First Project

Now that you guys have learnt how to collect, clean, wrangle, analyse and visualize data, it's time to put it all together! Presentations are a major part of the Data Science process, be it to your boss, investors or colleagues.

They're a crucial piece of the puzzle and often your project funding, approval or investment will be completely dependent on how you present your findings. Yes, they are that important!

Timelines

Today: Choose a dataset and get cracking. We are going to give you a repository of pre-vetted datasets to choose from so you don't waste time on looking for them.

Thursday: Presentations will be held in class, with the structure specified below.

Guidelines

1) **Select a dataset** from the below suited to your interests, what areas of the class you want to practice more, and your level of expertise in the domain.

2) **Select a specific audience** you will address for your presentation, which is dependent on the 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? Average citizens?

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. This is a vital skill if you want to use machine learning for practical use cases, so it's important to begin to develop it as early as possible.

3) **Explore the data**, looking for something that is worth focusing on. This section may not be included in your final analysis or 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 and/or conclusion** that has a tangible value and is actionable. 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 "lives can be measurably improved by taking these X steps". For financial data, it might should like "here is how can make above market risk adjusted returns." For general data, it might sound like "we can reach more people and 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, "since the data shows that a poor diet is strongly correlated with poor health, educating people about proper diets before they get sick is likely to be a more 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 making 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 "are rates of sickness correlated to

anything else? 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 how its conclusion and your actionable insight will improve something worth improving.
>
> 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

Structure

All presentations will follow the following structure:

1) 5 minutes to present all of point 7) above
2) 3 minutes for Q&A

We will time everyone to make sure that everyone abides by the time limit, and so that we don't overrun class time. We will also provide feedback on your presentations and our comments on how they can be improved.

Pre-Vetted Datasets

1) McDonald's Menu Nutrition Stats
   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 Sales
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

Health/Medicine:

8) 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

Law/Gov:

9) 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

10) 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

11) 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:

12) European Football Team Match Stats
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

13) 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