Part 1 - Exploring and Understanding Data

Chapter 2
Data

Obj - SWBAT describe a variable in terms of its *Who, What, When, Where, Why,* and *How* and be prepared to remark when that information is not provided.

Statistics is the science of data. The volume of data available to us is overwhelming.

- The Census Bureau's American Community Survey collects data from 3,000,000 housing units each year.
- Astronomers work with data on tens of millions of galaxies.
- The checkout scanners at Walmart's 6500 stores in 15 countries record hundreds of millions of transactions every week.

What Are Data?

- Data can be numbers, record names, or other labels.
- Not all data represented by numbers are numerical data (e.g., 1 = male, 2 = female).
- Data are useless without their context…

The "W's"

- To provide context we need the W's
- Who
- What (and in what units)
- When
- Where
- Why (if possible)
- and How of the data.

When collecting and organizing data, there are many aspects that must be taken into account:

Case: an individual about whom or which we have data.

Population: all the cases we wish we knew about.

Sample: The cases we actually examine in seeking to understand the much larger population.

Variable: holds information about the same characteristic for many cases.

Units: a quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.

Data Tables

- The following data table clearly shows the context of the data presented:

Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
Katharine H.	Ohio	10.99	440	Nashville	N	B00000I5Y6	Kansas
Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Garbage

Who

- The Who of the data tells us the individual cases for which (or whom) we have collected data.
- Individuals who answer a survey are called respondents.
- People on whom we experiment are called subjects or participants.
- Animals, plants, and inanimate subjects are called experimental units.

Who (cont.)

- Sometimes people just refer to data values as observations and are not clear about the Who.
- But we need to know the Who of the data so we can learn what the data say.

What and Why

- Variables are characteristics recorded about each individual.
- The variables should have a name that identify What has been measured.
- To understand variables, you must Think about what you want to know.

What and Why (cont.)

- Some variables have units that tell how each value has been measured and tell the scale of the measurement.

The International System of Units links together all systems of weights and measures by international agreement. There are seven base units from which all other physical units are derived:

• Distance	Meter
• Mass	Kilogram
• Time	Second
• Electric current	Ampere
• Temperature	Kelvin
• Amount of substance	Mole
• Intensity of light	Candela

Categorical and Quantitative Variables

Categorical variable: a variable that names groups or categories.
examples:

Quantitative variable: numbers that act as numerical values. This variable always has units.
examples:

What and Why (cont.)

- A categorical (or qualitative) variable names categories and answers questions about how cases fall into those categories.

- A quantitative variable is a measured variable (with units) that answers questions about the quantity of what is being measured.

What and Why (cont.)

- The questions we ask a variable (the Why of our analysis) shape what we think about and how we treat the variable.

What and Why (cont.)

- Example: In a student evaluation of instruction at a large university, one question asks students to evaluate the statement “The instructor was generally interested in teaching” on the following scale:

1 = Disagree Strongly; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Agree Strongly.

- Question: Is interest in teaching categorical or quantitative?

Counts Count

- When we count the cases in each category of a categorical variable, the counts are not the data, but something we summarize about the data.
- The category labels are the What, and
- the individuals counted are the Who.

Shipping Method	Number of Purchases
Ground	20,345
Second-day	7,890
Overnight	5,432

Counts Count (cont.)

- When we focus on the amount of something, we use counts differently. For example, Amazon might track the growth in the number of teenage customers each month to forecast CD sales (the *Why*).

Month	Number of Teenage Customers
January	123,456
February	234,567
March	345,678
April	456,789
May	. . .
. . .	. . .

Identifying Identifiers

- Identifier variables are categorical variables with exactly one individual in each category.
- Examples: Social Security Number, ISBN, FedEx Tracking Number
- Don’t be tempted to analyze identifier variables.
- Be careful not to consider all variables with one case per category, like year, as identifier variables.
- The Why will help you decide how to treat identifier variables.

Where, When, and How

- We need the *Who, What,* and *Why* to analyze data. But, the more we know, the more we understand.
- *When* and *Where* give us some nice information about the context.

Where, When, and How (cont.)

- How the data are collected can make the difference between insight and nonsense.
- Example:
- The first step of any data analysis should be to examine the W's—this is a key part of the Think step of any analysis.
- And, make sure that you know the Why, Who, and What before you proceed with your analysis.

What Can Go Wrong?

- Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.
- Just because your variable's values are numbers, don't assume that it's quantitative.
- Always be skeptical—don't take data for granted.

What have we learned?

What have we learned? (cont.)

Ex. 1: A *Consumer Reports* article about 98 HDTVs lists each set's manufacturer, cost, screen size, type (LCD, plasma, or rear projection), and overall performance score (0-100). Are these variables categorical or quantitative?

Ex. 2: Jake is a car buff who wants to find out more about the vehicles that students at his school drive. He gets permission to go to the student parking lot and record some data. Later, he does some research about each model of car. Finally, Jake makes a spreadsheet that includes each car's model, year, color, number of cylinders, gas mileage, weight, and whether it has a navigation system. Who are the individuals in Jake's study? Identify the variables measured and whether they are categorical or quantitiative.