Capstone Project

Combine data on a Chicago’s crime patterns with other data sources [geo-spatial polygons, Mass transit station locations, police stations locations, demographics, and weather data]. Then, use the resulting data to examine correlations and [hopefully] find enough evidence to suggest new station locations and/or policing patterns. Note: Proximity to transit/police stations will likely only correlate with certain types of crime.

**What is the problem you want to solve?**

Find significant patterns in crime that could be used to educate officers/guide resource deployment and contribute to improved police infrastructure/behavior.

**Who is your client and why do they care about this problem?**

I will have a hypothetical client in the Police Department [PD] of the city. The purpose will be to explore a number of crime characteristics in the hopes of recommending how the force might better deploy their resources. Based on the geo-spatial characteristics discovered, I may also provide a framework for determining where to open future police departments. In theory this will be similar to business supply chain efficiency problems. I will need the following information to solve this:

* Which types of crimes are correlated with proximity to points of interest [transit stations and police stations]. The obvious assumption/intuition being that crime reduces in proximity to a police station and may increase near mass transit stations. This may not bear out in the data.
* Demographics of a neighborhood may suggest the need for another station. For the sake of simplicity, I will assume that each station has the same # of police officers.
  + E.g. Community demographics may correlate with higher crime and thus require more enforcers.

**What data are you going to use for this? How will you acquire this data?**

I would like to have the following data sources, but I am not yet sure where to get all of them:

* Crime data: CHICAGO crime data
* Neighborhood/precinct polygons
  + Source: https://data.cityofchicago.org/.
* Mass transit locations [x&y]
* Police station locations [x&y]
  + Source: https://data.cityofchicago.org/.
* Demographics mapped to polygons mentioned above
  + Source: https://data.cityofchicago.org/.
* Basic Weather data
  + Source: [Weather Underground](https://www.wunderground.com/history/airport/KMDW/2016/3/5/CustomHistory.html?req_city=Chicago%20Midway&req_state=IL&reqdb.zip=60499&reqdb.magic=11&reqdb.wmo=99999)

**Initial Data Variables [Features] to be used**

* Categories and classifications of crime
* Neighborhood/precinct
* Crime coordinates
* Dates and times parsed out into requisite features
* Weather categories
* Minimum [Euclidian] distance to points of interest
  + I think I could use something like this to calculate these, but I am not positive:
    - [r - Calculating min distance between to geo points](http://stackoverflow.com/questions/27782488/r-calculating-the-shortest-distance-between-two-point-layers)

# Data Exploration

May use some other crime data, but the base Chicago crime data has the following variables/features:

1. Dates and Times
2. Category
3. Descriptions (subset of Category)
4. DayOfWeek
5. Police department District
6. Resolution
7. Address
8. X and Y Coordinates

I will have to find the other data on my own. Other data I hope to use for examining correlations and eventually building a machine learning model:

1. Neighborhood polygons
   1. To group geographically
2. Demographic data mapped to neighborhoods [or other geographical grouping set]
   1. To examine correlations and to use as a feature within prediction model
3. Mass Transit station coordinates
   1. As a proximity factor within prediction model and a possible correlation exploration for certain types of crime.
4. Police station coordinates
   1. As a proximity factor within prediction model and a possible correlation exploration for certain types of crime.
5. Weather data
   1. To examine correlations and to use as a feature within prediction model

## Data Munging

The following are obvious steps I will need to take to adjust the existing data:

1. Separate date/time metric into
   1. Year
   2. Month
   3. Day of Month
   4. Day Name
   5. Hour of Day
2. Join multiple data files together
3. Grouping data
4. After exploring some graphs, create new variables that group hours into categories such as:
   1. 2 Hour increments
      1. Early AM 5-7 am
      2. Morning Commute 7-9 am
      3. MidMorning 9-11 am
      4. Midday 11-1 pm
      5. Early Afternoon 1-3 pm
      6. Late Afternoon 3-5 pm
      7. Afternoon Commute 5-7 pm
      8. Evening 7-9 pm
      9. Etc.
   2. Broad Categories
      1. Commute Times
      2. Dark Hours
      3. Light Hours
      4. Etc.
5. Create broader crime classifications such as
   1. By classification
      1. Felonies
      2. misdemeanors
   2. by meta-category (maybe using UCR categories)
      1. Financial Crime
      2. Violent Crime
      3. Property Crime
      4. Other
      5. Etc.
6. Using loops and/or dplyr functions to create geographical stdev ellipses and density metrics.
7. Etc.

## Produce visualizations and observations on patterns found, correlations, etc.

Include Boxplots, histograms, scatterplots, maps, regression models, etc. This will be fleshed out during the above data exploration.

# Create final visualizations and write up

I will use various methods for visualization. After the presentation, I will also use this project to learn Shiny.