Project: Transcribing Data

Daniel T. Kaplan

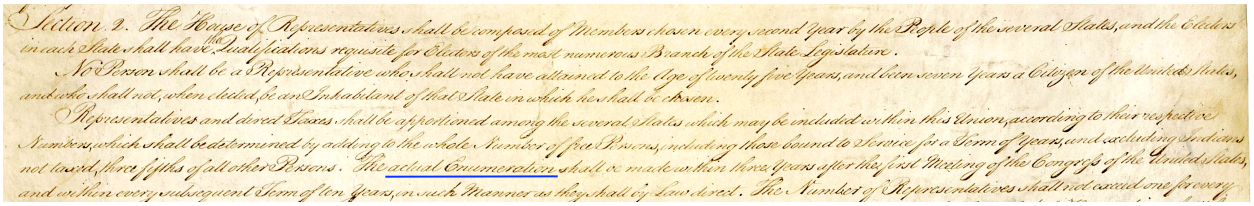
2023-01-12

In order to use data wrangling and graphics techniques, the data frames themselves must be properly formed. Unfortunately, most people are unfamiliar with the basic principles behind data-frame organization. Consequently, even machine-readable documents such as spreadsheets look organized but often contain flaws that prevent their full use as data.

This project is meant to help you internalize principles of proper data organization so that they become second nature to you. Another goal of the project is to develop your skills in recognizing spreadsheet pitfalls that interfere with wrangling. Finally, the project will introduce you to the concept of a *relational database*. We will not be making much use of relational databases in Math 300, but anyone literate with data needs to be aware of this critical and ubiquitous tool and why it is so important to working with data.

The project context: the US Census

As you may know, Article 1 Section 2 of the US Constitution requires that an "actual Enumeration" of all residents of the US be made every ten years.



Section 2. The House of Representatives shall be composed of Members chosen every second Year by the People of the several States, and the Electors in each State shall have Qualifications requisite for Electors of the most numerous Branch of the State Legislature.

No Person shall be a Representative who shall not have attained to the Age of twenty five Years, and been seven Years a Citizen of the United States, and who shall not, when elected, be an Inhabitant of that State in which he shall be chosen.

Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct. The Number of Representatives shall not exceed one for every

Figure 1: ARTICLE 1, Section 2 of the US Constitution, requiring a census every 10 years. The section also contains the infamous "three-fifths" provision for counting enslaved people.

A basic input to the census enumeration is the "population schedule," which lists individual persons as rows in a spreadsheet. By law, the Census Bureau must keep private the information on individuals, and is not even able to share it with other government agencies. Only the summary tabulations made via data wrangling can be published.

However, 72 years after each census the population schedules can be released. In 2022, the population schedules for the 1950 census were released. We are going to use population schedules from the 1940 census, which are more accessible than the newly released 1950s sheets.

@fig-pop-schedule-small shows a population schedule from North Nevada Street in Colorado Springs, CO. This is a spreadsheet in the original sense of the word: a broad sheet of paper used for accounting.

Spreadsheet structure

We call a document like that in @fig-pop-schedule-small a "spreadsheet" because it is not yet organized appropriately for a data frame. (This is understandable, because the concept of relational databases originates in the 1970s. Today's Census Bureau, like every data organization, uses relational spreadsheets.)

@fig-pop-schedule-portion zooms in on the upper-left side of the sheet to make the structure more apparent.

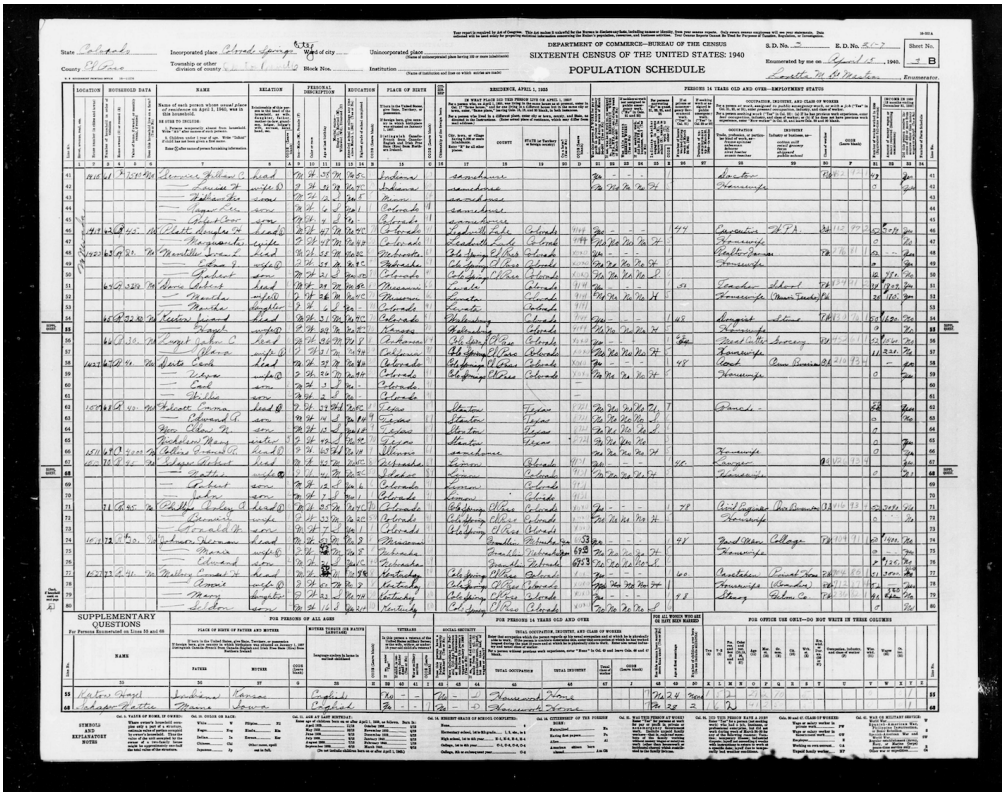
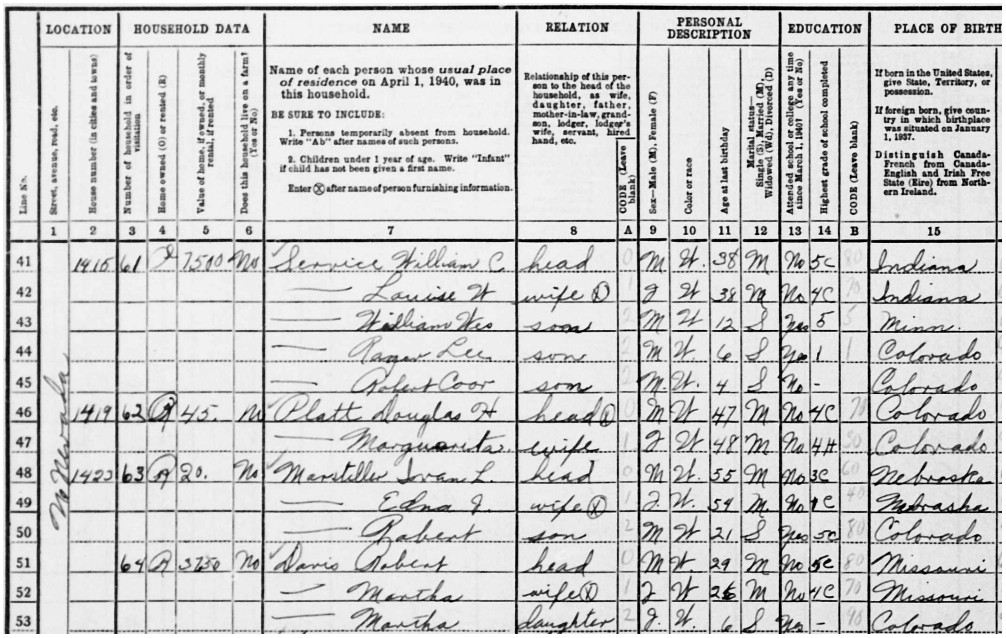


Figure 2: One of the millions of population schedules from the 1940 census.



Line No.	Street, avenue, road, etc.	House number (in cities and towns)	Number of household in order of visitation	Home owned (O) or rented (R)	Value of home, if owned, or monthly rental, if rented	Does this household live on a farm? (Yes or No)	Name of each person whose *usual place of residence* on April 1, 1940, was in this household. BE SURE TO INCLUDE: 1. Persons temporarily absent from household. Write "Ab" after names of such persons. 2. Children under 1 year of age. Write "Infant" if child has not been given a first name. Enter ⊗ after name of person furnishing information.	Relationship of this person to the head of the household, as wife, daughter, father, mother-in-law, grandson, lodger, lodger's wife, servant, hired hand, etc.	CODE (Leave blank)	Sex—Male (M), Female (F)	Color or race	Age at last birthday	Marital status: Single (S), Married (M), Widowed (Wd), Divorced (D)	Attended school or college any time since March 1, 1940? (Yes or No)	Highest grade of school completed	CODE (Leave blank)	If born in the United States, give State, Territory, or possession. If foreign born, give country in which birthplace was situated on January 1, 1937. Distinguish Canada-French from Canada-English and Irish Free State (Eire) from Northern Ireland.
1	2		3	4	5	6	7	8	A	9	10	11	12	13	14	B	15
41		1416	61	R	7500	No	Lennice William C.	head		M	W.	38	M	No	5c		Indiana
42							— Louise H	wife ⊗		F	W	38	M	No	4c		Indiana
43							— William Wes	son		M	W	12	S	Yes	5		Minn.
44							— Roger Lee	son		M	W	6	S	Yes	1		Colorado
45							— Robert Coor	son		M	W.	4	S	No	-		Colorado
46		1419	62	O	45.	No	Platte Douglas H	head ⊗		M	W	47	M	No	4c		Colorado
47							— Marguerite	wife		F	W.	48	M	No	4H		Colorado
48		1423	63	O	20.	No	Merrille Ivan L	head		M	W.	55	M	No	3c		Nebraska
49							— Edna J.	wife ⊗		F	W.	54	M	No	8c		Nebraska
50							— Robert	son		M	W	21	S	Yes	5c		Colorado
51			64	R	3750	No	Davis Robert	head		M	W	29	M	No	5c		Missouri
52							— Martha	wife ⊗		F	W	26	M	No	4c		Missouri
53							— Martha	daughter		F	W.	6	S	No	-		Colorado

Figure 3: Detail from the upper-left portion of @fig-pop-schedule-small.

A key aspect of data-frame organization is the idea of a "unit of observation": the kind of thing represented by a row. Looking carefully at @fig-pop-schedule-portion, it is evident that each row corresponds to an individual person. For example, line 42 is about Louise W. Service, a 38-year-old married woman.

However, for many of the individuals, there are blank cells. Many of the names start with a long dash, as in  . The point of such elisions is to avoid duplicating entries. For instance, Louise W. shares the last name of her husband, William C., as well as the same address, 1415 N. Nevada Street.

Avoiding unnecessary duplication is a fundamental principle in the organization of data.[1] In 1940, avoiding duplication reduced the amount of writing needed for data entry. In the 2020s, avoiding unnecessary duplication is still important as a way of avoiding possible inconsistencies and making it clear that two or more cells must be identical.

The use of blanks or dashes in the spreadsheet is a common-sense way of indicating that information from one row is shared by an adjacent row. But it is not consistent with the proper organization of data frames and databases.

A fundamental, though unexpected and non-intuitive principle of data-frame organization is that the *order of the rows* does not matter. Instead, all information about relationships is represented by the data itself, regardless of order. This principle has important consequences for the organization of data and a major part of the motivation for relational databases.

A modern way to interpret what's going on in the population schedule spreadsheet is that there are actually *two* different *units of observation* involved: i) the individual person and ii) the household. Since a data frame can have only a single unit of observation, representing the spreadsheet in a modern format will require two different data frames: one for the individual persons and another for the individual households.

The requirement that every data frame have its unit of observation provides many advantages. For instance, suppose it was realized after the data entry that the rent on the household's home is incorrect. Fixing it can be accomplished by changing a single number in a single data frame. Or, consider what happens when it is discovered that a person was missed in the enumeration of a household. Adding this new person can be accomplished by appending a new row to the data frame containing individuals.

Tasks for the project

Note that you **do not need** to do any work in R for this project.

You will need two URLs and the `pop_schedule_ID` from this roster.

1. Population schedule URL: links to an image of a population schedule. Download this to your laptop.
2. Data entry URL: links to a Google sheet. You will edit the Google sheet from a browser on your laptop (or any other machine.) The Google sheet contains two tabs, for the *household* and the *persons* data frames.
    - Fill in the *household* data frame first. You will need it for the *persons* data frame.
    - Some of the population schedules are continued from a previous sheet. Start your transcription with the first *complete* household.

Data-entry tips

1. When entering data for a categorical variable, decide on a set of levels before you start entering the data.
2. There is a `household_ID` variable in both the *households* and *persons* data frames. Similarly, there is a `pop_schedule_ID`. These must be absolutely consistent across the two data frames.
3. Each population schedule contains a column (#3) labeled "Number of household in order of visitation." Use this as the `household_ID` in your data frames. It is unique within the population schedule. Together with the `pop_schedule_ID` it is unique universally, since every population schedule has a unique `pop_schedule_ID`.

[1] An example of *necessary* duplication appears in lines 52 and 53, where two people happen to have the same name.

4. In the *persons* spreadsheet, we have collapsed columns 21-25 into one multi-level categorical variable named `workplace`. Read the column 21-25 headers to determine a set of levels that will properly encode the information in those columns.

Notes in draft for instructors

1. Google sheet for distributing population schedules and data-entry spreadsheets: https://docs.google.com/spreadsheets/d/1ZxbiNCKrGeCAqYTd1R9QLSRBLia9HBk4k8DxHxKqiok/edit?usp=sharing
2. An example data-entry spreadsheet: https://docs.google.com/spreadsheets/d/1nUMP1VQfTmhQy1q1sKGSlJ825GpnyXto8M8tkUhtUOU/edit?usp=sharing