How does San Francisco Crime Change when Population Characteristics Change?

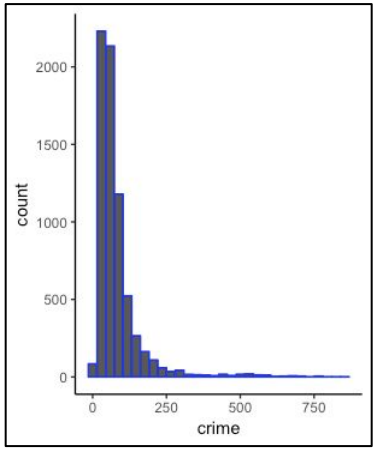
Andrea Bonilla, Colin Kou, Malvika Rajeev, Yulun Wu

Our objective is to fit and interpret models on major categories of crime.

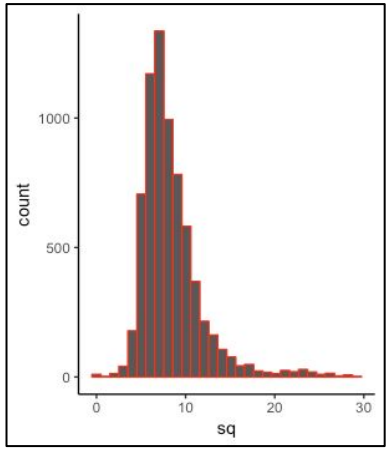
- **Goal** is to understand features that impact the following categories of crime
  - Assault
  - Robbery
  - Theft
  - Burglary
  - Vehicle Theft
- Models are trained with 2010-2016 data and evaluated with 2017 data

$$\sqrt{CrimeCounts_{region}} = f(\overrightarrow{X_{region}})$$

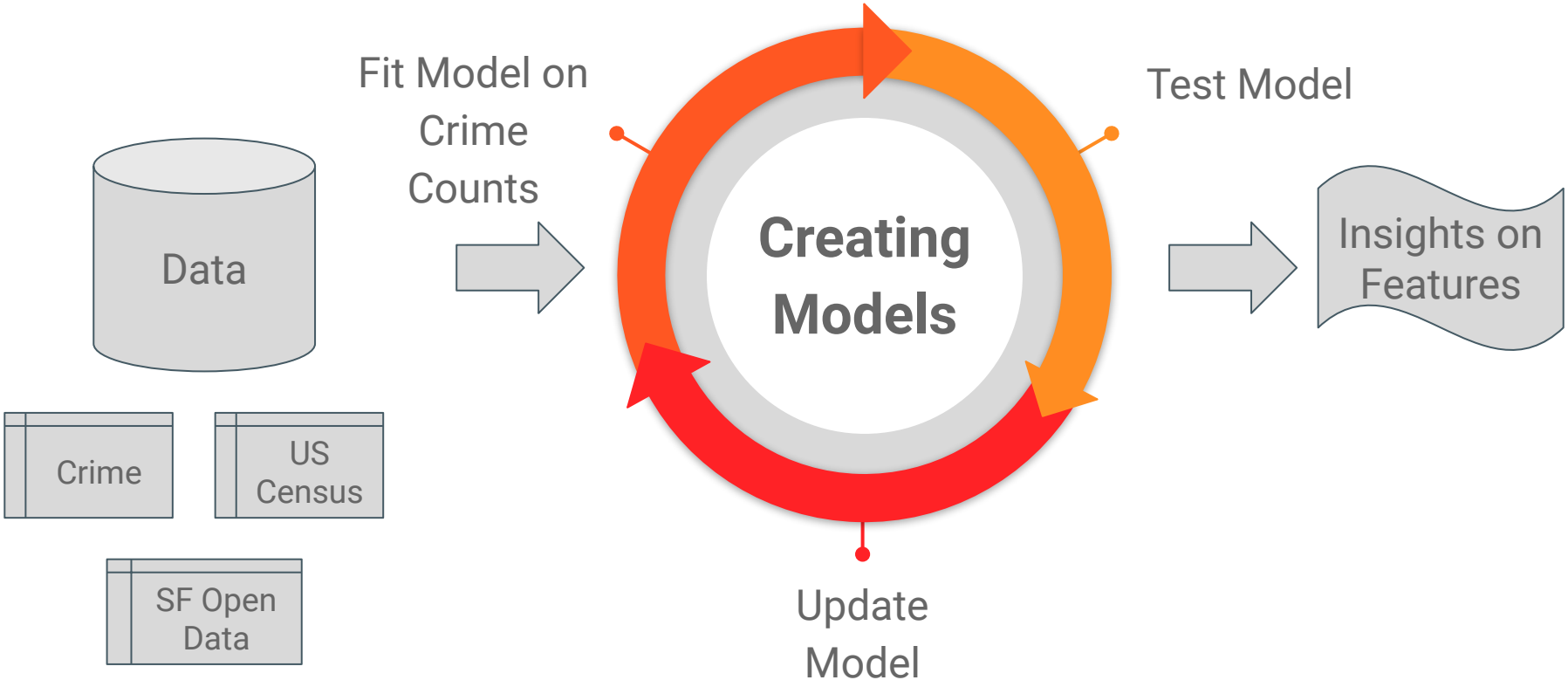
Counts



Square Root of Counts



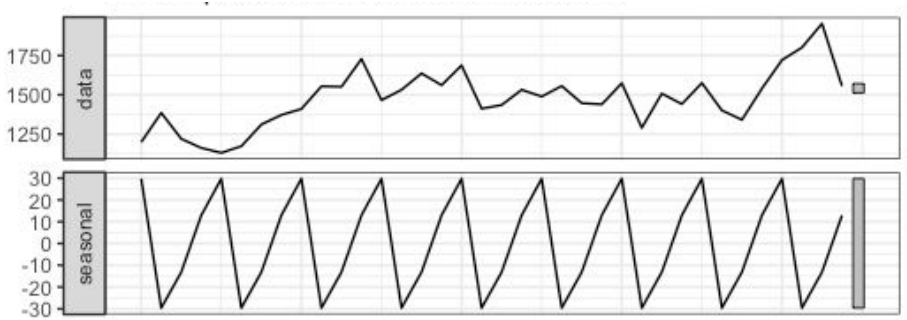
We identified models to understand feature impact on crime.



Simplifying Data

Chose to predict quarterly crime counts with additional spatial temporal data.

- We turn daily crime counts into quarterly crime counts to capture seasonality.



Following additions to census data improve the model performance

- Neighborhood Activity: 311 Calls
- Neighborhood Infrastructure Activity: Housing Inventory Data
- Number of Schools
- Number of Bus Stops

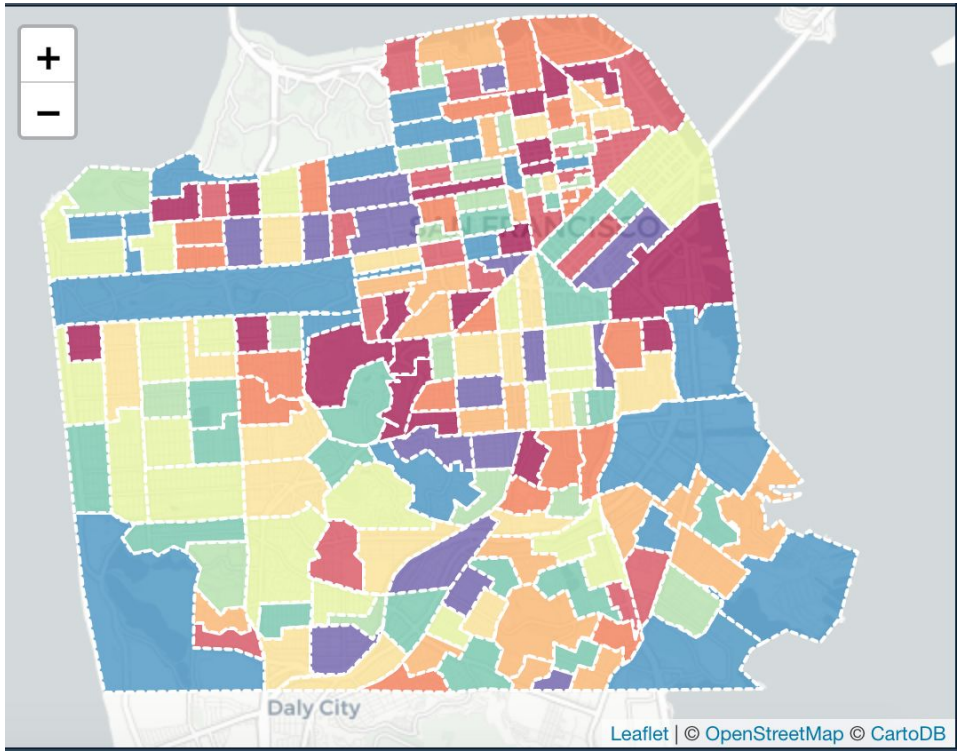
Use clustering to group similar Census Tracts

Grouping Census Tracts

1. Use Same-Sized K-Means to create 10 clusters
2. Develop individual models for each cluster of similar regions

195 ⇨ 10

Reason: tradeoff
data quantity vs. model quality



Reduce our feature space to address multicollinearity.

Feature Space Reduction

1. Use Hierarchical Clustering to create 20 clusters
2. Within each cluster, pick the highest correlated variable with crime count

189 ⇨ 20

Example: Variables in Cluster 8

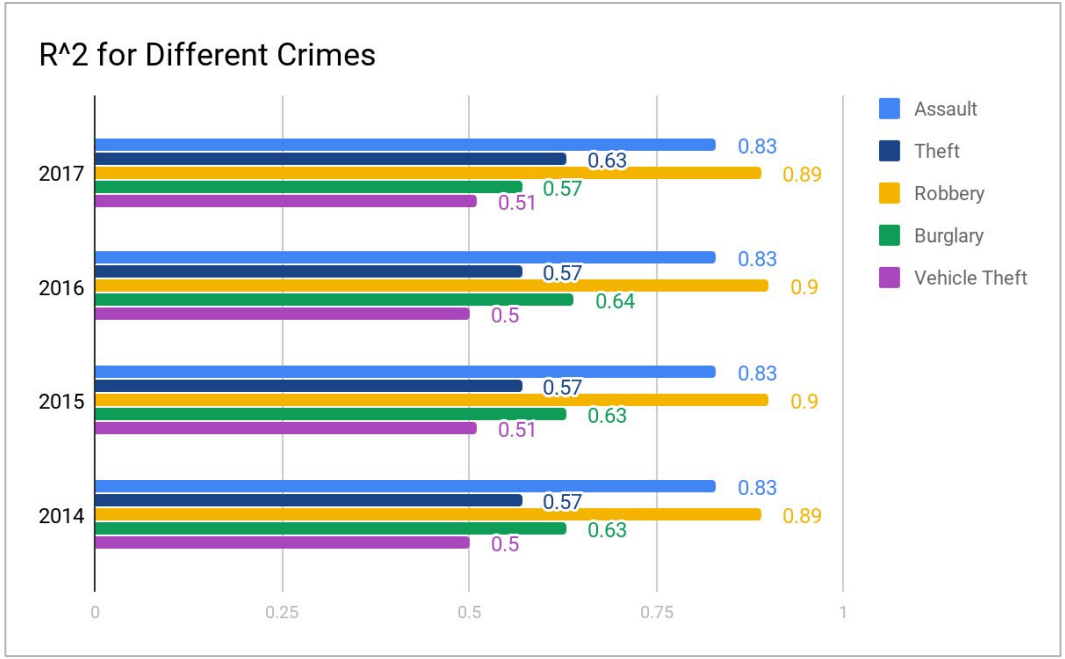
Female and Male Education

- Up to 4th grade,
- Up to 5th and 6th grade
- 9th grade
- no schooling completed,
- Less than High School
- Bachelor's degree or more

Models

Our baseline OLS model works considerably well for theft and assault crime categories.

- We use OLS as baseline model with our clustered variables
- Variable coefficients and their significances provide approximations of feature importance
- Vehicle theft is more rare and may be more of an opportunistic crime as opposed to theft



MSEs by Clusters

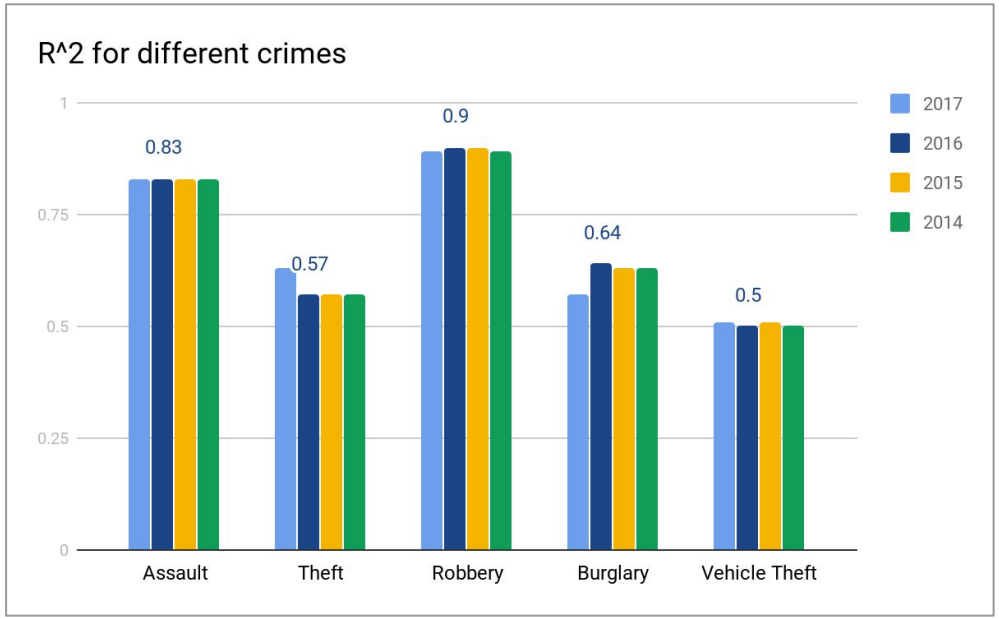
	Assault	Burglary	Theft	Robbery	Vehicle Theft
C1	0.848	**0.39**	4.036	0.667	0.953
C2	1.082	0.689	4.608	0.697	**0.561**
C3	1.361	0.692	10.611	0.777	0.925
C4	0.739	0.609	**2.68**	0.541	0.662
C5	1.505	0.708	4.704	0.768	0.682
C6	1.186	0.831	6.6	0.625	1.017
C7	1.066	0.499	3.543	0.636	0.681
C8	1.634	0.594	4.123	1.015	1.03
C9	**0.713**	0.395	2.916	**0.468**	0.801
C10	2.564	0.632	8.251	0.897	0.863

Crime-Wise Important Features

Assault	Burglary	Theft	Robbery	Vehicle Theft
Private School	SFCCD Schools	Housing Activity	Private School	Hispanic Population
Citizen Activity	Male Population	Transience of Female Population	Transience of Neighbourhood	Educated Male Population
Transience of Female Population	Transience of Female Population	Senior Citizens Male	Male 75 to 79 years	Under 18
Unemployed Poor Males	Renter Occupied Residences	Proportion Under Poverty	Area of Tract	Number of SFUSD schools
Unemployed Female	Uneducated Male Population	Neighborhood Activity	Number of SFUSD Schools	Renter Occupied Residences

Like our baseline model, Random Forests work considerably well for robbery and assault crime categories, and has overall better metrics.

- Our models have varying predictive explanatory power per category and has lower MSE compared to our baseline
- Our RF models have similar R^2 when performing time series cross validation.



MSEs provide hints into clusters.

	Assault	Burglary	Theft	Robbery	Vehicle Theft
C1	0.635	0.551	1.159	0.531	0.551
C2	0.557	0.582	1.212	0.526	0.598
C3	0.614	0.529	1.248	0.574	0.559
C4	0.576	0.536	**0.62**	0.547	0.554
C5	0.571	0.584	0.886	0.544	0.508
C6	**0.541**	0.551	0.833	0.494	0.54
C7	0.602	**0.519**	0.922	0.618	0.564
C8	0.615	0.572	0.853	0.6	0.531
C9	0.57	0.582	1.03	**0.474**	**0.517**
C10	**0.85**	**0.662**	**1.732**	**0.674**	**0.645**

Crime-wise feature importances aligns with criminology theory.

Assault	Burglary	Theft	Robbery	Vehicle Theft
Number of Schools	Proportion of Vacant Houses	Male Population	Number of Schools	Male Population
Number of Subway Stops	Owner Occupied Residences	Proportion of Vacant Houses	Male Population	Racial Index
Racial Index	Working Class population	Renter occupied residences	Proportion of Vacant Houses	Income earned by Males
Male Population under poverty	Male Population	Number of Subway Stops	Racial Index	Population in 35-64
Male Population	Area of Tract	Racial Index	Population under poverty	Area of Tract

Takeaways

We do not expect models to fully capture crime variability because crime is not nearly a scientific phenomenon.

Models Agree

- Busyness of Neighborhood and Poor Males are a high contributor to assault
- Renter Occupied Residences and Males are related to burglary
- Schools and males impact Robbery

Models Differ

- Each model for theft had different feature importances but share variables related to homes, males, and poverty.
- Each model for vehicle theft had different feature importances but share variables related to males

But adding spatial temporal data, improves our predictions. We inspected Clusters with high MSE

- Spatial features were included in feature spaces
- Cluster 10 had a sharp increase in males with no schooling completed, a lower number of schools, more people in 35-64 age group, more population density, and more vacant houses

Assault

- Mission Tract 176 (Most Crime)
  - MTA stops, prop male, schools, owner occupied residence, racial index
  - MSE 0.65
- Dogpatch Tract 226 (Highest median income in 2017)
  - MTA stops, schools, age index, racial index, 18-34
  - MSE: 0.56

A Note On Bias

Any statistical inference we makes on this data does not disguise the fact that the data itself has reporting bias.

- **Dark figure of crime:** We assume that crime incident reports act as stand in for crime in general though this is not true. Many crimes are unreported. Sexual assault cases is a primary example of that. Victim blaming and general trust in police influence self reporting, which accurately cannot be numerically captured.
- **Police bias:** We assume that police forces are acting unbiasedly, which is a major assumption that leads us to believe that all crime that is witnessed by police forces is recorded. Training models on biased historical data and having police focus on certain communities, will lead to even more arrests of minorities, but will not lead to solving the crime problem.

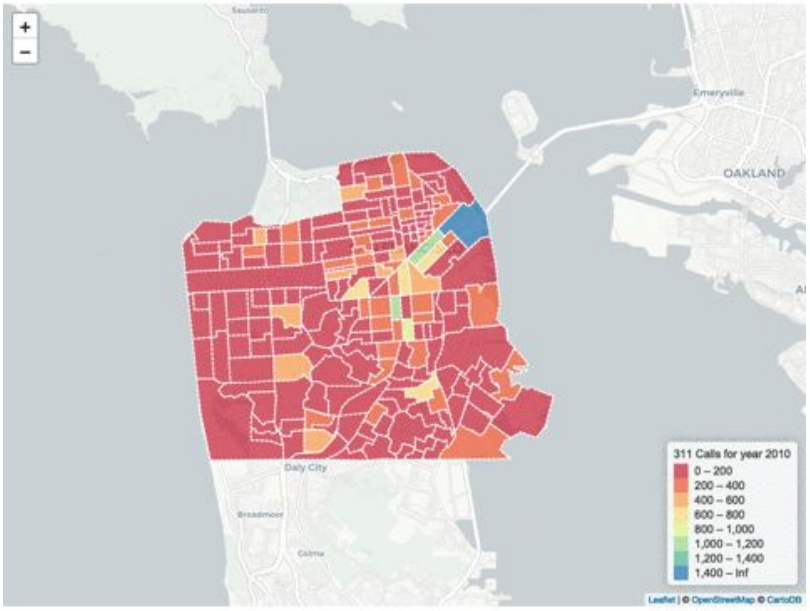
Social Stigmas + Distrust in Police = Unreported Crimes

Appendix


SAN FRANCISCO, CA
ADD COMPARISON
POPULATION
884,363
1.55% GROWTH
MEDIAN AGE
38.3
MEDIAN HOUSEHOLD INCOME
$110,816
6.76% GROWTH
POVERTY RATE
11.7%
NUMBER OF EMPLOYEES
534,557
4.54% GROWTH
MEDIAN PROPERTY VALUE
$1.1M
7.82% GROWTH

Incorporating additional spatial and temporal data will improve predictions.

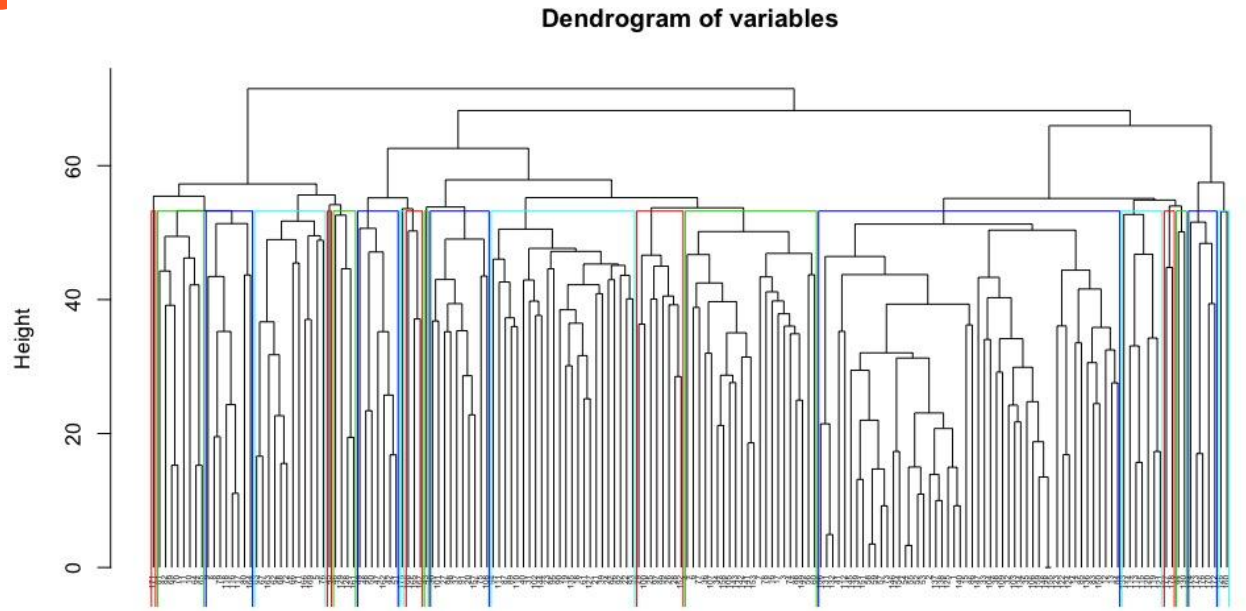
- A recent study has improved crime prediction by incorporating additional spatial temporal data, in addition to census variables



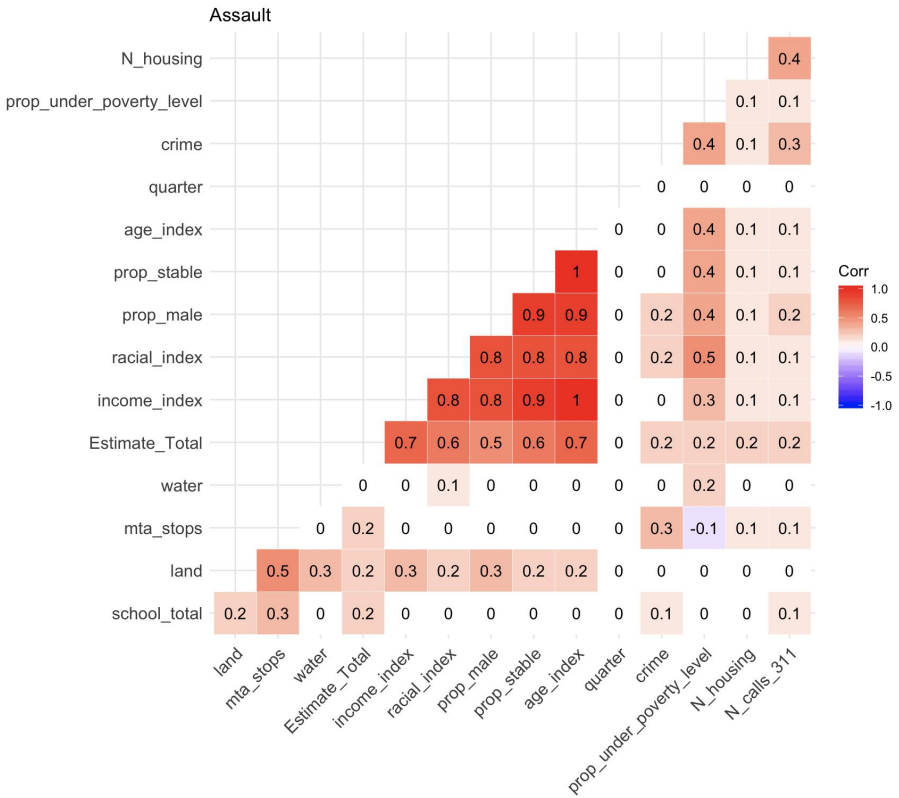
Variables in Our Models

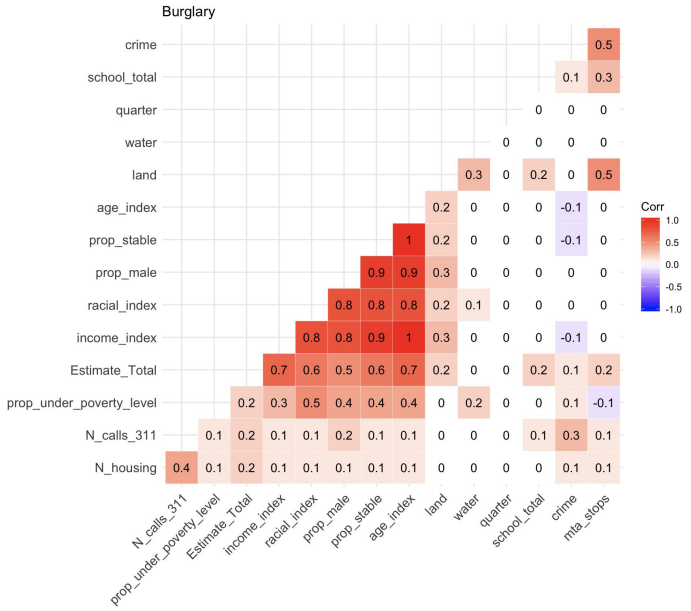
- Census Data
- Number of 311 Calls (Nbd activity)
- Number of Evictions
- Number of Schools
- Number of Bus Stops

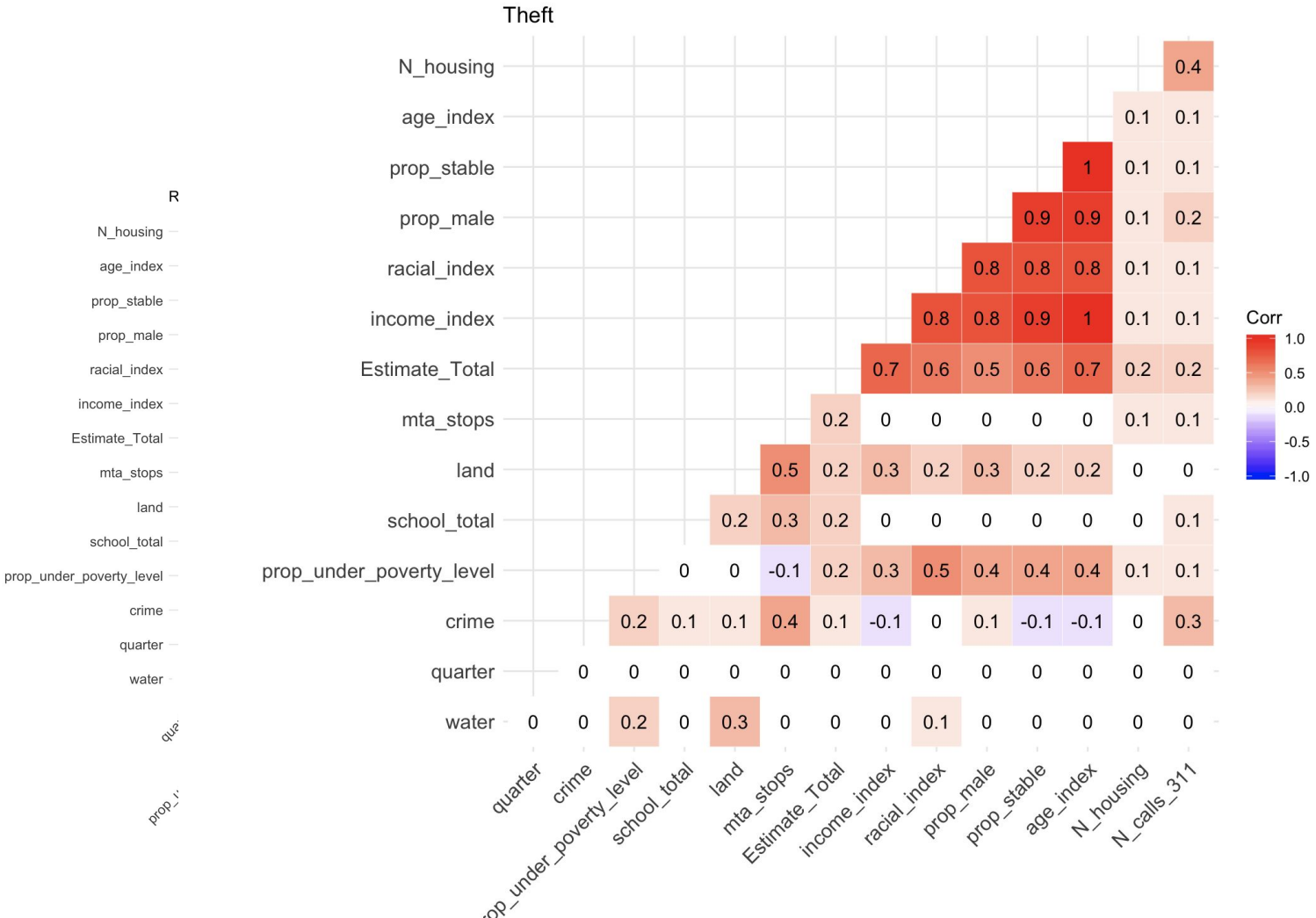
We used K-means and Hierarchical Clustering and correlation coefficients to reduce feature space in our mo

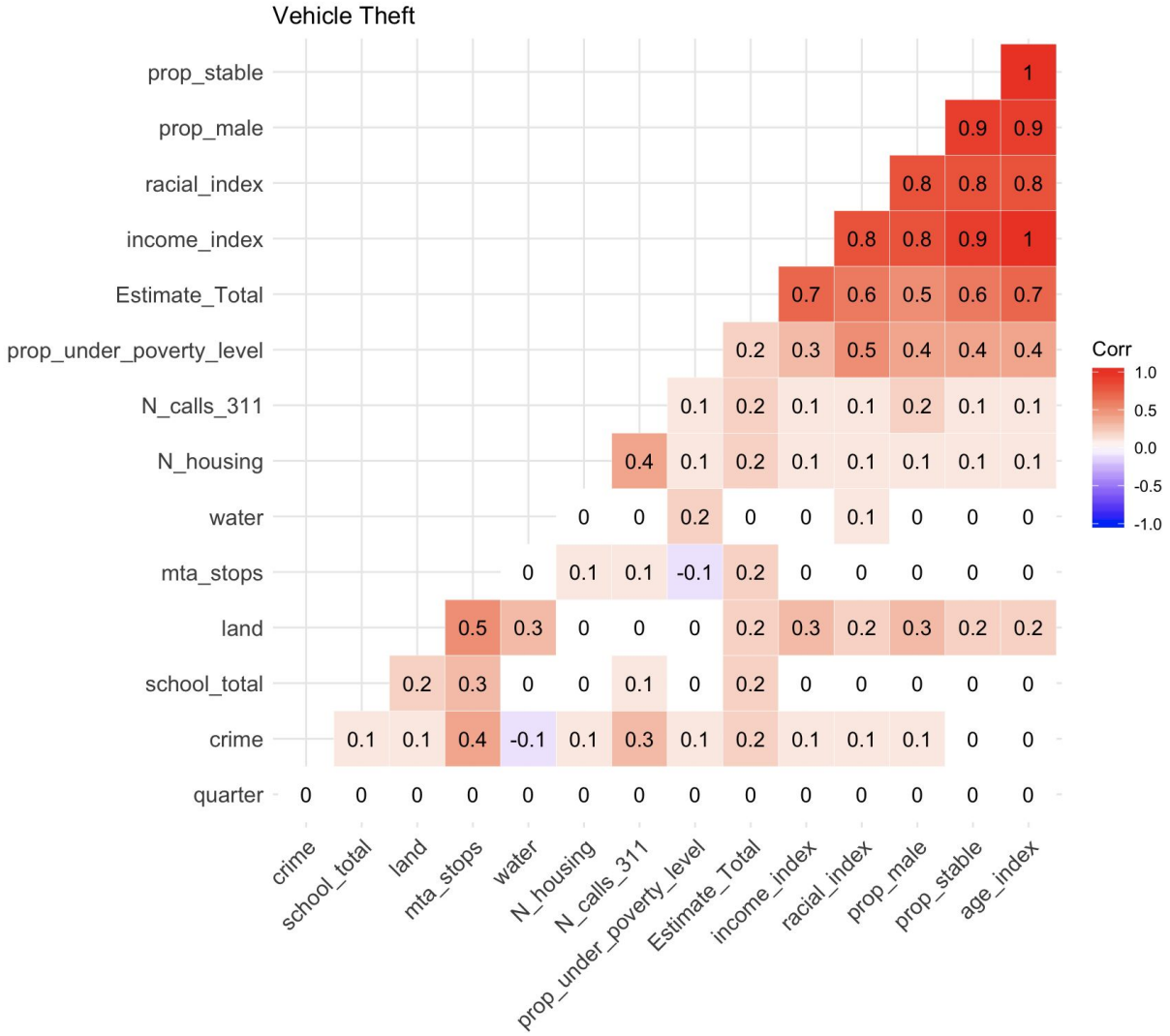


Correlation with Crime






Theft
N_housing
age_index
prop_stable
prop_male
racial_index
income_index
Estimate_Total
mta_stops
land
school_total
prop_under_poverty_level
crime
quarter
water
quarter
crime
prop_under_poverty_level
school_total
land
mta_stops
Estimate_Total
income_index
racial_index
prop_male
prop_stable
age_index
N_housing
N_calls_311
Corr
1.0
0.5
0.0
-0.5
-1.0


Vehicle Theft
prop_stable
prop_male
racial_index
income_index
Estimate_Total
prop_under_poverty_level
N_calls_311
N_housing
water
mta_stops
land
school_total
crime
quarter
1
0.9 0.9
0.8 0.8 0.8
0.8 0.8 0.9 1
0.7 0.6 0.5 0.6 0.7
0.2 0.3 0.5 0.4 0.4 0.4
0.1 0.2 0.1 0.1 0.2 0.1 0.1
0.4 0.1 0.2 0.1 0.1 0.1 0.1 0.1
0 0 0.2 0 0 0.1 0 0 0
0 0.1 0.1 -0.1 0.2 0 0 0 0 0
0.5 0.3 0 0 0 0.2 0.3 0.2 0.3 0.2 0.2
0.2 0.3 0 0 0.1 0 0.2 0 0 0 0 0
0.1 0.1 0.4 -0.1 0.1 0.3 0.1 0.2 0.1 0.1 0.1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0
crime
school_total
land
mta_stops
water
N_housing
N_calls_311
prop_under_poverty_level
Estimate_Total
income_index
racial_index
prop_male
prop_stable
age_index
Corr
1.0
0.5
0.0
-0.5
-1.0

References

Criminal Justice Research

- Graif, C., Sampson, R.J.: Spatial heterogeneity in the effects of immigration and diversity on neighborhood homicide rates. Homicide Studies 13, 242–260 (2009)
- Lee, B.A., Iceland, J., Sharp, G.: Racial and ethnic diversity goes local: Charting change in american communities over three decades key findings. Technical report, Brown University (2012)
- https://www.journals.uchicago.edu/doi/abs/10.1086/655357

Methodological Approach

- Mining large-scale human mobility data for long-term predictiton https://arxiv.org/pdf/1806.01400.pdf