Homework 02

Submit Assignment

Due Feb 9 by 11:59pm **Points** 100 **Submitting** a file upload
Available Jan 27 at 9am - May 5 at 11:59pm 3 months

Instructions

You must use a Jupyter notebook for the assignment. You must turn in your .ipynb source code and the rendered HTML file.

Problem 01

The Tidy Data paper by Hadley Wickham (**paper link here (https://vita.had.co.nz/papers/tidy-data.html)**) has multiple examples of "tidying" messy data. The examples each focus on a different theme for how the data are "messy". Section 3.2 discusses handling data sets with multiple variables stored in one column with an example data set based on Tuberculosis (TB) data from the World Health Organization. **You must perform the necessary steps to TIDY this messy data set. You must state your decisions for how you are "tidying" the data, based on the discussion in the paper.** Once you have prepared the data you must state:

- How many rows and columns does your final "tidy" data set consist of?
- How many unique `country` values are there? Is this consistent with the number of unique `country` values in the original data set?
- How many unique `year` values are there? How many rows are associated with each year in the final "tidy" data set?

The original data set is larger than that described in the paper. If you would like to see it, it is the tb.csv file in the **paper's github repo (https://github.com/hadley/tidy-data/tree/master/data)** .

However, you will work with an already prepared data set, that is more similar to the one presented in the paper itself. That data set is linked below. The file name is in reference to the fact that you are working with a data set consistent with Table 9 in the Tidy Data paper. Note that the column names are slightly different from those shown in the paper.

tidy_data_table_9_use.csv

Problem 02

You will now practice grouping and aggregating the TB data set from the previous problem. If you do not feel comfortable with the "tidy" data set you created in Problem 01, you may manipulate the original data set to answer this question.

- Which year has the most missing values? Which year has the fewest missing values?
- Which year has the most total cases (across all age groups) for males in the countries with codes AO and AR?
- Which year has the most total cases (across all age groups) for females in the countries with codes AF and AM?

Problem 03

Five CSV files containing stock market data are provided at the end of these instructions. Four of the files contain company stock prices. The rows correspond to days and the column names give the company stock symbol. The column names are consistent with the file names, and describe what the price relates to. The `daily_open_prices.csv` file contains the open prices for each day for the companies, while the `daily_close_prices.csv` file contains the closing prices for each day for the companies. The `daily_low_prices.csv` file contains the lowest prices observed for each day for the companies, and the `daily_high_prices.csv` file contains the highest observed prices for each day for the companies. The last file, `stock_info.csv` provides the company names, stock symbol, and several other pieces of information.

You must TIDY all data sets together in a single LONG-FORMAT data set. That data set must clearly provide the open, low, high, and close price for each company, following the rules described in the Tidy Data paper. Your long-format data set should include columns for the additional pieces of information present in the `stock_info.csv` file as well.

After tidying the data, you must state:

- How many unique companies are in the long-format data set? Is that number consistent with the original separated data files?
- How many rows and columns are in your final tidy data set?
- How many rows are associated with each day in the final tidy data set?

Problem 04

You will now practice grouping and aggregating the stock data from the previous problem. If you do not feel comfortable with the "tidy" data set you created in Problem 03, you may manipulate the original data sets to answer this question.

- How many unique days are associated with each MONTH in the data set?
- Calculate the difference between the daily close and daily open prices for each stock.

  - What is the maximum daily close-to-open difference for each company?
  - What is the maximum daily close-to-open difference for each company in each MONTH?
  - What is the average daily close-to-open difference for each company in each MONTH?
- Which DAY had the maximum daily high price for each company in each MONTH?

Stock related data for Problems 03 and 04

daily_open_prices.csv

daily_close_prices.csv

daily_low_prices.csv

daily_high_prices.csv

stock_info.csv