CICS 397A – Fall 2019

Final Project

Due Thursday, December 19th at 11:59pm

For the final project, you will be conducting exploratory data analysis and writing up your findings. The goal for the final project is to give you an opportunity to apply some of the techniques we've covered over the course of the semester while asking and answering your own questions from the data.

The project is intended to be open-ended, in that it's up to you to decide where to focus your efforts. The data set(s) you work with are your choice: you may use the data you collected for Homework #2, download something publicly available, or generate something new. Your exploration should consist of applying one or more of the techniques we studied over the course of the semester to the data. For instance, you may choose to cluster your examples, and offer some interpretation of the output. Alternatively, you could build a predictive model to try and classify your data, or compare the results of multiple classifiers. The goal is to find a question (or set of questions) about the data that interest you, and attempt to answer them through your analysis.

Format

You may work individually, or collaborate with 1-2 other students on your project. You must submit all code, along with a 3-6 page writeup describing your work. For group projects, the scope of your work and writeup is expected to be more substantial than if you're working by yourself.

Code

All code used in your exploration should be turned in with your writeup. You free to code up your own learning algorithms, but are encouraged to leverage existing tools (pandas, sklearn) as well. You are free to base your code off of any of the examples covered in class or homework assignments. While your code doesn't have to be pristine, you should include comments that helps a reader to understand how your code functions.

The Writeup

Along with any code, you should include a writeup describing your efforts. How you structure the writeup is up to you, but you should include:

- The names of all team members, along with a brief overview of how each person contributed
- A description of the data set, including how you obtained it and any preprocessing you did to get the data into a usable format

- A task writeup, summarizing the techniques you used, as well as any conclusions you were able to draw
- Any relevant plots or visualizations you were able to produce from the raw data or the outputs of supervised or unsupervised learning algorithms you used
- Ideas for future exploration of the data, including interesting questions raised by your analysis
- Description of challenges you encountered when working with the data, and how you were able to overcome them (or not!)
- Descriptions of any insights into the data or domain that you obtained through your work
- An overview of the code you wrote and existing tools you used, along with instructions on how to run the code

Grading

Your project will be graded holistically, taking into account effort, creativity, degree of difficulty, technical proficiency, quality of the writeup, etc. Note that "negative results" are totally acceptable for this assignment (for example, "here's ten things we did along with a theory of why none of them worked"). In terms of scope, your efforts should be equivalent to 1-1.5 homework assignments (not counting the first one, which everyone hated).

What to Submit

You should submit single zip file containing:

- A file writeup.pdf, with your writeup (described above)
- All code used
- Data files used in the project (if they are > 5MB in size, you can include a text file called data.txt containing links to the data sets)