SI 330 Fall 2019 Project

Goal

The purpose of this project is to have you demonstrate your ability to acquire, manipulate, and clean real-world data both as an individual data analyst as well as in a larger team environment. The intention is that through this project you will demonstrate the skills you have learned in this course, and the rubric for grading is aligned roughly with those skills.

Overview

You're a fresh data analyst out of the BSI and you just landed a gig to work at a skunk works media group which runs a website and a couple of television channels for seeing sports and athletic events. They've employed a team of fresh new BSI's (45 of them, which is a ~ $2.7M investment per year) to help build out content for the company. The company believes there is a market for data driven sports reporting, where expert personalities are replaced with insights from data. The vision articulated by the CEO was that each person would work with different data related to sports/athletics, and the company would build a single portal around that data, then leave dark mode and start to raise funds using this as the seed of the business model. You and the other 44 new hires have one month to build something interesting.

The CEO's Director of Quality Control (me) is going to evaluate your efforts across three dimensions, data gathering, data cleaning, data preparation, and data reporting. At the end of the probationary period (coincidentally this happens to be the end of this course!) you'll be given a grade. The criteria for interesting data by the company is actually quite broad, since part of the appeal is to go beyond the big money maker sports (e.g. NFL, MLB, NBA, and of course, NHL) and hit the "long tail" of people's interest. Badminton? Check. Finnish Dressage? Sounds cool! Little league baseball? Why not!? However, the CEO's vision is that for the sport you choose you must have at least the following attributes:

1. Must contain events which cover a time period
2. Must contain entities (e.g. teams) which have a geography
3. Must have some kind of competition leading to outcomes
4. Must have variation in the above of at least 3 distinct units.

This last item means that there must be at least 3 different time periods (e.g. seasons or events), 3 different geographies (e.g. locations or routes), and 3 different outcomes (e.g. championship games).

Delivery #1 - Pitch

The Director of Quality Control is nervous, this is significant capital outlay for new hires and he's new on the job and wants to make sure everything is looking good. You need to provide him with an 800 word document on the dataset you are planning to look for and how it meets the four criteria above. In addition, the Director recently read a book on building community and decided that everyone is going to have to show up on **October 16th at 8:30am** to share their pitches with one another. Text copies of the pitches should be sent to him via email by midnight the same day with the subject line **OMG I had this great idea!**.

Delivery #2 - Notebook

The Director of Quality Control is expecting you to provide the final project in the form of a Jupyter notebook along with associated data files. There will be a format for this notebook and datafiles released at a later date. The **deadline for this is December 12th, 2019**. There will be no final graded project presentations.

Evaluation Rubric

The Director of Quality Control will grade each of the submissions across four dimensions, with each dimension being awarded a grade out of three points indicating the level of sophistication of analysis you have performed. For instance, a 1 would be awarded if the supplied code was minimal but sufficient, a 2 would be awarded if the notebook was good but

could be easily done by some other company or analyst, and a 3, the highest value, would be awarded for outstanding sophistication. Thus the grade which will be assigned to each analyst will be out of 12 points, with partial points possible. A list of the four dimensions of evaluation and examples of different levels of sophistication follow.

Dimension 1: Data gathering

- Top marks: Novel datasets scraped from the web using beautifulsoup, selenium, scrapy, or other libraries, and those datasets are sufficiently interesting and complex to aid in your analysis. Use of "live" sources where you connect to web APIs.
- Middle marks: Use a breadth of dataset formats such as CSV, XML, JSON, or databases. These may be created through preprocessing of data or download of data directly
- Bottom marks: Use a series of canned CSV files that you have downloaded from Kaggle with pandas read_csv().

Dimension 2: Data cleaning

- Top marks: A mixture of methods used to clean data columns, including significant use of regex with capture groups, apply(), time series features such as resample(), or other mechanisms to align and preprocess data (e.g. changing coordinate systems in geographical data). A use of an ETL library would also classify as going above and beyond and fall into this category.
- Middle marks: A few regexes and a few data preparation functions which are significant and suitable but not particularly sophisticated.
- Bottom marks: Very basic regex or string split() functions at a level less than addressed in this course (e.g. an SI 206 level).

Dimension 3: Data manipulation

- Top marks: Data is joined from various places, weaving SQL and pandas work together to answer specific questions posed by the author. Use of SQL subqueries and joins with pandas data is demonstrated. Appropriate use of an object-relational model (e.g. SQL Alchemy) would also fall in this category, along with specific design of database architecture (e.g. inclusion of an ER diagram outlining how data was manipulated into a coherent form).
- Middle marks: Only one of SQL/pandas is used, and joins are mostly straight forward merge() functions or basic queries. No significant attempt to create a structure to be reused by other kinds of questions.
- Bottom marks: One or two merges of data, or poor merging (inefficient or broken merges).

Dimension 4: Data reporting

- Top marks: Data is presented in a coherent and meaningful manner for the question being used, and a toolkit like matplotlib, seaborn, bokeh is used to visualize and demonstrate findings in the data, or advanced statistical techniques (regression, time series correlations, etc.) are used appropriately.
- Middle marks: Jupyter notebook contains a mixture of markdown and code, runs without error, and is presented as a computational narrative.
- Bottom marks: Jupyter notebook has minimal narrative and is presented mostly as if it were a computer program or coding exercise.

Important Notes

The Director of Quality Control (again, me!) will use his sole judgement in determining the level of sophistication you have reached in your project for each of the areas above. For instance, if you scrape one tiny bit of data and then just use canned CSVs for everything else that's probably not going to get awarded top marks in that category. Similarly, points may be deducted if you do not show some of the competencies at the middle level (e.g. notebook runs without error and

is a computational narrative) but you went beyond. Thus each dimension of your work is graded holistically, and the examples above are not exclusive ways to show you learning.

Philosophically, with this assignment it is up to you to demonstrate to me what you have learned in this course. To this end the top category for each dimension goes above and beyond the lecture material, and my examples above are intended to help you demonstrate that you have achieved an ability to extrapolate and engage in independent learning in the topic of data manipulation broadly.