DATS 6101 Introduction to Data Science (E Lo)

Project 1 Outline (Fall 2018)

Goal: Better understand the initial stages of a data focused project by conducting background research and completing Exploratory Data Analysis (EDA).

A. Development of a research driven question (SMART) focused on a dataset either inside of R or one of your choosing from any online sources. Acceptable dataset for this “big data” class requires 5000+ observations (that is, 5000+ rows of data for the data frame).

B. Provide an R-markdown file, knitted into one of the three formats among html/pdf/docx, which shows explanations and rationale of the Exploratory Data Analysis of your project. This document shows a technical person the math/stat/codes that you used in your analysis. It should include:
   a. Summary of the dataset
   b. Descriptive Statistics
   c. Graphical representations of the data
   d. [When applicable] Measures of Variance / sd
   e. [When applicable] Initial correlation / Chi Square tests / ANOVA analysis / Z-interval / T-interval etc.

C. Write a roughly 10-page (definitely no more than 4000 words) summary of the research and EDA process of your project. You can use R-markdown to knit your summary report in html/pdf/docx formats, or directly in any of these three formats from other authoring tools. You may use part of part B (such as graphs and results) in here. They can overlap. This summary is to-be presented to your boss, your client, or to-be submitted for publication in journals. Potential area of topics to address in this summary include:
   a. What do we know about this dataset?
   b. What are the limitations of the dataset?
   c. How was the information gathered?
   d. What analysis has already been completed related to the content in your dataset?
   e. How did the research you gathered contribute to your question development?
   f. What additional information would be beneficial?
   g. How did your question change, if at all, after Exploratory Data Analysis?
   h. Based on EDA can you begin to sketch out an answer to your question?
   i. References

D. Develop a 15-20 minute presentation that effectively communicates the results of these initial stages of a data science project to be presented during class.

Grading:

A. 25%
B. 25%
C. 25%
D. 25%