Final Project: Guidance Document

David Chen & Deric Liang

Due April 30, 2019

Purpose

This document is required to indicate where various requirements can be found within your Final Project Report Rmd. You must **indicate line numbers as they appear in your final Rmd document** *accompanying each of the following required tasks. Points will be deducted if line numbers are missing or differ signficantly from the submitted Final Rmd document.*

Final Project Requirements

Joins (at least) two different data sources for analysis

Description: Report includes meaningful contributions from (at least) two different data sources that are JOINED for the analysis

(A) .Rmd Line number(s) for join operation: 185-189

Data Wrangling (5 of 6 required)

Description: Clear demonstration of proficiency should include proper use of 5 out of the following 6 topics from class: (+) general purpose data wrangling–e.g., filter, select mutate, summarise, arrange, group_by, etc. (+) spread & gather to stack/unstack variables (+) regular expressions (+) user-defined functions (+) loops and control flow (+) vectorized functions like the `apply` *family,* `purrr::map` *or* `dplyr::do`

(A) .Rmd Line number(s) for an example of general purpose data wrangling: 106

(B) .Rmd Line number(s) for a spread & gather (or equivalent): 220

(C) .Rmd Line number(s) for use of regular expressions: N/A

(D) .Rmd Line number(s) for use of user-defined functions: 48-86

(E) .Rmd Line number(s) for use of loops and/or control flow: 71-85

(F) .Rmd Line number(s) for use of vectorized functions: 212

Data Visualization (all 4 required)

Description: Clear demonstration of proficiency should include at least FOUR distinct and useful data visualizations that are relevant to purpose of the analysis. Include at least one effective visualization with layered data from distinct data structures–e.g., a `geom` *that plots data from a secondary data frame AND another must effectively display many–3 or more–variables. All plots must demonstrate good plotting practices.*

(A) .Rmd Line number(s) for visualization with layered data from distinct data structures–e.g., a `geom` that plots data from a secondary data frame: 262-272

(B) .Rmd Line number(s) for visualization displaying many–3 or more–variables (A, B, C, & D must be four **different** plots): 250-256

(C) .Rmd Line number(s) for a third visualization: 277-284

(D) .Rmd Line number(s) for a third visualization: 435-442

Data Analysis (3 of 5 required)

Description: Clear demonstration of proficiency should include proper use of 3 out of the following 4 topics from class: (+) statistical modeling/supervised learning (+) unsupervised learning (+) user-defined simulations (+) analysis of text data (+) R tools for "big data"

(A) .Rmd Line number(s) for statistical modeling/supervised learning: 287-410

(B) .Rmd Line number(s) for unsupervised learning: 412-541

(C) .Rmd Line number(s) for user-defined simulation(s): N/A

(D) .Rmd Line number(s) for analysis of text data: N/A

(E) .Rmd Line number(s) for R tools for "big data": 27-44

Miscelaneous (Nothing for you to report in this Guidance Document)

- *Overall GitHub contributions* will be verified upon the due date using the "contributors" page in your project Repo. (Nothing to report in Guidance Document)
- *Overall Reproducibility of analysis* will be verified upon the due date by cloning your project Repo and independently executing your analysis
- *Overall STAT 380 for social good* will be verified based on the subject of your final project submission
- *Code quality* will be verified based on alignment to the STAT 380 R style guide
- *Narrative quality* will be evaluated based on the written descriptions of your analysis, motivation, reasoning for each step, and conclusions in the final project report submission
- *Extra credit* will be awarded if the final project report is submitted to Canvas as a URL to a functioning website hosting your analysis (not the URL of the GitHub Repo itself).