Project 1

Telling a Story with Data*

The Data Science Pipeline

Learning Objective: At the highest level, the goal of this project is to (a) pick a data set of your choice (b) in tandem identify a set of interesting questions addressable by the data set (c) tell a nice story intertwining the data set with the questions. By choosing a topic of your own, you will gain exposure to various stages of the data science pipeline we have discussed in class: *Question, Wrangle, Explore, Model, Communicate*.

You will work in teams of 2 or 3 people. Following are various stages and deliverables for the project:

1. *Project Proposal (20% of project credit)*
   Due: 11:59 pm Friday, February 14 (week 5)
   How: Canvas

Provide an overview of the project goals and the *motivation* for it. Discuss any related work. Did a news article, blog post, web site, something we discussed in class trigger your interest? Briefly discuss the data set you plan to use. Where did you find it? What is its structure? What questions would you like to answer using the data set? Enumerate and briefly discuss at least three interesting and non-trivial questions.

There are a huge number of data sets on a wide range of topics available for free on the web. As with all projects, the proposal is will lay the foundation of work to follow. The course instructor and TAs will provide feedback on your proposal and assist with right-sizing it. Our concerns at this stage will be (a) the "interestingness" of the questions and (b) the amount of effort your project will require; neither too little nor too much.

Your project proposal needs to be 2 to 4 pages.

Once you have identified a candidate data set your team would like to explore, please enter a 1-2-line description and the URL(s) where that data set is located in the Google sheet available in the project folder. Nowadays, there huge number of data sets are freely available. So, unless there is a compelling reason, I would like each team to explore a unique data set to avoid overlap and a richer distribution of project themes.

2. *Intermediate Demonstration of Progress (20% of project credit)*
   Due: Before 5pm Monday, February 24 (week 7)
   How: ~15 minute in person meeting with TAs or Instructor

Work in data "science" often follows the well-established "scientific method": form a hypothesis, perform an experiment, validate or repudiate the initial hypothesis and if needed formulate a new hypothesis. Setup a time to have a brief meeting with the TAs or course instructor to discuss and demonstrate what has been done to date and what remains to be done. The exact details of this will vary from project to project. We will be looking for a good faith effort demonstrating progress towards project completion.

** or … Telling the Story of the Data

3. *Class Presentation (35% of project credit)*
   Due: In class, Monday, March 2 or Wednesday, March 4 (week 8)

Your presentation in class describes your problem domain, dataset(s), methods of analysis and key findings. Present your project in a way that those unfamiliar with the domain or data will understand. Presentations will be limited to 9+1=10 minutes each (presentation + Q&A). Any team volunteering to present on March 2 earns 5% bonus credit; all other teams will be randomly assigned for presentation on either day.

4. *A Jupyter notebook (15% of project credit)*
   Due: 11:59 pm Friday, March 6 (week 8)
   How: Canvas

   You are welcome to perform your work using any *Python based tool* and/or Tableau. Report all your work in a Jupyter notebook. You should detail your problem domain, datasets, methods, assumptions and approaches you have used in your analysis. Your report should also detail your findings in appropriate 'business' language and should include all useful supporting code, charts, graphs, or summaries. There should be enough detail so that the teaching assistants and course instructors can clearly understand, recreate your analyses and evaluate the credibility and soundness of your approach.

5. *Project Video (10% of project credit)*
   Due: 11:59 pm Friday, March 6 (week 8)
   How: Submit a text file with the URL of a web-hosted short video (e.g., on YouTube) to Canvas

   The end result of a data science project is to effect some form of action. Clear communication of the results of the investigation are critical for success. In this component of the project you will produce a short 1-3 minute screencast of your work. Your video will show a narrated demo of your application and/or some slides. We will strictly enforce the three minute time limit for the video, so please make sure you are not running longer. Focus the majority of your video on your main contributions rather than on technical details. What do you think is the coolest part of your project? What insights did you gain? What is the single most important thing you would like to show your audience? Make sure it is up front and center rather than at the end.

Assessment:

These following terms are used to describe your work on this project:

A. Outstanding. Deliverables exceed requirements in all respects. Quality of work / reports are outstanding in terms of content, analysis, thoroughness, clarity of thought and expression, as well as quality and depth of insights. Notebook, presentation, screencast are all very well prepared and clearly presented.
B. Good. Deliverables meets requirements in all respects and may exceed requirements in some respects. Content, analysis, clarity of thought are good and reports and demonstrations provide some insights into subject matter. Deliverables are well organized, well written and presentation is clear.
C. Satisfactory. Deliverables meets requirements in some respects but may be inadequate in some respects. Reports and demonstrations demonstrate basic effort in terms of thought, expression, or analysis. Quality and

depth of insight or research is acceptable, but results or analysis are apparently thin or minimal. Conclusions, details of project plan or supporting documentation and argumentation may be questionable or not well supported. Appearance, lines of argument and/or mechanical details are adequate, but attention to detail is needed.

D. D = Unsatisfactory. Project work generally does not meet requirements. Deliverables are shallow, unconvincing and/or poorly written or presented and there is little to commend it.

Peer Evaluations:

This is a team project. It is expected that each member will contribute equally and effectively towards all aspects of the project (research, development, deliverables, presentation preparation etc.). Peer evaluations will be used to adjust for individual contributions. Please see the course instructors early if there are any unresolvable concerns.