CS418 Gold Team Rules: A Look Into Chicago

UIC CS 418, Spring 2019

Name	NetID	GitHub Handle
Osama Ahmad	oahmad7	oahmad7
Maxwell Dausch	mdausc2	mdausch
Taj Atassi	tatass2	tajatassi
Mohamed Saeed	msaeed7	MohamedSaeed99
Abdullah Kidwai	akidwa2	akidwhy

Problem

Our intention is to look for correlation between Chicago's historical crime data, as well as permit data. We feel that this can aid in investment opportunities when combined with Chicago's government owned land (Where plots of land are sold for $1) One might be able to make a reasonable purchase, that can have a higher return on investment in the future.

Our initial ideas are based around wondering if an area has a decreasing crime rate, while having an increasing "investment" area with rehabs, and new buildings, the area might be good for a future investment. On the other hand, an increase in crime, might not be so good for investment, which we would then stay away from looking into properties in the area, and such investments might not turn a profit. Some factors that might play a role in the increase/decrease of crime rate is the status of the area(poor/rich neighborhood).

Data

We are currently using a mix of Chicago data sets, each offering a different look into the city's past and present. Make sure to download the data sets as TSVs and place them in the root of the project.

Data Set	Rows	Features	Description	Download
Crime	6,805,668	2001 - Present, Location, Crime type	See how crime has changed over time, and the most common areas for crime	Download (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2)
Permit	564,828	2006 - Present, Location, Crime type	See how much certain areas are being invested in, and the types of renovations.	Download (https://data.cityofchicago.org/widgets/ydr8-5enu)
Land	15,932	Location, Zoning Information, Square Footage	See which Government owned land is for sale, while also being cheap.	Download (https://data.cityofchicago.org/Community-Economic-Development/City-Owned-Land-Inventory/aksk-kvfp/data)

What we hope to achieve

To start off, we first would want to plot each of the data sets onto a map, that way we can compare them side by side, and check for inconsistencies, or correlations. Our next step would be to plot them all together, and adjust the time frames to see how the data has evolved.

We envision that as a final product, the project will be dynamic, allowing us to see a full overview of the data, or by selecting a timeframe. Our goal is to make it visually easier for people to see which areas are available to buy and how that area is with crime. We also would like to predict which areas is going to be infested with crimes to prevent any bad investment choices.

Set Up Project Dependecies

You will first need to install the google maps library

Install it with pip

```
pip install gmaps
```

Now run this next cell to set up all of the modules and functions

In [ ]:
```
%run './base.py'
```

Build the databases

This assumes all the tsv files have been downloaded to the project's root folder. Running the script below will generate the SQL databases.

In [ ]:
```
makeDB()
```

Exploratory Data Analysis

In [ ]:
```
print(getYears('Crimes_-_2001_to_present.db'))
```

Analysis

- Structure: The data is not very rectangular.
  - Some entries from City-owned_land_inventory are blank or just contain a PIN and nothing else
  - Positive: No Nested Data, so it becomes a little easier to manipulate
- Granularity: The data is very fine
  - They get very specific in terms of what buildings are available
  - Gives many details about the crime comitted such as Primary and secondary descriptions, location, and more
- Scope: The Data is mostly complete
  - Some entries are left blank in City-owned_land_inventory
  - Everything else is complete
  - It has everything we need (Land, permits, and crimes) and not much extra
- Before Cleaning, the Data has different date ranges for each database, but has been truncated after cleaning
  - Crime Data: 2001-2019 -> 2006-2019
  - Permit Data: 2006-2019 -> 2006-2019
  - Land Data: Unknown
- The data has been captured and retrieved directly from the City of Chicago

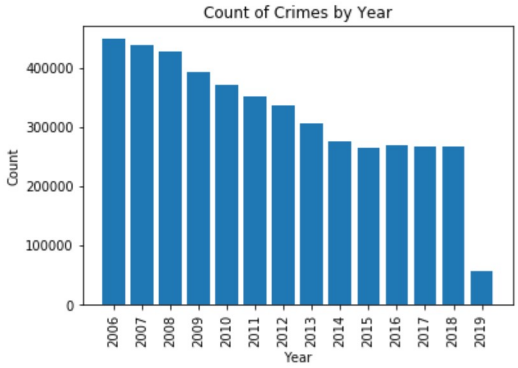
Other observations

- Observation: the fetchall and fetchone functions return a list of tuples even if you are retreiving one item, so we need to extract the the proper items and place them into lists
- After cleaning the data, the year of crimes spans 2006-Present.

Visualizations

Figure 1: Count of crimes throughout years

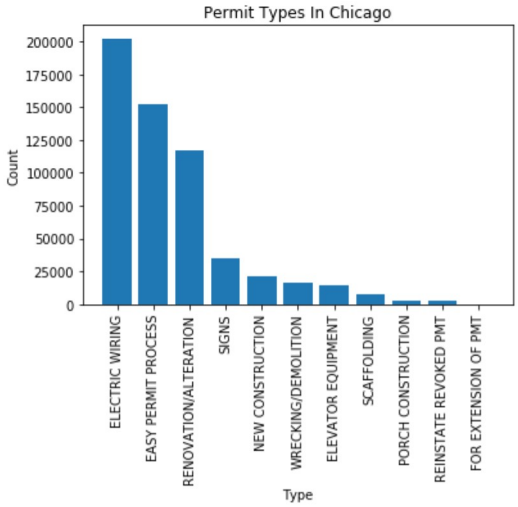
This visualization shows the count of crimes added throughout the years. Our goal was to track the trends of crime per year. We chose a bar graph because it allows us to separate each year and see the trend as the big picture. The number of crimes per year has decreased since 2006. Because 2019 isn't over yet, the number appears very low. Crime in Chicago has been on a steady decline since the records were tracked. It's important to keep in mind that not all crimes are reported. Gang violence is something that is unfortunately common in the South Side of Chicago, so some crimes committed in the South Side might not be documented throughout the years. However, cameras were installed in order to prevent and record these crimes.




```
In [ ]: crimesVis() #Run this to view the graph directly
```

Figure 2: Permit Types in Chicago

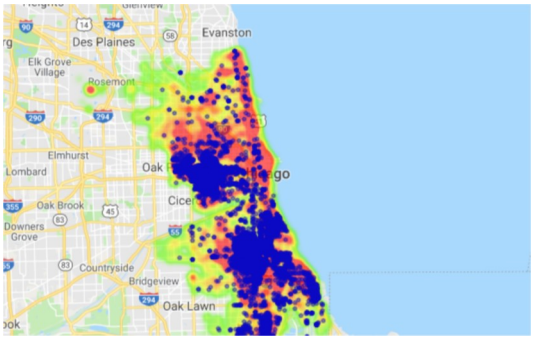
This visualization shows the different types of permits that were granted in Chicago. The most common type of permit is for electric wiring. The permit that was least prevalent was the one for porch construction. It is important to keep in mind that many people do not even apply for permits when doing renovations, so those renovations are not accounted for. For residential properties, homeowners hire private companies that do not require a permit from the City before starting their work. Also, the reason porch permits are very low is because firstly, there are a lot of apartment buildings here in Chicago. Secondly, porch renovations are very easy to do. Many times, they do not require any electrical work and there are few risks of damaging any sewer, gas, or water lines. Lastly, porch renovations do not take a long time and can be completed in less than a weeks' time.



In [ ]:
```
permitVis() #Run this to view the graph directly
```

Figure 3: Heatmap

This heatmap represents the crime rates and government owned land throughout the Chicagoland area. Crime rates and government owned lands north of North Ave are very low. What is very interesting is that for the rest of the places where crime rates are either moderate to high, the number of government owned lands are very high. Another peculiar observation is the North Loop area. Crime rates are relatively high here, however the number of government owned lands are very low. This is a very expensive and rich area, so theft and burglary is very high here.



In [ ]:
```
heatmapVis() #Run this to view the interactive heat map
```

Permits by Zipcode

We filtered the results that are displayed by only showing zipcodes that have more than 750 permits. This was so that we would only show results that had enough data that we can analyze properly. We have then selected a few of these results to analyze, as we can still see the varying results and rapidly different data coming from the selected zipcodes.

In [ ]:
```
df = zipcodeDF()
```

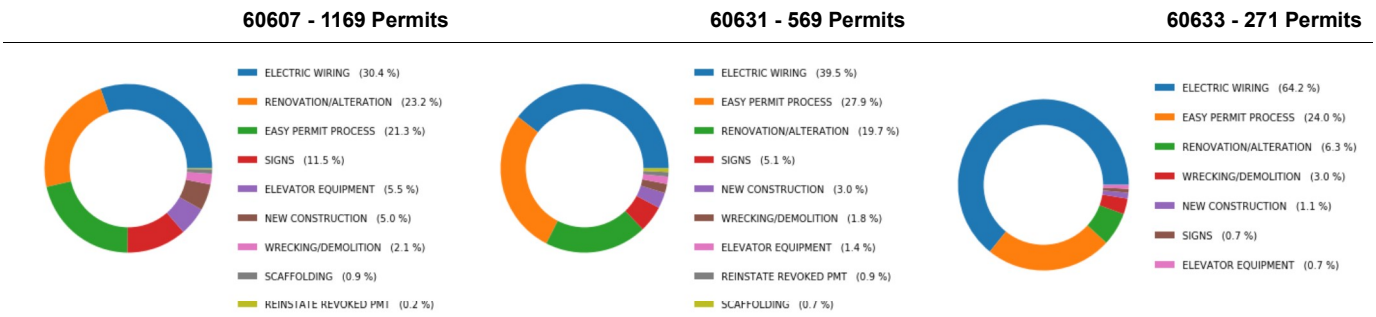


Figure 4: Analysis of permit types in 60607 Zip Code

This visualization shows the different types of permits in the Zip Code 60607. The donut chart is used to compare each group of permits with one another so one can see the most common occurrences. The Zip Code 60607 covers the West Loop of Chicago going as far south as 16th Street and as far north as Kinzie Street running from Ashland Ave to Wells St and includes UIC's east campus. The most common type of permit is for electric wiring and the least common was reinstating a revoked permit. Because of the constant construction going on in the West Loop, many permits are needed for Electrical Wiring and Renovation/Alteration, especially with the recent expansion of UIC's east campus and the many new places opening around the area.

Figure 5: Analysis of permit types in 60631 Zip Code

This visualization shows the different types of permits in the Zip Code 60631. The donut chart is used to compare each group of permits with one another so one can see the most common occurrences. The Zip Code 60631 covers the Norwood Park and Edison Park, just east of O'Hare International Airport. The most common type of permit is for electric wiring and the least common was scaffolding. A major intersection in the area is the intersection of Devon Ave, Harlem Ave, and Northwest Hwy. Devon Ave is known to have many small South Asian shops and restaurants and is located in a majority-residential area further north. Many houses and shops are purchased and remodeled and new places have electrical wiring installed.

Figure 6: Analysis of permit types in 60633 Zip Code

This zip code covers a southern area of Chicago as well as part of the city of Burnham. It has a relatively small population size of 13000. There is not a lot of development happening. However, recently the Ford motor company recently went under a billion-dollar expansion for their assembly plant in that area. The 64.2% of electric wiring permit and 24% of easy permit process reflects that billion-dollar expansion.

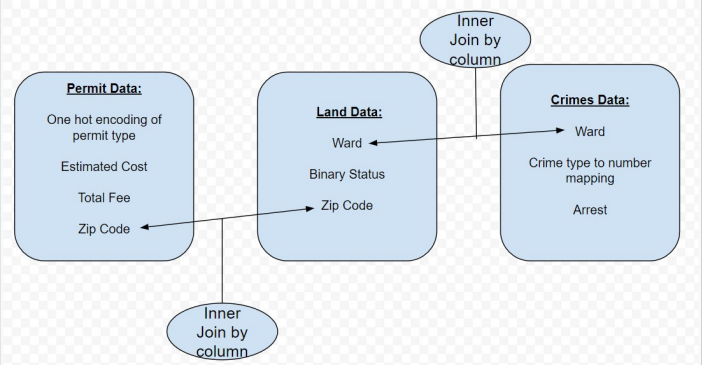
In [ ]:
```
#Run these to generate the visualizations
zipcodeVis(df, 60607)
zipcodeVis(df, 60631)
zipcodeVis(df,60633)
```

Machine Learning

Below is our attempt at machine learning. Our goal was to try and predict whether or not one of the properties from Chicago's land plots is sold or not sold. While at first glance, this might not seem like a big problem to solve, we feel it would be instrumental for someone looking to invest in the area. This is due to the fact that if we are able to predict sold and not sold well, it would allow someone to see the areas that might be considered a "hot commodity".

The way one might be able to apply this is, by giving a piece of land, if the model predicts that that that piece of land is sold, and the actual piece of land has not been sold, it means that the land will likely be bought soon, so it would be good to act fast and determine if one would like to purchase the property. Combining this with the permit investment data, and crimes data, one could make a more informed decision about what to do.

In [ ]:
```
df = getMLTable()
```



In [ ]:
```
features,status = mlSplit(df,.001)
```

Neural Net

Our initial attempt at solving this problem was through the use of a neural net. We attempted to preform a 'brute force' approach to this by passing in a mixture of the ammount of hidden layers, as well as the number of nodes in each layer. Our neural net uses relu as its activation function, 10,000 max iterations, epsilon of 0.001 and an alpha of 0.

We then took this, fit it to our data and then preformed 10-fold cross validation on the model and our data to find the expected error of the model. This would be done for each permutation of the layer size and number of nodes.

We found that the optimal configuration produced an error of 0.7037720483670696 with 2 hidden layer and 5 nodes. If you would like to verify this, run the next cell, but be warned that this will take quite a while. We knew that this was not an optimal solution, so we looked to trying other types of models to find a better result.

In [ ]:
```
runNeuralNet(features,status)
```

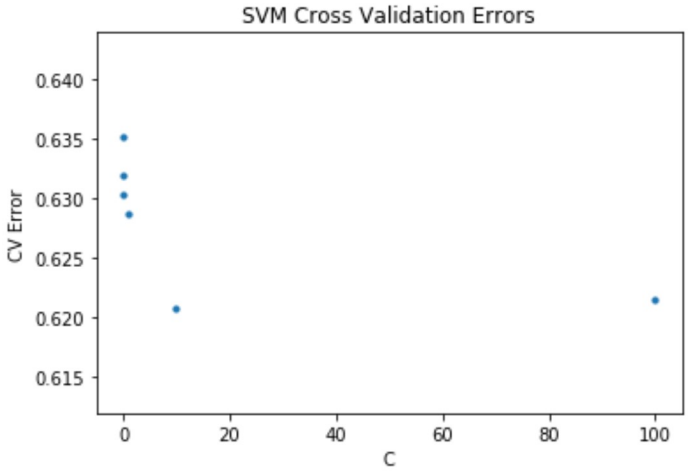
SVM

Our next attempt at solving this problem was through the use of a SVM. We went with a radial kernel and worked through a list of C Values to try and find the optimal 10-Fold cross validation error.

When trying to run this, it was taking absolutely forever, so we went with a bagging approach to handle the svm as a way to speed this process up. This uses ensemble learning, by training multiple SVMs on subsets of the data to reduce the entries per model. This is a way to reduce training time, while speeding up the process, and giving the same (or even better) scoring for the model.

This did turn out better than the neural net, though it was still not a good solution to the problem. We would hope to find a better accuracy than 0.6206257666824331 which was found with a C of 10.In the future it would be a good idea to try other classifiers to see how their models fit the data, and the accuracy they provide.

If you would like to run the svm, run the following cell. **Be warned though, that this will take an extremely long time. It ran overnight and still had not finished on one computer, while throwing memory errors on others.** Below is the cross validation errors that we had found for the different values of c.



In [ ]:
```
runSVM(features,status)
```

Takeaways

We discovered some substantial takeaways from the visualizations, machine learning models, and analysis we completed in the project. We discovered that the overall crime rate in Chicago is decreasing. While this is certainly good for the city and investment, our analysis does not tell us what parts of the city have reduced crime. Additionally, we discovered there is much more investment in rehabilitation and renovation than new buildings and facilities.

Another critical takeaway we discovered in our analysis is that there is high overlap between high crime rate areas and government owned land. This is obviously not ideal as locations with high crime rates are not well-suited for investment. There are, however, a small amount of prime candidates for investment available. These specific locations are not within these high crime areas, and may prove to be worthwhile investments.

Reflection

What is the hardest part of the project that you've encountered?

The hardest part of the project so far has been cleaning the data, specifically with the Crime Dataset. At a whopping 6.8 million rows, the Crime Dataset takes an incredibly long time to clean. The extremely long time it took for us to group the crime data hindered our ability to continue to work on later portions of the project, because they were all reliant on the data cleaning to be done first.

We also ran into issues with the Google Maps API. We quickly ran out of funding to make enough requests. Our visualizations suffered because of this. Getting responses for the requests also took a tremendous amount of time, taking over eight hours.

The two machine learning algorithms proved to be a problem as well. Both models took hours to run, often running overnight. Because they took so long to run, we had to wait an incredibly long time to complete the analysis and results for the project.

What are the next steps you could take to improve your project?

If we had more time with the project, we would use scraping sites like Zillow or Realtor to find a more accurate representation of housing data. Additionally, we would request more credit in the Google API to make more requests and improve our visualizations. Next we would find a way to lower the error of our neural network by testing with new parameters.

With our current time constraints, we were only allowed to run our machine learning models on a small percentage of the overall dataset but given more time, we would run our algorithms on the entire dataset. Finally, we would break up our predictions and predict on each neighborhood instead of predicting with respect to all of Chicago. That way we could find more specific prediction for each individual neighborhood.