**Capstone Project Proposal**

**Author: Chris Leisner**

**Business problems we want to solve**

1. Score banks based on their propensity for receiving complaints.
2. Find factors that are driving this complaint score (e.g., location of the branch receiving the complaints).

**Who is the client?**

Both banks and consumers who are shopping for banks and bank services. The complaint propensity score(s) could be used by consumers who are shopping for a bank or a particular bank product (such as mortgages or auto loans). The predictors of this score could be used by the banks themselves to improve their customer service. For example, if number of bank employees (normalized by bank size) is a strong driver of number of complaints, the bank could respond by adding employees, or at least by making existing employees more effective. Maybe number of branches in a zip code (normalized by population density) is a driver of complaints as well.

**Dataset to be used:** I will start with the consumer complaints dataset below, which can be downloaded:

<https://catalog.data.gov/dataset/consumer-complaint-database#topic=consumer_navigation>

I’ll merge on data from other sources, e.g. data on bank sizes as measured by market cap and number of branches. The ranked bank data at the two links below must be cut and pasted into .csv files.

<http://www.usbanklocations.com/bank-rank/total-equity-capital.html?d=2016-09-30>

<http://www.usbanklocations.com/bank-rank/number-of-branches.html>

Some other bank characteristics are at the link below. Again, this will have to be cut and pasted into .csv files.

<http://www.usbanklocations.com/bank-rank/>

Here are zip code populations and areas as of 2015. They can be downloaded, merged, and used to find zip code population densities:

<https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk>

https://www.census.gov/geo/maps-data/data/gazetteer2015.html

**Approach to solving the problem:**

1. Merge datasets together
2. Clean data
3. Decide on a target variable.
   1. This could be either a weighted sum of complaints (at the company level) where the weights would be a function of the complaint category or the product. However, there are 95 distinct issues, so I'd probably have to think of some way to group the issues. There are only 12 distinct products, so I probably won't have to group those. In addition, the weighted sum of complaints would have to be scaled by the size of the company. Some ways to measure a company's size are market capitalization, sales, total assets, and enterprise value.
   2. Alternately, I could make several different target variables by considering separately complaints in each complaint category and product. This data set has many different categories of complaints (these categories are called "issues" in the data).  Some of these issues are "communication tactics", "APR or interest rate", and "billing disputes". In addition, each complaint is connected to a type of product offered (e.g., vehicle loans or mortgages).
4. Do some brainstorming to select candidates for predictor variables.
5. Do some simple univariate analysis to find relationships between independent variables and target.
6. Build predictive model on training set, test on test set.
7. Do some research to find out what people are looking for in a bank (e.g. low fees, investment services…) Use this to plan interactive dashboard; we can anticipate consumer questions and decide what the dashboard should be able to drill down into.

**Deliverables**:

1. Code
2. Interactive dashboard constructed using “flexdashboard” package in R
3. List of drivers of complaints, with weights (larger weights for more powerful drivers)
4. White paper
5. Slide deck