Project Proposal (Due Before Class on Thu. Feb 14)

Learning Objectives:

- Articulating questions of interest before an analysis and establishing the relevance to the associated business.
- Assembling a data set and describing it at a high level using summary statistics and plots.
- Clearly describing the steps needed for an analysis and explaining how they address the questions of interest.
- Communicating effectively the proposal to a technical analysis to a non-technical business audience.

Description

Each team will submit on Blackboard a project proposal in the format of a Jupyter notebook. The proposal describes a dataset you have found and articulates questions of interest that can be addressed by analyzing the dataset, as well as a plan of analysis. The target audience is a senior manager of a business or organization who might be interested in the result of the analysis, and you should communicate in appropriate and convincing language why your analysis is relevant to the business or organization.

The proposal should include the following four components:

A. Motivation: a list of questions that can be (at least partially) addressed by analyzing the dataset and an explanation of why these questions are relevant to a business or organization.

B. Description of Dataset: a description of what is in the dataset and what each field means, along with code that showcases the data using basic summary statistics and plots.

C. Methodology: a description of how you propose to analyze the data, outlining the steps in order and giving as much detail on each step as you think is appropriate so that a trained analyst would be able to conduct the analysis exactly as you would by following your description.

D. Discussion: a critical examinination of why your methodology is appropriate for your chosen questions of interest, and which parts of the analysis you think are the weakest links. If you had more time or resources, what additional data would you collect and why would they be helpful?

Your Jupyter notebook should be formatted in a way that looks professional, with appropriate headers and bullets, and free from spelling or grammar mistakes. You should limit tables, figures, and code outputs only to those that help you make your point, and avoid long and unnecessary outputs that may be distracting to a non-technical audience.

Finding Datasets

You may use any data from your prior experience or from the Internet. A few datasets are posted on Blackboard which you may use, but you are encouraged to find your own data if possible, as this would make your final project unique and look better to future employers.

If you have trouble finding a dataset, the following websites are good places to start looking.

- Kaggle Datasets (from data analytics competitions): https://www.kaggle.com/datasets
- Los Angeles Open Data: https://data.lacity.org/
- New York City Open Data: https://opendata.cityofnewyork.us/
- US Government Open Data: https://www.data.gov/
- Inside AirBnb: http://insideairbnb.com/get-the-data.html
- Federal Reserve Economic Data: https://fred.stlouisfed.org/
- World Bank Open Data: https://data.worldbank.org/

- Lahman's Baseball Database: http://www.seanlahman.com/baseball-archive/statistics/
- IMDB Movie Data: https://www.imdb.com/interfaces/
- Movie Lens Data: https://grouplens.org/datasets/movielens/
- UCSD Recommender Systems Datasets: https://cseweb.ucsd.edu/~{}jmcauley/datasets.html
- Socrata Open Data: https://opendata.socrata.com/

The following websites list additional data sources, organized by topic:

- An automatically generated collection of public datasets organized by topic: https://github.com/awesomedata/awesome-public-datasets
- 19 Free Public Datasets: https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- Flowingdata listing: https://flowingdata.com/2009/10/01/30-resources-to-find-the-data-you-need/
- Financial data: https://quant.stackexchange.com/questions/141/what-data-sources-are-available-online
- Development economics data: https://sites.google.com/site/medevecon/development-economics/devecondata

For this project, you should look for data on the order of kilobytes or megabytes. (Some of above data sources contain datasets on the order of gigabytes, which you are welcome to analyze but I recommend that you focus on a smaller subset to start off.)

Grading Rubric

The project proposal will have a total of 30 points, with 3 points coming from each of the following 10 categories. The below description outlines what it would take to obtain full grade from each category. There will be a 0.5 point deduction for every minor issue and a 1 point deduction for every major issue.

1. Questions of Interest: Clearly describing questions that may be addressed by analyzing the data and phrasing each question in as precise language as possible.

2. Establishing Relevance: Clear and convincing explanation of why the above questions of interest are relevant to the business or organization. The explanation should demonstrate an understanding of the business application and what are the most important issues for the business.

3. Description of Dataset: Clearly and systematically describing what is contained in the dataset and what is not contained. If the data has many files or column headings, you should describe what each file and column contains.

4. Summary Statistics and Plots: Code that correctly loads the data and show appropriate basic statistics about the dataset, such as how many entries is in the data, what time periods does the data cover, what are the averages of the numerical columns, and how many unique elements is in each column. You should also give at least 3 appropriate plots that give a high level view of the data.

5. Clarity of Methodology: Breaking down the analysis into steps and giving a clear description of each step, with as much detail as would be needed for a trained analyst to follow the steps.

6. Correctness of Methodology: Your methodology addresses the question of interest, with no signficant errors or gaps in the logic.

7. Critical Examination: Your discussion section examines your methodology and justifies why it is appropriate, while conceding the weak points.

8. Additional Data: A clear description of what additional data would be helpful for the analysis and why.

9. Appropriate Communication: Avoiding technical jargons and explaining technical words if you choose to use them. Keep in mind the target audience is a senior manager with a non-technical background.

10. Polish: Appropriate use of headers, bolding, and bullets inside Markdown cells. Proposal is organized in a easily readable way and not overly lengthy. Proper spelling and grammar. Avoiding long tables or code outputs that hamper readability.