Ideas for Data Science

1. Can we look at trying to predict the rise of housing prices in an area from Yelp reviews?
   1. https://www.yelp.com/dataset/download
   2. Would likely have to merge data from Yelp and NYC (Census Data for example): http://maps.nyc.gov/census/
   3. Can draw some inspiration from here: <https://towardsdatascience.com/predicting-wealth-in-nyc-53b854c0a8a0>
   4. It would be cool if we could combine this with walkability scores, the idea being that housing prices will rise as amenities like public transport and grocery stores become denser within cities (gentrification)
      1. I’ve emailed Walk Score to ask for student access to data, but it could take up to two days for a response and I’m not sure about the weekend: <https://www.walkscore.com/cities-and-neighborhoods/>
   5. If we can combine walkability data, yelp review data, and census data, I think we can potentially create something awesome. Not sure how advanced will be able to get (geomapping and such for example), but I think it could be a cool project.
2. Use Google Trends Data (search terms) by country to predict changes in macroeconomic indicators (2004 to present)
   1. <https://trends.google.com/trends/explore?date=all&q=google>
3. Can do the kind of Spotify correlation matrix using online retail data to suggest customers buy another product based on what they’ve already bought: <http://archive.ics.uci.edu/ml/datasets/Online+Retail>
   1. Would be relatively simple, but probably not terribly novel or unique
4. Pricing model for cocaine: https://www.kaggle.com/everling/cocaine-listings
   1. Inspiration : <https://medium.com/thought-skipper/dark-market-regression-calculating-the-price-distribution-of-cocaine-from-market-listings-10aeff1e89e0>
5. We could somehow invent an always win strategy for Monopoly (and ultimately for life): <https://data.world/dataremixed/monopoly-board-frequencies-and-economics/workspace/file?filename=Monopoly+Simulator.xlsx>

Could be good if we find a good thing to merge with

1. Affordable housing developments in Chicago: <https://data.cityofchicago.org/Community-Economic-Development/Affordable-Rental-Housing-Developments/s6ha-ppgi>
2. Avg. rent by class/area in Hong Kong: <https://data.gov.hk/en-data/dataset/hk-rvd-tsinfo_rvd-property-market-statistics>
3. Something about how close offices or apartments are to transit?
4. It’s old 2007-2013, but it could be cool to use pornhub data to predict things. Kind of like what the guy who write Everybody Lies uses. Would be fun and controversial for the class too. Only problem is the data does not seem to be geotagged in any way, which would definitely enhance our ability to use it for something cool. <http://sexualitics.github.io/>
5. Business Dynamics: <https://www.census.gov/ces/dataproducts/bds/data.html>
   1. The Business Dynamics Statistics (BDS) is a public use data set of annual aggregate statistics describing establishment openings and closings, firm startups, job creation and destruction by firm size, age, industrial sector, and state.
6. Brookings Global Metro Monitor: https://www.brookings.edu/research/global-metro-monitor/
   1. Richard Florida’s Urban Productivity Ratio
   2. Not really sure how to download though
7. Financial Well-being data: [https://data.world/cfpb/financial-well-being-survey-data#](https://data.world/cfpb/financial-well-being-survey-data)
   1. Depending on what we can match geographically, we can maybe compare it to urban vs. rural, millennial vs. octogenarian, and another of things in developing a predictive model
   2. Could potentially look at

Other Datasets I’ve found

1. Hate speech identification from tweets: <https://www.cooldatasets.com/>
   1. Could create a filter for key words and terms to flag as hate speech
   2. Might be able to find another dataset containing random tweets that we could merge and test on to see if it’ll flag those inappropriate tweets
      1. It’s actually quite possible to scrape some twitter data by hashtag – it’d be cool to look through certain hashtags and filter hate tweets
      2. Examples of datasets with random tweets by hashtag or subject: <https://docs.google.com/spreadsheets/d/1uhAR5PDb-1OnTrvotfN3XiXc5Ivf2CioTVb2-oFDqR8/edit#gid=724681126> and <https://archive.org/details/twitter_cikm_2010>
2. Database on plane crashes between 1920 and 2018: <http://www.planecrashinfo.com/database.htm>
   1. Provides information on the date, time, airline, reason for crash, number of fatalities, etc.
   2. Could predict most common causes of plane crashes and how this has changed over the past decade
3. Information on all proposed amendments to the US constitution between 1787-2014: <https://www.datazar.com/file/f5d4b5cb5-8a4e-4ff0-905e-fe2f1acbd5e0>
   1. Could find and predict the most common phrases and amendments to the Constitution, how frequently they come up, and test whether certain ones will reoccur
4. Hate crimes by state: <https://github.com/fivethirtyeight/data/tree/master/hate-crimes>
   1. Fivethirtyeight used this dataset to discuss how hate crimes might be tied to inequality. Might be cool to do a train/test model to predict hate crimes by state