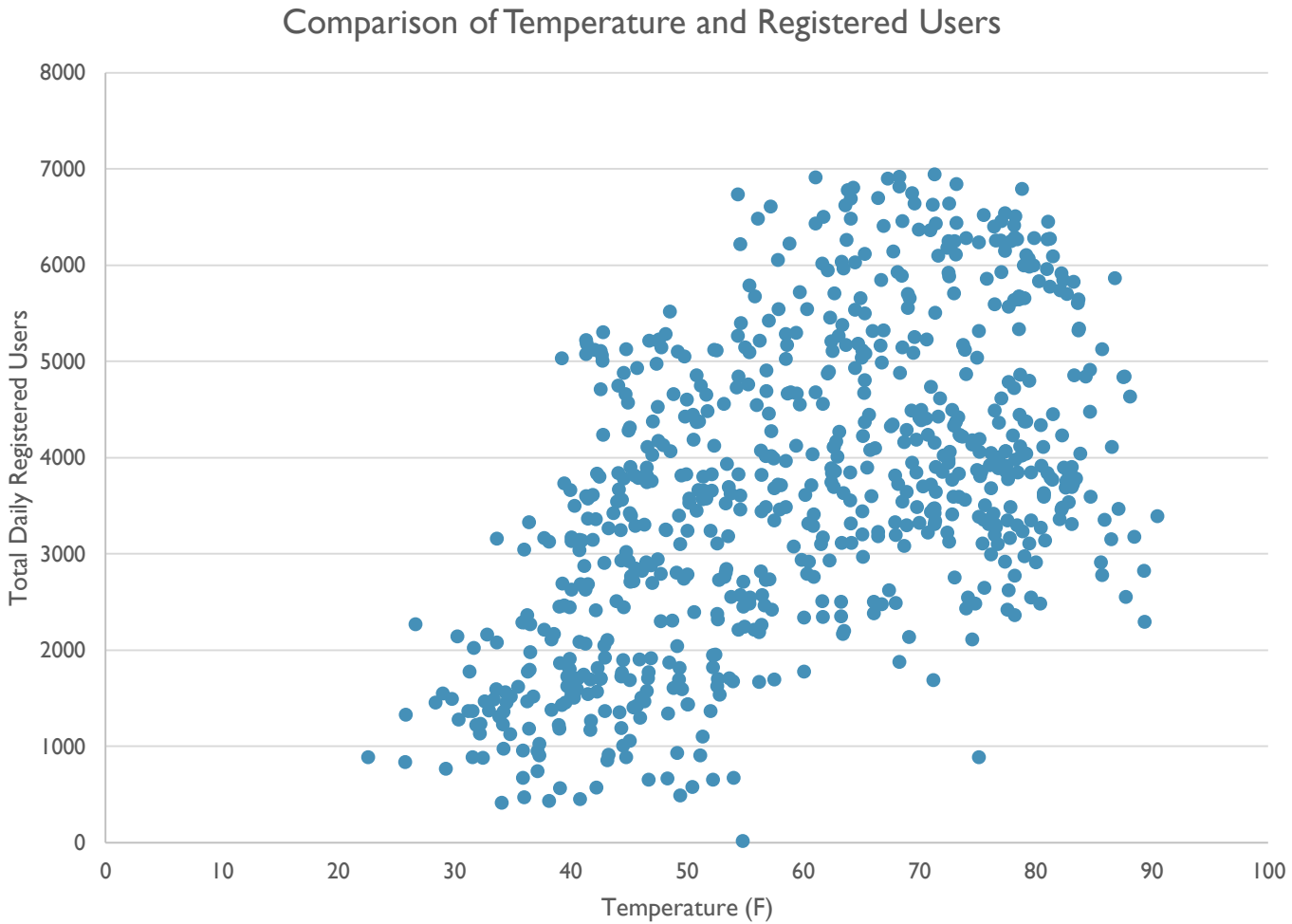
EXAMPLE – BIKE RENTAL DATA

- Very intriguing!
- Tons of things revealed by data.
- Why do customers use bike rentals less in winter? Temperature?
  - Can look at **scatterplot** of temperature and user count to see possible relationship.

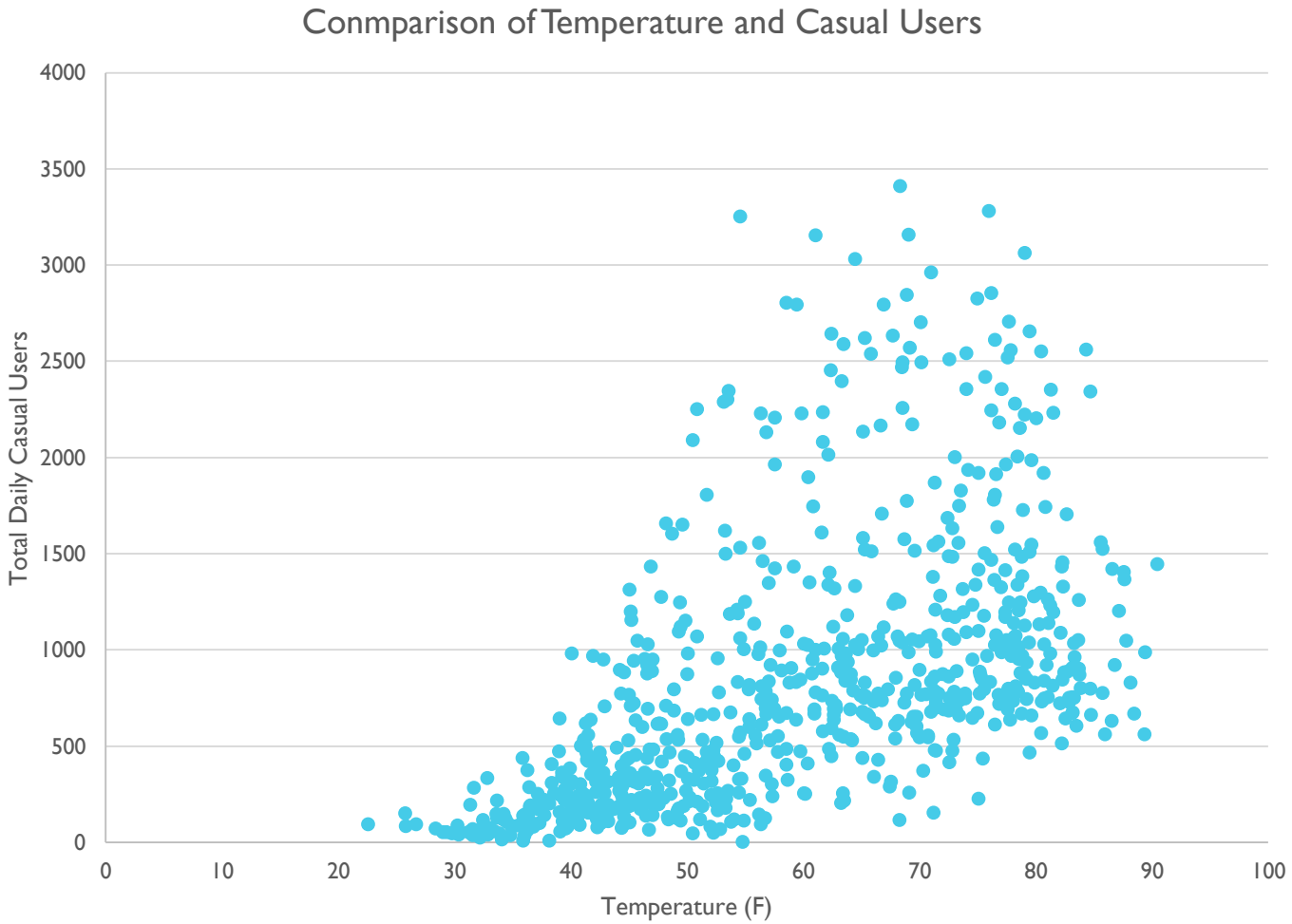
EXAMPLE – BIKE RENTAL DATA

- Very intriguing!
- Tons of things revealed by data.
- Why do customers use bike rentals less in winter? Temperature?
  - Can look at **scatterplot** of temperature and user count to see possible relationship.
  - What about registered or casual users?

EXAMPLE – BIKE RENTAL DATA



EXAMPLE – BIKE RENTAL DATA



SUMMARY

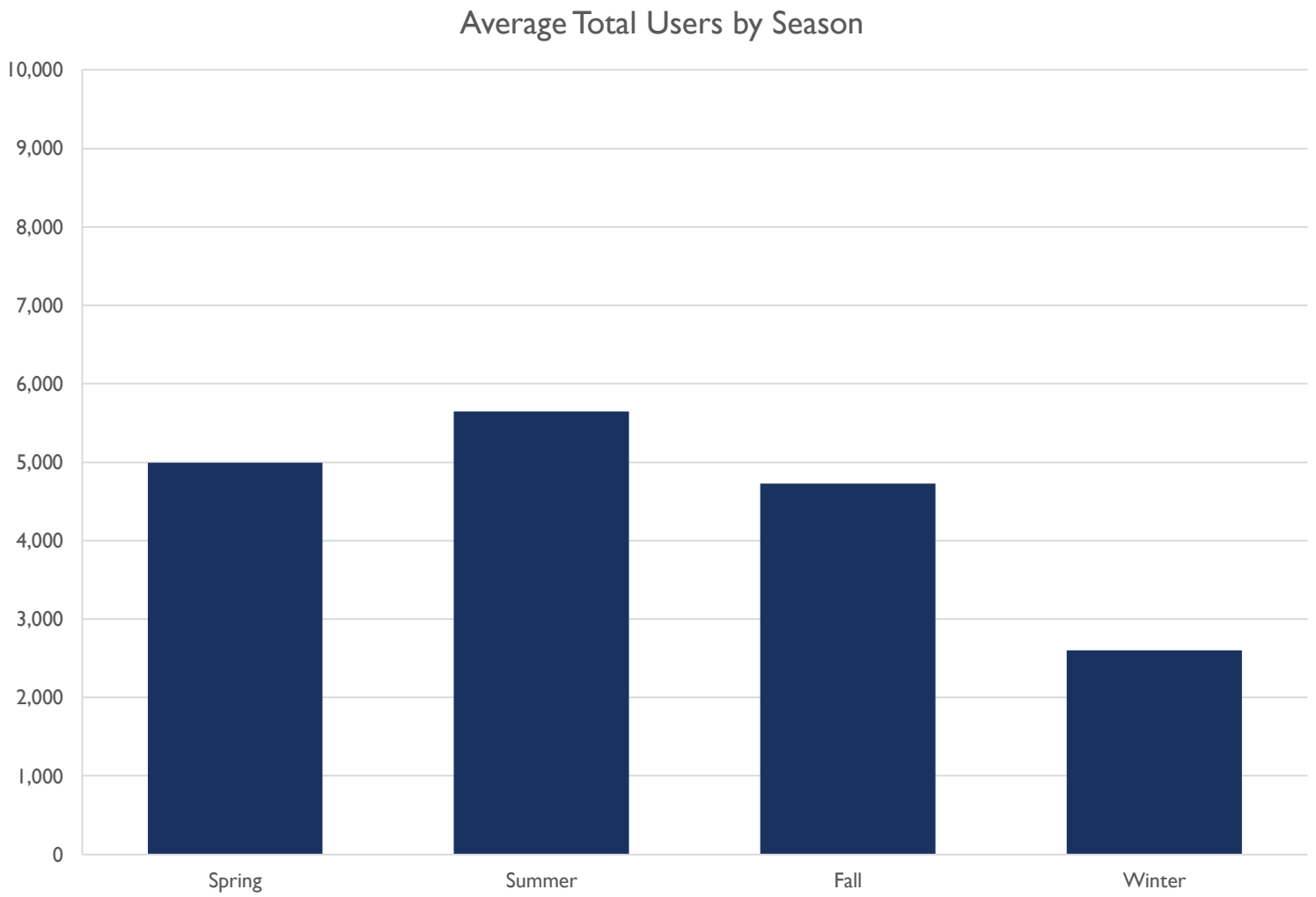
- Data by itself is just information.
- Exploring data reveals potential insights and valuable uses of that information.
- Visuals help explore data.
  - Distributions, bar charts, stacked bar charts, scatterplots, etc.

ASSOCIATION AND CORRELATION

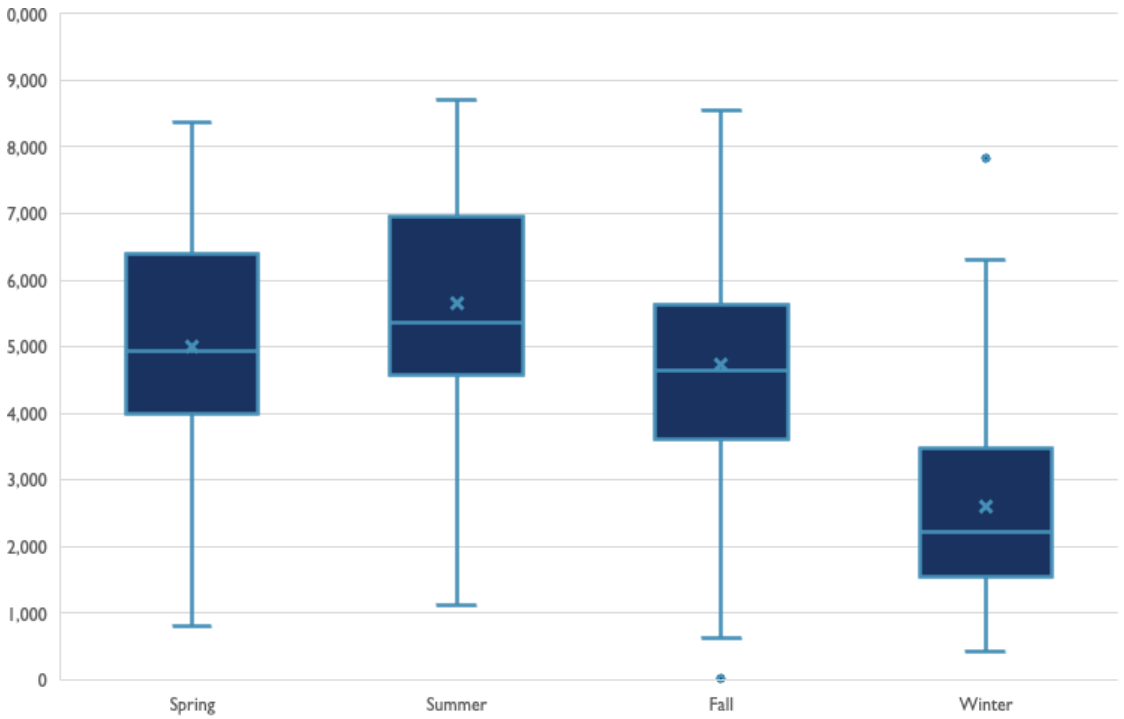
WHAT IS DATA?

EXPLORING DATA RELATIONSHIPS

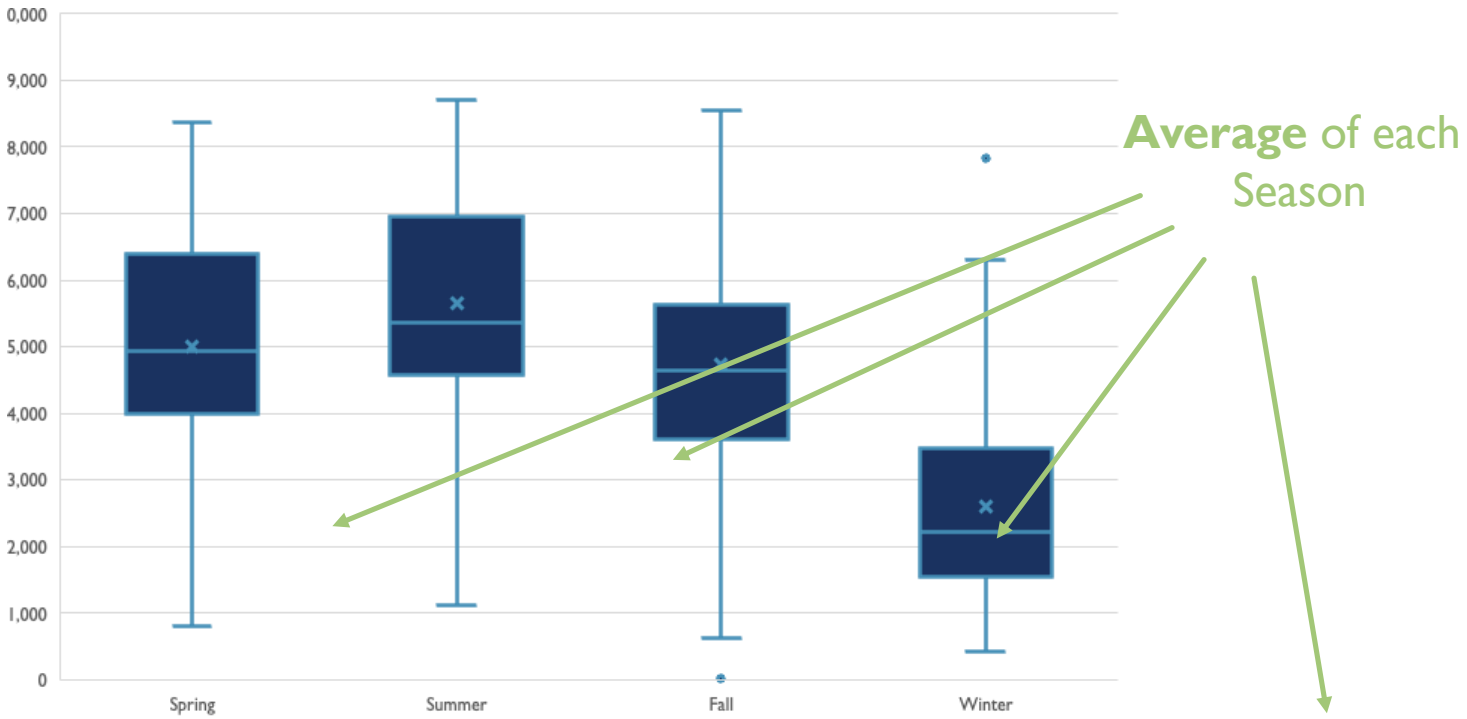
- Exploring data reveals potential insights and valuable uses of that information.
- Visuals help explore data.
  - Distributions, bar charts, stacked bar charts, scatterplots, etc.
- Are visuals enough?



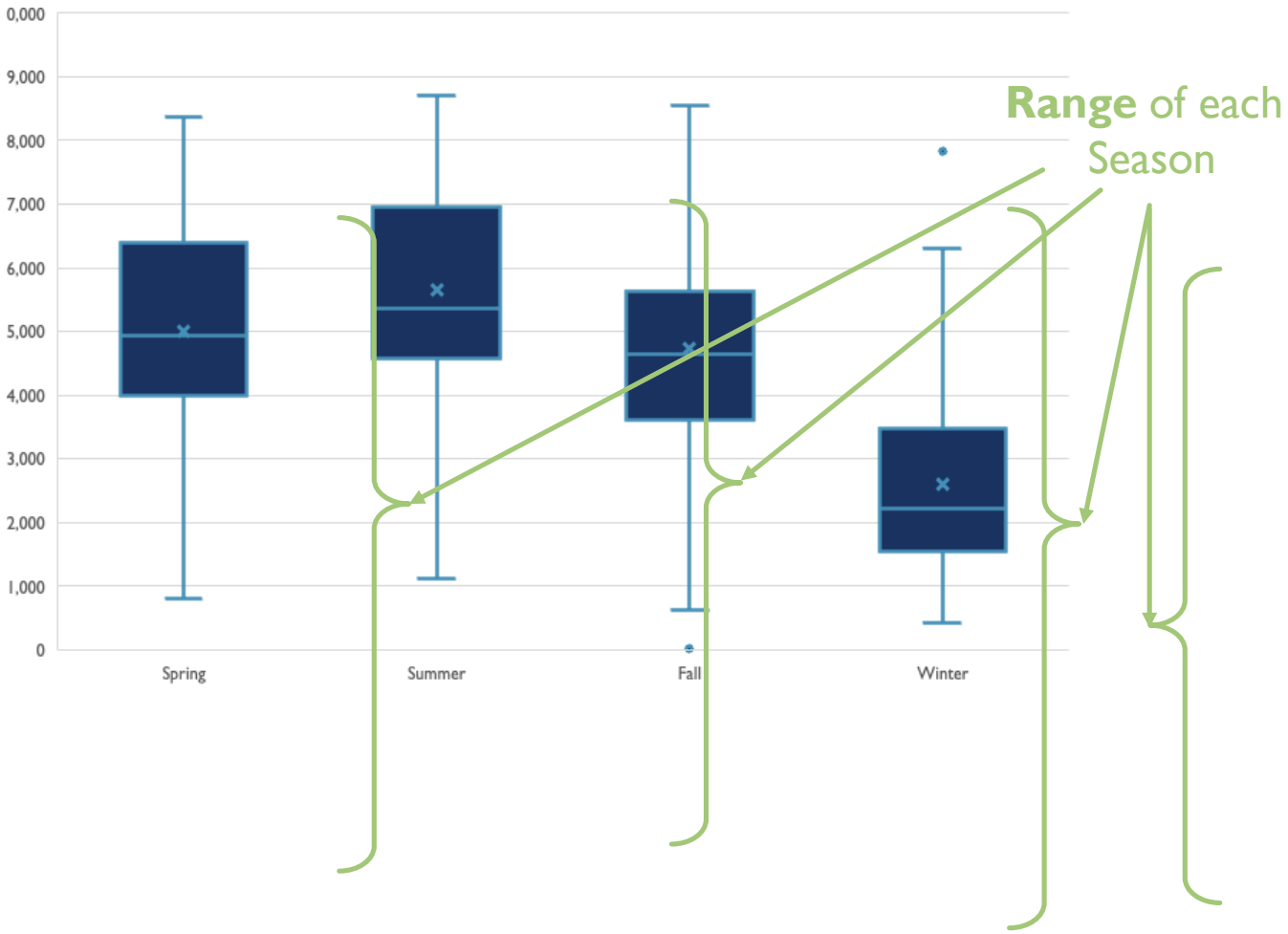
EXAMPLE – BIKE RENTAL DATA



EXAMPLE – BIKE RENTAL DATA



EXAMPLE – BIKE RENTAL DATA



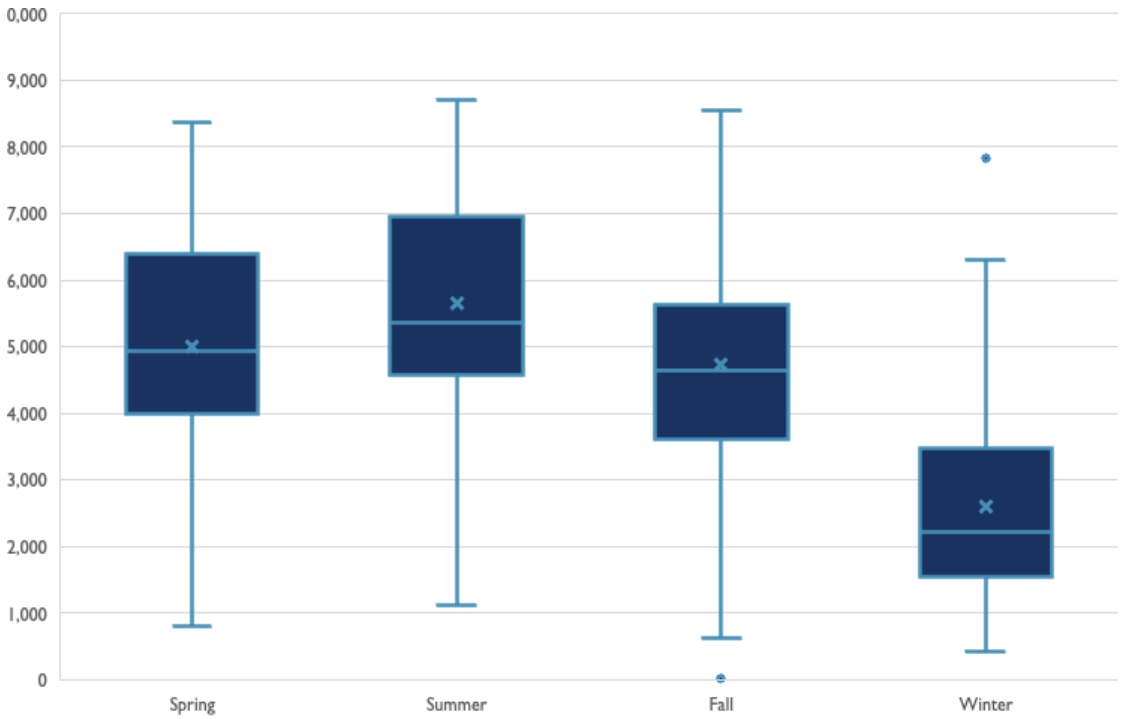
EXAMPLE – BIKE RENTAL DATA

VARIATION

- One of the most important concepts in statistics is **variation**.
- Data points vary from one to another and that is expected:
  - Don't see everything.
  - Measure things imperfectly.

VARIATION

- Why does one day in summer different than another? What if temperature is the same?
  - Can't be perfectly sure why days are different.
  - Differences are expected by apparent **randomness**.
- Are the seasons truly different? Or could there be some expected random variation?



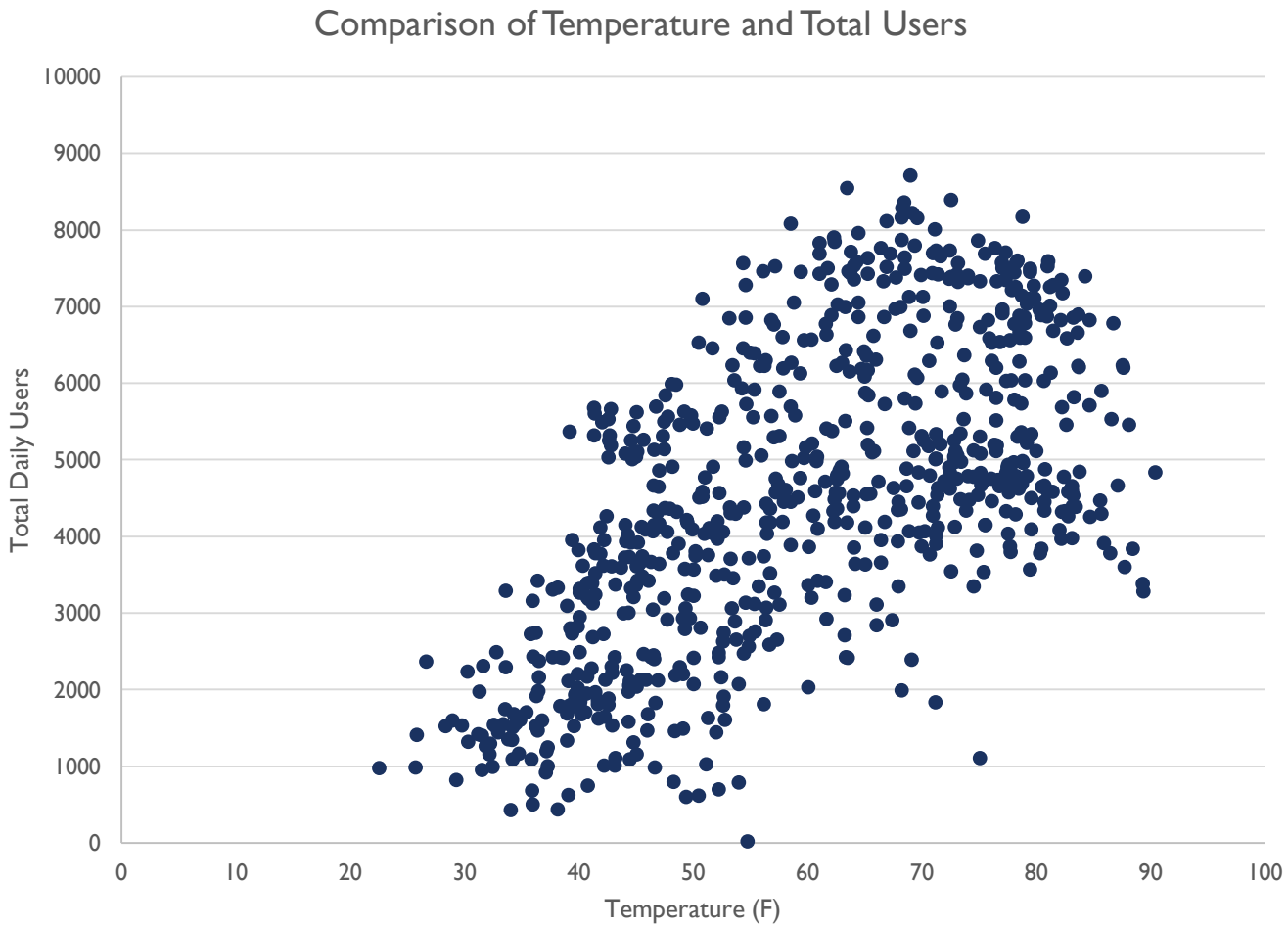
EXAMPLE – BIKE RENTAL DATA

STATISTICAL TESTING

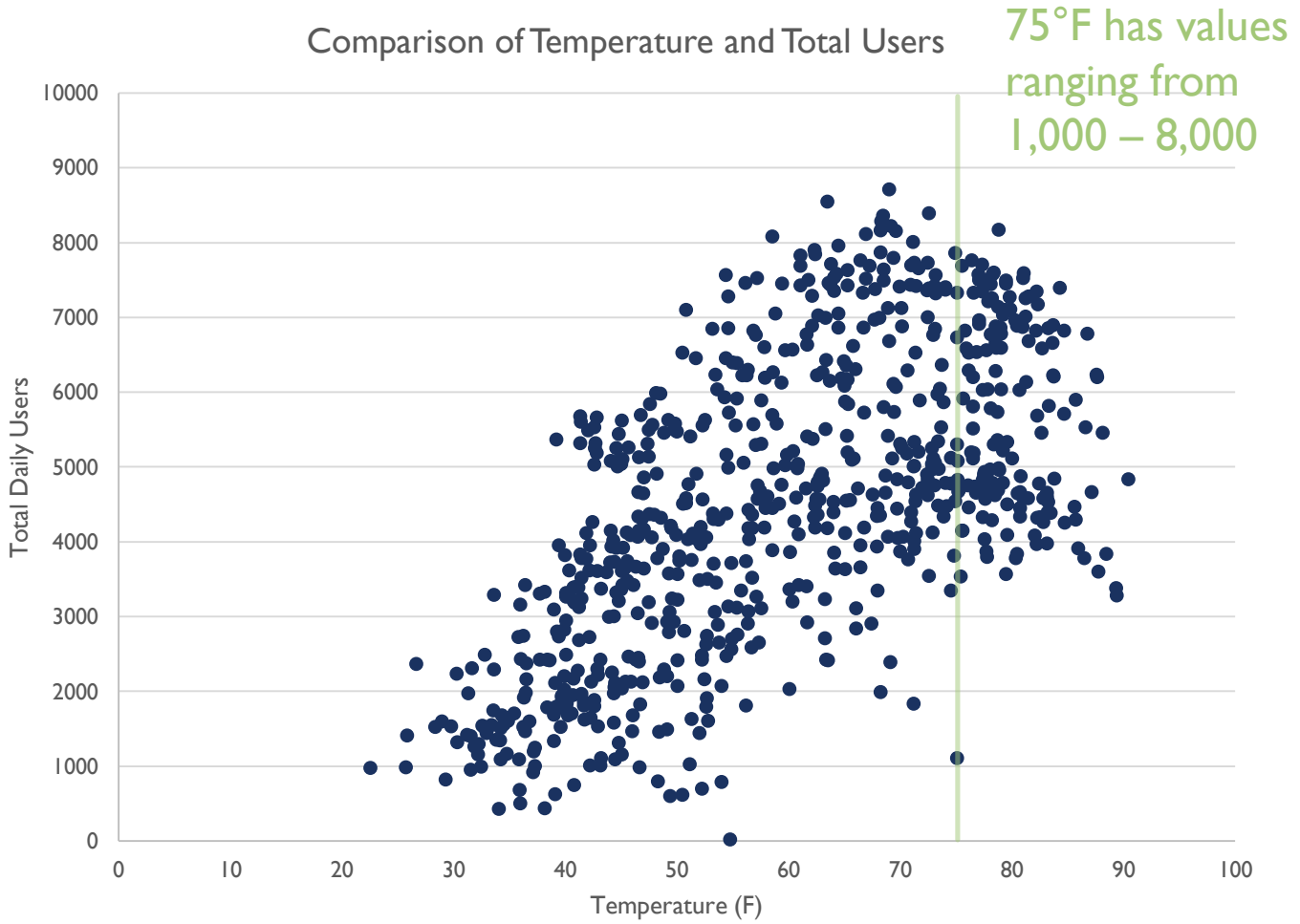
- Statistical **hypothesis testing** can help answer these questions.
  - Use the data available to see if the differences we see are expected due to just random variations or if we can say there is an **association** between season and number of users.
- An **association** is a statistical relationship between a qualitative and quantitative variable.

STATISTICAL TESTING USEFULNESS

- Examples:
  - Did the previous marketing campaign bring in more customers?
  - Did that drug treatment help the patient?
  - Did our program for veterans help them find jobs after they left the service?



EXAMPLE – BIKE RENTAL DATA



EXAMPLE – BIKE RENTAL DATA

STATISTICAL CORRELATION

- Variation occurs when looking for relationships between two quantitative variables as well.
- Use the data available to see if the differences we see are expected due to just random variations or if we can say there is a **correlation** between temperature and number of users.
- A **correlation** is a statistical *linear* relationship between two quantitative variables.
  - Stronger the correlation → stronger the *linear* relationship.

SUMMARY

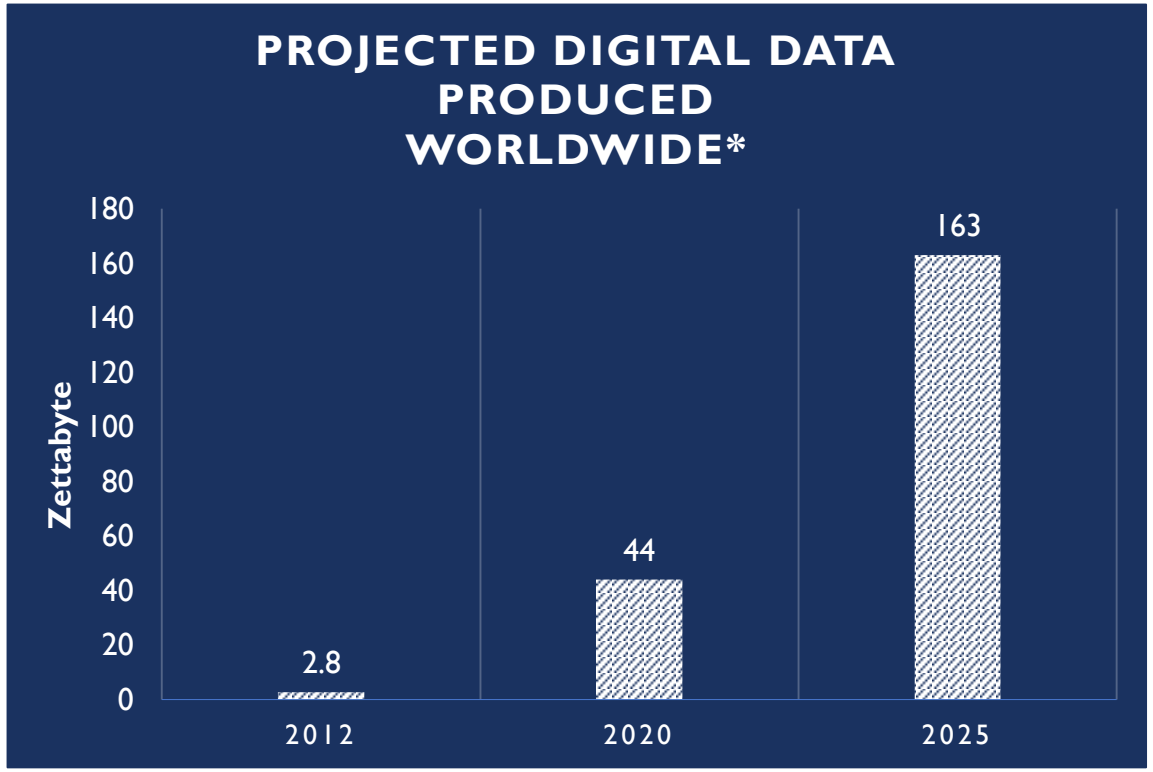
- Data has natural and expected variation.
- Some of this variation could be due to apparent randomness.
- Statistical testing can help evaluate if the variation is random or intentional.
  - An association is a statistical relationship between a qualitative and quantitative variable.
  - A correlation is a statistical *linear* relationship between two quantitative variables.

DATA IN THE WORLD AROUND US

WHAT IS DATA?

WORLDWIDE DATA PRODUCTION

- Data is everywhere.
- 1 zettabyte = 1,000,000,000,000 GB
- If you were to fill the latest smart phone full of data…
  stacked them end to end…
  they would go to the moon…
  and back to Earth…
  and back to the moon again!



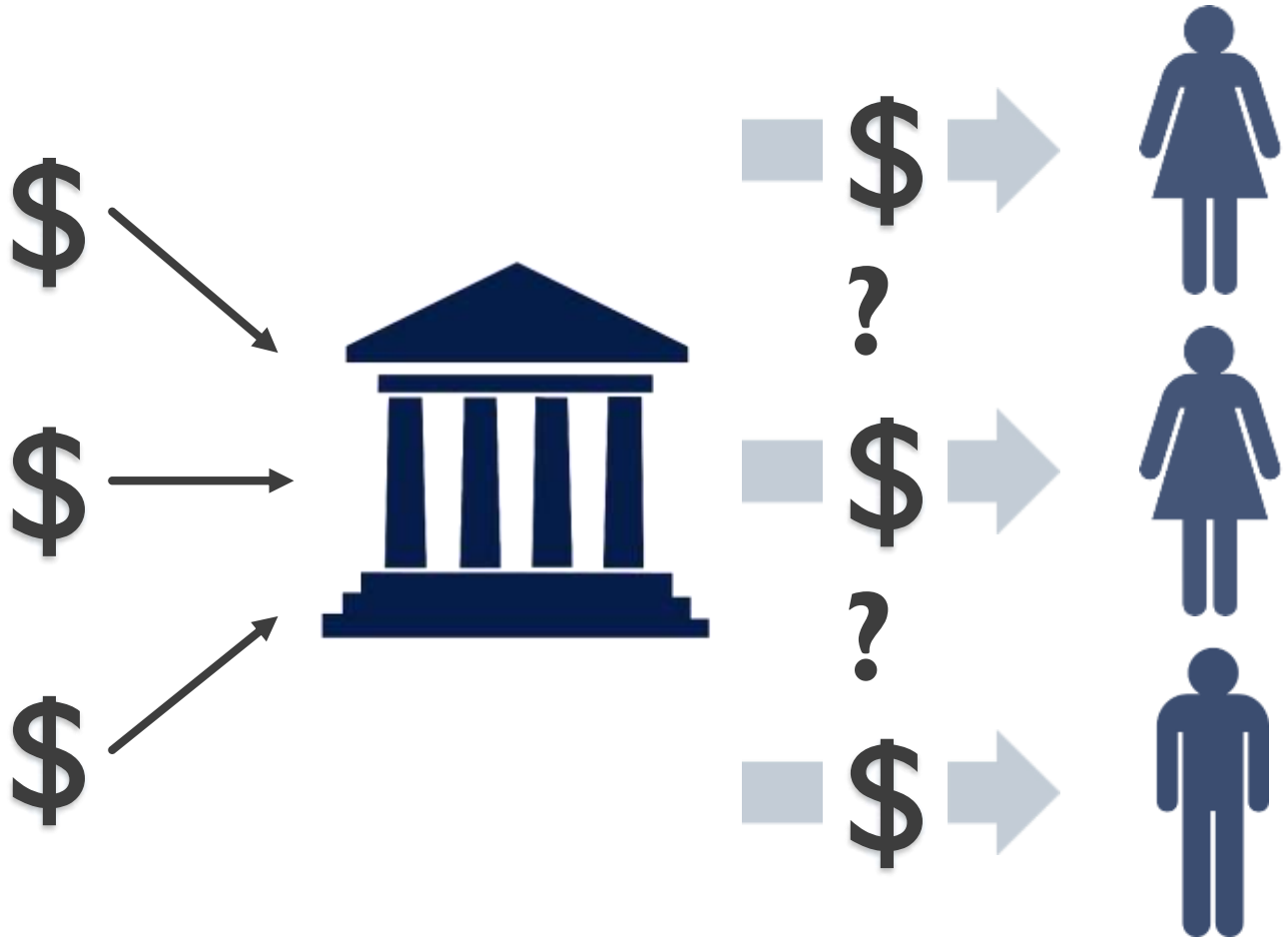
*IDC *Digital Universe*

DATA IS EVERYWHERE

- Data exists in all types of industries – only depends on how it is used!
- Banking / Finance
- Marketing
- Healthcare
- Supply Chain / Agriculture

BANKING / FINANCE

- Who do banks give loans to?
- Microfinance banks help impoverished and developing nations through small business loans.
- Use data modeling to help find which clients would be best to loan money to at the least risk.



MARKETING

- Who do you advertise to?
- How do you advertise to them?
- Marketing companies use statistical hypothesis tests to compare effectiveness of different campaigns.
- Data about customer purchases helps companies group customers into similar buying habits.

Social influencer vector created by storyset - www.freepik.com

HEALTHCARE

- Healthcare costs are increasing.
- How can we make healthcare more efficient without losing quality of care?
- Hospitals and medical agencies use data to help identify onset of disease sooner, determine who is at higher risk for hospital readmission, and provide more specialized care for patients.

Health data vector created by katemangostar - www.freepik.com

SUPPLY CHAIN / AGRICULTURE

- How do we use food more efficiently?
- Agriculture companies use data to efficiently track food from their seed, to the field, to the store, and to the table.
- This helps keep food fresher longer.

SUMMARY

- Data exists in all types of industries – only depends on how it is used!
- Knowledge of data and its usefulness is helpful in all aspects of life, not just in the career you are going.