THE UNIVERSITY OF CHICAGO

Statistics for Data Science
Final Project

As a culmination of your experience in this course, you will submit a final project that applies a data science approach to derive insights regarding a problem statement(s) of interest. This project is designed to replicate a real-life experience of working within a data science context. You should plan to make progress on this project throughout the course and you can consider the following schedule as a guideline:

Weeks 1-2: Find data set and frame problem statement. Perform any necessary data cleaning.

Weeks 3-4: Perform exploratory analysis with the help of unsupervised modelling. Plot and visualize your data. Calculate descriptive statistics. Perform clustering algorithms on your data. Consider if your data is a good candidate for dimension reduction.

Weeks 5-6: Begin supervised analysis on your data. Perform multiple bivariate and/or multivariate linear regression models to determine the most informative solution. Perform any relevant hypothesis testing.

Weeks 7-8: Consider if classification via logistic regression or group mean comparison via ANOVA can improve the depth and quality of your findings. Prepare and submit PowerPoint presentation. Begin work on final paper write-up.

Week 9: Submit final paper.

Data

You are free to use any data set you find interesting for this project. There are many built-in available datasets in R. In addition, https://www.kaggle.com/datasets is an excellent resource for finding data for modelling. Alternatively, you are free to find your own data or use any previously collected data that interests you. While pursuing intricate data sets on topics of personal interest is encouraged, keep in mind that by at least halfway through the course (after week 4) you should have a cleaned and well-structured dataset at your disposal that you can begin modelling with.

PowerPoint

In the final week of the course (week 8), you will submit and possibly present a PowerPoint presentation summarizing your problem statement and findings. Without any additional context, this PowerPoint should convey the primary purpose and conclusions of your work. In other words, if somebody opened this presentation without any additional explanation, they should be able to understand the problem you set out to solve and what you discovered about that problem.

As you have one additional week to complete your project, it is acceptable to describe work that you are still planning on completing for your project. For example, you might share that you have discovered a certain effect using linear regression and are planning to apply ANOVA to answer another question.

Final Paper

Your final paper should provide a detailed description of your project in its entirety. This paper will be submitted one week after the completion of the course (week 9) and will be evaluated to determine your grade for this project. This paper should not exceed 1,000 words, excluding any charts or figures. Consider the following sections for your final paper:

Declaration of problem statement/purpose

Dataset Description

Unsupervised Analysis

Description of methods used and relevant results.

Supervised Analysis

Description of methods used and relevant results.

Insights and Significance of Results

Share your main conclusions and consider if they have any practical applications.

Improvements and Future Work

Discuss questions you were not able to find answers for, consider how this might be improved, and speculate what the next steps of this project might be.

Code

In addition to your final paper, you are required to submit a file displaying the code used to perform your models and visualize your results. This can be submitted as a .R file, a .Rmd file, or a knitted R markdown file in either Word or PDF format.