DA5020 Homework 4: Strings and Factors

2017-09-28

Preparation

Download US Farmers Markert Directory data from the website of USDA (click on "Export to Excel"). Rename the file as *farmers_market.csv*.

Download the Know Your Farmer, Know Your Food Projects dataset and name it as *kyfprojects.xls*. Put it into the same folder.

Read the data:

Warm Up

This dataset stores city and state in different columns, what if you want to print out city and state in the format "City, State"?

Questions

Please edit this file and add your own solutions to these questions. Make your output as readable as possible. Next time you would need to create this file on your own. Feel free to try out other templates (e.g. Tufte Handout) if your are familiar with LaTex. But for whatever template you choose, you should always include a link to your GitHub repo at the first page of your PDF.

1. (20 points) Cleanup the `Facebook` and `Twitter` column to let them contain only the facebook username or twitter handle name. I.e., replace "https://www.facebook.com/pages/Cameron-Park-Farmers-Market/97634216535?ref=hl" with "Cameron-Park-Farmers-Market", "https://twitter.com/FarmMarket125th" with "FarmMarket125th", and "@21acres" with "21acres".
2. (20 points) Clean up the `city` and `street` column. Remove state and county names from the `city` column and consolidate address spellings to be more consistent (e.g. "St.", "ST.", "Street" all become "St"; "and" changes to "&", etc...).
3. (20 points) Create a new data frame (tibble) that explains the online presence of each state's farmers market. I.e., how many percentages of them have a facebook account? A twitter account? Or either of the accounts? (Hint: use the `is.na()` function)
4. (20 points) Some of the farmer market names are quite long. Can you make them shorter by using the `forcats::fct_recode` function? Create a plot that demonstrates the number of farmers markets per location type. The locations should be ordered in descending order where the top of the graph will have the one with the highest number of markets.
5. (20 points) Write code to sanity check the `kyfprojects` data. For example, does `Program Abbreviation` always match `Program Name` for all the rows? (Try thinking of your own rules, too.)

Submission

You need to submit an .Rmd extension file as well as the generated pdf file. Be sure to state all the assumptions and give explanations as comments in the .Rmd file wherever needed to help us assess your submission. Please name the submission file LAST_FirstInitial_1.Rmd for example for John Smith's 1st assignment, the file should be named Smith_J_1.Rmd.