2020-06-03

Predicting crime rate in Chicago and Identify the factors that cause crime

CAPP 30254 Machine Learning

Final Project Presentation

Team Members

01 Ruize Liu

02 Yimeng Qiu

03 Wenjun Shi

04 Jiaqi Yang

Agenda

1 Problem Foundation

Problem

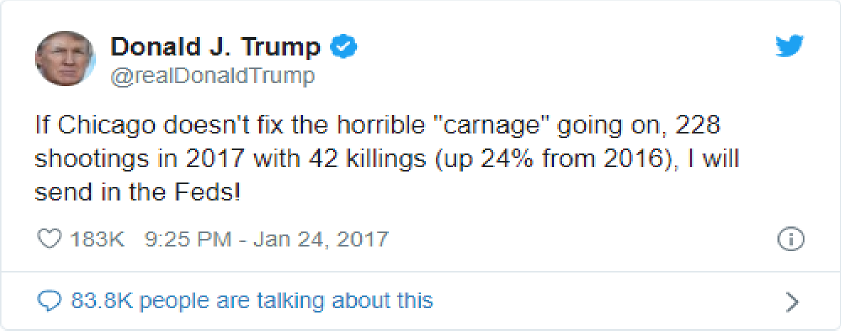
Illinois and Chicago have a highest homicide rate among States and Metropolitans with similar demographic features. We want to explore reasons behind this from the perspective of city infrastructure.

Goal

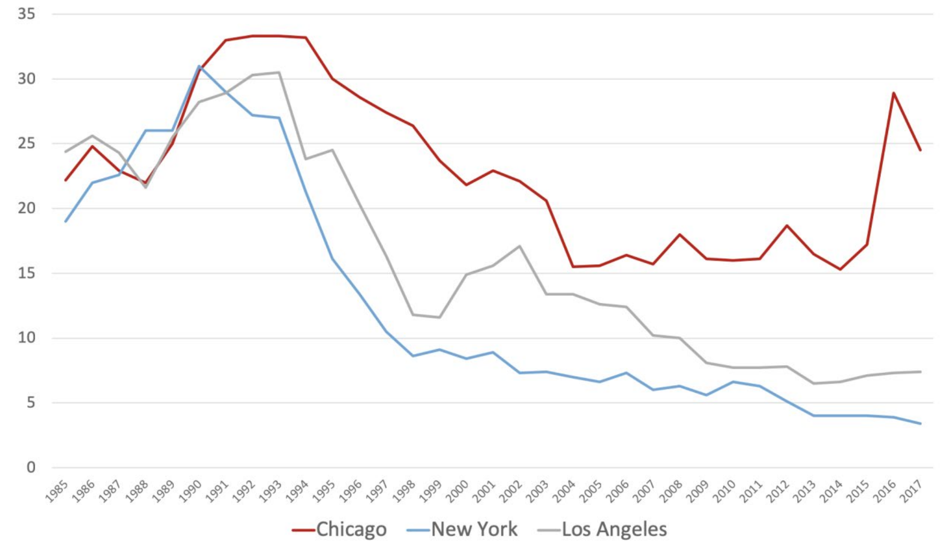
Identify key features beside traditional demographic factors that result in higher crime rate in metropolitan areas.

Potential Policy

Provide recommendations for new urban renewal projects that could revitalize less developed blocks and reduce crime rate simultaneously.



Chicago has a much higher number of homicides per 100,000 population than New York and Los Angeles.



Source: Uniform Crime Reporting data compiled by John Hagedorn.

2 Data Resources

Overview

Analytical Data

Crime data from FBI

State-level open dataset including 2011~2018

City of Chicago Open Data

Infrastructure and Point of Interest (POI) data

Census Data

Demographic features including median income, race, age, and gender

Supportive Data

Gentrification in America Report

Contains ten Chicago-like cities

State-County-City Match

Matching city with county, state and FIPS code

Property Value

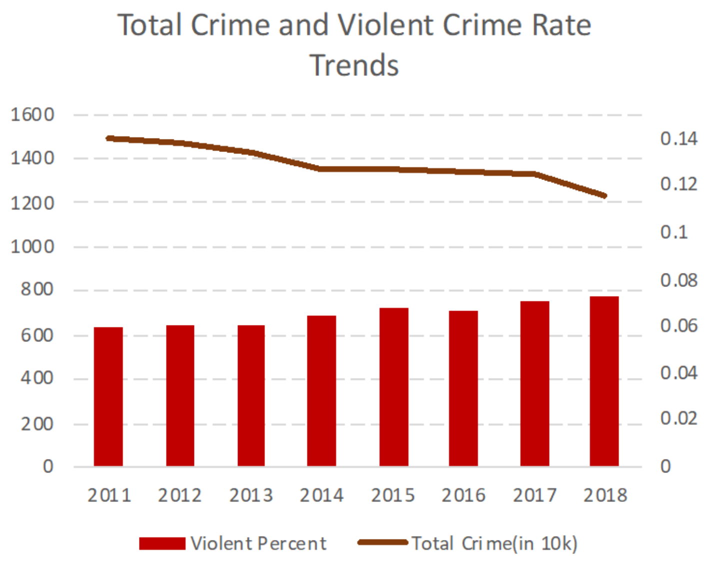
State-level median property value

2 Data Visualization

Crime Rate and Historical Trends at State Level

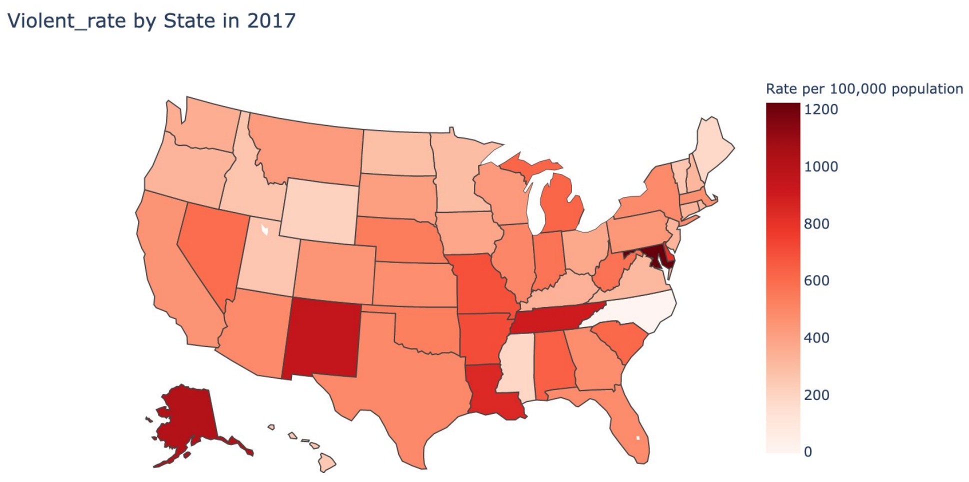
Total Crime

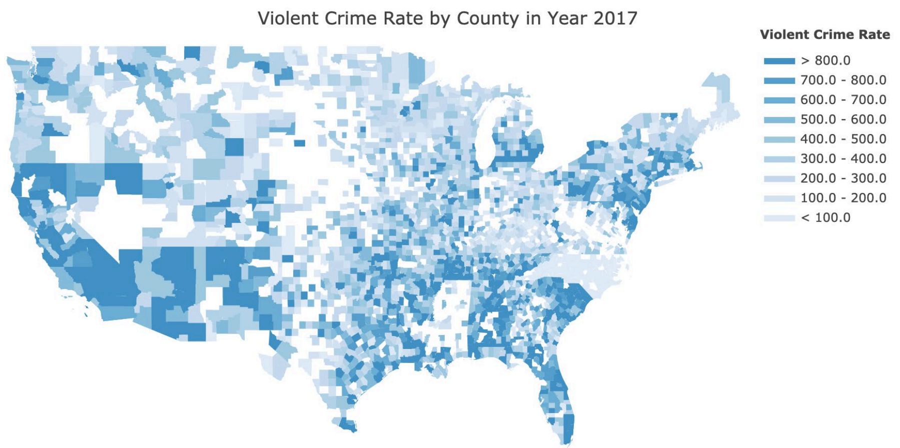
Total number of crime within US has been decreasing steadily in the past decade, but the percent of violent crime is increasing.



Break down into State and County level, we found there're great variations among States and Counties.

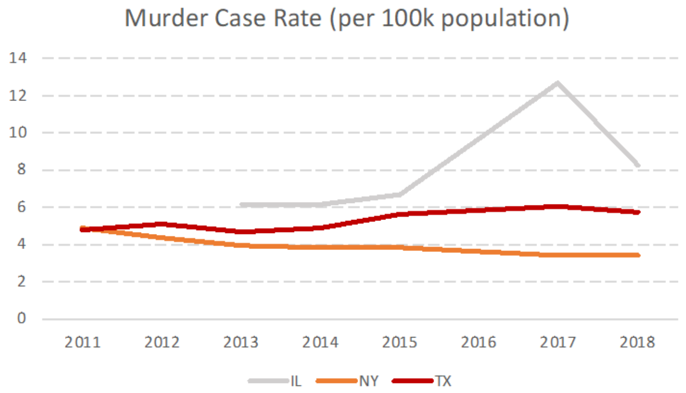
Violent Crime within Total Crime

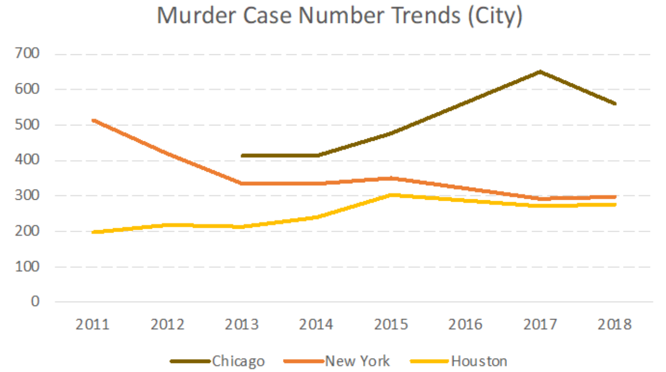




Murder Cases within Violent Crime

Illinois and Chicago has a remarkably high rate and absolute number in murder cases compared with NY and TX.





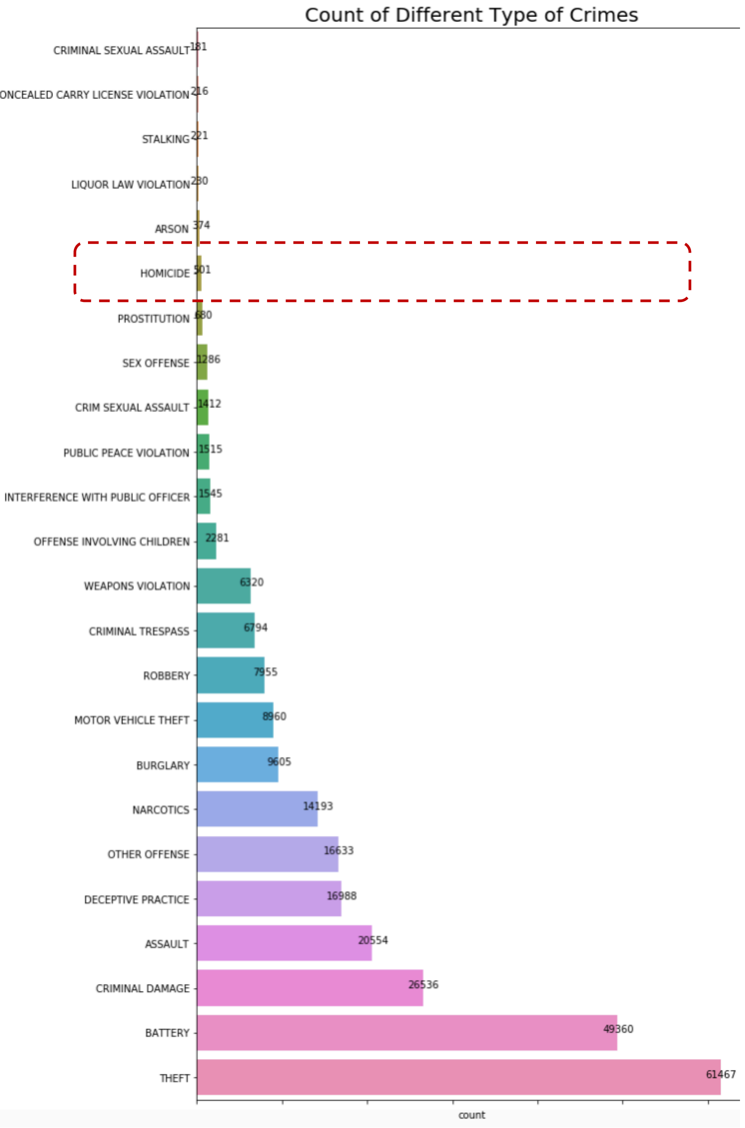
- We use 2017 FBI Crime data because data of a substantial portion of county breakdown data in 2018 & 2019 is missing
- Data Source: Census API, FBI Crime

CHICAGO DATA PORTAL

THE UNIVERSITY OF CHICAGO Harris Public Policy

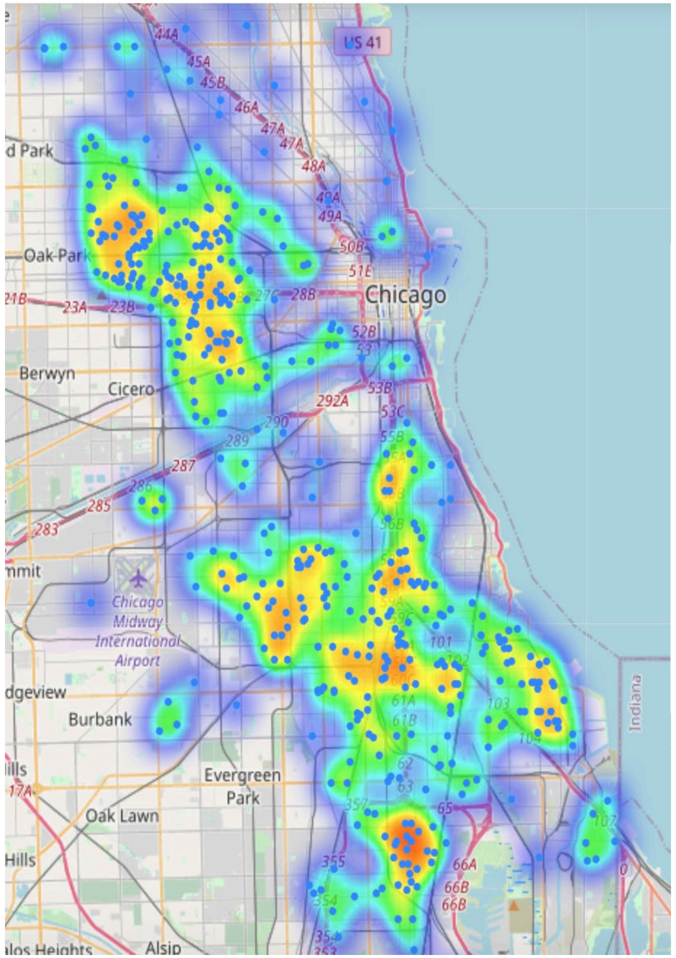
2 Data Visualization

Crime Rate, Trend and Infrastructure in Chicago



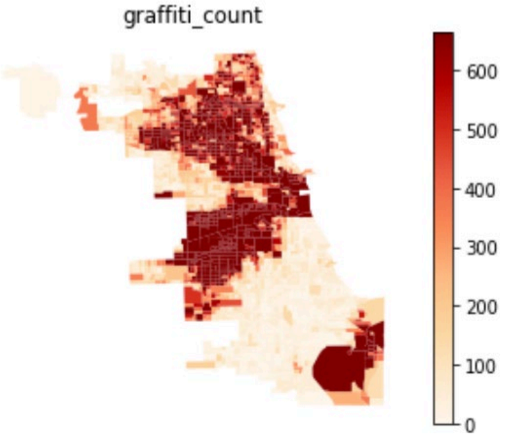
Target Variable – Homicide

Heatmap of Homicide Cases in 2019

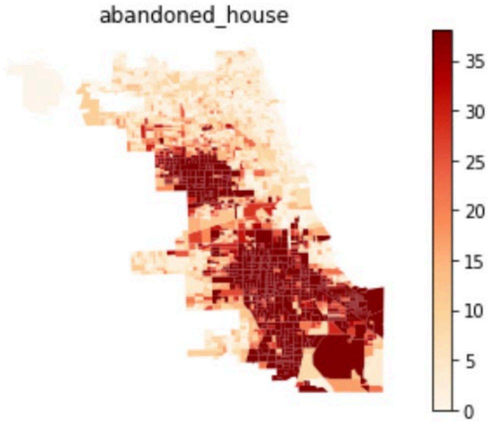


Feature Variables – Infrastructure

8 variables in total: total units of affordable housing by block, count of grocery stores, count of library, count of public arts, count of police station, count of abandoned houses, count of graffiti, and daily sum of traffic flow



count	2194.000000
mean	479.293984
std	707.085165
min	0.000000
25%	48.000000
50%	182.000000
75%	663.750000
max	6642.000000



count	2194.000000
mean	29.636281
std	43.480852
min	0.000000
25%	3.000000
50%	11.000000
75%	38.000000
max	343.000000

- Data Source: Census API, FBI Crime, ACS API

CHICAGO DATA PORTAL

THE UNIVERSITY OF CHICAGO Harris Public Policy

3 Machine Learning Methods

Overview

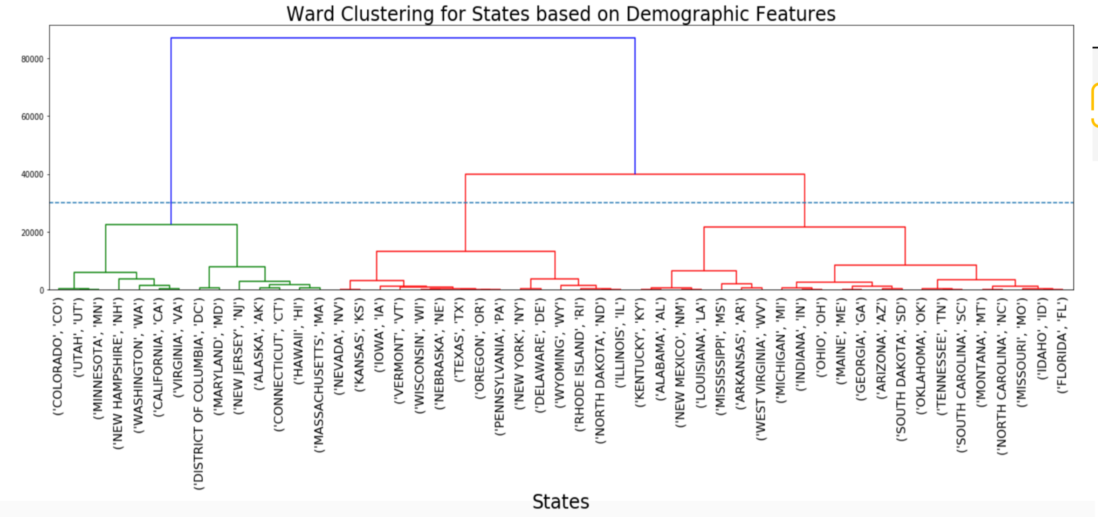
	Unsupervised	Supervised	Synthetic Control
Purpose	Select States with similar demographic features with Illinois	Analyzing relationships between infrastructure features and homicide rate within a block under scope of Chicago	Combining results from unsupervised and supervised learning, implementing a possible policy and testing outcomes
Data Source	• Unit: State • Source: Census API • Variable: Median Income, Race, Age	• Unit: Block • Source: City of Chicago API • Variable: 8 in total, including number of abandoned house, graffiti, sum of traffic flow…	• Unit: City • Source: ACS API
Algorithms	1. Hierarchical Clustering - Ward Method 2. K-means - Divide into 3 groups	1. Linear Regression 2. Decision Tree (to be implemented)	Synthetic Methods as described later
Validation	Use two classification methods and select States in the same group with IL in both ways	1. Training and Testing dataset 2. Bootstrapping 3. Regularization	Nonparametric permutation test

3 Unsupervised

Hierarchical Clustering and K-means

Hierarchical Clustering

- Methods: Ward
- Advantage: flexible group size and number of group
- Disadvantage: sensitive to outliers



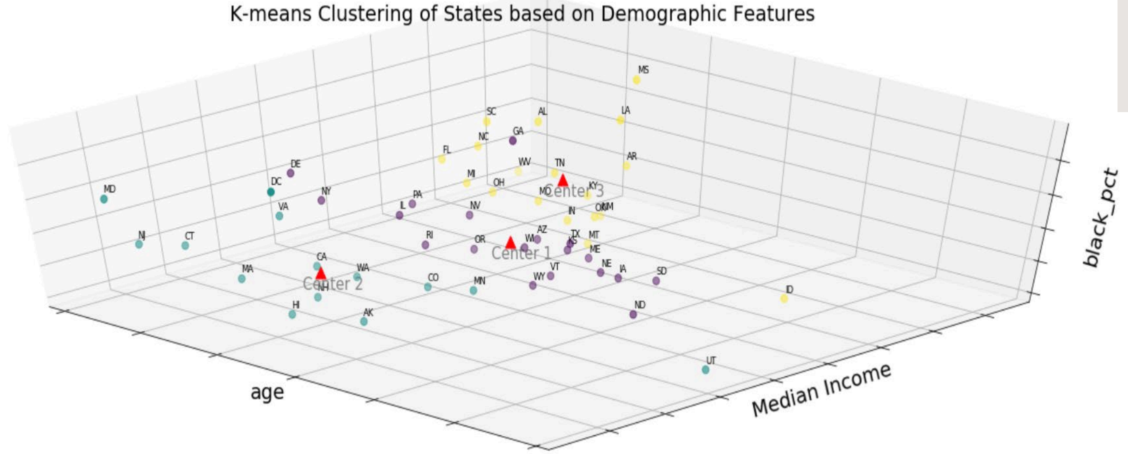
group	state_id
1	HI, DC, CA, MA, WA, CO, UT, NJ, AK, MD, NH, VA, CT, MN
2	OR, NY, NV, RI, DE, WY, VT, ND, TX, WI, IL, PA, NE, IA, KS
3	ID, AZ, MT, FL, ME, NM, SD, GA, NC, TN, SC, MO, MI, IN, KY, LA, OH, AL, MS, AR, OK, WV

After applying two methods, we cross search in the two groups IL belonging to, and find 14 States fall into the same group as IL in both ways:

TX, NV, WI, NE, PA, IA, ND, KS, OR, RI, WY, NY, DE, VT

K-means

- Advantage: less computational complexity
- Python function: KMeans()



group	state_id
0	VT, AZ, DE, WY, ND, RI, TX, NV, NY, SD, GA, OR, WI, IL, PA, NE, IA, KS, ME
1	MN, HI, NJ, CT, VA, MA, CA, NH, MD, AK, WA, UT, DC, CO
2	OH, FL, AL, MS, AR, LA, KY, TN, MI, MO, SC, MT, NC, NM, OK, ID, IN, WV

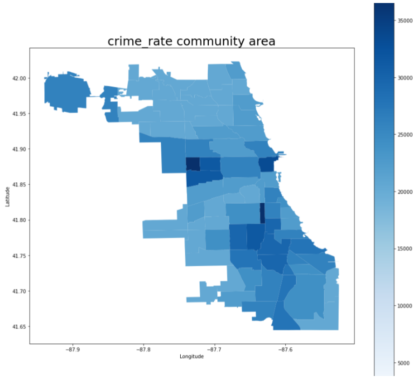
- Data Source: Census API
- For better visualization, we plot only three variables in K-means, but all are used in actual model.

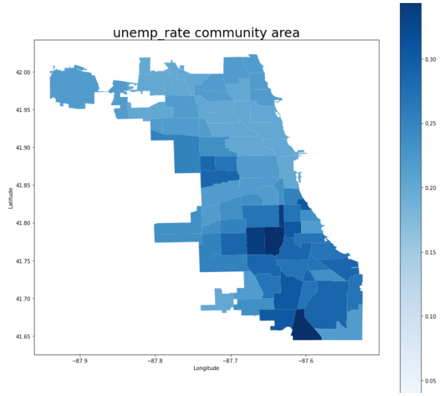
Linear Regression – Demographic Features

We run regression on two sets of features separately: demographic features and infrastructure features

Regression Results

We first tried demographic features and found they're indeed positively correlated with homicide rate, but cannot fully explain difference between blocks.

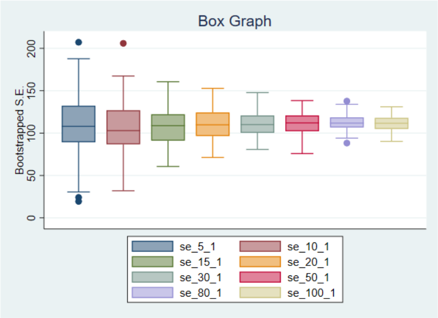


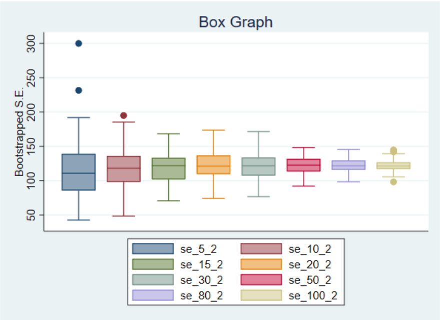


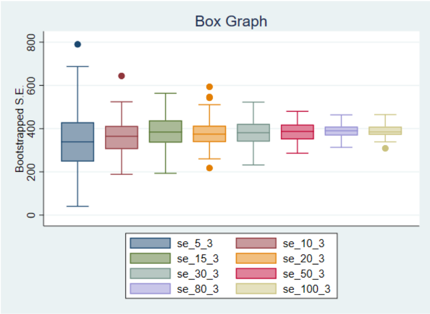
crime_rate	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
black_pct	1281.51	113.2512	11.32	0.000	1059.419	1503.602
white_pct	-363.6456	123.3259	-2.95	0.003	-605.4943	-121.797
unemp_pct	481.9305	386.4257	1.25	0.212	-275.871	1239.732
bachelor_pct	360.8975	206.1161	1.75	0.080	-43.30706	765.1021
median_income	-4.12e-07	1.83e-07	-2.25	0.024	-7.70e-07	-5.31e-08
_cons	688.3121	91.64183	7.51	0.000	508.5976	868.0266

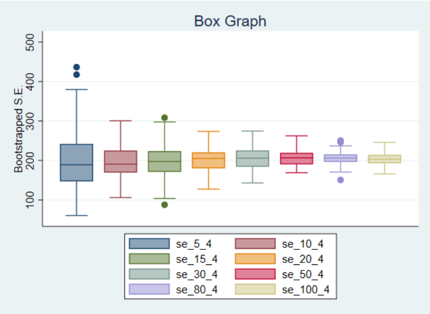
Bootstrapping for Validation

From the Box graph, clearly the mean and standard deviation for the estimated standard errors become stable as we increase the number of bootstrapped samples. It looks like the outliers are suppress after we have sufficiently large number of samples.







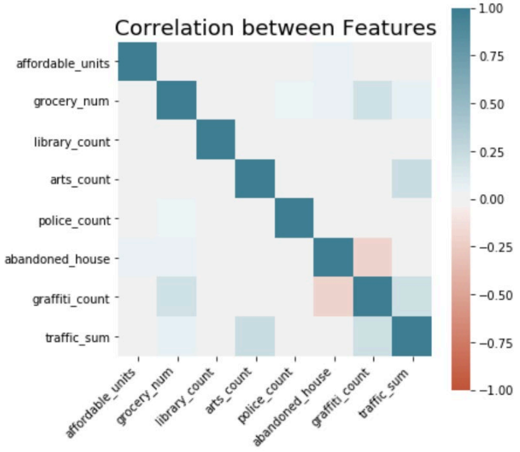


3 Supervised

Linear Regression – Infrastructure Features

Feature Variable Check

We use eight variables in linear regression. Before running model, we conduct sanity check and compute correlation, in order to eliminate error and multicollinearity.



	affordable_units	grocery_num	library_count	arts_count	police_count	abandoned_house	graffiti_count	traffic_sum
count	2,118.00	2,118.00	2,118.00	2,118.00	2,118.00	2,118.00	2,118.00	2,118.00
mean	10.73	0.23	0.04	0.09	0.01	30.63	492.52	11,400.19
std	43.41	0.52	0.19	0.96	0.10	43.93	714.87	26,273.37
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00	0.00	4.00	51.25	0.00
50%	0.00	0.00	0.00	0.00	0.00	12.00	198.00	0.00
75%	0.00	0.00	0.00	0.00	0.00	39.00	682.00	16,500.00
max	648.00	5.00	1.00	27.00	1.00	343.00	6,642.00	537,500.00

Model Construction Flow

Random Split Dataset: 80% Training + 20% Testing

Train Model and Calculate Scores

Apply Model to Test Set Calculate Testing Scores

Repeat this process (# of trials) for multiple times for a stable results

Results and Insights

The model achieves stable R2 around 0.27 after trial number increased above 60.

```
With 100 trials:
 Training Data Stat:
    Mean R2 = 0.2668
    Mean Variance Score = 0.2668
  Testing Data Stat:
    Mean MSE = 11672088.7888
    Mean MAE = 1849.8447
    Mean R2 = 0.2571
```

Interesting Findings:

1. Infrastructure plays an important role in crime rate and thus large urban renewal project can reduce crime in less developed blocks.
2. Compared with library or arts, renovating abandoned houses and removing graffiti are more effective methods

CHICAGO DATA PORTAL

THE UNIVERSITY OF CHICAGO Harris Public Policy

3 Synthetic Control Method

(Plan to implement in next week)

Methodology

Synthetic control method is "matching" based on variables in pre-treatment periods, including outcome variables. Which is specified using a factor model:

$$
\begin{aligned}
Y_{jt} &= \alpha_{jt} D_{jt} + Y_{jt}^{N} \\
&= \alpha_{jt} D_{jt} + (\delta_t + \theta_t \mathbf{Z}_j + \lambda_t \mu_j + \varepsilon_{jt})
\end{aligned}
$$

Using these variables, we find "optimal weight" for each control unit. Then, we can obtain a synthetic control estimate:

$$\tau^{s} = Y^{T} - \sum_{j=2}^{J+1} w_j^{*} Y_j^{C}$$

Key Reference:

Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californias Tobacco Control Program

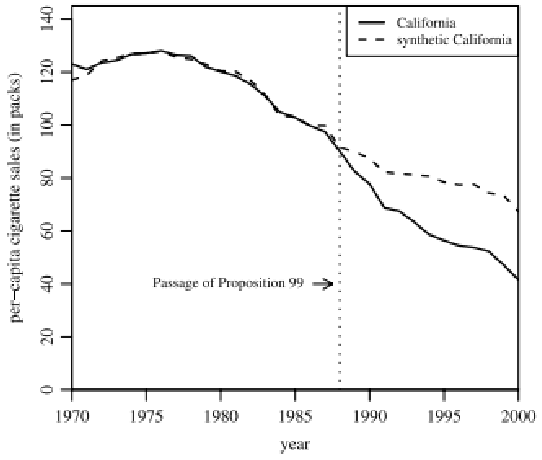
(Abadie, Diamond, & Hainmueller, 2010)

Expected Outcomes

1. State weight in synthesis

Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0



2. Predicting change after implement the policy after a certain point

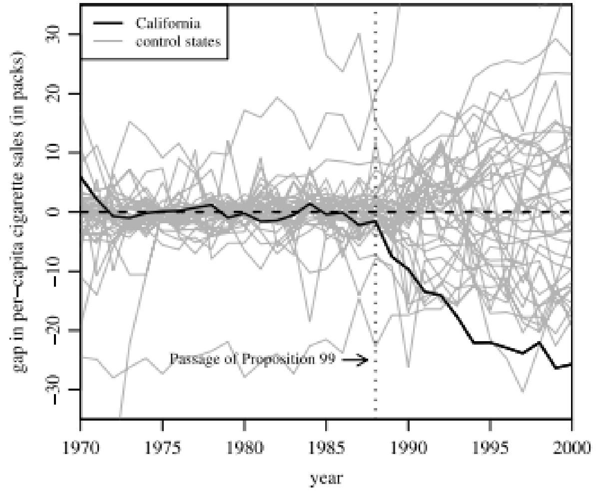


Figure 4. Per-capita cigarette sales gaps in California and placebo gaps in all 38 control states.

3. Testing performance using a nonparametric permutation test

CHICAGO DATA PORTAL

THE UNIVERSITY OF CHICAGO Harris Public Policy

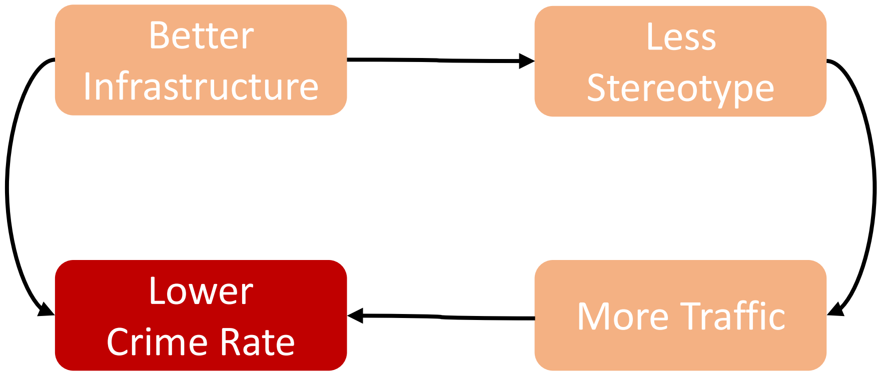
4 Summary and Policies

Identifying Key Features

- Instead of focusing on demographic features only, our model suggests paying more attention on infrastructure features
- Abandoned housing, graffiti, and traffic flow are the three key features that correlate with regional homicide rate
- Yet, number of libraries and number of public arts have less significant relationships with regional crime rate.

Policy Implications

- Based on our findings, we suggest policymakers to initiate large-scale urban renewal project to revitalize less developed blocks.
- Focus of such project should be paid on renovating outdated buildings, rather than building libraries or art institutes.
- Positively publicizing those blocks and attracting traffic also helps



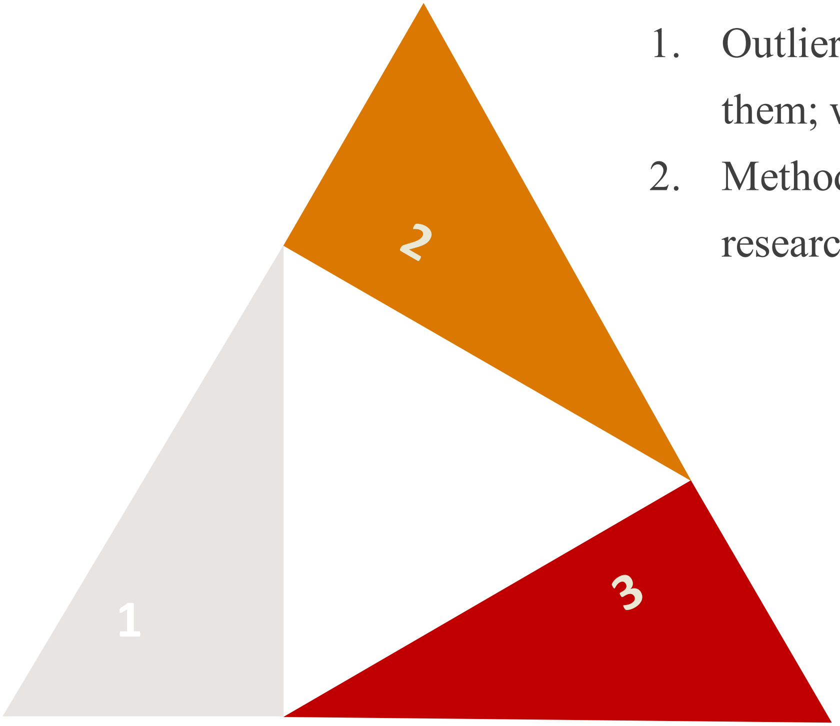
Research Plan for Next Week

- Collect data from other cities within the listed similar states, and complete the Synthetic Control Method.
- Try more machine learning models in the Supervised part, including Decision Tree (after grouping crime rate into several categories) and so on.
- Develop a more comprehensive tool in Python for crime rate prediction with certain urban renewal project data input for policymakers

5 Limitations

Data Sources

1. Different cities provide different variables on Open Data, thus hard to incorporate into one large model, which leads to difficulties in Synthetic Control Part.
2. Missing data is a serious problem in FBI Crime Data, and that's why we give up on county-level classification



Machine Learning Applications

1. Outliers always exist and normalization did not eliminate them; we're still looking for better methods for them.
2. Methods have advantages and disadvantages; we're still researching in finding the most suitable algorithms

Prediction Power

1. Crime rate is a synthetic outcome of complex socioeconomic features, though we tried our best efforts, it's still impossible to capture all reasons for this issue.
2. Historical cannot guarantee future.

Thanks!

References

Abadie, A., Diamond, A., & Hainmueller, J. (2007). *Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californias Tobacco Control Program*. doi: 10.3386/t0335

Alves, L. G., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. Physica A: Statistical Mechanics and its Applications, 505, 435-443.

Fajnzylber, P., Lederman, D., & Loayza, N. (2002). What causes violent crime?. European economic review, 46(7), 1323-1357.

Gupta, S. D., & Garg, V. (2018). Crime Patterns and Prediction: A Data Mining and Machine Learning Approach.

Han, B., Cohen, D. A., Derose, K. P., Li, J., & Williamson, S. (2018). Violent crime and park use in low-income urban neighborhoods. American journal of preventive medicine, 54(3), 352-358.

Keyvanpour, M. R., Javideh, M., & Ebrahimi, M. R. (2011). Detecting and investigating crime by means of data mining: a general crime matching framework. Procedia Computer Science, 3, 872-880.

Nath, S. V. (2006, December). Crime pattern detection using data mining. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops (pp. 41-44). IEEE.

Stretesky, P. B., Schuck, A. M., & Hogan, M. J. (2004). Space matters: An analysis of poverty, poverty clustering, and violent crime. Justice Quarterly, 21(4), 817-841.