**Capstone project 1 milestone report**

This project makes use of crime data of Chicago and the questions I am trying to answer are: where do crimes happen and what affects them? Do different types of crimes show similar patterns or different patterns? Where is the dangerous zone or safe zone in Chicago? These are important questions that many people in Chicago would like to know the answers to.

Regarding crime analysis, the clients should include law enforcement, people living in Chicago and maybe even the government. For law enforcement, the geographic and temporal pattern of different crimes in the city can better help them allocate their resources to the most important spots instead of randomly patrolling in the city. For average residents in the city, maybe they can try to avoid areas (or the time) where certain types of crimes are high and better protect themselves. For the government, the reason some places have high crime rates may be related to racial, economical, and even educational reasons. Trying to find ways to mitigate the situation could be an important task for the government and law makers.

First the most important data file, the crime data can be acquired from Chicago data portal. I will also be using weather data (from NOAA website), US Census data regarding social, education, household income and poverty status (from US census website). From the same Census website, the tracts data can also be downloaded to use as location to visualize all the data on the map. Chicago transportation data will be used as well to analyze the relationship between public transportation lines and crime locations.

This crime data set has very rich information. It contains the time stamp, location, block, crime type, community area, arrest made or not, and geographic coordinates (latitude and longitude). The time can be used for time series analysis. Crimes can be visualized by year, month, day of the week and even hours of the day. These graphs will clearly show temporal patterns of different crimes. For instance, which month of the year has the most incidents of murders? What time of the day, do burglaries happen at a high rate? Knowing this will help law enforcement be more prepared. The community column is useful when combined with other statistics of Chicago. For example, if we combine Census data such as income level, unemployment, poverty level and education with the crime data, we can explore the relationship between total crimes or specific type of crimes and the Census statistics. How does education level in a community correlate with its crime rates? Another important feature this data set provides is the geographic coordinates of each crime. This comes in handy when people want to visualize crimes on a map. And hot spots can be easily identified when crimes are plotted on a map either as a heatmap or as dots. They also show spatial patterns and clustering of different crimes. If combined with temporal data, one can even track how crime hot spots change over time. By combining all the features here, one can try to build a classifier to predict the type of crimes based on information such as location, month, time, community, coordinates and even temperature. A good classifier could potential help the law enforcement decide how many officers to send when they receive a reported crime.

Even though the crime data set contains very much useful information, it also has limitations. First, these crimes were entered based on the reported incidents. Whether the status was later changed due to further investigation is not guaranteed. Second, the Chicago Police Department does NOT guarantee 100% accuracy of all the data, so inferences made based on this data set should be treated carefully. Third, all the coordinates and addresses are modified to protect privacy of the victims so data visualization on the map is only an approximation. Fourth, we may not be able to answer questions such as why a certain type of crimes increases in certain months or in certain communities just based on this data set. Other information would be needed to account for the change. Fifth, for homicides, we have no information about the cause of each incident. So we can’t correlate the homicides with weapon violations.

This data set is relatively clean but still needs cleaning and wrangling. There were many missing entries for Ward and Beat. These two columns are not very useful either for visualization or predicting so they were dropped first. Then any other NAs were dropped or filled based on the context. The final data table contains 6.5M rows. Then I converted the Date column to a datetime column and set it as index for time series analysis. I also used a weather data set to find out weather conditions for each day from 2001 to 2016. This data set contains many useful statistics such as temperature, precipitation, and weather type. The weather type only contains numbers and they correspond to certain weather types. I used a dictionary from NOAA website to convert the number system to actual weathers. Then I converted the date column to data time object. The crime data and weather data were merged on their index.

By exploring the data set, I did a time series to see the trend in total number of crimes from 2001 to 2016. Total crimes decreased. By plugging in the population data for each year, I also plotted crime rate from 2001 to 2016. The crime rate also showed a decrease. These indicate that in Chicago total crimes are going down and it is getting safer over the years. But when I zoomed in on different crimes on the same time series between 2001 and 2016, I found that certain crimes had some unusual peaks. For instance, weapons violations started increasing in 2014 and homicides had a historical high in 2016 in Chicago. Gun control has always been a problem in Chicago and it has gotten worse these years. Is there a relationship between weapons violation and homicides? Without further information, we can’t draw the conclusion. Also, what caused the sudden increase in homicide in 2016?

A very useful tool when doing crime analysis is to plot the data on map. When I plotted the homicides of 2015 and 2016 on google map, I could see that most incidents clustered in the same regions but there are some places where they did not overlap. Did this reflect some gang activity changes? Also when I plotted different types of crimes on google map, they did not always show the same clustering on the map. For example, most crimes concentrated in the west and south but burglary and narcotics seemed to have some hot spots in the north.

Besides the map visualization, I also investigated the temporal distribution of different crimes. By plotting crimes by month, day of week and hour of the day, I could get a lot of insights too. Monthly plot revealed a pattern where most crimes happened most frequently in hot months such as June, July and August with the exception of narcotics which peaked in colder months of January, February and March. Temperature was confirmed by plotting average temperature in each month in 16 years to be highest in the summer months but lowest in winter months. Furthermore, by plotting histogram of crimes based on temperature distribution, I saw most crimes happened when temperature was between 70 and 80 degrees. Then why does narcotics not follow the same pattern? The figure where I plotted different crimes that happened in different days of the week also gave valuable insights. While burglary peaked during weekdays, battery and homicide seemed to peak on the weekend. Hourly plots showed that burglary peaked around 8am while battery mostly happened around 3pm. Homicides and weapons violation peaked at late night (11pm and 12am) and so did robbery.

Based on these findings, I am going to use other data to try to get a plausible answer to some of my questions. For instance, I will combine income, education and poverty data for each community and then try plotting choropleth maps to correlation different crimes with these factors. Some may be intuitive, and some may give a surprise. I will also do some statistical testing to see if some correlations are significant or not. I will also try to combine temporal and spatial distribution to see how certain crimes change spatially at different times. Lastly, I will also try to build a classifier to predict crime types given a number of features.