Analyzing Chicago Crime Data to Inform Resource Deployment

The purpose of this analysis was to examine the spatial characteristics of Chicago crime data and to explore relationships between the density of that crime, proximity to locations of interest (Police and Transit Stations), and community demographics in order to inform further investment in police infrastructure. We also examined some of the temporal characteristics of crime to inform resource management/deployment. Our initial hypothesis was as follows:

1. *Crime rates (as measured by* count/km^2*) decrease as one moves closer to a police department*
    a. *This assumes that proximity equates to higher police presence which acts as a deterrent.*
    b. *Alternatively, police may be more likely to make an arrest for a crime where they most regularly patrol, resulting in increased activity as one gets closer to a police station. This would cause us to reject this section of our hypothesis.*
2. *Crime rates increase as one moves closer to a transit station.*

The data from 2013-01-01 through 2014-12-31 actually indicated that crime rates do not decrease but actually increase as one moves closer to a police station, causing us to reject part 1 of the hypothesis. However, crime rates do increase exponentially as one approaches a transit station, supporting section two of our initial hypothesis. Our final recommendation as it concerns police infrastructure can be found in the conclusion.

1 Table of Contents

2 Introduction

The “client” has requested analytic support to determine how to best invest in new police infrastructure. The purpose of this analysis is to determine significant spatial, temporal, and demographic crime patterns that could be used to guide investments and/or improve policing patterns. We will examine Chicago Crime Data from several angles and utilize multiple visual methods including the following:

- Histograms; Frequency of crime by type relative to
    - Police presence [proximity to PD’s]
    - Proximity to Transit stations,
    - Hours of the day
    - Days of the week,
    - Community
- A choropleth map showing crime density by community
- Point analysis for the city as a whole that will include spatial standard deviation ellipses.
- Crime heat-maps
- Scatterplots and linear model analysis to determine the efficacy of a basic linear model in predicting crime density. This model could be used to inform policing patterns and/or future station locations.

Our initial, spatial hypothesis:

1. Crime rates (as measured by $count/km^2$) decrease as one moves closer to a police department
    - This assumes that proximity equates to higher police presence which acts as a deterrent.
        - *Note: for the sake of simplicity we assume that each police department employs the same number of officers. This could be tightened with actual weights if we wish to extend our engagement.*
    - Alternatively, police may be more likely to make an arrest for a crime where they most regularly patrol, resulting in increased activity as one gets closer to a police station. This would cause us to reject this section of our hypothesis.
2. Crime rates increase as one moves closer to a transit station.

Assuming both parts of the hypothesis prove true and the spatial relationships are strong, we will likely suggest building small, permanent outposts in the city’s most crime dense communities, at their most affected transit stations. Such a strategy could reduce crime rates and place our force in the best locations for their communities.

3 Data

3.1 Background on Data

The superset of data is a combined set of files including Chicago crime data, geo-spatial polygons, Mass transit station locations, police department locations, demographics, Categorization tables, weather data, and more. All requisite data files can be found in the appropriate folder on Github.

While the crime data currently covers dates from 2013-01-01 to 2014-12-31, this could be easily updated at will. Currently, I have intentionally left 2015 data out, as I am using it as test data for a separate analysis.

3.2 Sources

Base crime data, shape files, coordinate files, and census data comes from the data repository at the City of Chicago website. Although not used in this iteration of analysis, weather data was downloaded from the Weather Underground. This analysis also uses a community-census tract mapping table courtesy of Rob Paral.

The base sources are then subsequently scrubbed, reformatted, joined, etc. in R. This can be viewed via the RMarkdown file “CreateSuperDataSet” on Github.

3.3 General Data Structure Notes

Each tuple in the base crime dataset equates to 1 crime. While the base crime data is robust, this analysis required a good amount of mutating, cleansing, etc. It does includes some null fields, duplications where a given crime may have had more than one victim, and some erroneous location data such as crimes occurring within a police station, outside of city limits, or within a different state, etc . Some of those characteristics, like duplications for more than one victim, are appropriate and left alone. Missing/Erroneous spatial data is dropped. You can see more detail and follow along via the appropriate file on Github.

Most of the **Demographic data** is provided by the city of Chicago's Department of Public Health as underlying pieces to their "Hardship Index." It includes the following:

- Percent of crowded housing
- Percent of households below poverty
- Unemployment rate, 16+
- Percent aged 25+ without a high school diploma
- Percent aged under 18 or 65+
- Per capita income

In order to keep the files robust, but manageable, I have created **categories for each primary type of crime**. These Meta Categories were created by grouping crime descriptions as closely to federal definitions as possible. Categories are as follows:

- Violent Crimes
    - Assault
    - Battery
    - Criminal Sexual Assault
    - Homicide
    - Etc.
- Non-Violent Property Crime
    - Arson
    - Burglary [ex Home invasion]
    - Motor Vehicle Theft
    - Etc.
- Financial Crime/Fraud
    - Deceptive Practices [embezzlement, etc.]
    - Intimidation - Extortion
    - Money Laundering
    - Etc.
- Other
    - Traffic violatons
    - Noise violations
    - Non-violent prostitution
    - Non-violent narcotic violations
    - Etc.

Note: The full mapping table is viewable in Github. This analysis displays all 4 types in initial plots but is trimmed to just Violent Crimes and Non-Violent Property Crime for all mapping exercises.

Hour Categories are split into 4 hour increments as follows:

- 7am - 11am is "AM Commute"
- 11am - 3pm is "Lunch"
- 3pm - 7pm is "PM Commute"
- 7pm to 11pm is "Evening"
- 11pm - 3am is considered "Night"

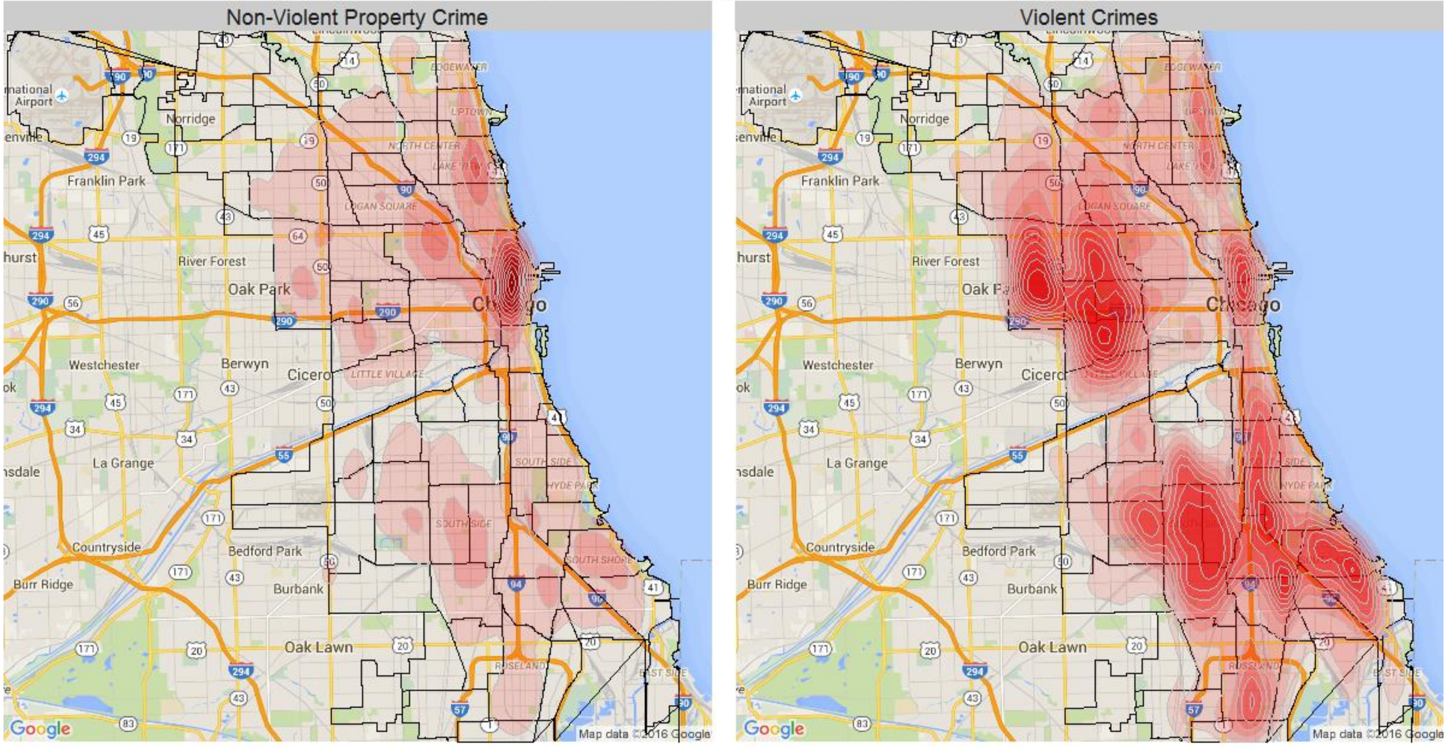
4 Overall View of Chicago Crime

Note: All spatial analysis will focus on Violent Crimes and Non-Violent Property Crime.

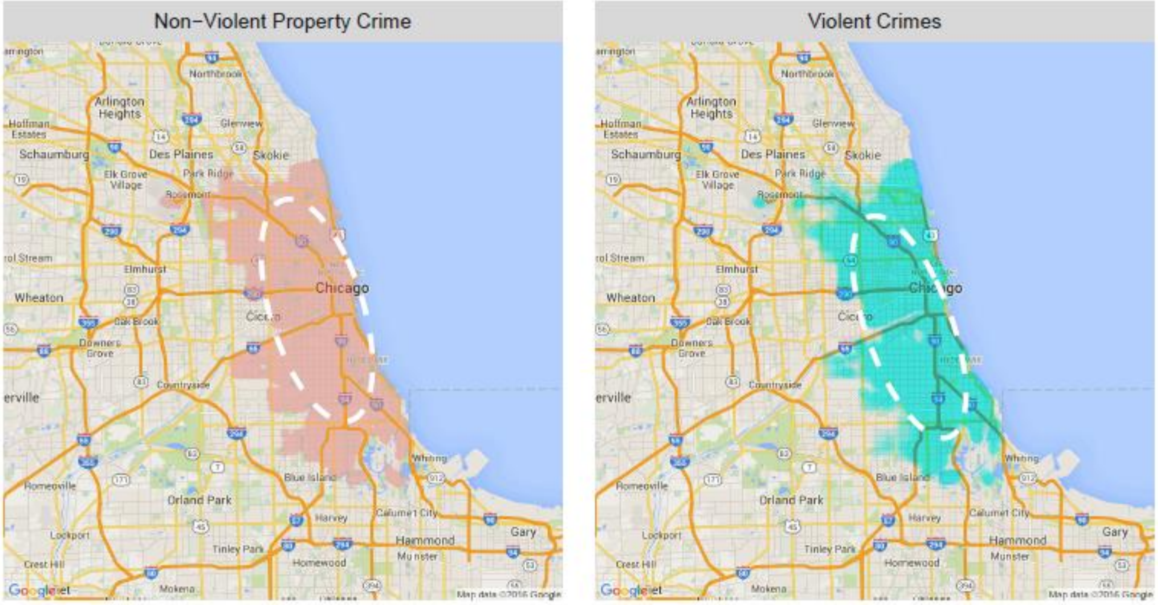
4.1 Heat-maps and dispersion

The following is an overall image of Non-violent Property Crime and Violent Crimes around Chicago. Property crime is highly focused downtown with a few moderate hot spots elsewhere. On the other hand, violent crime is widely dispersed with multiple hot zones.

Crimes in/around Chicago



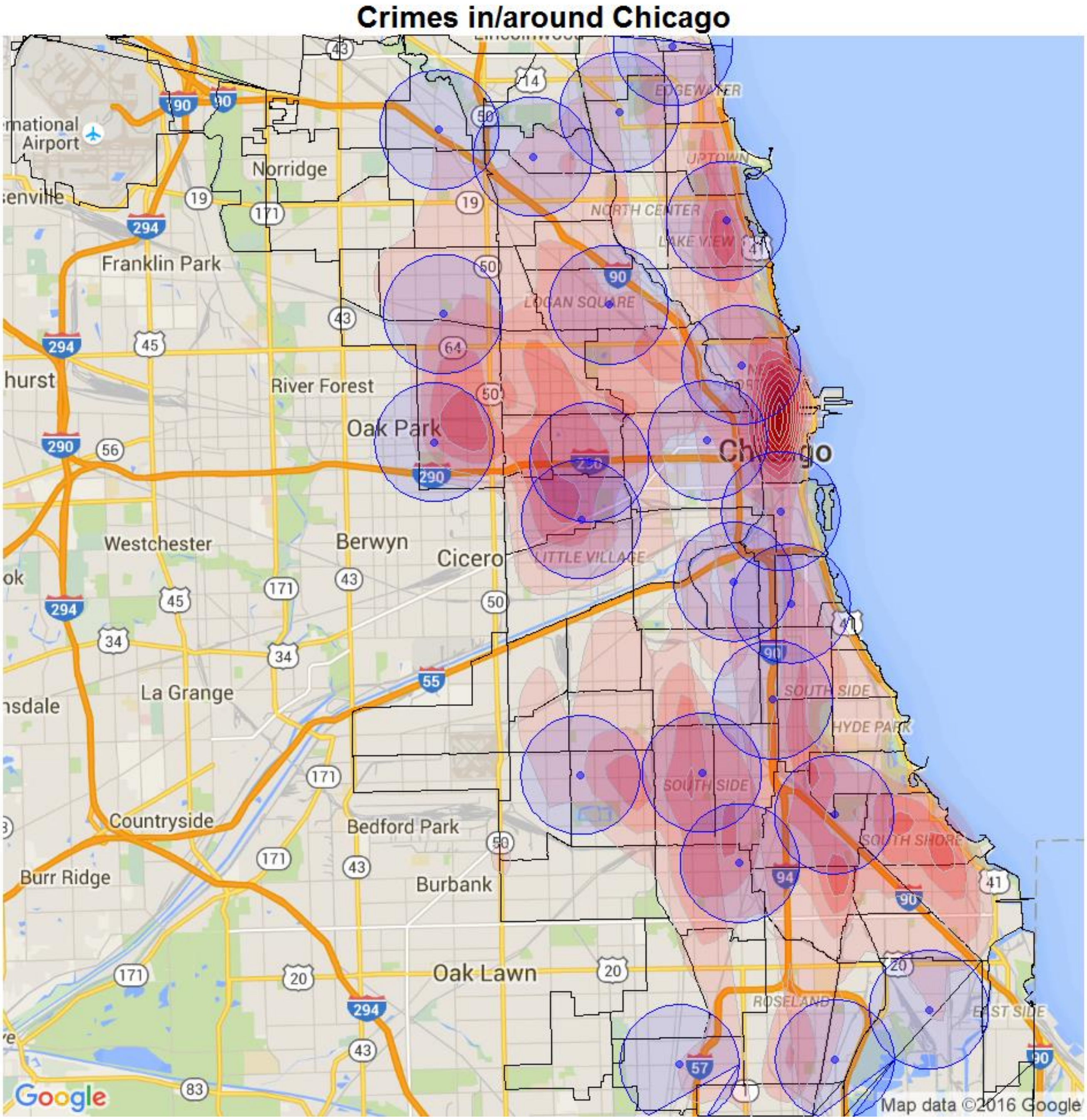
This leads to very large, geo-spatial standard deviations, and a multi-front battle against crime. Just to give an idea of that dispersion, the dotted lines on the following graphs encompass just one standard deviation [68%] of all Chicago crime:



4.2 Heat maps with PD and Transit location overlays

If our null hypothesis regarding proximity to police departments is correct, we should see dips in a localized heat map around police markers. The following heat map shows police departments as markers with 2km radius zones drawn around them. Close inspection reveals that PDs are typically not located at peaks. However, they also don't exhibit any strong reductive effects on crime rates in their immediate, surrounding areas. This map isn't detailed enough to reject the first piece of our null hypothesis, but it does not look promising.

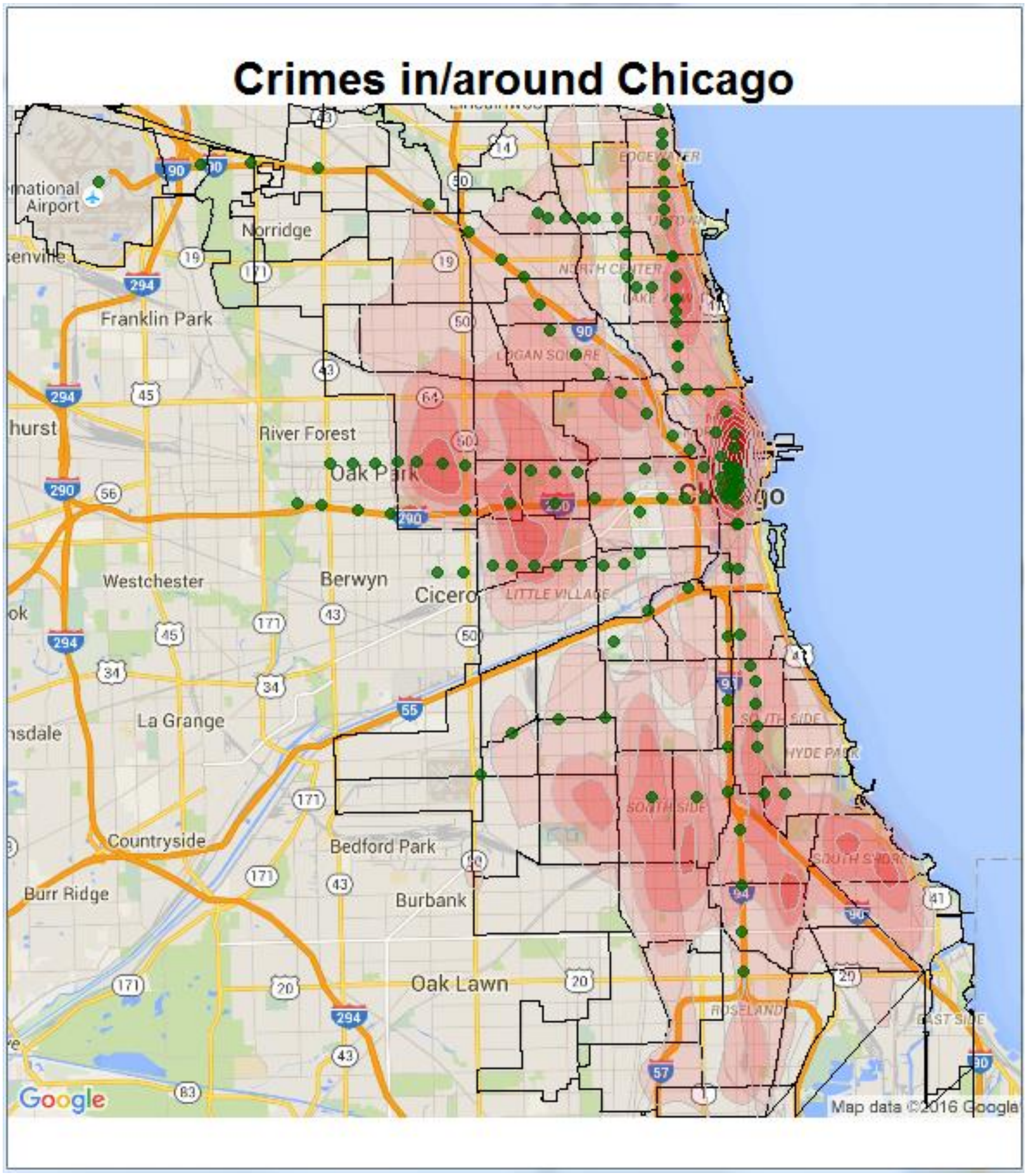
Note: We used a 2 km radius assuming that any proximity effects would be de minimis beyond that range [if there are any at all].



Chicago Crime Density shown with Police Department markers using 2km circles to denote patrol zones.

A transit station overlay (as represented by the L) shows a great deal of overlap between common transportation avenues and crime rates. Again, this is not enough to confirm the hypothetical relationship between transit proximity and spatial crime rates, but it does provide a promising initial picture.

Note: Although we do not have the data to confirm it, we suspect that this relationship would only be reinforced by adding bus routes:



Chicago Crime Density shown with Transit Station markers

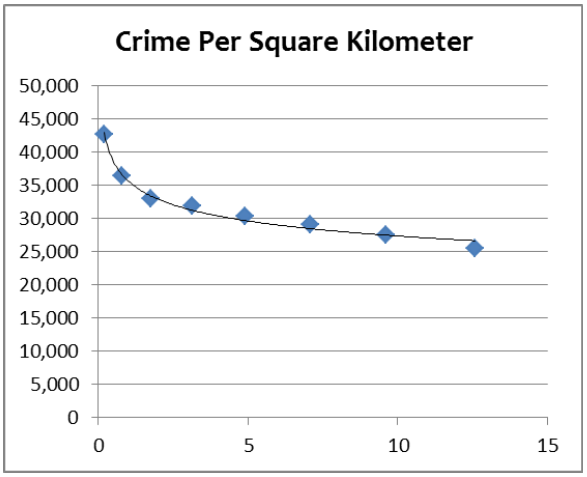
5 Proximity Analysis

5.1 Proximity to Police Departments

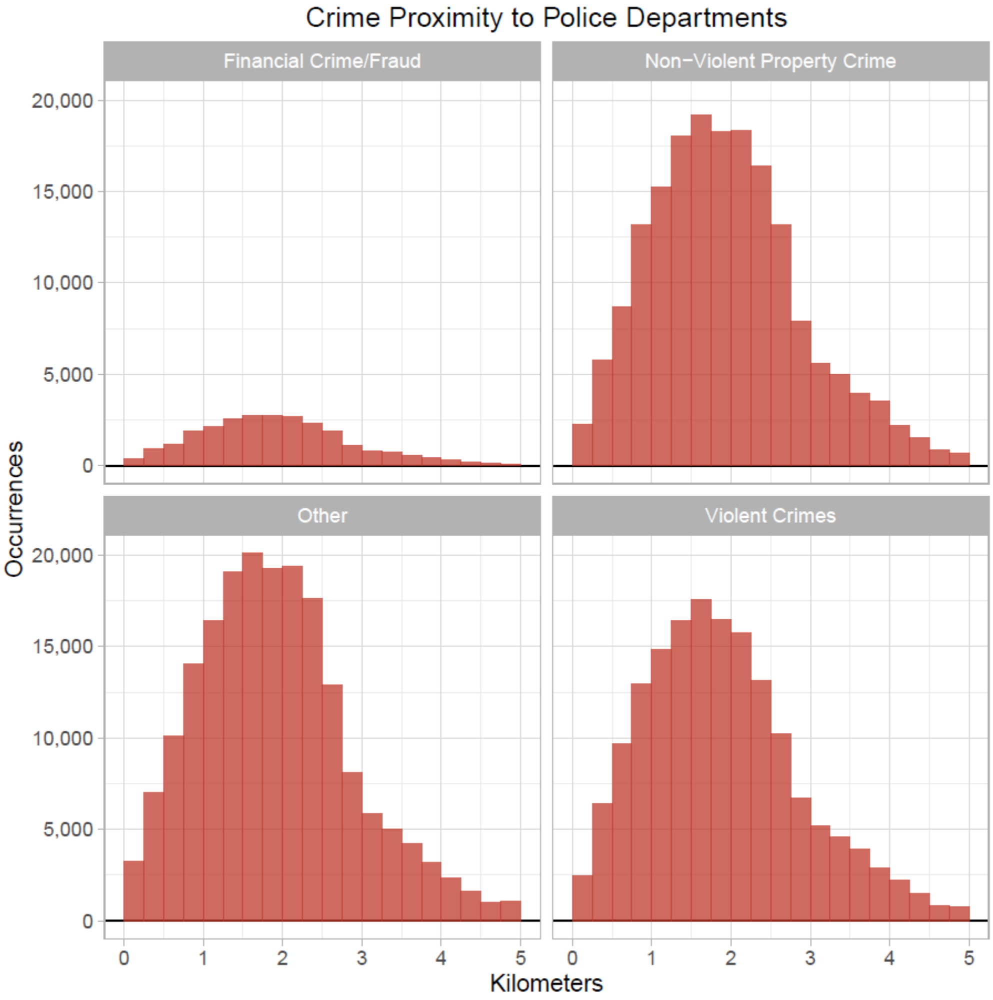
Contrary to our initial hypothesis, crime rates do not decrease as one gets closer to a given police department. While the effects are not extreme, crime rates actually appear to increase:

Crime Proximity to PD

Radius in km	Count	Area in km^2	Crime Per km^2	Change in Crime Density
0.25	8,381	0.196	42,684	17%
0.50	20,171	0.785	36,354	10%
0.75	29,665	1.767	32,944	3%
1.00	42,118	3.142	31,938	5%
1.25	48,635	4.909	30,348	5%
1.50	56,106	7.069	29,012	5%
1.75	59,536	9.621	27,503	8%
2.00	56,692	12.566	25,569	8%



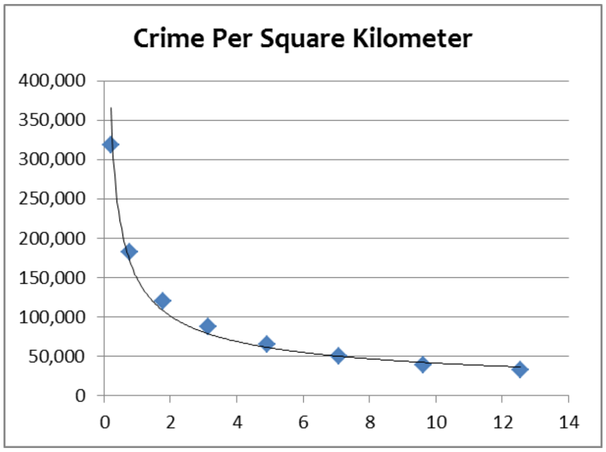
Some of the increase in the first ¼ kilometer may be due in part to incorrect location entries such as crimes occurring within police stations. However, the overall slope is meaningful. While area shrinks 98% from a 2km radius to a ¼ km, density increases by 67%. This would seem to support the alternative hypothesis that police presence does not have a dampening effect on crime, but that arrests actually increase with police patrols.



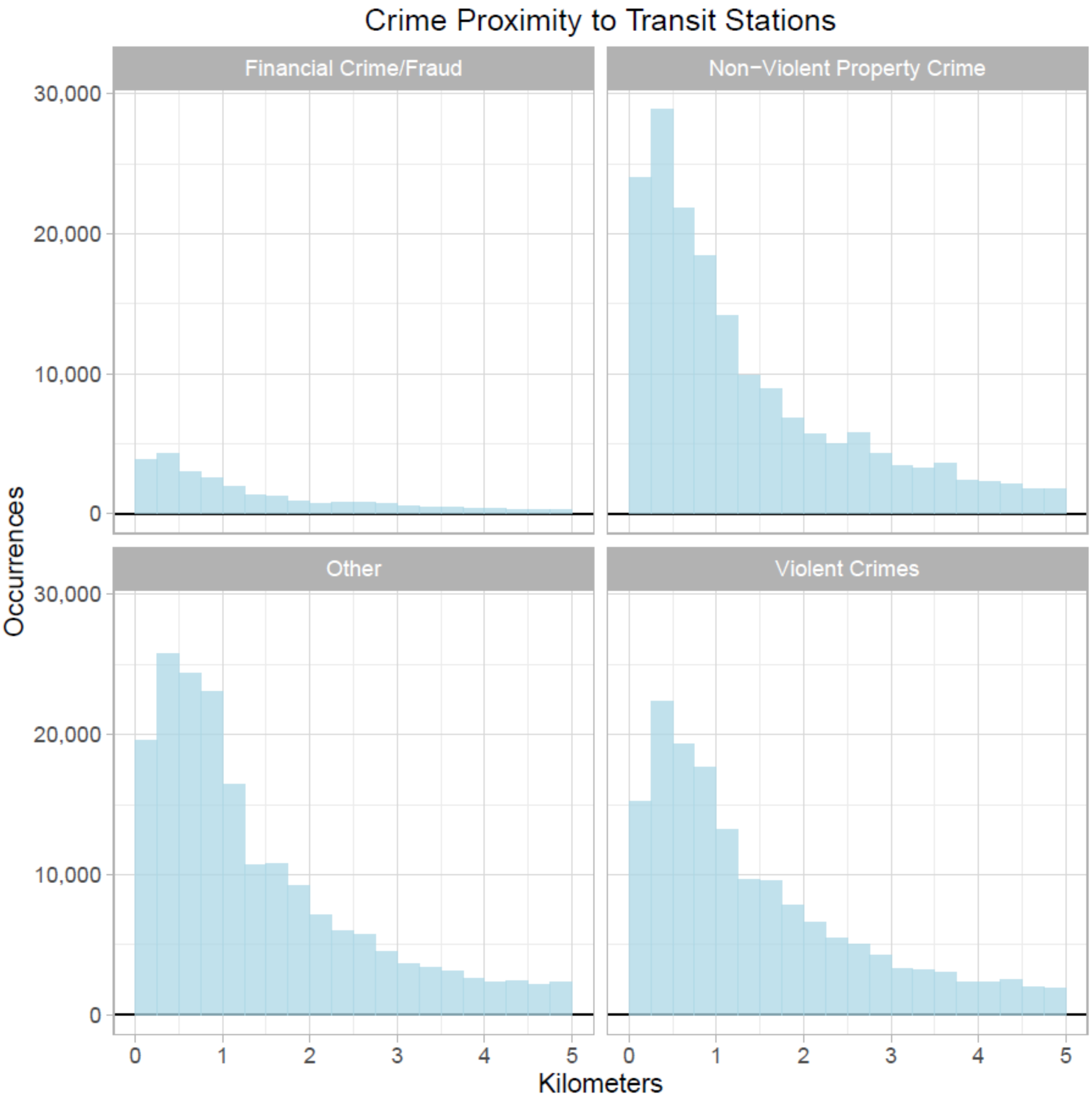
5.2 Mass Transit Locations [as represented by the L]

Here the initial hypothesis appears to be correct. Crime rate (density) clearly increases as one moves closer to a given transit station. Again, while area decreases 98% from a 2km radius to just a ¼ km, density increases by 985%. This effect starts to flatten out beyond 2 square kilometers.

Crime Proximity to Transit				
		Area in	Crime Per	Change in
Radius in km	Count	km^2	km^2	Crime Density
0.25	62,642	0.196	319,033	74%
0.50	81,328	0.785	183,308	52%
0.75	68,526	1.767	120,248	38%
1.00	61,716	3.142	87,284	34%
1.25	45,771	4.909	65,186	31%
1.50	31,563	7.069	49,734	25%
1.75	30,547	9.621	39,714	23%
2.00	24,756	12.566	32,376	21%



This could be further refined with transit station volume, but that data is not currently available. Regardless, it is clear that any future infrastructure investments must account for a city's transit characteristics.

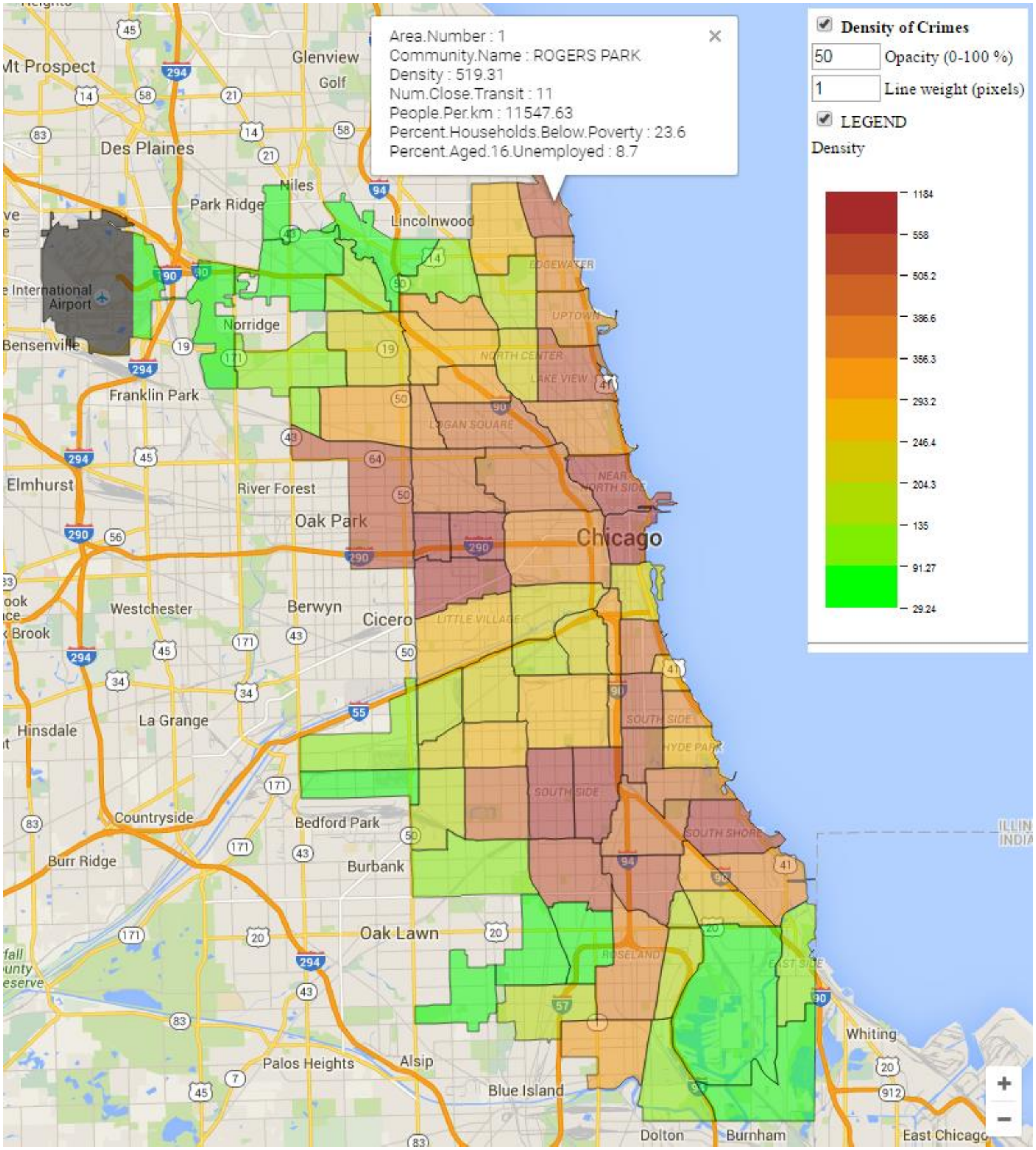


6 Community Level Analysis

6.1 Density Choropleth

So far, we have demonstrated that crime rate [density] increases as one moves closer to both police stations and transit locations, but given the structure and political influence of community organizations, it may prove equally useful to view the problem and any resource deployment recommendations bounded by community borders. The following choropleth displays Chicago communities according to increasing crime "density" as measured by annualized crime rate per square kilometer.

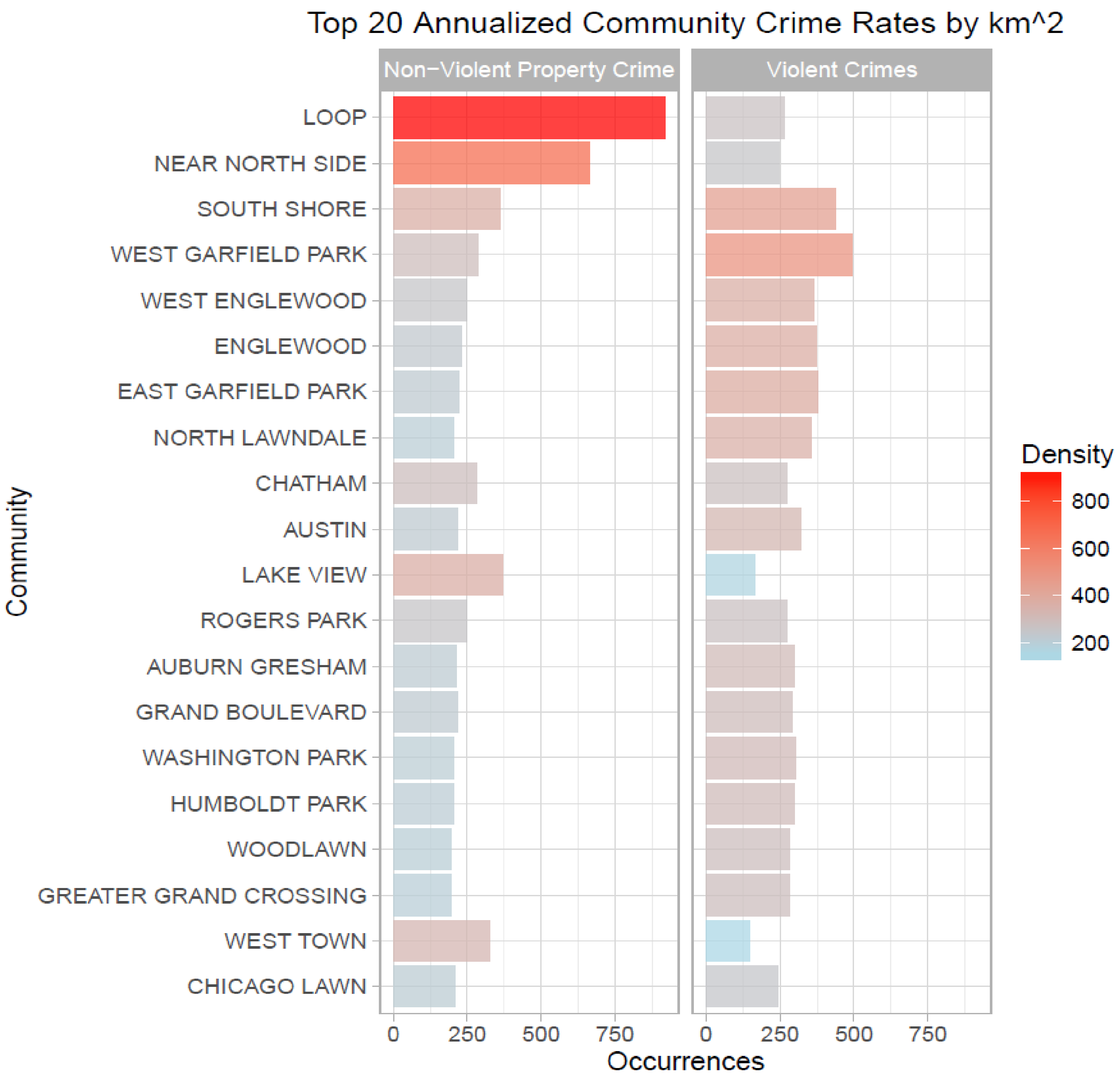
Note: This is an image of the interactive html file available via the project's Github folder.



Area.Number : 1
Community.Name : ROGERS PARK
Density : 519.31
Num.Close.Transit : 11
People.Per.km : 11547.63
Percent.Households.Below.Poverty : 23.6
Percent.Aged.16.Unemployed : 8.7

☑ Density of Crimes
50 Opacity (0-100 %)
1 Line weight (pixels)
☑ LEGEND

As an addendum to the above map, the following bar graph shows the 20 communities with the highest density ratings. It is ordered by a community's overall crime density and then split out to show Non-Violent Property Crime separate from Violent Crimes.

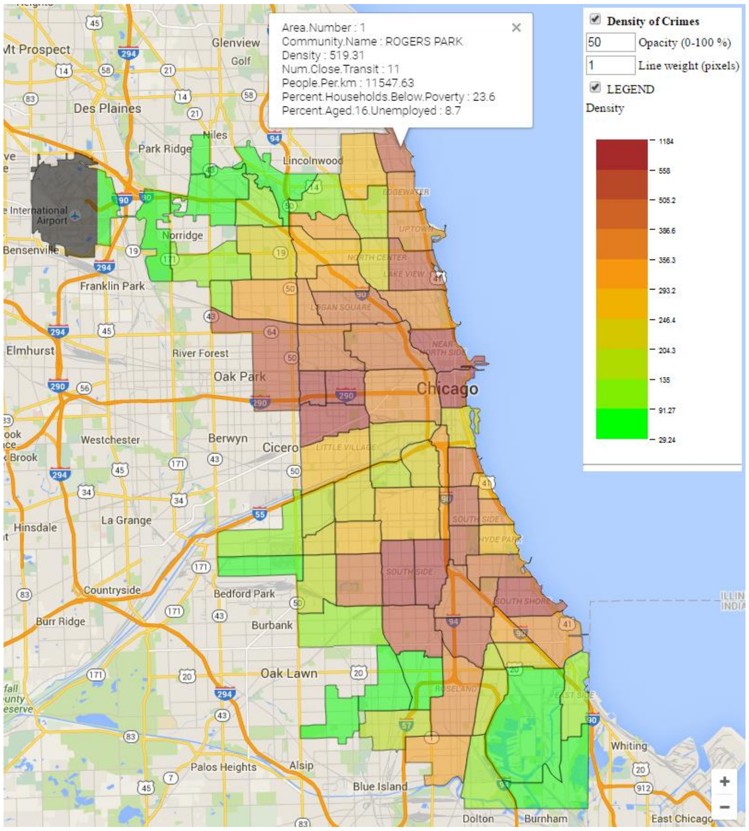


Note: Some communities [like the Loop and the Near North Side] are very dense but markedly one-sided in the make-up of their crime. Both of these communities are relatively affluent with per capita incomes of $65,526 and $88,669, respectively. Demographic based analysis might see their crime rates as anomalous; however, they have a very large number of accessible mass transit locations within their communities. As we will see later, this at least partially explains their very high crime rates.

Community	Per Capita Income	Number of Closely Accessible Transit Stops
Loop	$65,526	41
Near North Side	$88,669	22
Median	$20,956	5

Using another metric, compare crime density as represented in the preceding map to the same map colored by community hardship below. [See the section on data for information regarding the hardship index.] Note how hardship is not a tremendous predictor of crime. We will go on to examine this and other demographic factors in the next section.

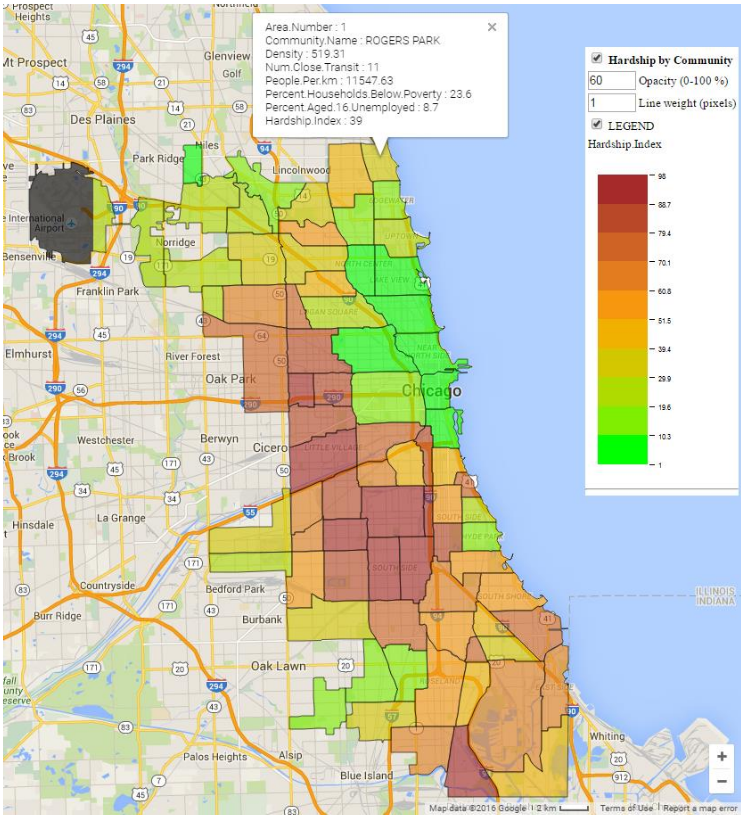
Crime Density



Area.Number : 1
Community.Name : ROGERS PARK
Density : 519.31
Num.Close.Transit : 11
People.Per.km : 11547.63
Percent.Households.Below.Poverty : 23.6
Percent.Aged.16.Unemployed : 8.7

☑ Density of Crimes
50 Opacity (0-100 %)
1 Line weight (pixels)
☑ LEGEND

Hardship



Area.Number : 1
Community.Name : ROGERS PARK
Density : 519.31
Num.Close.Transit : 11
People.Per.km : 11547.63
Percent.Households.Below.Poverty : 23.6
Percent.Aged.16.Unemployed : 8.7
Hardship.Index : 39

☑ Hardship by Community
60 Opacity (0-100 %)
1 Line weight (pixels)
☑ LEGEND

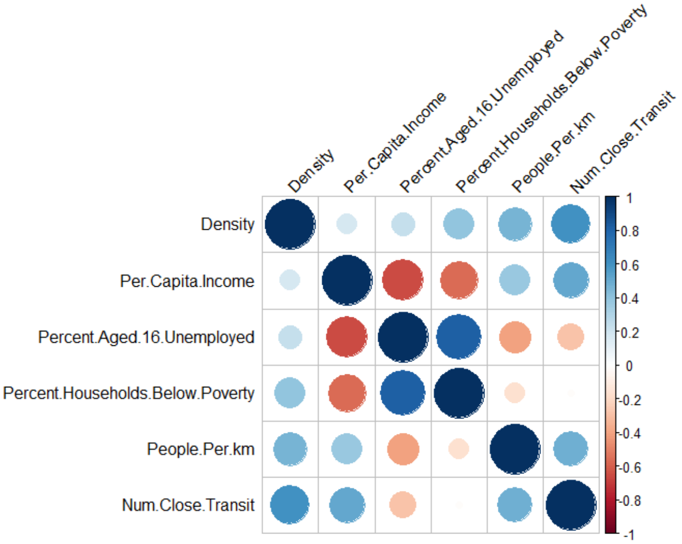
6.2 Demographics and Other Community Level Characteristics

6.2.1 Factor correlations at the community level

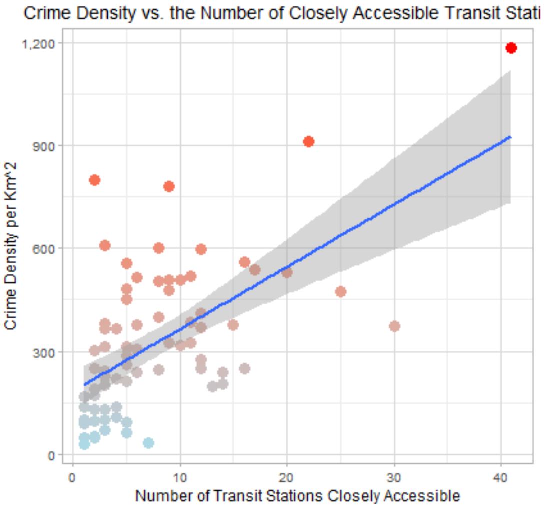
We chose a number of community level demographics to explore for our regression models. To summarize, we found that only 4 of the chosen factors were significantly correlated to a community's crime rate [density]. See the table and graphical correlation matrix below:

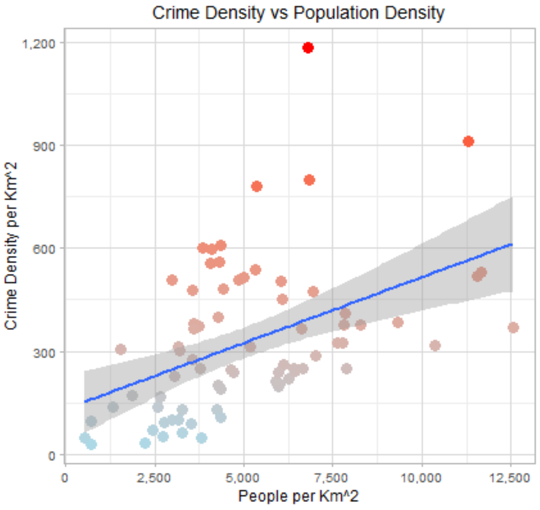
Note: We will use the number of transit stations accessible to a given community as a proxy for individual crimes' proximity figures in our regression models. Also, note the co-linearity between "% Households Below the Poverty Line" vs. "% Aged 16 Unemployed" [corr = 0.81] and "#of Transit Stations vs. "Population Density" [corr = 0.48]. This will dictate choosing one of each pair.

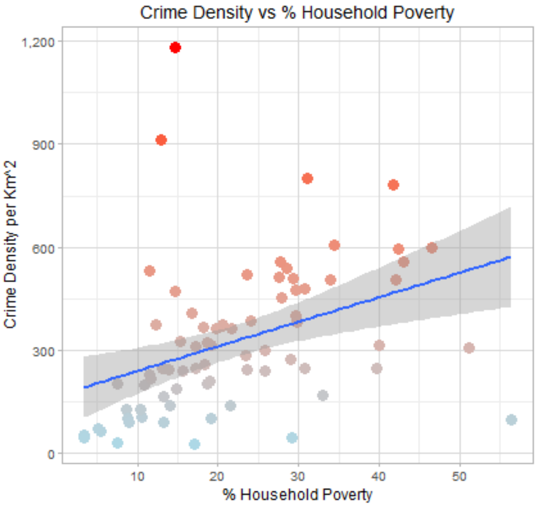
Community Characteristic	Correlation to Crime Density	P-value
# of Transit Stations Closely Accessible	0.598	0.0000019%
Community Population Density	0.459	0.0039%
% Households Below Poverty Level	0.386	0.069%
% Aged 16 Unemployed	0.231	4.8%
Per Capita Income	0.171	14.6%
Hardship Index	0.121	30.4%
% Aged 25 Without High School Diploma	-0.074	52.9%
% of Crowded Housing	0.042	72.5%

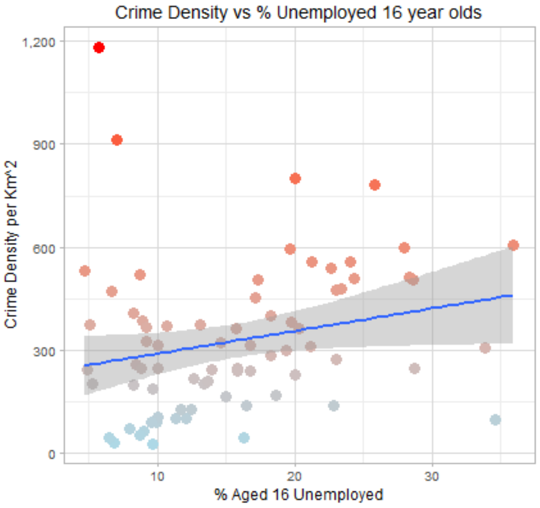


The following graphs illustrate these relationships in order of decreasing correlation and significance:

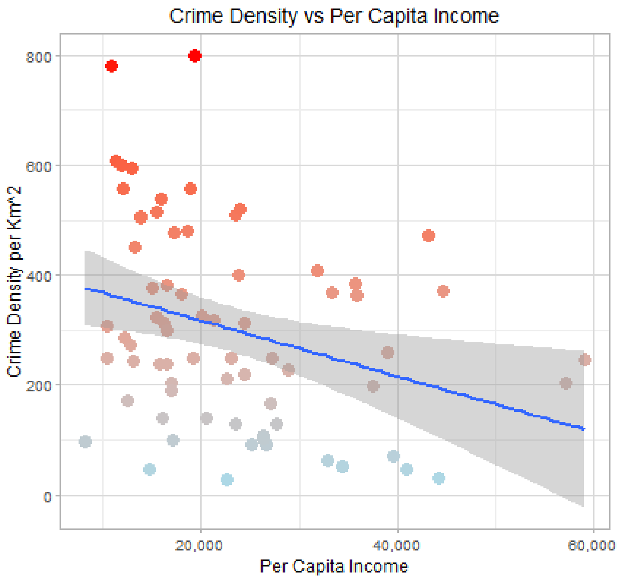

Crime Density vs. the Number of Closely Accessible Transit Stati

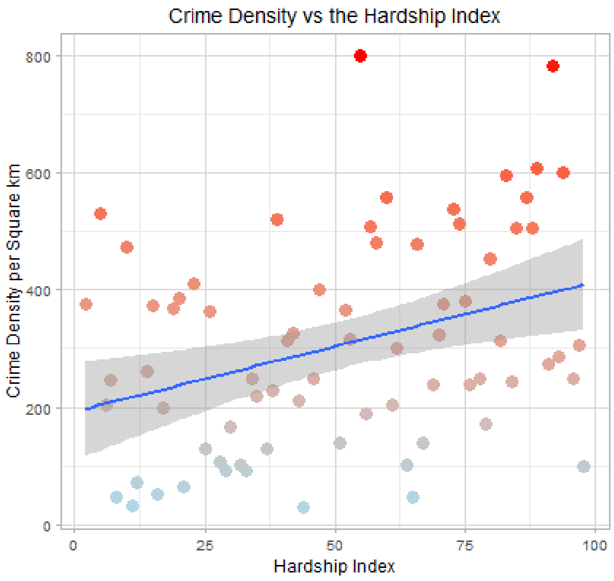

Crime Density vs Population Density

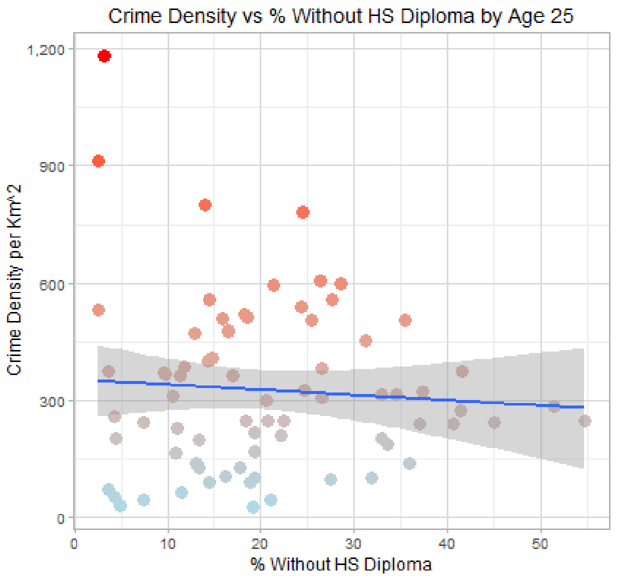

Crime Density vs % Household Poverty

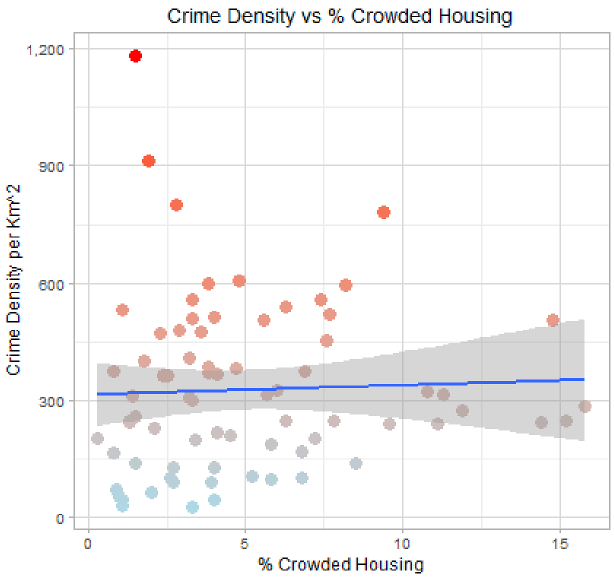

Crime Density vs % Unemployed 16 year olds

Demographic Factors (Continued)


Crime Density vs Per Capita Income


Crime Density vs the Hardship Index

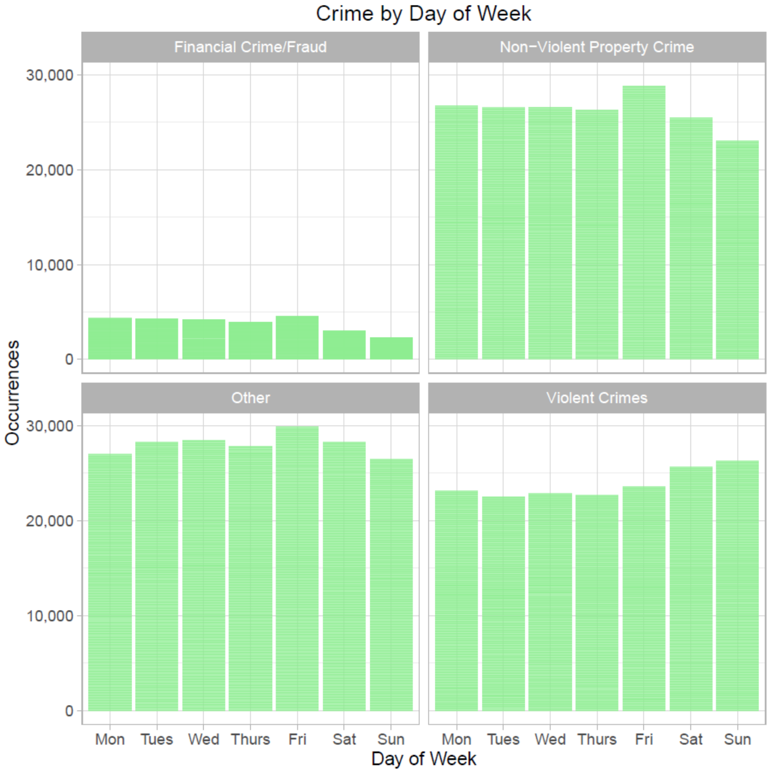

Crime Density vs % Without HS Diploma by Age 25


Crime Density vs % Crowded Housing

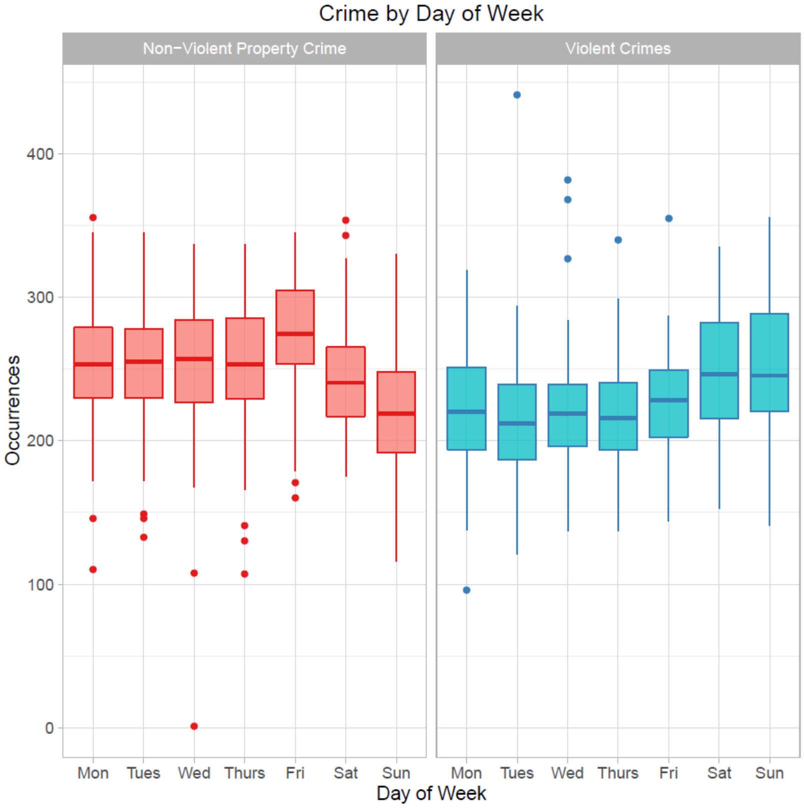
7 Temporal Analysis

7.1 Day of Week

Before we get to regression models, it may also be useful to examine some of the temporal characteristics of Chicago crime. First, it does not appear that crime substantially differs by day of week. Here are graphs showing all 4 categories. Note: The lack of statistical significance was corroborated by linear models not shown here.

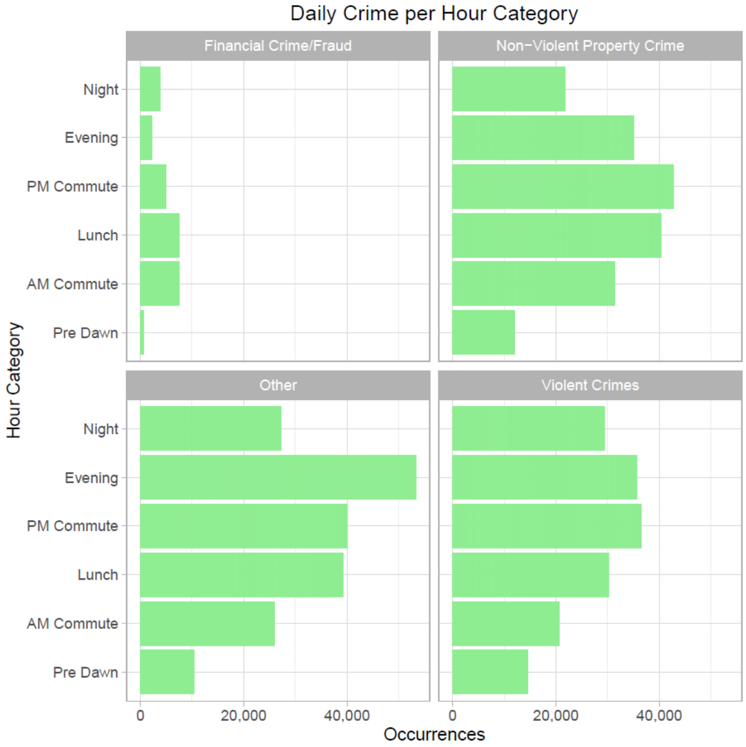


The following boxplots display greater detail for the **Violent Crimes and Non-Violent Property Crime** categories. It is interesting to note that property crimes fall on Saturday and Sunday while violent crimes rise.

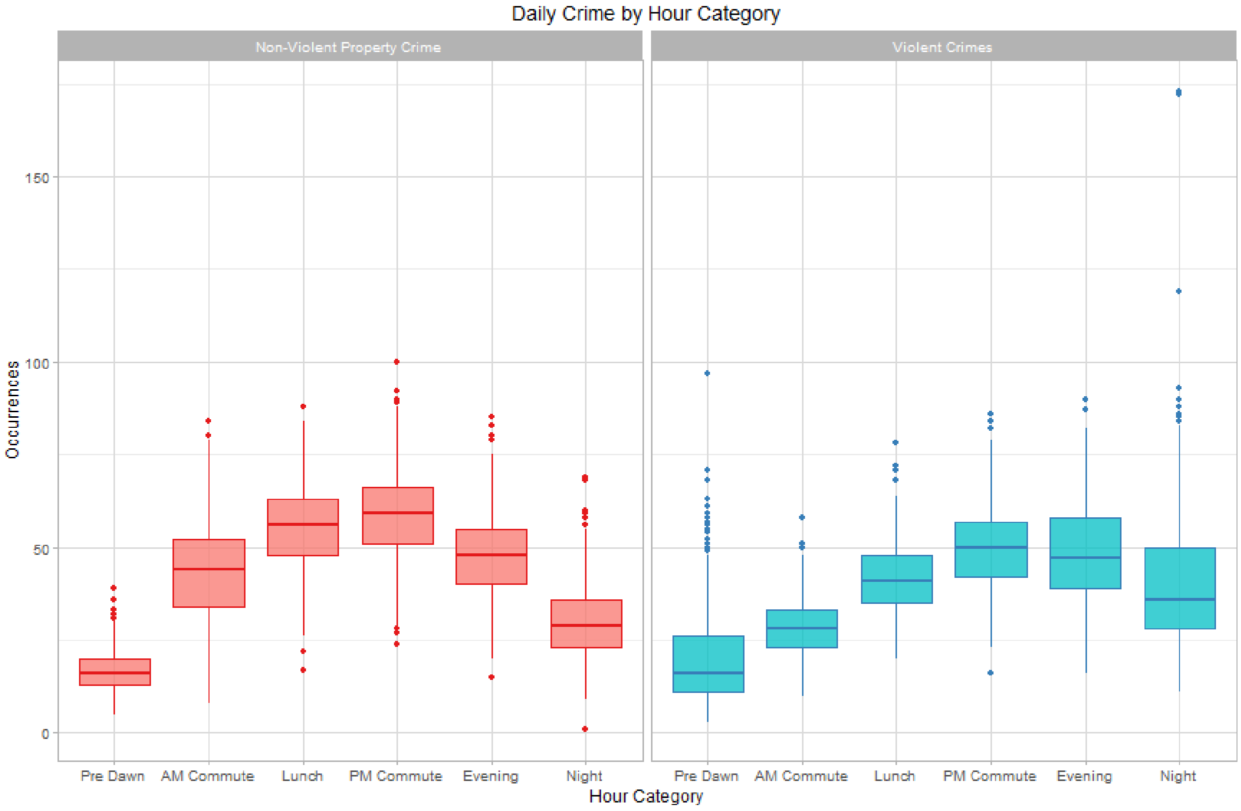


7.2 Distribution by Hour

Unlike day of week, there are large variances in crime frequency between hour categories:



The following boxplots display greater detail for the **Violent Crimes and Non-Violent Property Crime** categories. Note how property crimes peak over lunch and the PM commute, while violent crimes are more common later in the day. This distribution is statistically significant and will be explored more in the regression models below.



8 Modeling

Given each of the above factors, we chose to model the linear relationships between a given communities overall crime count and crime density in separate linear models against those factors from the above analysis that seem most significant taking co-linearity into account. These are 'Hour Category,' the '# of Transit Stations in Close Proximity to a Given Community', '% Households Below the Poverty Level,' and 'Minimum Distance to a transit Station' [as a proxy for access to any public transportation].

8.1 Model against count

A model based on these few factors explains nearly 50% of the variance in overall community crime counts. Note: Minimum Distance is not a significant factor in this model.

```
lm(formula = Count ~ Hour.Category + Num.Close.Transit + Percent.Households.Below.Poverty + 
    Minimum.Transit.Dist, data = crime_by_community3)

Residuals:
    Min      1Q  Median      3Q     Max 
-755.23 -161.67  -56.18  111.33 1403.43 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      -162.091     32.580  -4.975 7.84e-07 ***
Hour.CategoryAM Commute           168.201     31.376   5.361 1.06e-07 ***
Hour.CategoryLunch                291.159     31.375   9.280  < 2e-16 ***
Hour.CategoryPM Commute           348.592     31.377  11.110  < 2e-16 ***
Hour.CategoryEvening              292.607     31.375   9.326  < 2e-16 ***
Hour.CategoryNight                163.950     31.375   5.225 2.17e-07 ***
Num.Close.Transit                  24.015      1.413  16.990  < 2e-16 ***
Percent.Households.Below.Poverty    7.131      0.785   9.084  < 2e-16 ***
Minimum.Transit.Dist               -7.464      5.812  -1.284    0.199    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 269.9 on 879 degrees of freedom
  (55 observations deleted due to missingness)
Multiple R-squared:  0.4136, Adjusted R-squared:  0.4083 
F-statistic:  77.5 on 8 and 879 DF,  p-value: < 2.2e-16
```

I.e. within a given year and hour category, the above variables can explain 41% of the crime count variance by community.

8.2 Model against density

Another model based on the same factors can explain 51% of the annual variance in community crime density.

```
lm(formula = Density ~ Hour.Category + Num.Close.Transit + Percent.Households.Below.Poverty + 
    Minimum.Transit.Dist, data = crime_by_community4)

Residuals:
   Min     1Q Median     3Q    Max 
-68.22 -16.76  -1.79  13.92 229.05 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      -21.06111    3.65400  -5.764 1.14e-08 ***
Hour.CategoryAM Commute           24.42560    3.51890   6.941 7.55e-12 ***
Hour.CategoryLunch                42.76953    3.51888  12.154  < 2e-16 ***
Hour.CategoryPM Commute           50.40680    3.51899  14.324  < 2e-16 ***
Hour.CategoryEvening              41.45542    3.51885  11.781  < 2e-16 ***
Hour.CategoryNight                22.28764    3.51884   6.334 3.81e-10 ***
Num.Close.Transit                  2.77874    0.15852  17.529  < 2e-16 ***
Percent.Households.Below.Poverty   1.19176    0.08804  13.536  < 2e-16 ***
Minimum.Transit.Dist              -2.81023    0.65185  -4.311 1.81e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.27 on 879 degrees of freedom
  (55 observations deleted due to missingness)
Multiple R-squared:  0.5165, Adjusted R-squared:  0.5121 
F-statistic: 117.4 on 8 and 879 DF,  p-value: < 2.2e-16
```

9 Conclusions

9.1 Infrastructure Proposal

Based on this analysis, it would appear that crime density actually increases as one moves closer to both police departments and transit stations.

While it may not have the hypothesized benefit of reducing crime rates, placing police departments near transit stations will still give police the best chance to affect crime in their areas. As such, the city may still be best served by building small, permanent outposts at its most affected transit stations as opposed to a single, large department. This strategy could reduce costs while having a targeted effect on demonstrated hot spots. Regular assignments to these posts could have the added benefit of improving community relations and participation through personal familiarity. I would recommend running a few trials in the densest communities to determine effect. Per the choropleth and bar graphs in the community section of this paper, I would recommend running separate trials in West and East Garfield for Violent Crimes vs. the Loop and the Near North Side for Property Crimes. Trials should help to isolate divergent effects on Violent Crimes vs. Non-Violent Property Crime.

9.1.1 Proposal for further analysis

As has been reported in recent [spring of 2015] news, Chicago is "off to [its] deadliest start in nearly two decades." Homicides and shootings are up by more than 70% year over year through the first quarter, while arrests and investigative stops have decreased [follow the link for the source]. The reasons for these shifts are beyond the scope of this analysis. However, a time lapse analysis could shed some light on growth patterns. If the client wishes to move forward, I would suggest the following:

- Add additional transit station locations and volume estimates for each location. E.g. bus routes
- Add Police Station employee data
- Add additional community and/or demographic data
- Perform temporal analysis showing acceleration/deceleration of crime rates per community to understand trends.

9.2 Temporal (Schedule) Conclusions

Other than relatively weak, weekend effects, patrol schedules do not need to adjust for day of week volumes. There are, however, strong shifts in hourly volume and staff should be scheduled accordingly.

9.2.1 Proposal for further analysis

As a follow up to this analysis, I would suggest the following:

- Perform an Erlang-C style analysis to determine workforce requirements on an hourly basis per police district.
- Examine Holiday effects on crime.
- Expand prediction models to include other variables that could assist with scheduling decisions such as weather.