Deliverable: Codebook

Due 28 February

Overview

Find and wrangle a dataset of interest to you, and then prepare it for "public" use. In addition to the coded, cleaned dataseet, you will create a codebook that serves as a reference guide for anyone who wants to use your data. You must complete the work in `R Markdown`. See below for details.

Deliverables

Post the following files in a public `GitHub` repository called `Codebook`:
1. Codebook in `.pdf` format.
2. Markdown script for the codebook.
3. Final data in .csv format (only numeric values for non-character variables).
4. Final data in .Rdata format with appropriate factor labels and orders.
5. Replication .R script showing how you wrangled the data.

Your data

Select a dataset of interest to you and with at least fifty observations. The dataest you choose for this assingment *must require wrangling* (cleaning, merging, appending, and/or reshaping); selecting pristine variables from a well-maintained dataset will not work.

The final "tidy" data must include:

- Minimum of ten variables (including necessary ID variables and at least two factor variables)
- Universal missing codes (implicit or explicit)
- Clear, consistent variable names
- Appropriate variable order
- Correct labeling and ordering for factor variables

Your codebook

The codebook should be a professional-quality guide to understanding and using your dataset. It should be structured as follows:

- Overview of data
- Sources and methodology
- Note on missing values
- Itemized presentation of variables

The sources and methodology section must cite the data appropriately. Then describe (briefly, in your own words) the composition of the data.

The itemized variable entries must include the name of the variable in the dataset, a brief description of the variable, and summary information. For character variables: it is enough to note that they are character variables. For numeric variables: present a table of summary statistics. For factor variables: present a table with values, labels, and frequencies for each category. NOTE: you may use (and adapt to suit your style) the functions we wrote in class to create summary and frequency tables.

Evaluation

I will evaluate the data in terms of its ease of use. I will open the data and, with reference to your codebook, poke around: making tables (e.g., `count(df, var1)`), checking summary statistics (e.g., `summary(df)`).

I will evaluate the codebook on its own merits, looking at the quality of the outputs (text and tables) and ease of use.

Note that I will take ambition (or wrangling burden) into consideration.

Example entries

For your consideration as you think about your own codebook style, some no-frills examples:

Respondent last name

Variable name: `rLastName`
Variable type: character
Description: Recorded last name of the respondent.

Respondent age

Variable name: `rAge`
Variable type: Numeric
Description: Respondent's age in years at time of interview.

	`rAge`
Min	24

	rAge
Mean	42
Median	37
Max	91
dk/nr (-9)	14

Party identification

Variable name: `rPID`
Variable type: factor
Description: Self-reported party identfication on 5-point scale from Democratic to Republican.

Value	Label	Freq
-9	dk/nr	21
1	Democratic	12
2	Leans Democratic	19
3	None/Other	24
4	Leans Republican	12
5	Republican	18