Lily McElwee
Foundations of Data Science

Capstone Project – Outline

What is the problem you want to solve? Who is your client and why do they care about the problem?

I've decided to examine shifting political sentiment in Silicon Valley. Historically, California is a blue state politically, and Republicans are at a disadvantage in many parts of the state; in the 2015-16 cycle to date, Democratic party and candidate contributions have raked in 58.4% of total contributions while the Republicans have received just 37.7%. Party leaders focused on the Silicon Valley area want to improve fundraising efforts by directing marketing expenditures to those areas in which they have, on average over the past three presidential election cycles, sourced the lowest amounts of funding (distribution of funding sources across the state) and for which funding #s declined in 2016 versus 2008 (for this latter element, I may consider only those contributions occurring before the party nominations, in order to account for the fact that the 2016 cycle is not complete yet).

What data are you going to use?

I'll be using the FEC's official data on presidential campaign contributions from a list of Silicon Valley zip codes for 2008, 2012, and 2016 election cycles. The data can be found by entering specific zip codes here: http://www.fec.gov/disclosurep/pnational.do#.

Since the list of Silicon Valley zip codes is pretty extensive (58 total), I'm going to begin my exploratory analysis with just one or two (e.g. Campbell, CA - 95008). I'll begin by downloading the individual datasets from the FEC website, and classifying the candidates represented by their party affiliations to get aggregate #s for funding to R/D/other for each zipcode in each election cycle.

Ideally, I'd ultimately add further dimensions. Since the datasets span contributions across the election cycle, I'd like to classify each candidate by whether they became the nominee (N/NN), whether they ultimately won the general election (W/L), and whether the contribution came before/after a candidate became the nominee (B/A). This will allow me to understand elements such as party loyalty (whether funding to Republicans improved/declined after the nominee was selected), but also distinguish funding to candidates overall from candidates that actually took part in the primary. As I've seen some

preliminary exploration of the dataset for Campbell, CA, for instance, there are candidates from both parties that received contributions but did not ultimately participate in the California primary (e.g. Jeb Bush in the 2016 cycle). I have been able to acquire some of this data already from Wikipedia, and have created a secondary dataset containing the following information on all participants in the 2008, 2012, 2016 primaries: cycle, name, votes received (in primary), share of votes received (in primary), party affiliation, whether they won the primary, date of primary, whether they received the party nomination, date of nomination, whether they became president, and the party affiliation of the incumbent president.

What are your deliverables?

My deliverables will include the two datasets I put together (the first combining zipcode-specific data into one dataset for all of Silicon Valley encompassing the 2008, 2012, and 2016 election cycles, and the second combining the list of participants in the CA primaries over those cycles and information on the party affiliation of each candidate, whether each candidate won the primary, was nominated by the party, and won the presidential election). Additionally, I will hope to learn in which zipcodes the Republicans are weakest in terms of funding on absolute terms and relative terms (versus Democrats), and in which cases funding declines after the primary takes place. While I may be able to reach further insights from the data, I will begin with this target.