Assignment: Extract, Transform, and Load Your Data

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

Instructor: Arthur Small

Version of 2021-02-08

This assignment focuses on the steps needed to acquire data from its source location, organize it, and prepare it for time series analysis. You are asked to:

1. Identify a time series data set that you might use as the basis for your class project.
2. Acquire the data from its source location, reproducibly.
3. Organize the data into a specific format, called a *tidy* data frame.
4. Convert this data frame into a particular type of object called a `tsibble` to render it ready for time series analysis.
5. Generate at least one table or graph based on the data (more if you like).

Create a new R Markdown document for this work, inside R Studio. Code and document all these steps in a single .Rmd file.

All steps should be fully *reproducible*. This means: If I, or anyone, rerun the code in your R Markdown file and reknit the .Rmd file to generate a new PDF output, I should be able to execute all your computational steps and regenerate your PDF essentially exactly.

To submit your work: knit your .Rmd file to generate a .pdf file. Push both the .Rmd and .pdf files to your Repo on Github, along with any supporting files. Submit your assignment on Collab, enclosing a link to your .pdf file on Github.

Here's some additional background and guidance on the assignment:

Overview: Extraction, transformation, and loading of data

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]

Step 1: Identify a time series data set you want to work with

Ideally, identify a data set that you that you might use as the basis for your class project. You need not necessarily commit at this time to using this data set for your project. However, bear in mind that you will in this assignment be investing time and effort in acquiring, cleaning, and organizing the data set. It is better for you if you invest that effort on the data set you will later analyze.

Please do not use a data set that already come packaged with `R` or that is otherwise already cleaned up. The point of this assignment is to learn how to use tools from the Tidyverse and tidyverts packages when working with data you acquire "in the wild".

Step 2: Acquire the data from its source location, reproducibly

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*.

Step 3 Organize your data into a *tidy* data frame

Hadley Wickham (2014) codifies the concept of tidy data:

> 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.

"Real datasets can, and often do, violate the three precepts of tidy data in almost every way imaginable. While occasionally you do get a dataset that you can start analysing immediately, this is the exception, not the rule." In real life, the systems people and organizations use to collect, manage, and store data are governed by many priorities: end-user convenience, clarity of presentation, storage costs, processing speed, etc. Making data fit for analysis by data scientists is typically a minor consideration, if it is thought of at all.

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

Expanding on the theme, Wickham identifies five of the most common problems with non-tidy, "messy" datasets:

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

The Tidyverse packages integrate a range of tools to help with transforming the messy data often encountered in the wild into tidy formats more suitable for analysis. See the Tidyverse website for an overview, or this Coursera course for more guidance.

Step 4: Convert your data into a `tsibble` object

In this step of the assignment, you will convert your tidy data frame into a `tsibble` object. Doing so in effect tells `R`: "These data actually form a time series. One column, which I designate the `index`, contains time values, in equal intervals." Taking this step enables the use of specific tools for data visualization, exploratory analysis, and forecasting for time series data. The tsibble package for R provides "a data infrastructure for tidy temporal data". Review the documentation for instructions.

Step 5: Generate at least one table or figure

This step simply confirms that you have completed the ETL process and that your data is ready to analyze. You could simply add a line of code: `print(name_of_your_tsibble)` to generate a simple table.

With your data organized into a `tsibble` object, you are positioned to do some *exploratory data analysis*, the topic of the next assignment. If you want to get a head start, check out the functions for data visualization and exploratory analysis of time series in the feasts package. Or work through the examples in Chapter 2 or Chapter 4 of Forecasting: Principles and Practice, 3rd ed.

References

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (1, 1): 1–23. https://doi.org/10.18637/jss.v059.i10.