Data: Wrangle and Display it With Relative Ease.

Contents

Data Cleaning

In this tutorial, we will be taking in two data sets from a fictional company, cleaning and reformatting them to make it usable, joining them together, and producing some graphics from them. Happily, all of this is actually pretty easy.

Dependencies and setup

Make sure you have the following packages installed * `openxlsx` * `reshape2` * `dplyr` * `magrittr` * `ggplot2` * `ggthemes` * `roperators`

I've attached an installation script that you can run to make sure the packages are all there:

```
pkgs <- c("openxlsx", "reshape2", "dplyr", "magrittr",
          "ggplot2","ggthemes", "roperators")

# Hackfix tip:
# Opening the CRAN mirror in a browser can help with some restricted networks
utils::browseURL("http://cran.stat.auckland.ac.nz/")
utils::browseURL("https://meta.wikimedia.org/wiki/List_of_countries_by_regional_classification")

# Another hackfix for restricted networks.... just in case
httr::set_config(httr::config( ssl_verifypeer = 0L ))

for (package in pkgs){
  if(package %in% rownames(installed.packages()) == FALSE)
    # Try will attempt to do what is in parentheses, but won't die if it doesn't work
    try(install.packages(package, repos = "http://cran.stat.auckland.ac.nz/"), silent = TRUE)
}
```

Load your required packages

```
require(openxlsx)
require(reshape2)
require(dplyr)
require(magrittr)
require(ggplot2)
require(ggthemes)
require(roperators)
```

Load the data

In our data folder, there are two datasets, a csv called employee_data.csv and an excel workbook called survey_results.xlsx

To read in the csv data, we can use base R's `read.csv` function, which is the same as `read.table`, which you might see in other scripts, only with different default arguments. What's nice aboput data stored in csv files is that because they're just plain-text flatfiles, they can be opened in any program or programming language.

The following code reads: * Create a variable called employees * Into that value place the output of `read.csv()` * Where `read.csv()` is going to go out one folder (../) and then look for the file in another folder called data

```
employees <- read.csv("../data/employee_data.csv")
```

Excel's files are a little trickier to read in which is why we loaded the `openxlsx` package to handle it. There are other, older packages to read in .xlsx doccuments, however, they often have difficult Java dependencies, hence we prefer `openxlsx`

```
employees <- read.csv("../data/employee_data.csv")
```