What Is Data?

Andy Grogan-Kaylor

2023-09-23

Table of contents

1 Introduction

A data set is nothing more than a series of rows and columns that contain answers to responses to a survey:

- *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”)

2 Simulated Data

person	neighborhood	age	income
1	A	34	49703
2	B	NA	52146
3	B	37	49524
4	B	40	49387
5	B	38	NA
6	B	45	50925

3 Both The Data And Documentation Are Useful

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.

> **Hypothetical Survey**
>
> **Question 1** What neighborhood do you live in?
> A - Neighborhood A
> B - Neighborhood B
> …
> -8 - Other Neighborhood (please indicate)
> -9 - Don't Know / Refused
> **…**
> **Question 3** What is your income?
> $_____ (annual number)
> -9-Don't Know / Refused

4 Missing Values

Some cells of the table above have 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 an "NA", so that they don't goof up our calculations.

5 Variable Names Should Be Short

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.

6 Variable Names Should Be Limited To a Single Row of the Spreadsheet

Generally, in spreadsheet software, the first row of the data should be used to list the variable names, as is seen in the example below.

person	neighborhood	age	income
1	A	34	49703
2	B	NA	52146
3	B	37	49524
4	B	40	49387
5	B	38	NA
6	B	45	50925