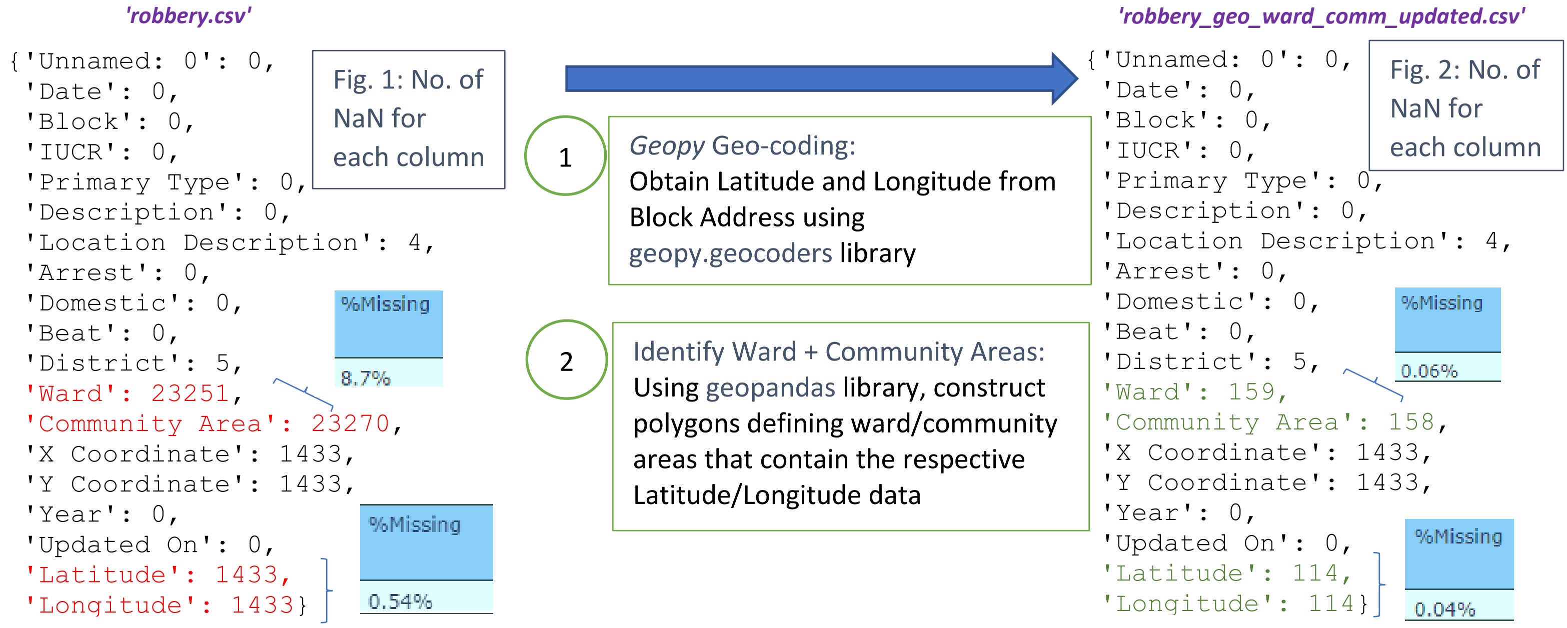
Report on Chicago Crimes

Bruce Wen, CMSC 119

(1) Data Cleaning and Recovery from *Robbery.csv*



'robbery.csv'

```
{'Unnamed: 0': 0,
 'Date': 0,
 'Block': 0,
 'IUCR': 0,
 'Primary Type': 0,
 'Description': 0,
 'Location Description': 4,
 'Arrest': 0,
 'Domestic': 0,
 'Beat': 0,
 'District': 5,
 'Ward': 23251,
 'Community Area': 23270,
 'X Coordinate': 1433,
 'Y Coordinate': 1433,
 'Year': 0,
 'Updated On': 0,
 'Latitude': 1433,
 'Longitude': 1433}
```

Fig. 1: No. of NaN for each column

%Missing
8.7%

%Missing
0.54%

1 *Geopy* Geo-coding: Obtain Latitude and Longitude from Block Address using geopy.geocoders library

2 Identify Ward + Community Areas: Using geopandas library, construct polygons defining ward/community areas that contain the respective Latitude/Longitude data

'robbery_geo_ward_comm_updated.csv'

```
{'Unnamed: 0': 0,
 'Date': 0,
 'Block': 0,
 'IUCR': 0,
 'Primary Type': 0,
 'Description': 0,
 'Location Description': 4,
 'Arrest': 0,
 'Domestic': 0,
 'Beat': 0,
 'District': 5,
 'Ward': 159,
 'Community Area': 158,
 'X Coordinate': 1433,
 'Y Coordinate': 1433,
 'Year': 0,
 'Updated On': 0,
 'Latitude': 114,
 'Longitude': 114}
```

Fig. 2: No. of NaN for each column

%Missing
0.06%

%Missing
0.04%

It is clear from the *.isnull()* method that there exists numerous *NaN* values within Robbery.csv, which have been summarized as a dictionary as in Fig. 1. Other checks including the sensibility (e.g. no strings in '*Latitude*' column) of each data point within the columns have also been conducted, which prove the data to be satisfactory. To recover the missing latitude/longitude and ward/community area data, thankfully, the 'Block' column has 0 NaN values: within each entry under 'Block', the street address is provided, which through careful manipulations generate the values for the other missing columns. To obtain *'robbery_geo_ward_comm_updated.csv'*, geopy.geocoders library is imported. We run a *http* query on the given address name from 'Block' column to generate corresponding latitudes and longitudes. This returns a dataframe which has significantly reduced NaN values as seen in Fig. 2. We also export two *GeoJson* files from the Chicago Data Portal website*, namely 'Boundaries - Community Areas (current).geojson'* and *'Boundaries - Wards (2015-).geojson'* . With these two Geojson files, for each crime record, based on the existing or recovered latitude/longitude coordinates, we are able to loop through the community regions and wards to find their IDs respectively if they are missing with the help of the geopandas library. The resulting dataframe has significantly reduced NaN values. X/Y coordinates need not be resolved as latitude and longitude values are sufficiently precise to give insights on location. **Subsequent analyses use the *'robbery_geo_ward_comm_updated.csv'* exported as in the columns represented by Fig. 2.**

(2) Temporal Trends in Chicago Robbery

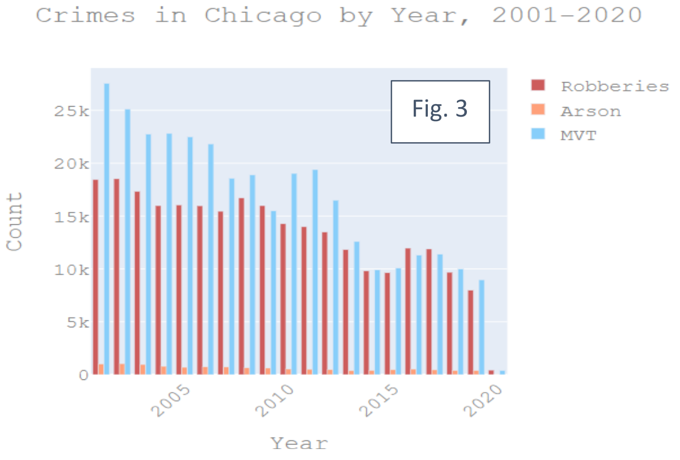


Fig. 3

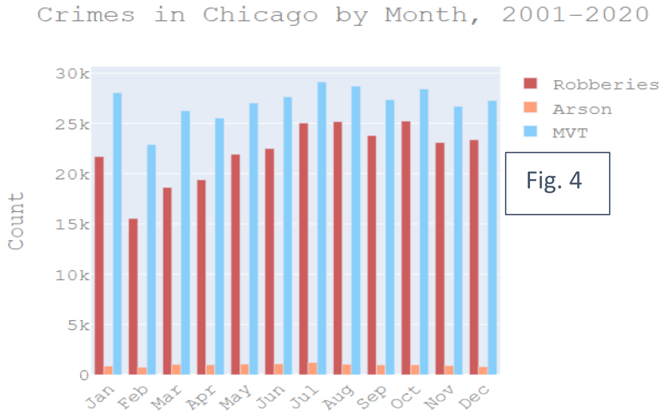


Fig. 4

To analyze the temporal trends, the "Date" column from *'robbery_geo_ward_comm_updated.csv'* is split into year, month, day of week, hour via *pd.to_datetime()* method, and then grouped respectively. The same is done for *'arson.csv'* and *'mvt.csv'*, and these data are together plotted in Figures 3-6.

First, it is evident from Fig. 3 that over the years, **all 3 forms of crime experience a general decline**. Relatively, **motor vehicle theft has decreased the most**, reaching almost the same level as robbery count (difference of ~1000) in 2019 despite a difference of ~10,000 in 2001. Next, considering the monthly data in Fig. 4, it is notable that **in February Chicago experiences the least crimes** in all 3 forms: robbery, arson, and motor vehicle theft. One reason could simply be that February has less days, 28 compared to the typical 30 or 31 days. But that cannot be the only reason: it is evident that in the summer and autumn months from June to October, all 3 forms of crime are in significantly higher numbers. This is especially evident for arson. Likely, the **winter cold is associated with a reduction in crime**.

While there is no clear trend for all 3 crimes in relation to the day of the week (Fig. 5), there are very significant differences in crimes by the hour of day (Fig. 6). Each of the 3 forms of crimes have different peak and trough periods. For robbery and MVT, the peak periods of crime are at 2200hrs and 2300hrs respectively. Notably, **while robbery and MVT experience lowest counts from 0700-0900**, this time period **for arson is virtually the entire day time**, from 0700-1800, with its peak from 2300-0500hrs. This is revealing: **Arson happens mainly in**

the night, and rarely during the day. Meanwhile, the peak periods of robbery/MVT are from 2100-0000hrs.

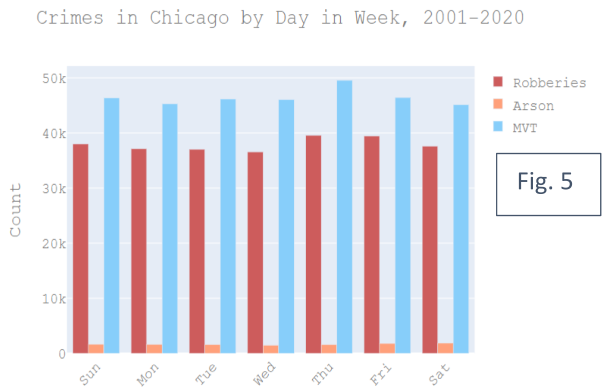


Fig. 5

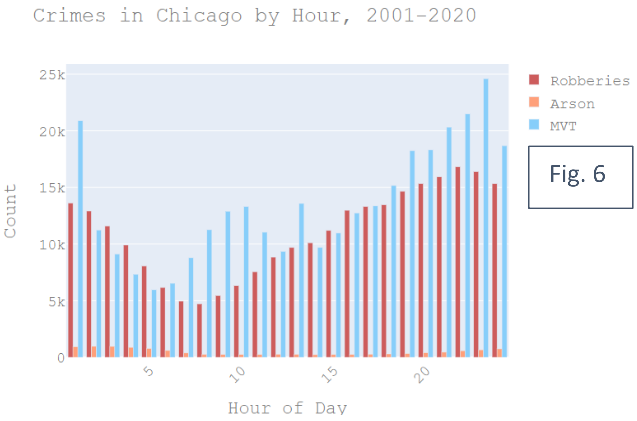


Fig. 6

Further analysis on temporal trends is conducted to incorporate spatial patterns in *Fig. 7, 8, 9*. These **choropleths** are plotted using *plot.ly* library, allowing us to view both the temporal and spatial patterns in the same 2D space. Recall that the boundaries of community areas are defined in Section 1 with *'Boundaries - Community Areas (current).geojson'*. Areas 8, 13, and 57 are labelled because of their interesting changes across the years. For area 13, while there was an 80% rise from 2002-2010, there was a subsequent -36.1% and -30.4% change in 2010-2015 and 2015-2019 respectively. Yet for Area 8, which is under a part of the Central 'side', there has been an especially large 104% increase in 2015-2019. The following study, *Understanding the Crime Gap: Violence and Inequality in an American City* from https://onlinelibrary.wiley.com/doi/full/10.1111/cico.12348, could give some insight into these changes: *"Neighborhoods that had higher levels of residential stability, with more homeowners and more long-term residents, and greater concentrations of immigrants experienced both lower levels of violent crime and more pronounced declines in it."* More in-depth study is required to draw the exact relations clearly and is suggested for future research.

Comparing the general North and South, it is interesting to note that **South Chicago is experiencing overall widespread decline** in robberies from 2010-2019, while **certain community areas in the North are experiencing increases**.

%Change in Robberies, 2002-2010

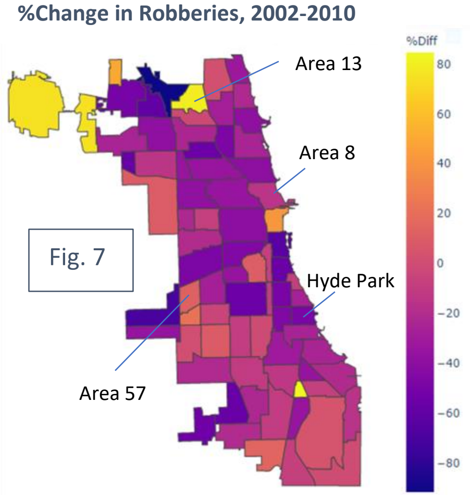


Fig. 7

%Change in Robberies, 2010-2015

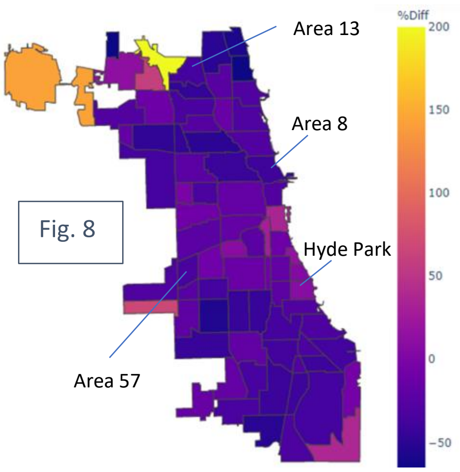


Fig. 8

%Change in Robberies, 2015-2019

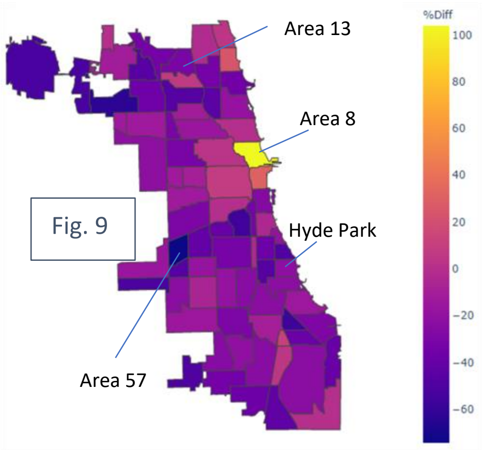


Fig. 9

Finally, we explore temporal trends of crime in **Hyde Park**. These graphs (Fig. 10, 11) are obtained by setting community area = 41 from ***'robbery_geo_ward_comm_updated.csv'***. Only robbery and MVT are considered because of the insignificant count of arson crimes (<=10 each year with large variance). Similar to the wider Chicago, Hyde Park is experiencing declining robberies and motor vehicle thefts. Furthermore, these crimes follow the general Chicago trend in that they are less frequent in the Winter months from December to April, and peak in the summer months.

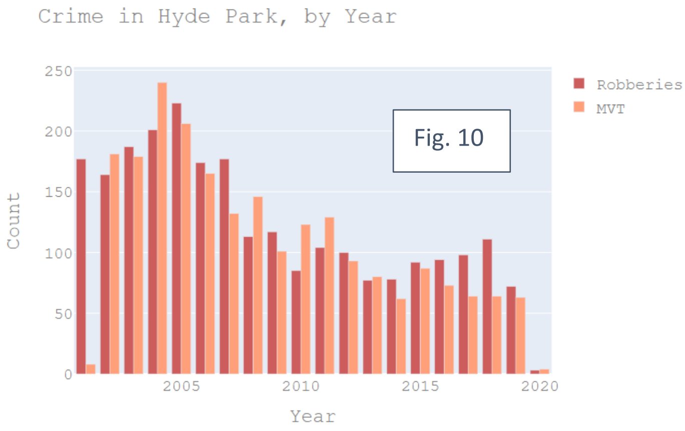


Fig. 10

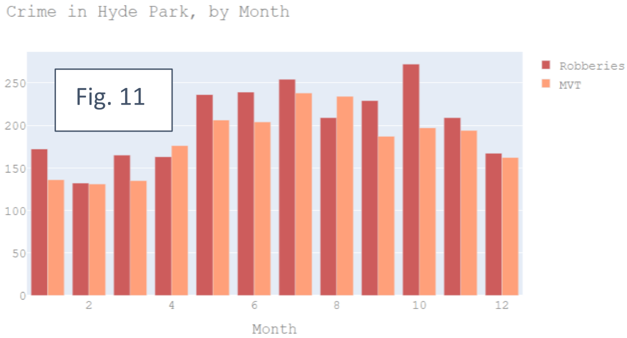
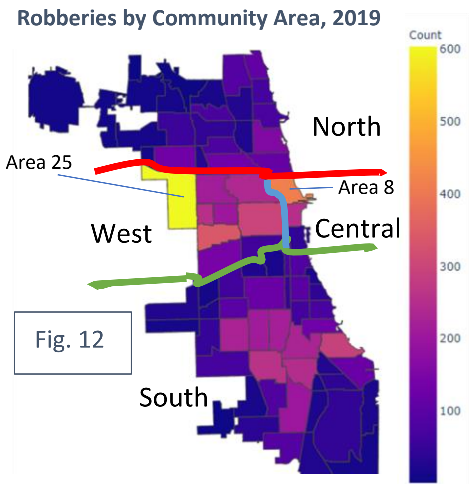


Fig. 11

(3) Spatial Patterns in Chicago Robbery

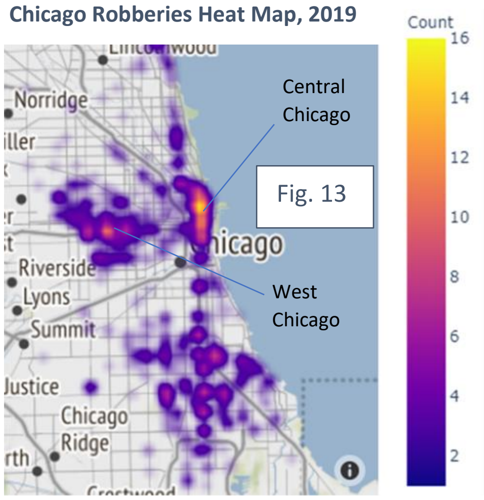
In this section we investigate spatial patterns beyond those depicted in Fig. 7-9. Here, we draw both a choropleth and a heat map to illustrate spatial patterns of robbery. The choropleth allows us to see clear contrasts between general community areas with different crime rates, while the heat map elucidates exact coordinates with highest robbery counts and also shows the areas that experience little robbery crimes.

Fig. 12 indicates that the West sector of Chicago experiences the highest robbery count in 2019. In particular, Area 25, **or Austin**, has significantly higher robbery count than the rest of the community areas, being the sole region with >=600



Robberies by Community Area, 2019

Fig. 12

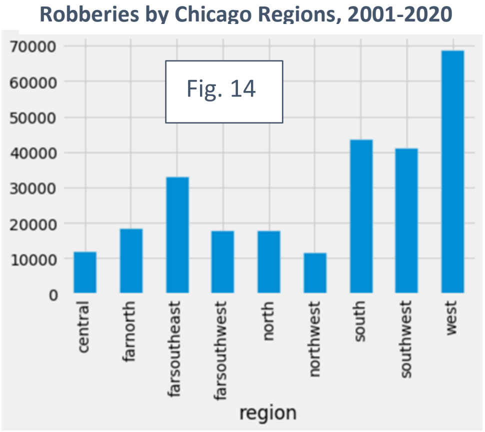


Chicago Robberies Heat Map, 2019

Fig. 13

robberies in 2019 (second highest being Area 8 with a count of 416), at a count of 604. **An investigation on Austin's demographics proves to be very telling on crime trends.** Austin is significantly different from other Chicago community areas: its median household income was $31,435, compared to $47,831 for Chicago (35% lower). Furthermore, Austin has a high proportion of young people, 30.5% under 19, and 20.8% from 20 to 34. Past studies have shown that the **2 most significant variables that influence crime rates are macro-economic factors and demographic factors**, that is, population within crime-prone years of 15-25 (*Predicting Crime*, Department of Justice Canada 2002 https://www.justice.gc.ca/eng/rp-pr/csj-sjc/jsp-sjp/rr02_7/rr02_7.pdf). Our investigation corroborates with the claims made in the report. Another distinguishing demographic of Austin is its racial makeup: 4.2% white, 84.2% black, 10.3% Hispanic. While it is imprudent to draw connections between the crime rate and racial distribution, some form of association could exist, considering the exceptionally high proportion of blacks in this Austin. Further analyses would still be required. Besides the West sector, the **South** sector seems to experience a fair amount of robberies, particularly in the center of the sector as seen in both Fig. 12 and 13. The Northern regions generally experience low robbery count in 2019.

Finally, Fig. 14 segregates the different regions defined as Chicago 'Sides' and depict the crimes from 2001-2020. Over the past 20 years, it is evident that in general, **the West and South 'Sides' experience the highest amount of robberies**. Further investigation should done to determine if there may need to be a heavier police presence in the West and South.

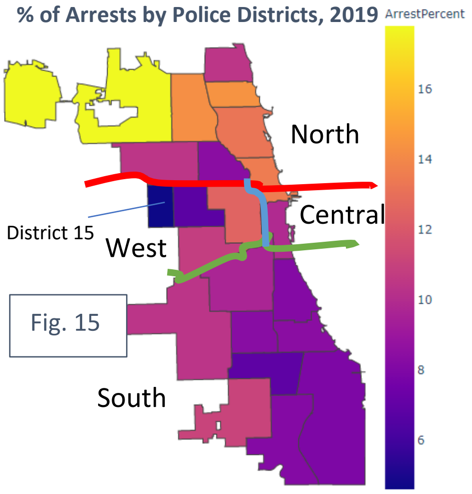


Robberies by Chicago Regions, 2001-2020

Fig. 14

(4) Patterns in Arrests Made

From *'robbery_geo_ward_comm_updated.csv'*, the **percentage of arrests out of total robbery crimes** are plotted in Fig. 15 as a choropleth, segregated by police district as defined in *'Boundaries - Police Districts (current).geojson'*, obtained from Chicago Data Portal. Evidently, in 2019, the North districts have relatively higher arrest percentages, while the South districts have in general lower arrest percentages.

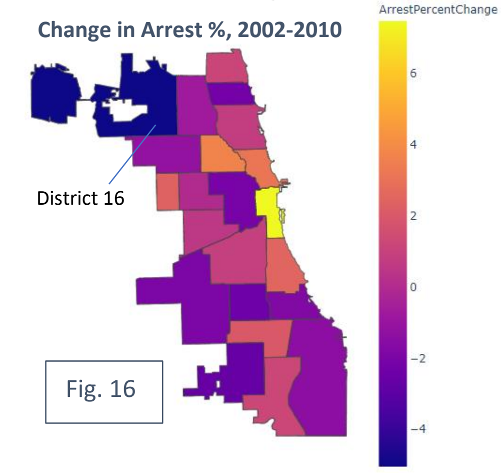


% of Arrests by Police Districts, 2019

Fig. 15

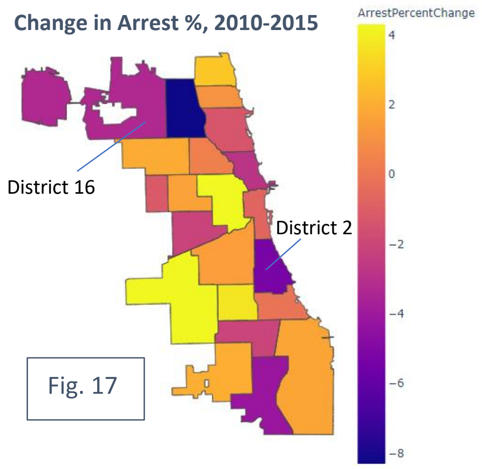
Temporal patterns in arrest % changes are plotted in Fig. 16-18, on the choropleths defined by the same districts. Notably, **District 16 (Jefferson Park) experienced the largest increase in arrest %** in the period 2010-2019 (Fig. 17, 18), resulting in the **highest arrest percentage in 2019 (Fig. 15)**. Possible reasons for this could include higher police efficiency with stronger leadership, but deeper study through communication with the Jefferson Park Police Department is required to thoroughly understand the reasons for this due to limited online resources. The lessons learnt could be used to help improve arrest % in other districts. Meanwhile, in **2019 District 15 (Austin) experienced the lowest % of arrests, at 4.5%.** This might be a factor for the crime rate in Austin as shown in Fig. 12.

Comparing Fig. 16-18 with Fig. 7-9, there **appears to be some association, albeit small, between change in arrest percentage and change in robbery count**. For 2015-2019, District 16 observed an arrest increase of +10.1%, while the robbery count fell across all community areas in District 16 (Fig. 9). However, for 2010-2015 in District 2 (Fig. 17), while arrests changed by -5.3%, community areas 42, 40, 38, 35 (Fig. 8) which lie in District 2 all experienced decreases in robbery. The association is hence unclear.



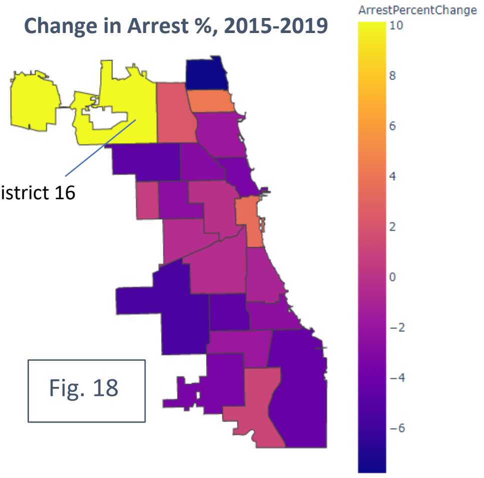
Change in Arrest %, 2002-2010

Fig. 16



Change in Arrest %, 2010-2015

Fig. 17



Change in Arrest %, 2015-2019

Fig. 18

(5) Hyde Park February 2020 Robberies Prediction

Fig. 20

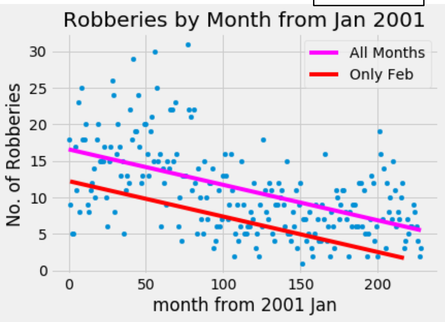
$$y = b_0 + b_1 x + b_2 D$$

b_0 = 16.58153377
b_1 = -0.0485592
b_2 = -4.3412148

Earlier analyses suggest that **as time progresses robbery count in Hyde Park has decreased** (Fig. 10), and that **February robbery count is the lowest** among all months (Fig. 11). These 2 variables are hence taken as regressors to construct a **multivariable linear regression model** as defined in Fig. 20, where x is the number of months from Jan 2001 (with Jan 2001 = 0, Feb 2001 = 1, … Jan 2020 = 228, Feb 2020 = 229). D represents the dummy variable for whether or not the month is February, with *D=1 if it is Feb* and *D=0 if not*. Notably, **arrest % is not taken as a regressor** because of 3 reasons: (1) the association is unclear, as discussed in section 4; (2) there is no arrest % data for Feb 2020 (3) there is some form of collinearity between variable x and arrest % that affects the prediction.

Using the CG method to minimize residuals, the b_0, b_1, b_2 **values are obtained as in Fig. 20**. Some **assumptions** made in this multiple linear regression are: (1) it follows a linear model, (2) 2020 Feb follows the previous monthly patterns (because of extrapolation), (3) February has lower crime rates, (4) there are no other factors that affect robbery count. The resulting linear regression graph along with original scatter points are plotted in Fig. 21. Using cross validation, the **mean error is 3.08, or 54%, far less** than the mean error of 4.71 crimes if the dummy variable for February were not included. In addition, a *95% confidence interval* for slope and intercept are calculated via 5000 bootstraps, which are [-0.0584, -0.0391] and [15.1, 18.1] respectively. Accordingly, with this model, we obtain a **prediction of 1.12 robberies for Feb 2020**, with 95% confidence interval of [-7.58, 12.1].

Fig. 21



(6) Proposal: Key Locations to Reduce Robbery Count

Fig. 22, Robbery Locations 2019

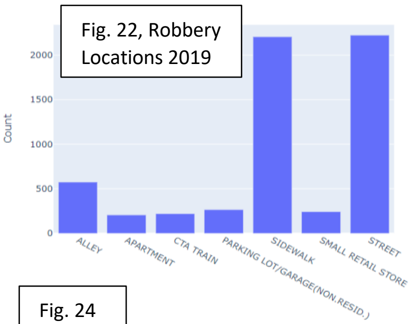


Fig. 23, Zoomed-In Heat Map, Robberies 2019

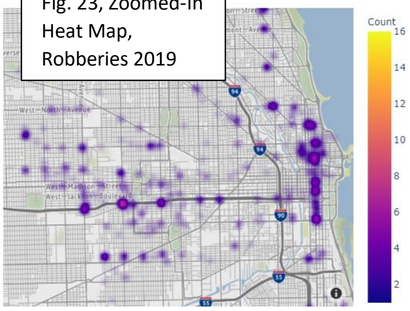
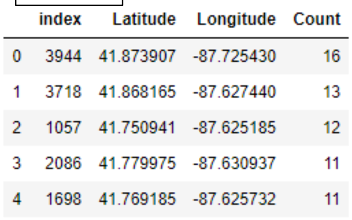


Fig. 24



	index	Latitude	Longitude	Count
0	3944	41.873907	-87.725430	16
1	3718	41.868165	-87.627440	13
2	1057	41.750941	-87.625185	12
3	2086	41.779975	-87.630937	11
4	1698	41.769185	-87.625732	11

Fig. 25

Rank	Robbery Count 2019	Address
1	16	530, South Pulaski Road, West Garfield Park
2	13	1133, South State Street, Printer's Row, Loop
3	12	79th, 15, West 79th Street, Chatham Fields, Chatham
4	11	63rd, 220, West 63rd Street, Becks Park, Englewood
5	11	69th, 15, West 69th Street, Park Manor, Greater Grand Crossing

In this section, we explore the implications of the "Location Description" column from *'robbery_ geo_ward_comm_updated.csv'* and propose intriguing ways to reduce robberies. Fig. 22 represents the grouped results of different location descriptions where robberies had occurred: clearly, **sidewalks and streets experience the most**. This notion is further supported by zooming into the Heat Map first seen in Fig. 13, near the hotspots; this zoomed in version is provided in Fig. 23. It is apparent from Fig. 23 that the robbery crimes occur mostly along street junctions, and certain junctions suffer from an especially high rate of robbery crimes. To better understand this, the hotspots with >= 10 robberies in 2019 are identified and tabulated in Fig. 24. To obtain the respective addresses, geopy library is used to conduct *reverse geocoding*, generating Fig. 25. Viewing these addresses on Google Maps generates interesting insights**: all 5 addresses are right beside a train station.** They are near the tracks or under the bridge of the tracks. Certain implications are thus clear. First, **there needs to be greater police presence at these respective addresses**. Next, it appears that the **vicinity of a metro station are hotspots for robberies**. Especially at night, where Fig. 6 has shown that robberies are more frequent, **policemen should be stationed near the station** if possible, in order to prevent such robberies.

Conclusion

With thorough data cleaning and recovery, missing latitude, longitude, community area and ward data are successfully recovered, providing more complete data for accurate analyses subsequently. The data visualized prove insightful. Encouragingly, crime rate has been decreasing over the past 20 years. *6 key findings are proposed, along with the corresponding recommendations*:

1) Peak periods of robbery, arson, and MVT crimes are identified. Arson is especially different in day versus night, with most cases in the night. Based on the results from Fig. 6, **police presence and readiness could be adjusted according to the respective hours**.
2) South and West regions of Chicago have **generally higher rates** of crime. **North** community areas tend to have **lower robbery** instances. Considering the **much higher arrest % in Northern districts** (Fig. 15), Chicago Police Department should consider **re-balancing the manpower ratio** from the Northern districts to other districts in Chicago.
3) Careful study of robbery count changes in ***Austin*** proves revealing. The high rates of robbery along with low incomes and young-skewed demographic further prove the **strong association between age demographics, economic hardship and crime rate**.
4) There is **no definite association between arrests made and reductions in robbery**. More study is required on this front.
5) A multivariable linear regression is used to predict the number of robberies in February 2020, at **1.12**.
6) It has been discovered that **streets and sidewalks are particularly prone to robberies**. This study goes further to identify the **specific streets and junctions** where these robberies occur most often and notes how each of these locations are *close to train stations*. Police forces should be deployed more heavily in these areas, and following the key finding (2), special care should be taken from **8pm-12am** where robberies are most frequent (Fig. 6).