Assignment: Tidy Your Data

SYS 5581 Time Series and Forecasting
University of Virginia Engineering, Spring 2021

Instructor: Arthur Small

Version of 2021-02-03

Before undertaking any data analysis project, you need to organize your data into a format to make it ready for analysis. Very commonly, the data you wish to work with will not come to you in a nice format that makes it ready to analyze. You typically will need first to *extract* your data from its original source (e.g., an Excel file, or cloud hosting service). Often it will be necessary to *transform* the data, applying a sequence of manipulations to get it into a nice format such as a single table. If you have saved your prepared table to a database or local file, your may finally need to *load* the data into memory on your working machine as a prelude to commence analysis. These steps together are the *extract-transform-load* (ETL) stage of a data analysis project.

The bad news is that working data scientists generally report that the ETL stage is the most time-consuming part of a data science project. The good news is that the `R` `tidyverse` packages offer a number of helpful tools to somewhat ease the pain of ETL work, also known informally as *data wrangling*.[1]

The ETL steps needed for a given project will depend on the nature of the data and on how they are originally organized. We can characterize how we want the data to look at the end of the ETL stage. To the extent possible, we want the data to be *tidy*.

Tidy data

Hadley Wickham (2014) codified the concept of tidy data as follows:

> Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types. In tidy data:
>
> 1. Each variable forms a column.
> 2. Each observation forms a row.
> 3. Each type of observational unit forms a table.

[1] "Wrangling" refers to work with cattle, sheep, and other livestock.

Wickham, Hadley. 2014. "Tidy Data." *The Journal of Statistical Software* 59 (10). http://www.jstatsoft.org/v59/i10/.