Homework: IMMERSE pre-training day 3

Synchronous session materials

Reproducibility Google slides

Day 3 wrangling key (.pdf)

Day 3 wrangling key (.Rmd)

Part 1: more wrangling practice code-along

`Immerse_day3_asynch.pdf:` This self-guided lesson continues from the synchronous session, to show you some more tools for data wrangling in R, focusing on working with categorical values.

In the R project folder we used in the synchronous session (or a new one), create a new R markdown, and then follow along with the script. Note: see Part 3 (optional) to download a zip with this pdf, the Rmd script, and another script for additional practice.

Part 2: wrangling categorical data on your own

We will use the CASchools dataset, available in the AER R package, to practice data wrangling, particularly with categorical variables.

The dataset

`AER::CASchools` dataset (Applied Econometrics with R):

> The data used here are from all 420 K-6 and K-8 districts in California with data available for 1998 and 1999. Test scores are on the Stanford 9 standardized test administered to 5th grade students. School characteristics (averaged across the district) include enrollment, number of teachers (measured as "full-time equivalents", number of computers per classroom, and expenditures per student. Demographic variables for the students are averaged across the district. The demographic variables include the percentage of students in the public assistance program CalWorks (formerly AFDC), the percentage of students that qualify for a reduced price lunch, and the percentage of students that are English learners (that is, students for whom English is a second language).

Your task

1. Create a new R Project in a location of your choosing. Start a new R Markdown document in the project root directory.
2. Set the chunk options to suppress messages and warnings, but show all code (echo = TRUE), so when we knit the R Markdown, we will see only the code and outputs, but not the various diagnostic messages sent by various functions.
3. In a new code chunk named "attach packages", attach the **`tidyverse`** package and the **`AER`** package. You may need to install AER using **`install.packages()`**.
4. In a new code chunk (name it something like "load and explore data"), load the CA Schools dataset from the AER package, using **`data('CASchools')`** - you should see it appear in the Environment pane. Explore the data a bit to understand the various columns, using

summary(), glimpse(), head(), and so on - though you might want to comment those out when you're done so they don't clutter up your final HTML when you knit the R Markdown.

5. In a new code chunk:
   a. Choose one or more of the continuous variables in this dataset, and turn them into categorical values using **`cut()`**, **`ntile()`**, **`ifelse()`**, and/or **`case_when()`**. Have at least one variable with three or more category values.
   b. Convert new categorical variables to factors, with levels in an appropriate order (e.g., low medium high).
   c. Use filter and select to narrow down the dataset to a subset of your choosing.
   d. Save out as a .csv file.
6. In a new code chunk
   a. Read in your saved .csv to a new object, using read_csv().
   b. Note that your carefully crafted categoricals are no longer factors! Fix that!
   c. Use **`pivot_wider()`** to convert a multi-level categorical into multiple binary dummy variables (use something along the lines of **`mutate(dummy = TRUE)`** to create a value that you will use as the value_from in **`pivot_wider()`**)

Part 3: more wrangling practice (optional)

This self-guided lesson will show you a lot of additional wrangling in R. There's also a brief intro to some spatial data wrangling in R - maybe you'll never work with spatial data, but it is pretty fun!

Download this .zip file - it contains an R project with code keys and some data. Unzip the file to create a new project in the same folder where you created your R project during day 3's synchronous session. (If you're already familiar with GitHub, you can fork and clone the repository from here: https://github.com/oharac/immerse_day3_more_wrangling - if not, don't worry about it!)

Go to the **`code_key`** folder and open the **`immerse_day3_hw.html`** file in a web browser.

In RStudio, in the R project, start a new R Markdown. Follow along with the HTML, coding each chunk step by step and run it line by line. You'll get the most out of this by:

- Making sure you understand what each line of code is doing
- Trying variations on the code to see how it works (or how it breaks - you learn a lot by breaking your code too!)
- If you find something cool, or frustrating, share it on the IMMERSE Slack #code-help channel

If you get stuck, you can also simply open the Rmd in the R project code_key folder…