Chapter 16 Data Module: Tidy Data Process and Guidelines

Guidelines (Wickham 2014)
In the module, you received an introduction to what tidy data is and why it matters. For further reading and understanding, we suggest you look to the piece titled "Tidy Data" from the Journal of Statistical Software by Hadley Wickham (2014).

Specific guidelines for tidying datasets

USAID Dataset
Lucky you! This dataset is already quite tidy, perhaps look at it to take inspiration from. Look at how the variables and observations are sorted. Take note of how, even though the dataset looks messy and overwhelming, if you take a step back and look at the organization it becomes much more manageable.

While tidying is not necessary, you will need to do a bit of elementary coding here. When looking at the vulnerability assessment, you will see that there are four silos of needs (education, housing, health and sanitation, and food security). Go through the USAID dataset and look at the variable "dac_sector_name" and create a new column following it called "silo_code". In this new column, determine which silo each observation fits into from the sector name and code a '1' for Education, a '2' for Housing, a '3' for Health and Sanitation, a '4' Food Security, and a '5' for Other. This will help you immensely to build visualizations.

ECHO
This dataset, unfortunately, is not quite as tidy. In order to make it workable, some changes must be made. First, you will need to put variables as column titles and observations as rows. Create a variable "observation" as a new column A. Number each observation up to 89. Additionally, you will need to convert the disbursement funds from euros to dollars. Create a new variable "Disbursement_USD" and convert each observation into USD using the exchange rate 1 Euro = 1.2 USD.

You will also need to do a bit of elementary coding here. When looking at the vulnerability assessment, you will see that there are four silos of needs (education, housing, health and sanitation, and food security). Go through the ECHO dataset and look at the variable "Disbursement_USD" and create a new column following it called "silo_code". In this new column, determine which silo each observation fits into from the sector description and code a '1' for Education, a '2' for Housing, a '3' for Health and Sanitation, a '4' Food Security, and '5' for Other. This will help you immensely to build visualizations.

Sida
Beware of duplicate data! 'Recipient-country-code' and 'recipient-country' have the same meaning. LB = Lebanon. If this variable isn't telling you anything new, then it might just be cluttering your dataset. What else can you do with this data? See all those blank cells in columns *O, P, U,* and *V*? Those can be deleted as their values have no bearing on this data when there is no information we can glean from them.

You will also need to do a bit of elementary coding here. When looking at the vulnerability assessment, you will see that there are four silos of needs (education, housing, health and sanitation, and food

security). Go through the Sida dataset and look at the variable “sector” and create a new column following it called “silo_code”. In this new column, determine which silo each observation fits into from the sector name and code a ‘1’ for Education, a ‘2’ for Housing, a ‘3’ for Health and Sanitation, a ‘4’ Food Security and ‘5’ for other. Again, use your own judgment to decide which sector corresponds to the given codes. This will help you immensely to build visualizations.

One last tip: if all of these variables are bothering you, you can hide them! Take the column ‘title-sv’ for example. Most of the information in this column is in Swedish, so it may not be of much use. Rather than deleting the column entirely and changing the integrity of the dataset, you can right-click (or control-click on Mac) the column and select ‘hide column.’