**Data Driven Approach to Crime and Traffic Safety Analysis**

W200 Fall 18 | Thurs 6:30 PM

Group Project - Chicago Crime Analysis

Presented on: December 13, 2018

Team Members: Eddie Zhu, Jonathan Hilton, Thanh Le

**The Purpose:**

The objective of this project is to analyze the correlations and potential cause / effect relationship of the variables between Chicago crimes data with Chicago traffic violations data and Chicago demographic data. The data will be cleaned up, investigated and grouped by a common variable, zip code, and further analysis will be done, as noted below.

**The Focus / Objectives:**

Our project will focus on providing an analysis to the following questions:

1. What are the various types of crimes in Chicago and what correlation it has to the other variables (traffic violations, demographics)?
2. Are there any correlation between Chicago crime and the level of income in the various neighborhood?
   1. Are there any correlation between Chicago crime and tickets and property tax?

**Data Sources (raw data):**

1. Chicago crimes data: [Data.gov link](https://catalog.data.gov/dataset?publisher=data.cityofchicago.org), [CSV file](https://drive.google.com/file/d/1BuCw_SsQmJGE0sJdIYnaWtGenaWo4zmH/view?usp=sharing)
   1. This dataset contains 6,749,651 rows and 22 columns, 1.59 GB.
   2. Columns of this dataset:
      1. ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location
   3. This dataset contains data from year 2001 to 2018.
2. Chicago traffic violation data: [Propublica.org link](https://www.propublica.org/datastore/dataset/chicago-parking-ticket-data), [CSV file (zip file)](https://drive.google.com/file/d/1iKgjaFVpmUwcSgrMke0kNB4aoLT26zzj/view?ts=5bfa1c50)
   1. This dataset contains 28,272,580 rows and 23 columns, 7.67 GB.
   2. Columns of this dataset:
      1. Ticket\_number, issue\_date, violation\_location, license\_plate\_number, license\_plate\_state, license\_plate\_type, zipcode, violation\_code, violation\_description, unit, unit\_description, vehicle\_make, fine\_level1\_amount, fine\_level2\_amount, current\_amount\_due, total\_payments, ticket\_queue, ticket\_queue\_date, notice\_level, hearing\_disposition, notice\_number, officer, address
   3. A data set collected based sticker tickets and has data from Jan 1, 2007 to May 18, 2018. The key data is the when and where of the tickets along with the amount and much more.
3. Chicago demographic data: [City-data.com link](http://www.city-data.com/zipmaps/Chicago-Illinois.html), [Excel file](https://drive.google.com/file/d/1ixxCrhJ1z-wbGo3DJWODnEuch9r8nqFR/view?usp=sharing)
   1. This dataset contains 55 rows and 24 columns, 13 KB.
   2. Columns of this dataset:
      1. Zip Code, Estimated Zip Code Populations by certain years, Mar index, Land area, Water area, Population Density (#/sq mile), Males, Females, Real Estate Property Tax by Paid for Housing Units by certain years, Property Tax % on Houses, Property Tax $ on Houses, Median Real Estate Property Tax Paid for Housing Units With Mortgage, Property Tax Paid for Housing Units Without Mortgage, Estimated Median House/Condo Value, Estimated Median Household Income, Median Monthly Owner Costs for Units With a Mortgage, Median Monthly Owner Costs for Units Without a Mortgage, Median Gross Rent, Median Price Asked For Vacant For-sale Houses / Condos, Unemployment, Lat, Lon
   3. The demographic dataset included much of the data that came from city-data.com. This website provides key data sets for the zip code. They key data use for this analysis was population, tax information, March 2016 index, unemployment rates and much more.

**Data Investigations / Explorations / Clean-up:**

Before starting our analysis, we perform a sanity check on the data and its variables to ensure its quality and integrity. For our sanity, the checks was limited to only set of data and variables that we will be using in our analysis:

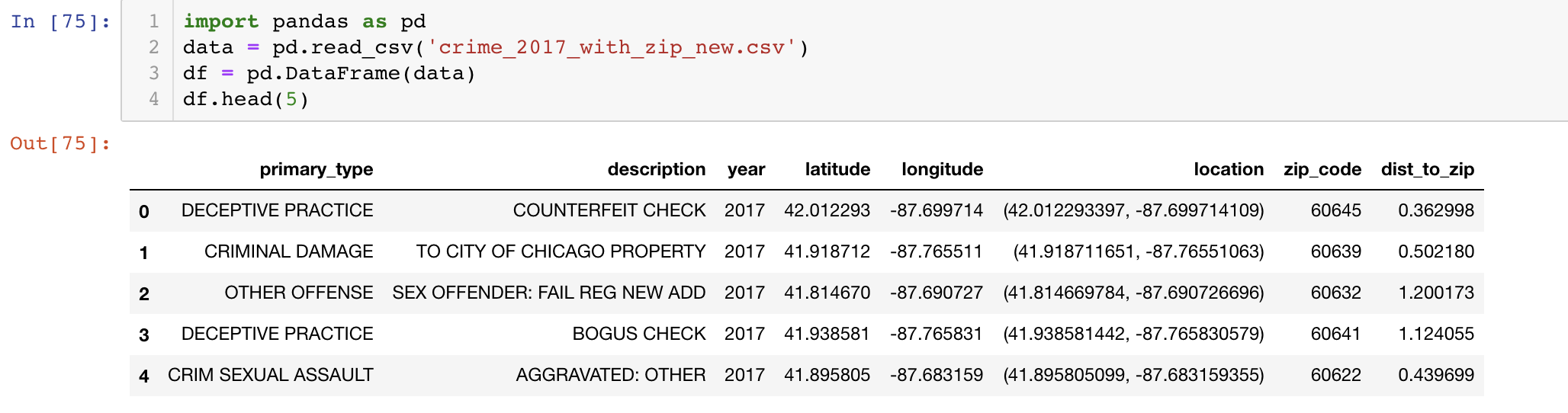
***Chicago Crimes Dataset:***

For this project, we filtered the dataset to only the following columns from the original dataframe: Primary Type, Description, Year, Latitude, Longitude, Location. And then filtered further to only data for year 2017 to perform our analysis.

We mainly checked for completeness of data for all the variables, using the len() function. The checks revealed to us that there are missing data (Latitude and Longitude) in various rows. The missing data was about 1% of the total dataset, 2,811 rows, of our 2017 data. We then use describe() function to perform other checks. Our conclusion is that this is a fairly usable and complete dataset. The rows with missing data will be removed from further analysis when zip code is considered.

Our initial hypothesis was that there should be correlation between traffic violation and non-traffic crime at any given location. In order to prove the hypothesis, we need to find the common field that can link crime and traffic datasets. However, we faced a challenge while examining the datasets. Crime dataset has latitude and longitude as location while traffic violation dataset only has zip codes. In order to convert the latitude and longitude to zip code, we used the trigonometric functions from the math model to calculate the distances from each crime location to the coordinates of every corresponding zip code and assigned a zip code to each crime location based on distance, which provided us a common field in both datasets to perform regression analysis and geo mapping, using Geopanda module.

The result looks great with the columns we need.



***Chicago Traffic Violations Dataset:***

These are the variables that were checked for this particular dataset: issue\_date, violation\_code, violation\_description, zip\_code, fine\_level1\_amount, fine\_level2\_amount, year

For the purpose of this analysis, we focused on 2017 to reduce the volume of the data and to ensure some comparability with the crimes dataset. This reduced the rows from 28M to 2,190,763. We also reduced the number of columns from 23 to these 8 because were mostly focused on the number of violations, the location, and the amount.

Another key cleanup on this data was to remove entries with no zip codes. There were a significant amount of zip codes that were either blank or were not in a Chicago based zip code.

***Chicago Demographic Dataset:***

No major data clean-up was required for the demographic data because it came in a structured form. However, we did have to do a lot of structuring to get the data in a consistent format since it did not come in a tabular format from the website.

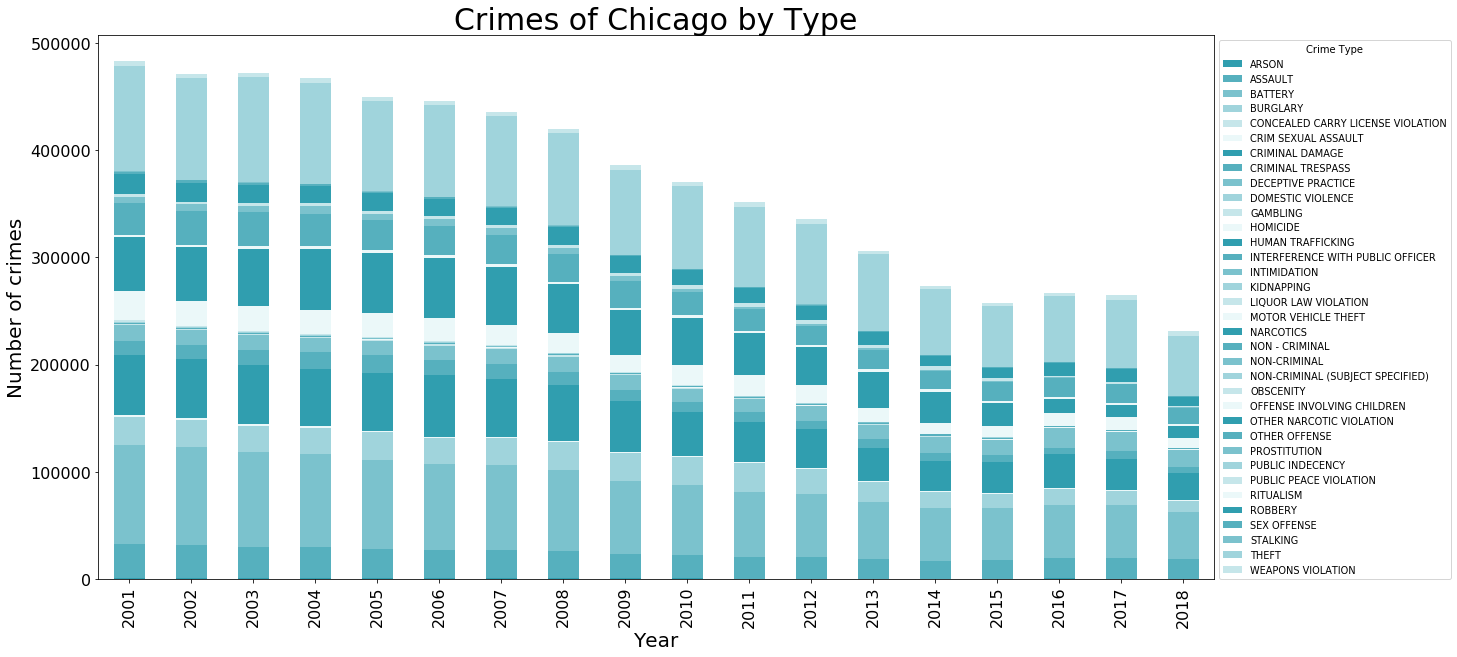
**Assumptions made:**

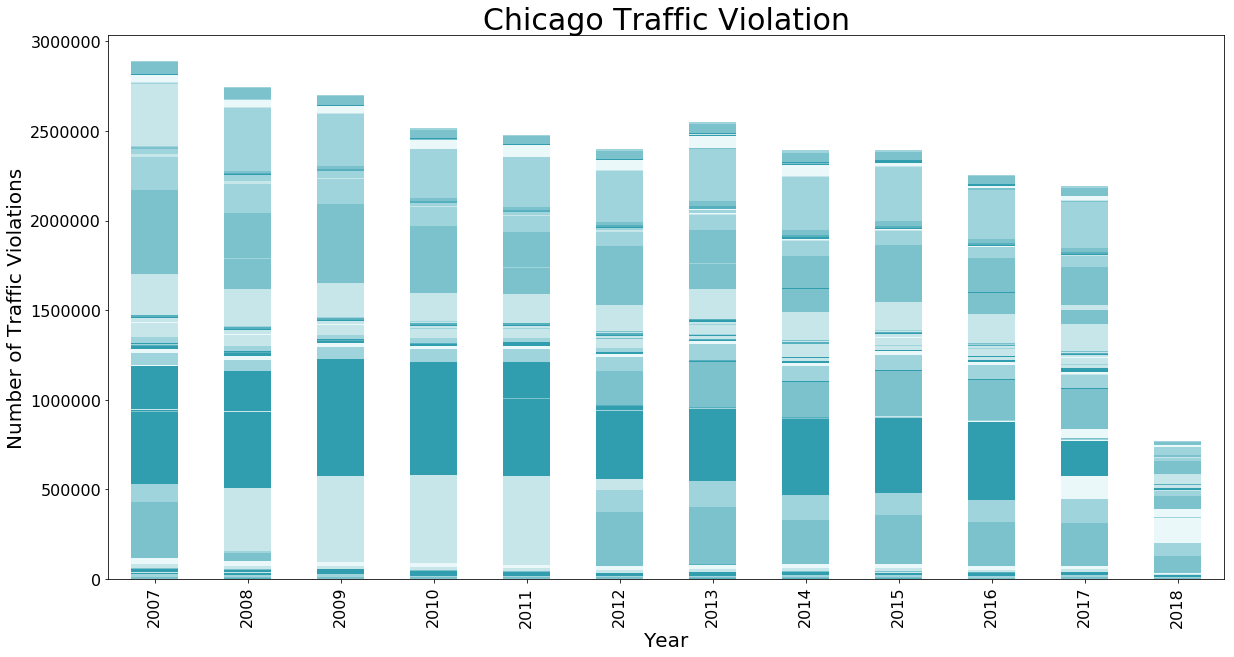
Here are assumptions we made while doing our analysis:

1. Data in our datasets were accurate representation of the truth
2. “High income” in Chicago = $100,000 and above
3. “Low income” in Chicago = $60,000 and below
4. Despite a high number of zip code entries being blank and excluded from this analysis, we assume that the remaining tickets with zip codes are representative of the total number of tickets per zip code.

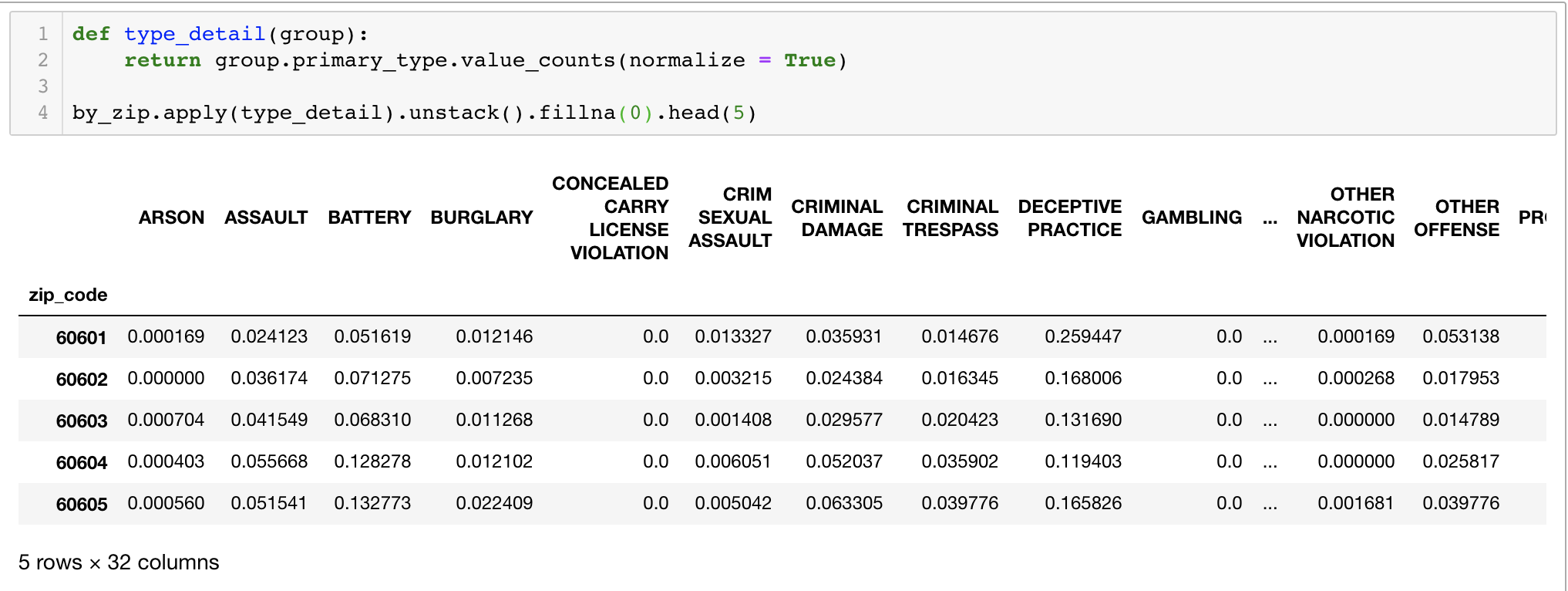
**Our Preliminary Findings and Analysis:**

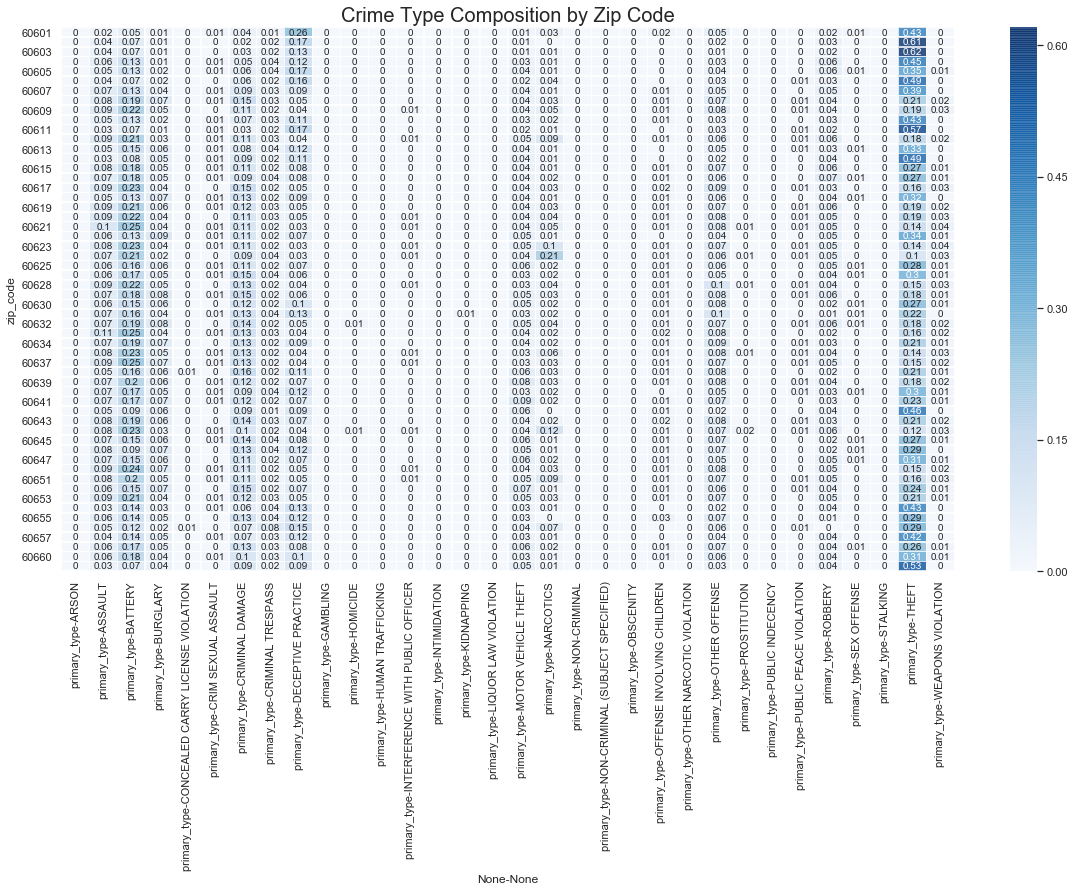
Our initial finding shows crimes in Chicago has been decreasing over the years; and coincidentally, traffic violations is also trending in that general direction over the years.



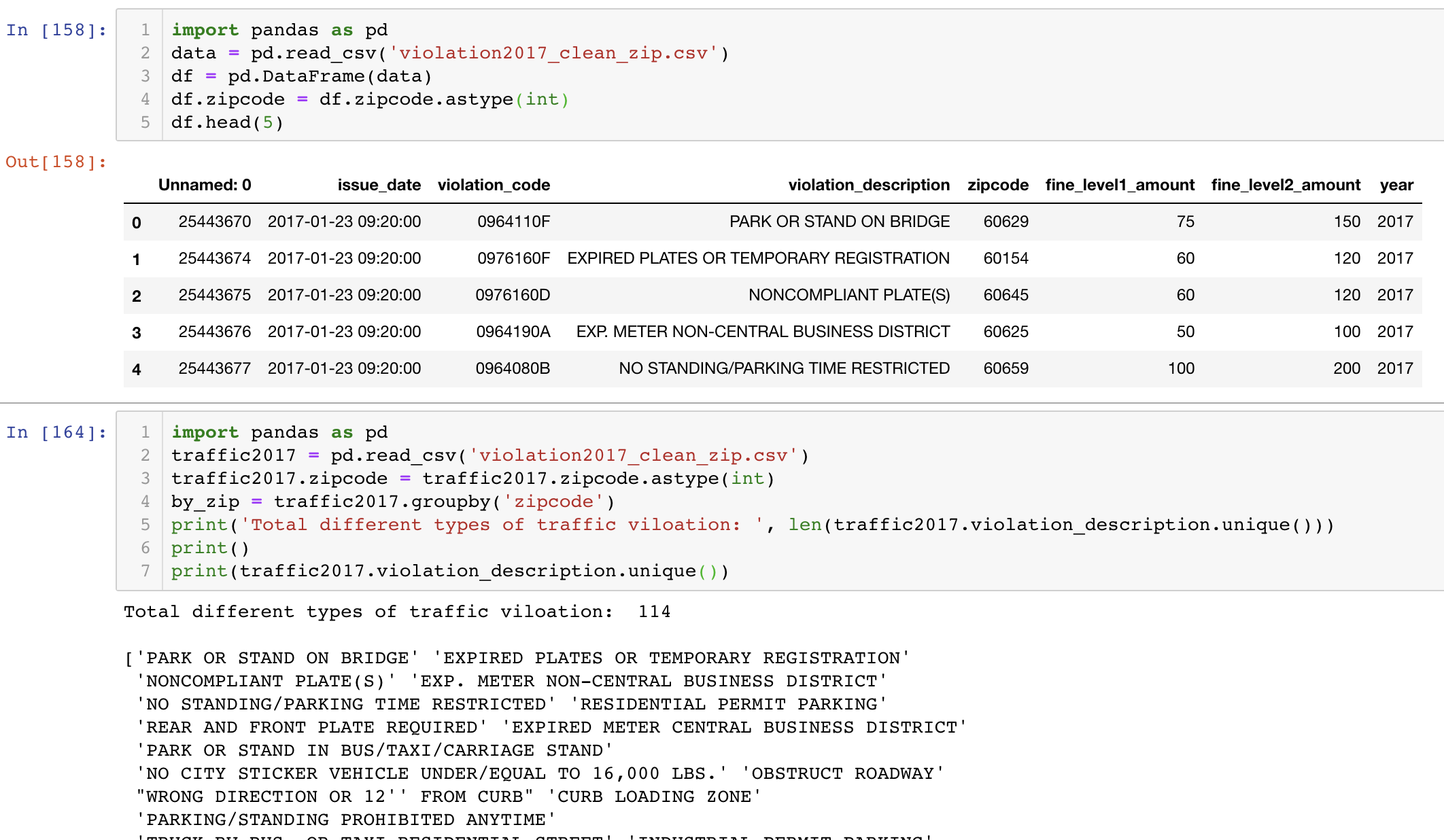
****

We looked at the different crime type composition in every zip code. Using pandas’ General Split-Apply-Combine methods, we calculated the percentage of every crime in every zip code area. The following table shows the crime type composition by zip code. We then created a heatmap, using seaborn. The result is very informative, indicating some areas have more concentrated crime types like theft while others are more spread across different types. This information can be helpful for police dept to design reinforcement allocation in different areas.

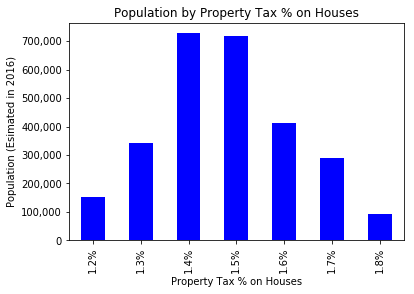


****

We attempted to create a heat map for traffic violation types by zip code. However, the traffic type descriptions appeared to be rather random with over 114 different types recorded for year 2017. It became evident that graph with this many variables won’t provide significant insights. However, as a recommendation to the law enforcement, we believe that standardized violation types will provide significant improvement on data analysis and workforce optimization.



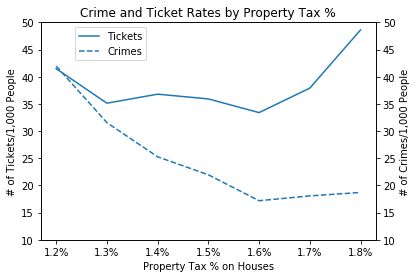
Another analysis was to determine if there was a relationship of tax rates to crime or ticket. Tax rates in chicago ranged from 1.2% to 1.8%. While this analysis cannot incorporate sales tax or other forms of income for the local government, it does represent a great deal of the revenue required for government operations.

In order to provide this analysis, we combined the demographic data which included population sizes, tax rates, and tax amounts based on median house price. This could be combined with the crime and tickets. For the crime data, this required grouping the data by zip code and summing the crimes by zip code. For the ticket data, this required grouping the number of tickets by zip code.

Once all of the data was in the same data frame by zip code, we sorted by tax rates. Next we needed to normalize data according to population. We did this because if we only look at the number of tickets or crimes by tax rate we may get undue representation based on population sizes under those tax rates. The distribution of population by tax percentage on houses can be seen here:

Therefore, we always wanted to look at the number of tickets or crimes per 1,000 people in order to see if the ticket or crime rates are related to tax rates.

Here are the results:



What we found from this analysis was there was actually a trend where more ticket issuance remained steady through the mid-tax rates and did have an inflection after 1.7%. What is more interesting is how the number of crimes per 1,000 people trends down with the increasing tax rates. This data could support a hypothesis that the higher tax rates push higher law enforcement efforts that are present in giving out tickets and also driving down crime. There are other socio-economic facts that likely play in higher tax areas having lower crimes but there is likely an element of that higher tax rate resulting in higher law enforcement.

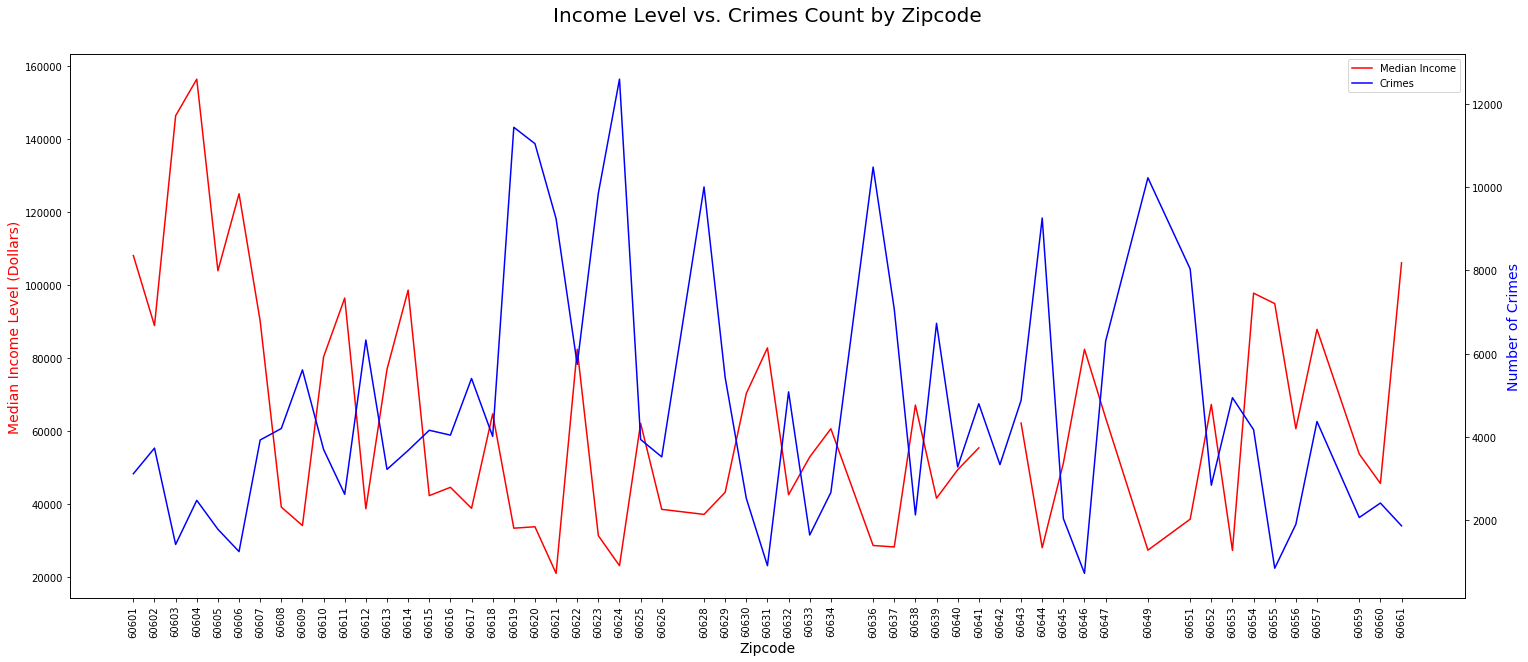
We also looked at this not only from a tax rate perspective but even from a range of the property tax amount from the median house value. This required a similar approach to the above as well as categorizing the amounts into bins to see if we could identify a trend.

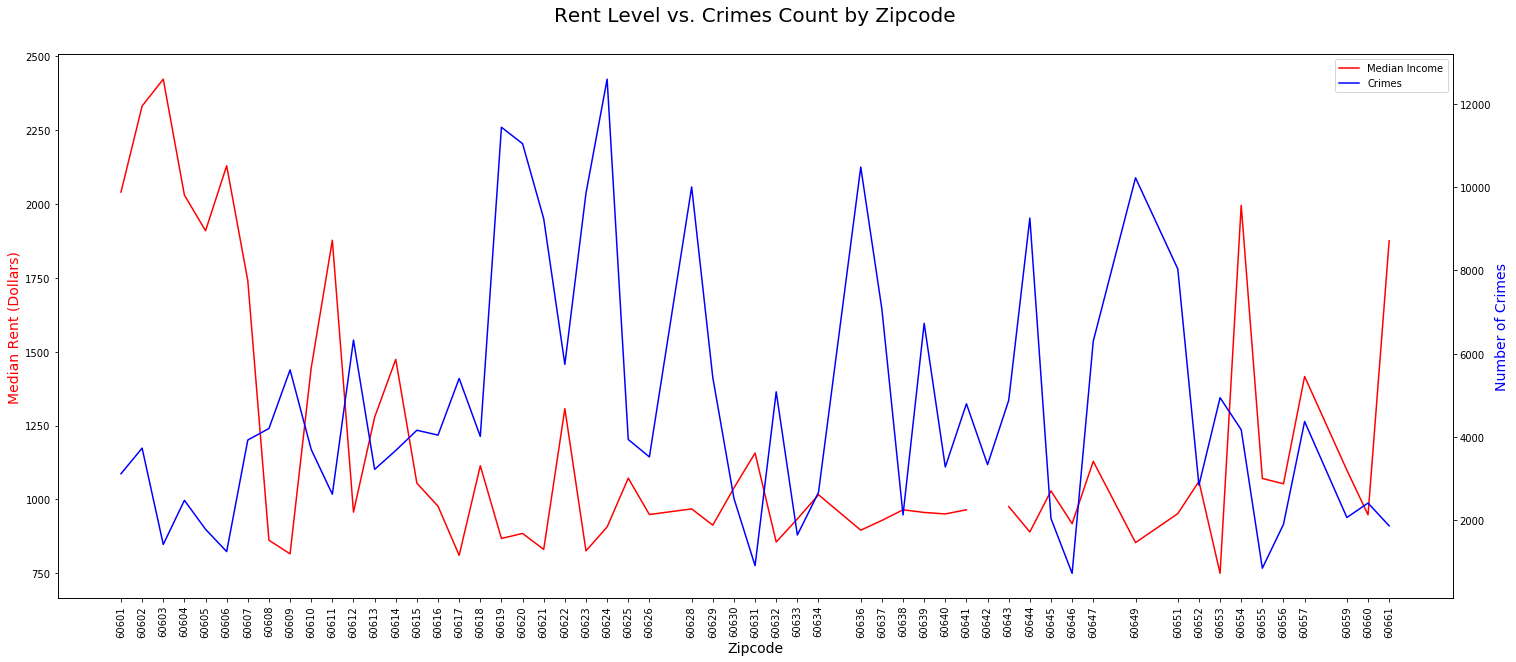
|  |  |
| --- | --- |
|  |  |

The trend here continues especially in regards to crime. As the tax amount increases, crime tends to go down. There is an anomaly in the $6-7k range which requires further investigation in those areas. As for tickets, the downward trend here is not as consistent as the previous graph where we saw the ticket issuance rate increase with higher tax rates.

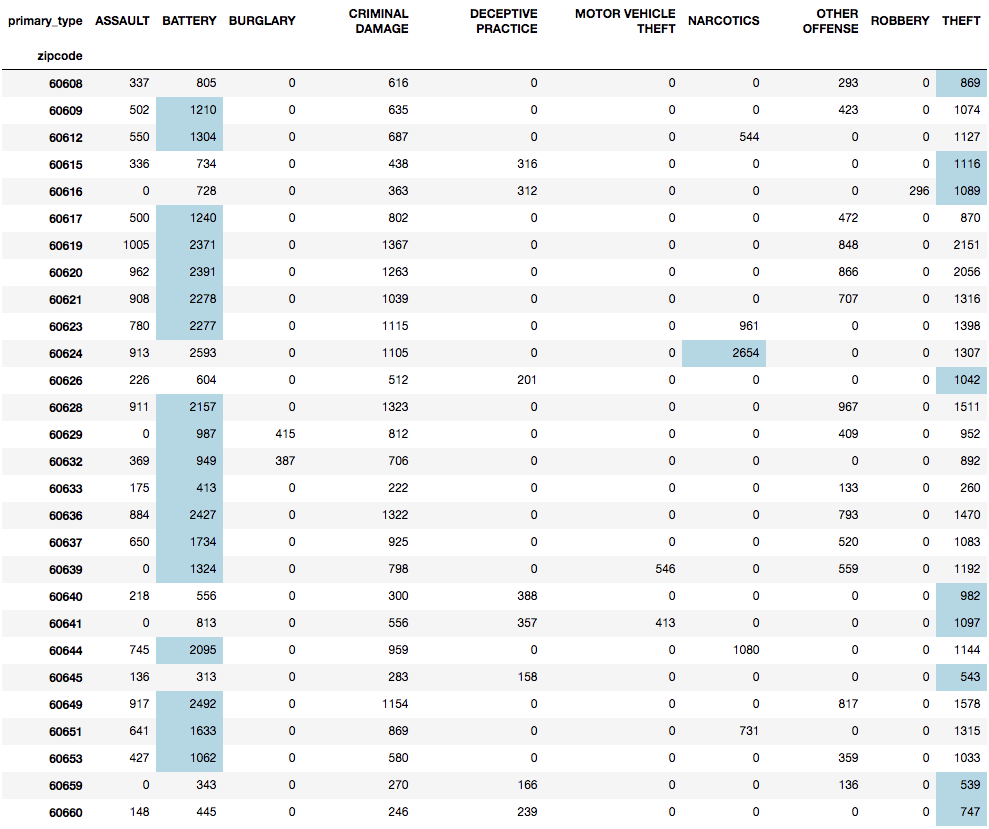
Overall, it does appear that crime rates have a stronger relationship to tax rates and amounts. Whereas ticket issuance is less strongly tied to average property tax amounts and slightly more tied property tax rates.

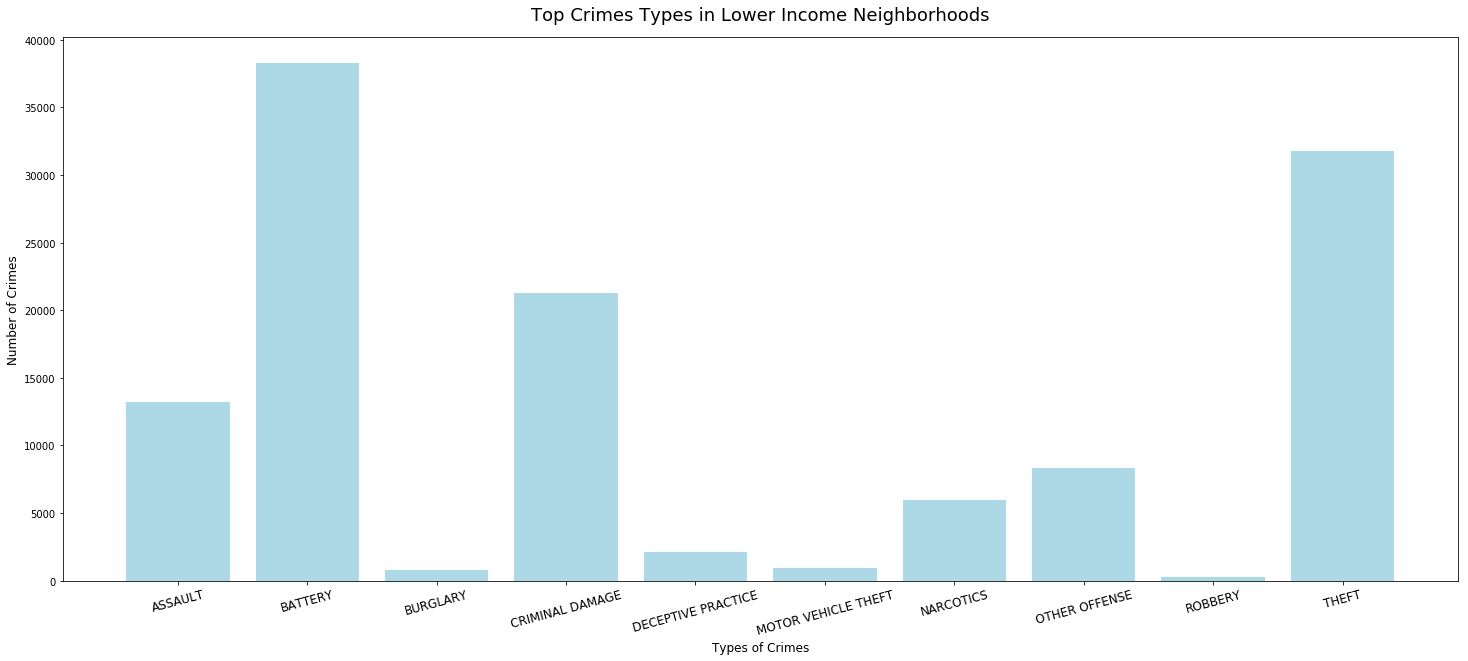
We perform another analysis to determine whether there is correlation between crimes and neighborhood median income and neighborhood median rent. We determined neighborhoods by its particular zip code. From the data and graphs, we concluded there is high correlation between crimes rate and the well-being (in term of income) of a neighborhood. Neighborhoods with relatively higher income (in this scenario, median income level >= $100,000), have less amount of crimes. And neighborhoods with relatively lower level of income (median income level <= $60,000) have higher number of crimes.



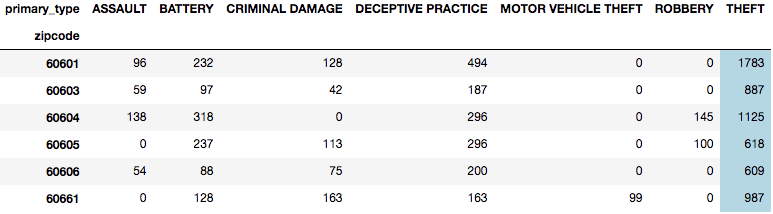


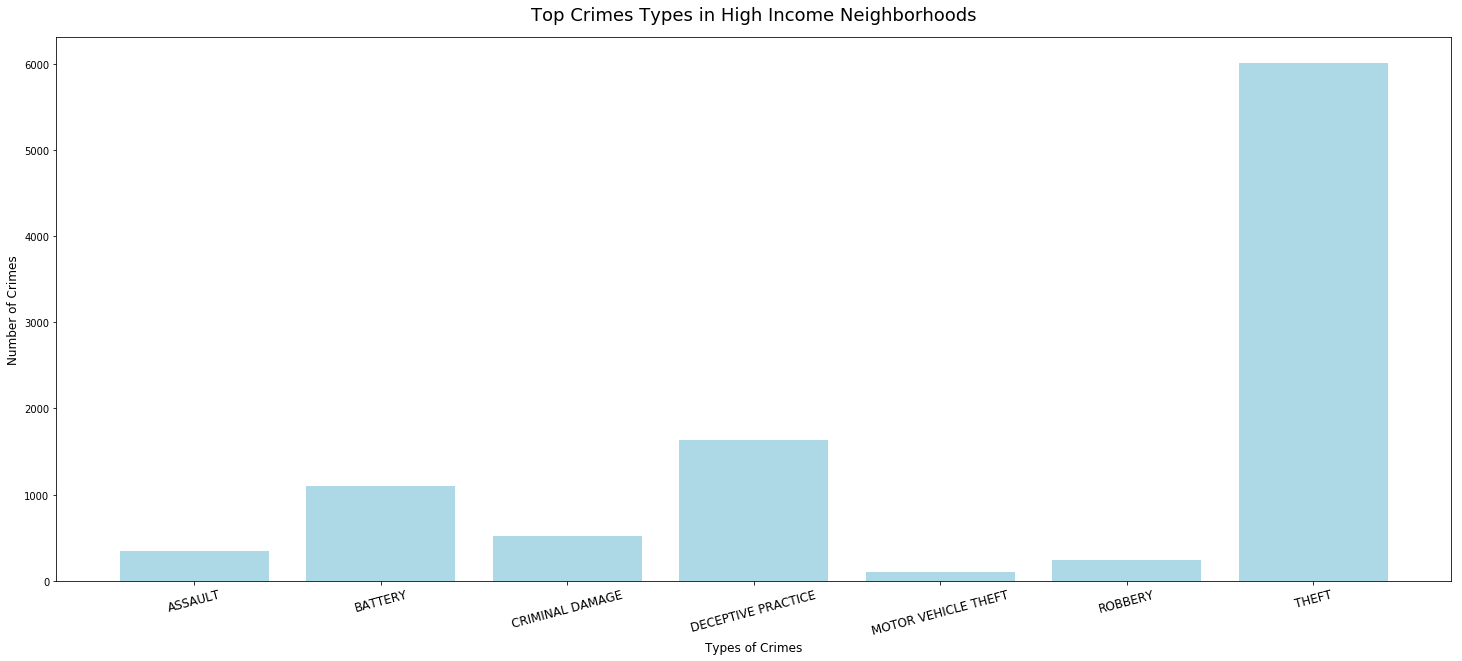
Digging further into this dataset, we extracted the zip codes for what we considered lower income neighborhoods, specifically median income level <= $60,000. This yielded 28 distinct zip codes / neighborhoods. From these zip codes, we filtered and pivot the original dataset to show the top 5 crimes per zip code, and the results are as followed: Assault, Battery, Burglary, Criminal Damage, Deceptive Practice, Motor Vehicle Theft, Narcotics, Robbery, Theft and Other Offense. The most common type crime is Battery in the lower income neighborhoods.





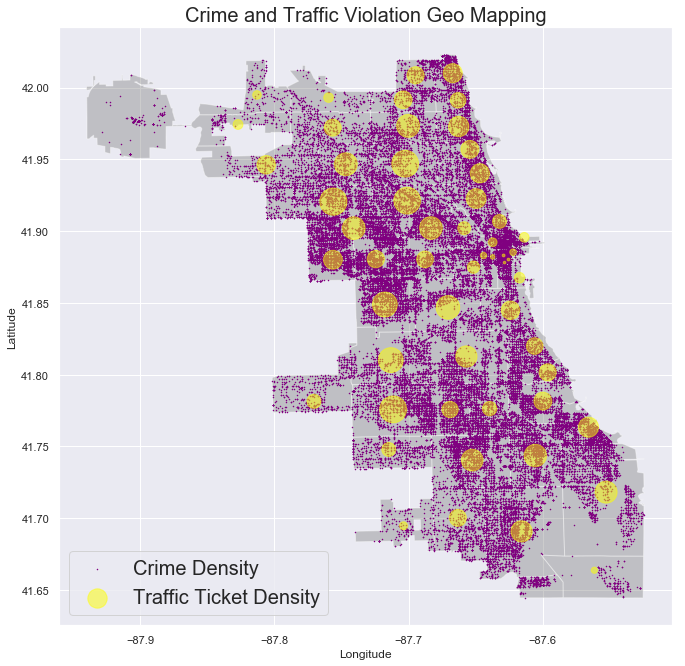
We then perform another filter to extract the zip codes for high income neighborhood, specifically median income level >= $100,000. This yielded 6 distinct zip codes / neighborhoods. From these zip codes, we filtered and pivot the original dataset to show the top 5 crimes per zip code as followed: Assault, Battery, Criminal Damage, Deceptive Practice, Motor Vehicle Theft, Robbery and Theft. The most common type crime is Theft in the higher income neighborhoods.



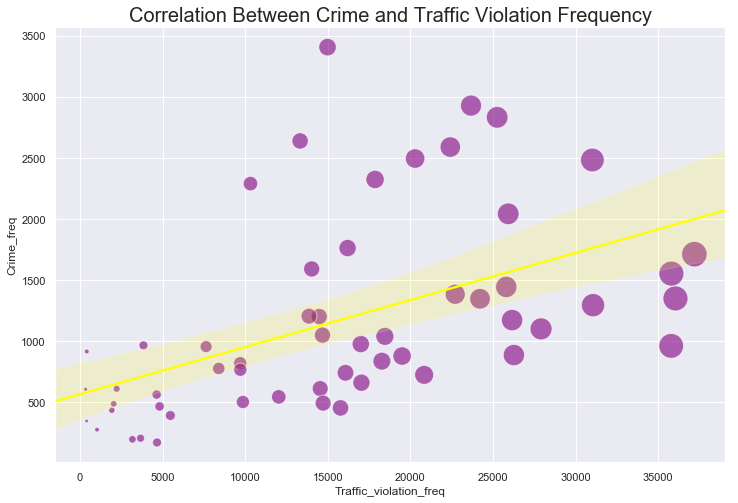


**Deeper Analysis:**

Using Geopanda module, we created a geo mapping of crime occurrence and traffic violations in Chicago area. Geo Mapping provided a visual proof of the correlation between traffic violation and non-traffic crime. In the map below, the purple dots represent the non-traffic crime occurrence in corresponding locations in 2017. The size of yellow bubble indicates the number of traffic violation occurrence in every corresponding zip code area. Except for a few outliers, we can visually conclude that, in general, areas with higher traffic violation occurrence tend to have higher crime rate, and vice versa.

****

In order to further prove the hypothesis, we ran a regression analysis, using Seaborn module. First, we used Pandas DataFrame groupby function to calculate the frequency of non-traffic crime and traffic violation occurrence. Then, we used Pandas DataFrame merge function to merge the two datasets together, with which a regression plot was created. The plot below shows us that there is a positive correlation between crime occurrence and traffic violation at any given location.

****

Another interesting finding from the approaches above was that the areas with extremely high crime rate tend to have low traffic violations. This phenomenon prompted us to further research the correlation between crime rates and demographics, with the following hypotheses.

1. Areas with high population density might have less people driving, therefore, less traffic violation, but higher crime occurrence.
2. Urban areas with lower average income and higher unemployment rate might present higher crime occurrence with lower traffic violations.

We used BoKeh to create an interactive geo mapping of crime occurrence, traffic violation frequency and population density by zip code. From the map below, we can conclude that our hypothesis has been proved. Crime occurrence is positively correlated with population density while traffic violation is negatively correlated with population density, knowing which can help law enforcement to better manage the allocation of the police force. For example, in areas with low population density, law enforcement can install more monitoring cameras, radar monitors and warning signs which would be sufficient to deter most traffic violations. In the areas where population density is high, law enforcement should allocate more police force to prevent hard crime. (open dynamic map [here](https://storage.googleapis.com/w200-berkeley.appspot.com/map_crime_violation_demo.html))

|  |  |
| --- | --- |
|  |  |

**Actionable Recommendations:**

Our findings are preliminary, but we believe it can assist Chicago Law Enforcements in the following areas:

1. Better manage the allocation of the police force
2. Technologies to monitor and deter traffic violations in rural areas
3. Police force to prevent hard crimes in urban areas
4. Standardized data entries provide significant improvement on data analysis and workforce optimization.

Bibliography

1. <https://catalog.data.gov/dataset?publisher=data.cityofchicago.org>
2. <https://www.propublica.org/datastore/dataset/chicago-parking-ticket-data>
3. <http://www.city-data.com/zipmaps/Chicago-Illinois.html>