What is Data?

Andrew Grogan-Kaylor, PHD
agrogan@umich.edu
www.umich.edu/~agrogan

A **data set** is nothing more than a series of rows and columns that contain answers to responses to a survey. ***Usually***, data is in *rectangular* format.

- Rows are *usually* used for **individuals** (although sometimes rows are larger social units like cities or states), while columns indicate the **questionnaire answers**, or **measures**, from those people.
- Answers to questions are often given numerical responses (e.g. "no" is frequently coded as "0" and "yes" is frequently coded as "1")

Person	Q1*	Q2	Q3
1	1	0	$100
2	2	0	$200
3	1	1	-9

Survey (documentation):

Question 1:

What neighborhood do you live in?

0 – Neighborhood A
1 – Neighborhood B
2 - Other Neighborhood:
____ (please indicate)
-9 – don't know/refused

…

Question 3:

What is your income? _____(value)
-9 – don't know/refused

In working through our research questions, we'll constantly be going back and forth between the **actual data** (to see the pattern of responses) and the **documentation**, to figure out the actual question asked as well as how the different responses are coded.

This cell of the table has a **negative number**. Frequently negative numbers are used to indicate what are called **"missing values"**. A missing value is a response like "don't know" or "refused to answer" or "did not answer". Before we start doing calculations with our data, we'll want to change negative numbers to true missing values (usually symbolized by a ".", or sometimes by "NA", so that they don't goof up our calculations.

* Often in a spreadsheet, you'll see the full text of a question written out (e.g. "What neighborhood do you live in"?) Most programs that work with data are going to want **abbreviations** (e.g. "Q1" or "neighborhood") for the questions. These abbreviations should usually have no spaces and be 8 characters or less.