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:

* Determine if types of crimes are correlated with proximity to points of interest [transit stations and police stations], the 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.
  + For the sake of simplicity, I will assume that each station has the same # of police officers.
* Demographics of a neighborhood may be a significant indicator of crime levels.
  + 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

# Data

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

The other data I hope to use for examining correlations is as follows:

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
   6. Etc.
2. Cleanse data of NA’s
3. Reformat data and create appropriate ordered factors
4. Reduce data down to valid spatial boundaries as represented by communities and/or police districts.
5. Appropriately transform spatial data per package
6. Join multiple data files into one superset of data to allow for varied exploration.
7. 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. Or broader Categories such as
      1. Commute Times
      2. Dark Hours
      3. Light Hours
      4. Etc.
8. 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.
9. Create functions to do the following
   1. Find closest PD and Transit station to each crime
   2. Calculate the distance between each of the above.
   3. Create spatial shapes [ellipses and circles] that correspond with the standard deviation of spatial dispersion and/or pertinent areas around each location of interest.
   4. Calculate crime /population densities, etc.
10. Group data in interesting ways to explore the data via histograms, linear models, etc.

# Data Exploration

Note: This is merely meant to provide color on my initial exploration process. All final graphs, etc. can be found in the final write-up file.

After setting up the superset of data, I initially focused on relationships between daily/monthly crime rates [initially defined as count/community population] by sub-category and the following:

* Various Demographics
* Location Descriptions
* Weather [temperature and precipitation]
* Proximity Metrics
* Temporal Charateristics

I also experimented with clustering and other techniques I might use for a predictive model, but in the end I decided to stick with the original spatial idea and leave the machine learning for the next project.

Demographics had much less predictive power than I initially expected, even at the annual level; although, I retained certain factors in the final write-up. Weather data [broudly applied by month] was significantly correlated as were some temporal factors such as hour of day.

I found a function capable of determining the closest PD and Transit Station to each crime, but it did not correctly calculate Euclidian distances, so I added a function that could and verified its output. With that in place, I explored proximity metrics and found that crime was not decreasing as much as I assumed it would near police stations. When I calculated the crime density of expanding circles, I actually found that crime density increased as one moved closer to a police department. Transit stations, however, followed expectations.

When my mentor and I met for one of our milestone discussions, we decided that the analysis had spread too thin and that I should focus back in on crime’s spatial characteristics. As such, I dropped the weather analysis and refined the demographics down to a few key stats. Also, assuming that communities might have an outsized political effect on enacting any proposed solution, I focused the final story on each community’s crime density defined as count/km^2. I found a way to summarize the crime/transit proximity data at the community level by creating a variable representing the distinct number of points “closest” to each of a community’s recorded crimes, e.g. the number of stations within close proximity to a community. Interestingly, that factor ended up having an even higher correlation with crime density than community population density.

In all, this was a very instructive exercise in working with geo-spatial data. I was able to learn a great deal about spatial visualization and regression models (both linear and non-linear). I hope you enjoy the end product.