DSCI 2000: Data Analysis Assignment

Data Pipelines, Generalizability and Adaptability of Code for Data, Tidy Data, and Data Visualization

Is there a correlation between unemployment and stock prices in the U.S.?

65 points
Fall 2020

Submission and Formatting Guidelines:

- Submit all of your code in an .Rmd file and include the .html file with R Markdown output to your GitLab portfolio **in a new, clearly-labeled subfolder** by the due date(s) specified. **Please include all files that you've used for this assignment in your GitLab repository as well**.
- Cite your sources. See the citation requirements under the "Plagiarism and Cheating" section of the syllabus.

Directions: The intent of this assignment is to apply the skills of tidying data, data manipulation, and data visualization to a current-day, real-world example so as to introduce you to how to write your own statistical analyses.

We will be using the seasonally-adjusted U.S. monthly unemployment rate data from the Bureau of Labor Statistics (posted in D2L as SeriesReport-20201006201227_5de657.xlsx), and stock prices from the S&P 500 from Yahoo Finance gathered via the `quantmod` package. Our goal is to see if S&P 500 closing stock prices on the days on which unemployment rates are reported are correlated with the corresponding unemployment rates. **Remember: correlation is not causation!** We do not have enough information to make any claims regarding causality.

Please section your .Rmd file so that it is clear how your report is organized by question using `#` or `##` as appropriate. See, for example, the Homework 4 solution from Spring 2020 at https://tinyurl.com/yxbnl6pz, code available at https://tinyurl.com/y5d57xgq. **You will have points taken away from your grade if it is difficult to discern how your document is organized**. I strongly recommend adding `message = FALSE` in any code chunks in which you are loading packages.

Text answers should be in Markdown text outside of R code chunks, with code to supplement your answers as appropriate. If you have any questions on this, please let me know.

You should expect this assignment to take **5-7 hours on average**. Remember that you should treat assignments as if they are take-home exams. Please do not procrastinate, and let me know if you run into problems as early as possible. You are expected to ask for help on this assignment. I will not accept late submissions due to last-minute asking for help. Keep in mind that this assignment is 20% of your grade in this class. **Please read the hints. They are provided to make your work easier for this assignment.**

Problems:

There are three data sets you will be working with:

- SeriesReport-20201006201227_5de657.xlsx, monthly unemployment rates from the Bureau of Labor Statistics. **Note there is a one-month lag to release (e.g., January 2020 unemployment rate is released on February 2020)**.
- Bureau of Labor Statistics release dates from 2007 to 2020, pulled via .csv file and website scraping using `BLS_Release_Dates.R` (requires the `rvest` and `dplyr` packages). The BLS_Release_Dates.R and BLS_Release_Dates.csv file should be in the same folder as the .Rmd file, and in the .Rmd file, you should execute

```
source("BLS_Release_Dates.R")
```

- Yahoo Finance stock prices for the S&P 500 (requiring the `quantmod` package), using the following code:

```
library(quantmod)
getSymbols("^GSPC", src = "yahoo")
SP_500 <- Cl(GSPC)
rm(GSPC)
SP_500 <- as.data.frame(SP_500)
SP_500 <- tibble::rownames_to_column(SP_500, var = "Date")
SP_500$Date <- as.Date(SP_500$Date)
```

1. **(25 points)** Read in all three data sets into R as indicated in the directions above. You should have three data frames in your R environment: `bls`, `SP_500`, and the data frame of unemployment rates.

   Tidy the unemployment-rates data frame. When tidying this data frame, one of the columns in this final data frame should be the date, e.g., "Jan 2010".

   Convert all dates, as well as year-month combinations, to have the `Date` data type. For year-month combinations, assume that you're working with the first day of the month. Make sure that all columns in all data frames have appropriate data types (e.g., numbers should be `numeric`).
2. **(25 points)** Reformat the data frames provided above so that you have only one data frame, consisting of the following columns:
   - The dates on which the employment numbers have been released
   - The closing stock prices of the S&P 500 on the corresponding dates
   - The unemployment rate announced on the corresponding dates

   All observations that do not occur on the dates on which the employment numbers have been released should be dropped.
3. **(10 points)** Using `ggplot2`, plot line graphs of the unemployment rates and S&P 500 data with release date on the x-axis from the data frame in the prior problem. Use appropriate axis labels and a black-and-white theme.
4. **(5 points)** Is there a correlation between the unemployment rate reported and the S&P 500 closing prices on these reporting dates? Interpret using Kendall's τ (tau). Do not make any claims regarding causality; you will receive a deduction if you do.

Hints:

1. Make sure you have all necessary packages installed. Don't forget to use `as.data.frame()` after using `read_xlsx()`. Use the strategy for tidying data to tidy the unemployment-rates data set.

   For year-month combinations, stick a "01" somewhere in the string to represent the first day of the month before converting to a Date.

   Use `as.Date()` with `?strptime` to convert to Dates as appropriate.
2. Before you start joining data frames, you should only show observations in the `bls` that have "Employment Situation for" in the "Release" column (these are the dates on which the unemployment rates are released). Recall that `grepl()` can be used for Boolean string matching.

   Don't forget that there is a one-month lag for the unemployment rates. If `df` is your unemployment-rates data frame, you can adjust the dates forward a month using

   ```
   df$Date <- df$Date %m+% months(1)
   ```

   from the `lubridate` package. Use the `lubridate` package and join using **both** month and year the `bls` data frame and the data frame of unemployment rates. (What type of join should you use?). Then, once these two data sets are linked, join the `SP_500` data set.
3. This problem does not require anything beyond what I taught in the `ggplot2` lecture if you've done the prior problems correctly.
4. Be sure that your three-column data frame from problem 2 has been arranged in ascending order by date (i.e., earlier dates first, most recent or later dates last).

   Like with the correlation coefficient r (review the Statistics Review video in D2L under Content $>$ Units $>$ Unit 1: Introduction if you don't remember what r is), values around -1 indicate negative correlation, values around $+1$ indicate positive correlation, and values around 0 indicate no correlation. Kendall's τ has the advantage of not relying on a linear trend and could be used to gauge a trend in general.

   If `df` is the name of your three-column data frame from problem 2, use

   ```
   cor(df$VALUE, df$GSPC.Close, method = "kendall", use = "complete.obs")
   ```

   to compute Kendall's τ.