Group Proposal

Group members

Alice Chang, Ben Thomas, and Olek Wojcik

Project proposals

College characteristics and post-graduate employment

- **Question:** What are the major determinants of the employment rate of graduates from a given college, 6 months after graduation?
- **Model type:** This would be a regression model with the graduate employment rate as the dependent variable. Inferential.
- **Data:** Data would be available through most rankings publications. U.S. News publishes data on most colleges and universities, which we could access through their "College Compass" tool. Additionally, we would use data from IPEDS (https://nces.ed.gov/ipeds/use-the-data).

Reddit comment toxicity

- **Question:** Can we build a model to predict whether and how a Reddit comment is "toxic"?

- **Model type:** Classification. Predicive.
- **Data:** We will train our model on a dataset of May 2015 omments released by Reddit (https://www.kaggle.com/reddit/reddit-comments-may-2015/download). The final training dataset would likely involve our own qualitative coding and classification of comments in the existing data from May 2015 into "safe"/different categorizations of "toxic." Categories may include "insult," "threat," "obscene," and "identity-based hate." Note also the Reddit api: https://www.reddit.com/dev/api/.

Corporate twitter content and quarterly earnings

- **Question:** Is there a relationship between the content/activity/engagement of a corporate twitter account and their quarterly earnings? Many brands on twitter have accounts that serve to promote their products, and they often take different approaches to how to run them. Some brands maintain a serious tone while others attempt to have a more personal, funny account. Looking at both the content of different corporate tweets, as well as measuring their engagement by likes, retweets, and replies, we could try to see a predictive relationship with quarterly earnings, change in quarterly earnings or another measure of company health.
- **Model type:** Regression. Predictive.
- **Data:** The twitter API could be one dataset, and the other could be the quarterly earnings reports of various large companeis with twitter accounts. https://developer.twitter.com/en/docs