Final Project

At this stage of the course, you guys know how to build an entire Data Science pipeline: collecting data with scraping or APIs, importing data, cleaning and processing data, applying Machine Learning techniques, validating models' significance and accuracy, and visualizing results. Congrats! That's a lot to learn in six weeks.

Sadly, absolutely none of this matters without the final, and in some ways most important piece, presenting the value of your new knowledge to non-technical decision makers.

For the final project, in addition to cementing all the lecture knowledge, you will practice the incredibly valuable skill of effectively convincing others to take actions from your data driven conclusions.

Timelines

Today: Choose a dataset from the list of pre-vetted datasets below so you don't waste time on looking for one, then begin your analysis.

Thursday: You will one hour to finish your analysis, then presentations will be held in class, with the structure described below.

Presentation Structure

All presentations will follow the following structure:

1) 5 minutes to present
2) 3 minutes for Q&A

We will time everyone to make sure that everyone abides by the time limit so that everyone has an equal chance to present. We will also provide feedback on your presentations and our comments on how they can be improved.

Project Guidelines

1) **Select a dataset** from the below list that is suited to your interests and your level of expertise in the domain.

2) **Select a specific audience** you will address for your presentation. This should be dependent on who you want to demonstrate your newfound ML knowledge to, as well as which dataset you choose. For example, if are you dealing with financial data, are you speaking to investors? Your boss? Regulators? For a social policy dataset, are you talking to elected decision makers or ctizens?

In real world situations, you will need to tailor your presentation to your audience's terminology, objectives, and familiarity with data science and machine learning.

3) **Explore the data**, looking for something that is worth focusing on. You may or may not choose to include this section in your final presentation, but it's what every data scientist starts with for a reason- you can't find or articulate actionable insights in the data if you don't know what the data describes.

4) Based on your explorations, **come up with a suggestion, hypothesis, or conclusion with measurable value.** For example, for social or government data your hypothesis might sound like "we should focus spending on these X factors to get the most benefit" or "life quality can be improved by taking these X steps". For financial data, it might should like "here's how we can make better risk adjusted returns." For general data, it might sound like "we can target customers more effectively with X," and so on.

5) **Find up to 3 data-centric pieces of evidence** which support your hypothesis. For example, in social data, "with 95% accuracy we can classify people into sick or healthy based only on their diet, which supports the hypothesis that educating people on proper diets is an efficient policy."

6) **Try to flip the tables and poke holes in your argument**. If you can do this well and not get emotionally attached to your hypothesis, you are well on your way to persuading people with data. Identify up to 3 counter arguments, and then try and use data to refute these arguments. To continue the previous example, a counter argument might ask "what other hidden correlated variables might account for this prediction? When we reduce dimensions and change input parameters to we get the same results? Maybe these people with a poor diet are all smokers, maybe that's what is killing them."

7) **Prepare a short presentation** in which you will attempt to persuade us of your hypothesis. Your presentation must include the following parts:

a) **Set the stage** by briefly telling the class and instructors which data set you chose and why, and which audience they should be listening as (ie politicians, investors, your boss etc).

b) **Explain your hypothesis** and WHY it matters. The best way to do it is to show you're your data driven conclusion will be better than existing solutions

c) **Explain how the data demonstrates** up to 3 points supporting your hypothesis

d) **Explain up to 3 counterpoints** against your hypothesis, and how the data does not demonstrate these

Pre-Vetted Datasets

1) **McDonald's Menu Nutrition S tats**
   https://www.kaggle.com/mcdonalds/nutrition-facts/data
   Difficulty - Easy
   Practice - basic pandas and charting
   Notes - menu nutrition stats per item, no (non-obvious) trends

2) **Youtube Video Views** **https://www.kaggle.com/datasnaek/youtube/data**
   Difficulty - Hard
   Practice - text data cleaning and pandas
   Notes - Video views and characteristics, text heavy, irregular data (tag lists)

3) **Twitter Metrics** **https://www.kaggle.com/rtatman/twitter-vs-newsletter/data**
   Difficulty - Easy

Practice - basic pandas and data cleaning, prediction/causation
Notes - Tweet stats with month, day and hour of publish, good for aggregating across hour of day for trend analysis

4) **Video Game S ales** https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/data
Difficulty - Medium
Practice - data cleaning, prediction/causation, time series
Notes - Video game titles with ratings, sales data, and categories, mix of text data. Good for data exploration by publisher, genre, etc

5) **College Tuitions vs Salaries**
Difficulty - Medium
Practice - advanced pandas (table joining), basic relationship analysis
Notes - Multiple tables of universities costs, and salaries by university, region, and by degree

Finance/Economics:

6) **Basic Income Survey Results**
https://www.kaggle.com/daliaresearch/basic-income-survey-european-dataset/data
Difficulty - Medium
Practice - data cleaning, categorization, tearing down late stage capitalism
Notes - Classified responses to survey questions and details of respondent, very text heavy, good for finding trends/relationships in responses

7) **Cryptocurrency Prices** https://www.kaggle.com/jessevent/all-crypto-currencies/data
Difficulty - Medium
Practice - time series
Notes - Prices for nearly every token/cryptocurrency out there (1080 in total) for 5 years

8) **Bitcoin Historical Data** https://www.kaggle.com/mczielinski/bitcoin-historical-data
Difficulty - Hard
Practice - time series
Notes – Historical prices for Bitcoin, the most famous cryptocurrency

9) **Credit Card Fraud Detection**
https://www.kaggle.com/dalpozz/creditcardfraud/data
Difficulty - Hard
Practice - time series, classification
Notes – Unlabelled features with no possibility of guiding intuition

Health/Medicine:

10) **Breast Cancer Data** https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data
Difficulty - Medium
Practice - prediction/causation
Notes- lots of quantified columns and classifications of malignant or benign, good for relationships and characteristics of malignant tumor

11) **Heart Disease Dataset**
http://archive.ics.uci.edu/ml/datasets/Heart+Disease
Difficulty - Medium
Practice – predictive/classification
Notes- precleaned data with lots of useable features.

Law/Gov:

12) **World Happiness Ranking** https://www.kaggle.com/unsdsn/world-happiness/data
Difficulty - Easy
Practice - basic pandas and charting
Notes - survey results by country for happiness, relationships already straightforward

13) **Los Angeles Crimes** https://www.kaggle.com/cityofLA/crime-in-los-angeles/data
Difficulty - Hard
Practice - data cleaning, geographical analysis, general pattern investigation, time series Notes- date and text intensive, will require a lot of data cleaning, good for map charts and detecting crime patterns

14) **Global Terrorism Database** https://www.kaggle.com/START-UMD/gtd/data
Difficulty - Hard
Practice - big data, data cleaning, investigation, time series
Notes - very large set of data, highly non-standardizes and irregular, requires lots of cleaning and investigation to find trends or insights

Sports:

15) **European Football Team Match S tats**
https://www.kaggle.com/jangot/ligue1-match-statistics/data
Difficulty - Medium
Practice - general pattern investigation, time series
Notes - game statistics by team (shots, tackles, fouls etc), good for analyzing/inferring team strategy from statistics

16) **FIFA 2017 (video game) Player Database**
https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global/data **Difficulty - Easy**
Practice - basic big data, pandas, data cleaning, basic algorithm

Notes - rating and personal stats per player, good for reverse engineering FIFI overall rating algorithm, not much else