Final Project: Guidance Document

Author Name(s)

Due December 11, 2020

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

Data Access

Description: (1) Analysis includes at least two different data sources. (2) Primary data source may NOT be loaded from an R package–though supporting data may. (3) Access to all data sources is contained within the analysis. (4) Imported data is inspected at beginning of analysis using one or more R functions: e.g., str, glimpse, head, tail, names, nrow, etc

(A) .Rmd Line numbers where at least two different data sources are imported:

(B) .Rmd Line numbers for inspecting data intake:

Data Wrangling (5 out of 8 required)

Description: Students need not use every function and method introduced in STAT 184, but clear demonstration of proficiency should include proper use of 5 out of the following 8 topics from class: (+) various data verbs for general data wrangling like filter, mutate, summarise, arrange, group_by, etc. (+) joins for multiple data tables. (+) spread & gather to stack/unstack variables (+) regular expressions (+) reduction and/or transformation functions like mean, sum, max, min, n(), rank, pmin, etc. (+) user-defined functions (+) loops and control flow (+) machine learning

(A) .Rmd Line number(s) for general data wrangling:

(B) .Rmd Line number(s) for a join operation:

(C) .Rmd Line number(s) for a spread or gather operation (or equivalent):

(D) .Rmd Line number(s) for use of regular expressions:

(E) .Rmd Line number(s) for use of reduction and/or transformation functions:

(F) .Rmd Line number(s) for use of user-defined functions:

(G) .Rmd Line number(s) for use of loops and/or control flow:

(H) .Rmd Line number(s) for use of machine learning (not "wrangling" but scored here):

Data Visualization (3 of 5 required)

Description: Students need not use every function and method introduced in STAT 184, but clear demonstration of proficiency should include a range of useful of data visualizations that are (1) relevant to stated research question for the analysis, (2) include at least one effective display of many–at least 3–variables, and (3) include 3 of the following 5 visualization techniques learned in STAT 184: (+) use of multiple geoms such as points, density, lines, segments, boxplots, bar charts, histograms, etc (+) use of multiple aesthetics–not necessarily all in the same graph–such as color, size, shape, x/y position, facets, etc (+) layered graphics such as points and accompanying smoother, points and accompanying boxplots, overlaid density distributions, etc (+) leaflet maps (+) decision tree and/or dendogram displaying machine learning model results

(A) .Rmd Line number(s) for use of mulitple different geoms:

(B) .Rmd Line number(s) for use of multiple aesthetics:

(C) .Rmd Line number(s) for use of layered graphics:

(D) .Rmd Line number(s) for use of leaflet maps:

(E) .Rmd Line number(s) for use of decision tree or dendogram results:

Other requirements (Nothing for you to report in this Guidance Document)

(A) *All data visualizations* must be relevant to the stated research question, and the report must include at least one effective display of many–at least 3–variables

(B) *Code quality:* Code formatting is consistent with Style Guide Appendix of DataComputing eBook. Specifically, all code chunks demonstrate proficiency with (1) meaningful object names (2) proper use of white space especially with respect to infix operators, chain operators, commas, brackets/parens, etc (3) use of `<-` assignment operator throughout (4) use of meaningful comments.

(C) *Narrative quality:* The narrative text (1) clearly states one research question that motivates the overall analysis, (2) explains reasoning for each significant step in the analysis and it's relationship to the research question, (3) explains significant findings and conclusions as they relate to the research question, and (4) is completely free of errors in spelling and grammar

(D) *Overall Quality:* Submitted project shows significant effort to produce a high-quality and thoughtful analysis that showcases STAT 184 skills. (2) The project must be self-contained, such that the analysis can be entirely rerun without errors. (3) Analysis is coherent, well-organized, and free of extraneous content such as data dumps, unrelated graphs, and other content that is not overtly connected to the research question.

(E) *EXTRA CREDIT* (1) Project is submitted as a self-contained GitHub Repo (2) project submission is a functioning github.io webpage generated for the project Repo. Note: a link to the GitHub Repo itself will be awarded partial credit, but does not itself qualify as a "webpage" of the analysis.