A Machine Learning Approach

PREDICTING CRIME RATES IN BALTIMORE

May 5, 2015

Mike, Sophie, Zoey, Krishna, Kyle

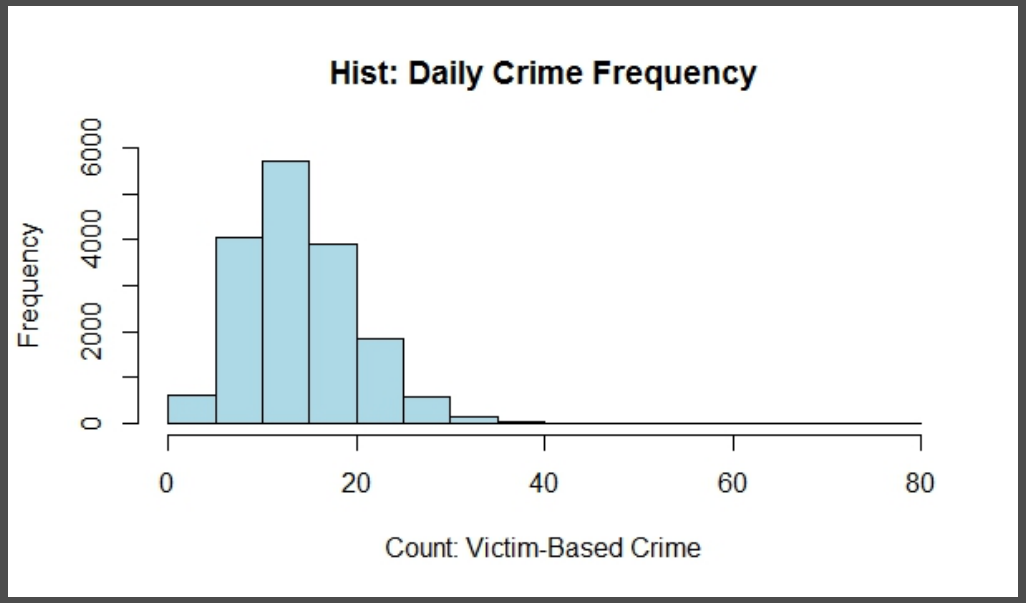
Motivations:

- Recent events in Baltimore
- Need for more creative ways to address chronic crime in certain areas
- More efficient use of police resources
- Predictive models help anticipate problems before they arise

The Data – an overview

- Crime:
  - Date, Type, Location
  - All victim-based crime 1/1/2010 – 2/28/2015
  - Includes exact geographic coordinates
- Neighborhoods:
  - Location
  - Demographics
- Weather Data:
  - Temperature & Precipitation

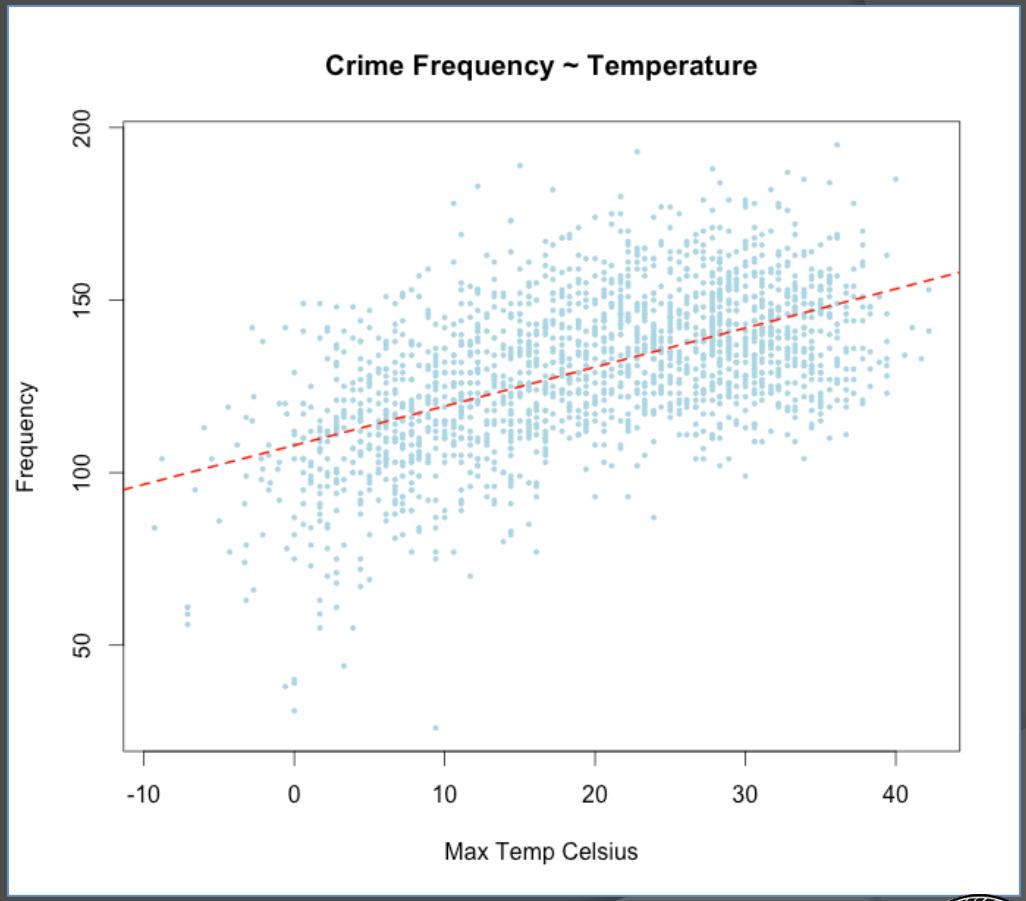
The Data - Continued



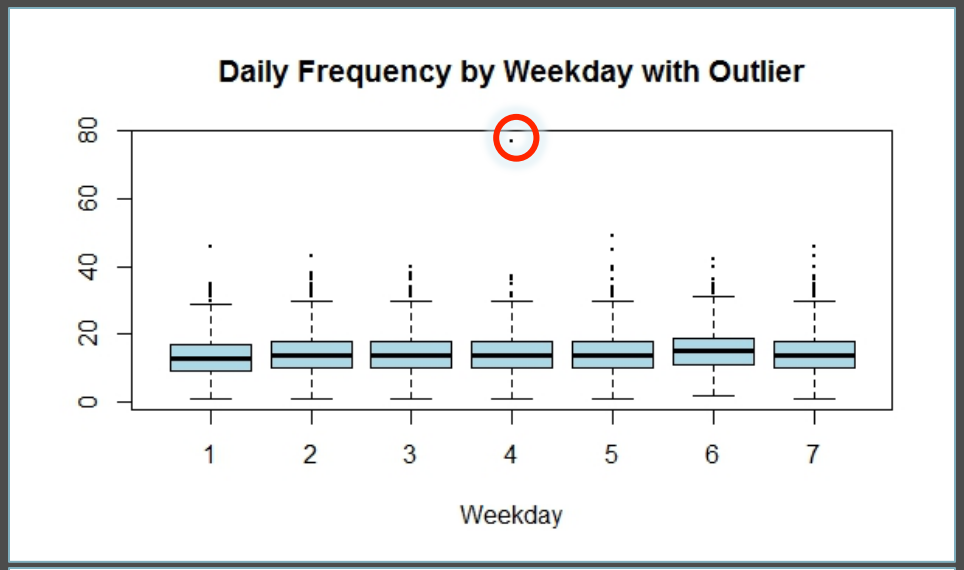
- **Daily Frequency**
  - Typical day between 10 and 20 incidents in one police district

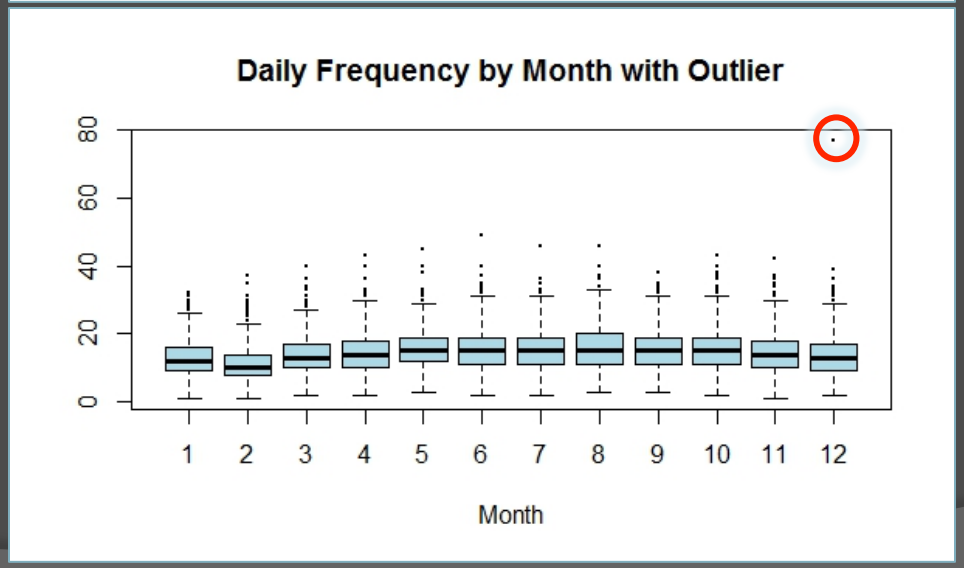
The Data - Continued

- **Weather & Crime**
  - Correlation : 0.65
  - Bulk of observations between 15 & 30 degrees Celsius
  - Sweaty people don't commit crime?



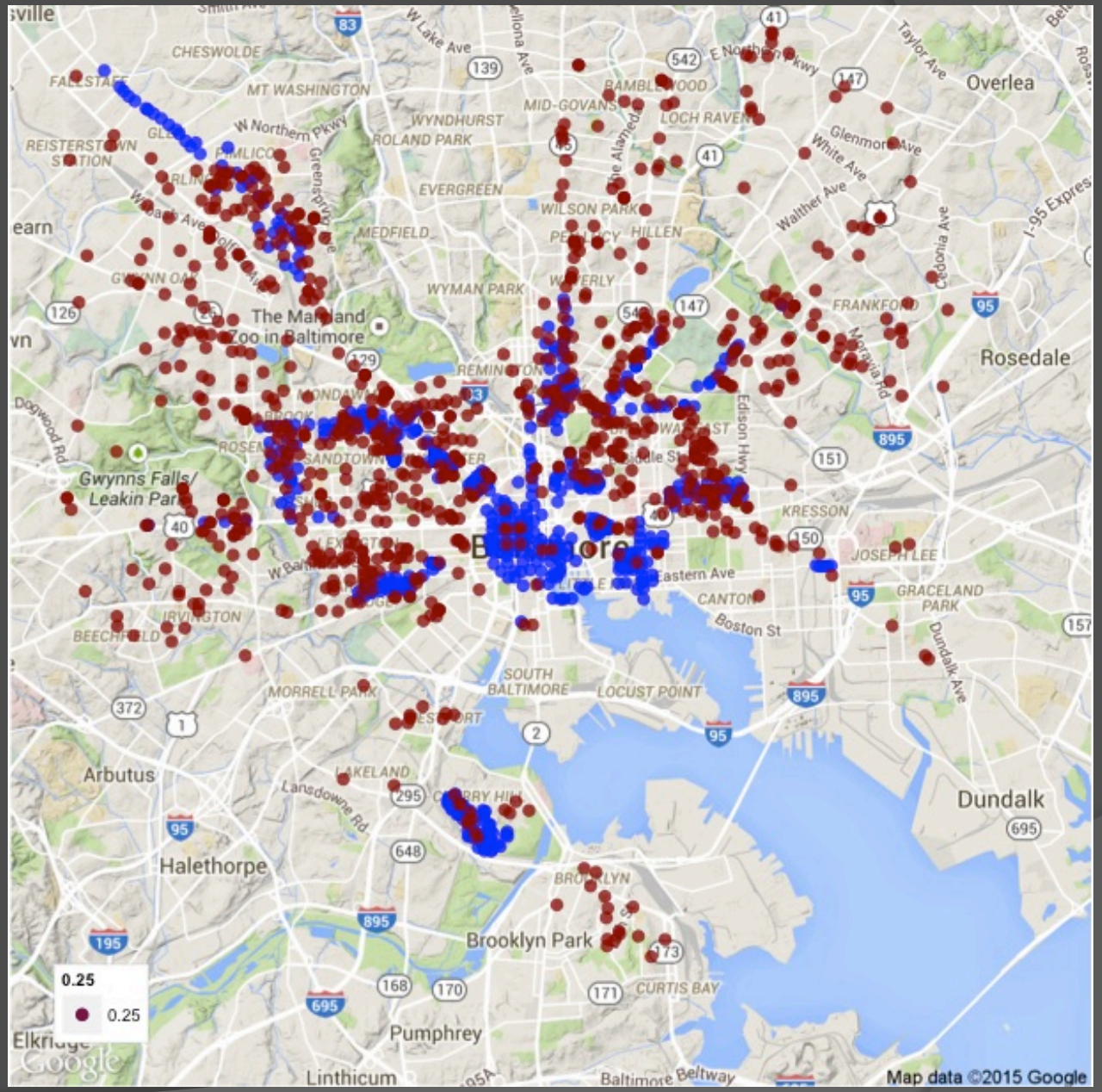
The Data - Continued



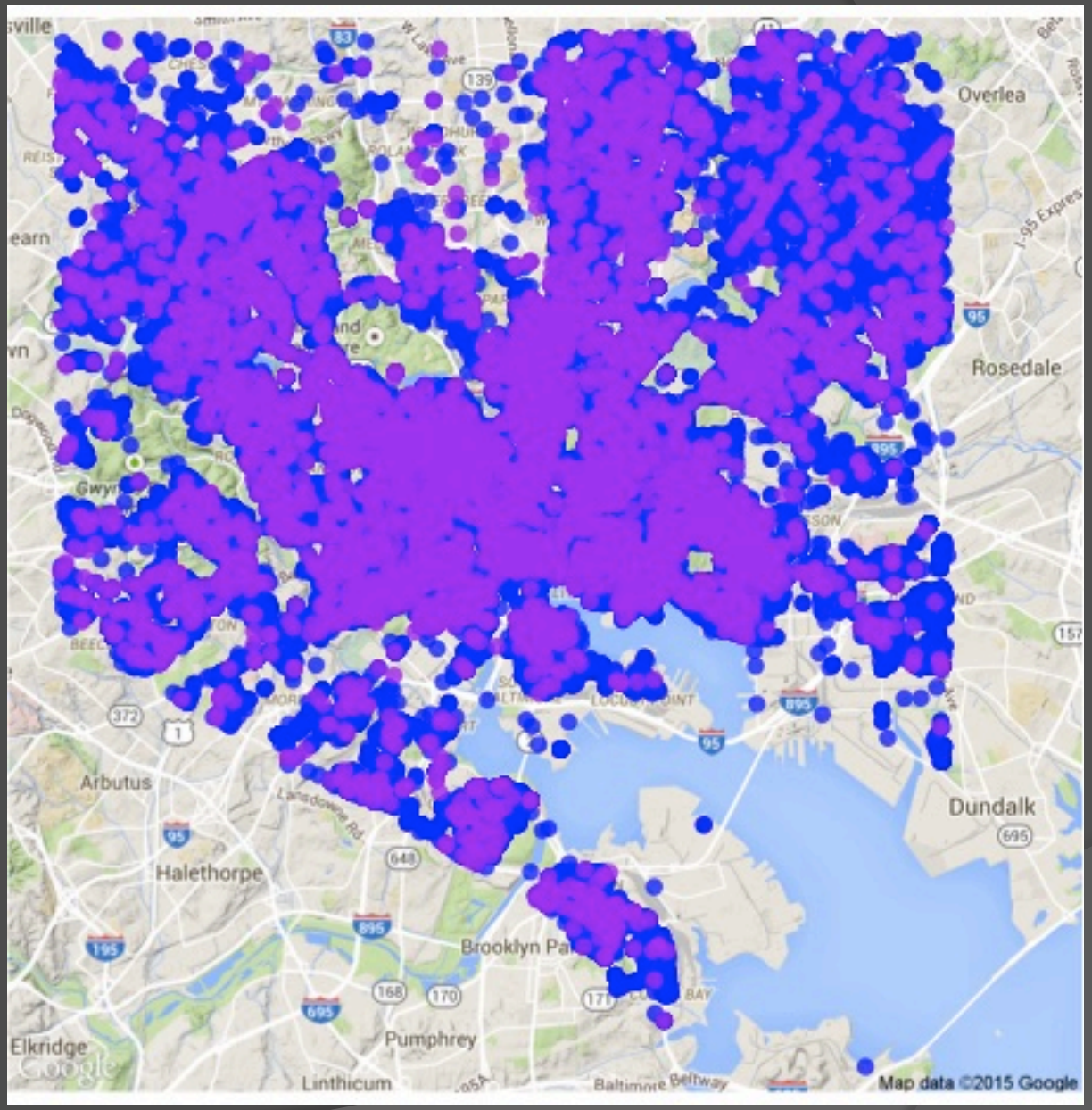


- **Seasonality**
  - Not much variation by weekday
  - Colder months have Less crime
  - **Outlier: 77** Christmas 2013 at the post office

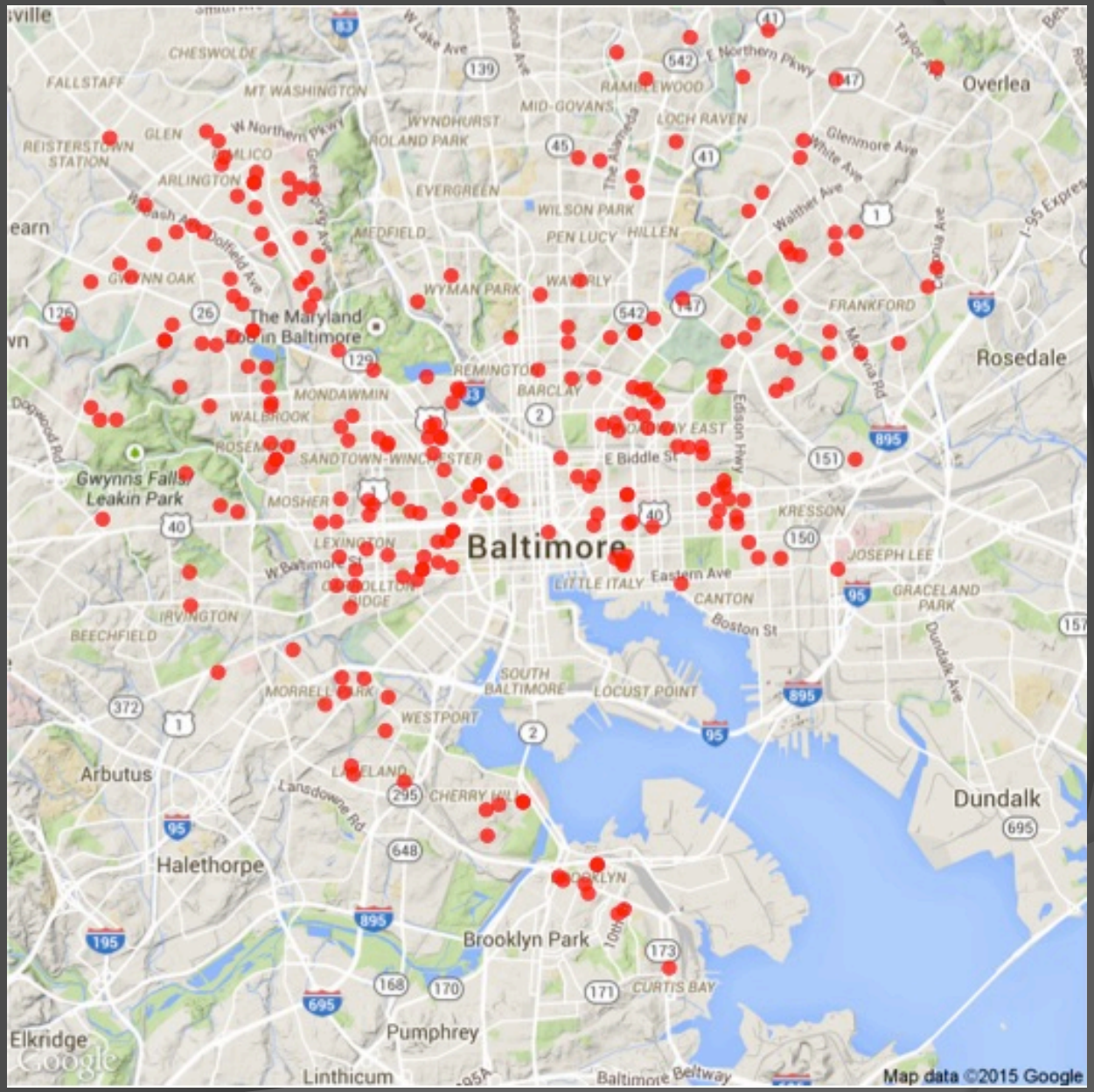
CCTV & Murder


0.25
0.25
Map data ©2015 Google
Google
The Maryland Zoo in Baltimore
Gwynns Falls/ Leakin Park
Arbutus
Halethorpe
Brooklyn Park
Pumphrey
Linthicum
Dundalk
Rosedale
Overlea
Baltimore Beltway
CHESWOLDE
MT WASHINGTON
WYNDHURST
ROLAND PARK
EVERGREEN
WILSON PARK
MEDFIELD
WYMAN PARK
HILLEN
FRANKFORD
SOUTH BALTIMORE
LOCUST POINT
CANTON
GRACELAND PARK
KRESSON
CURTIS BAY
Eastern Ave
Boston St
W Lake Ave
W Northern Pkwy
E Northern Pkwy
Taylor Ave
White Ave
Glenmore Ave
Walther Ave
Edison Hwy
Dundalk Ave
Lansdowne Rd
Dogwood Rd
Morgan Rd
I-95 Expres

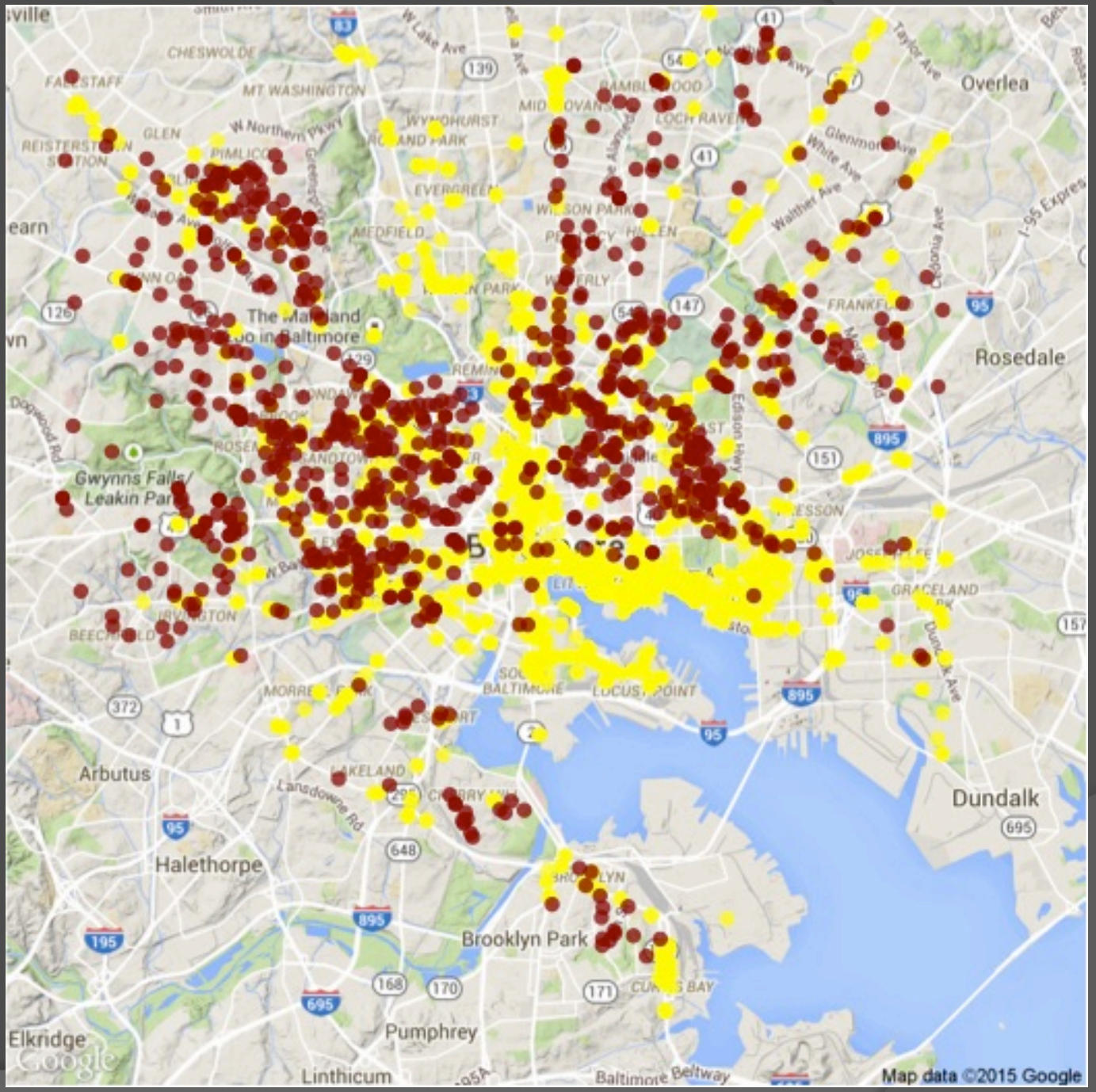
Assault & Muggings



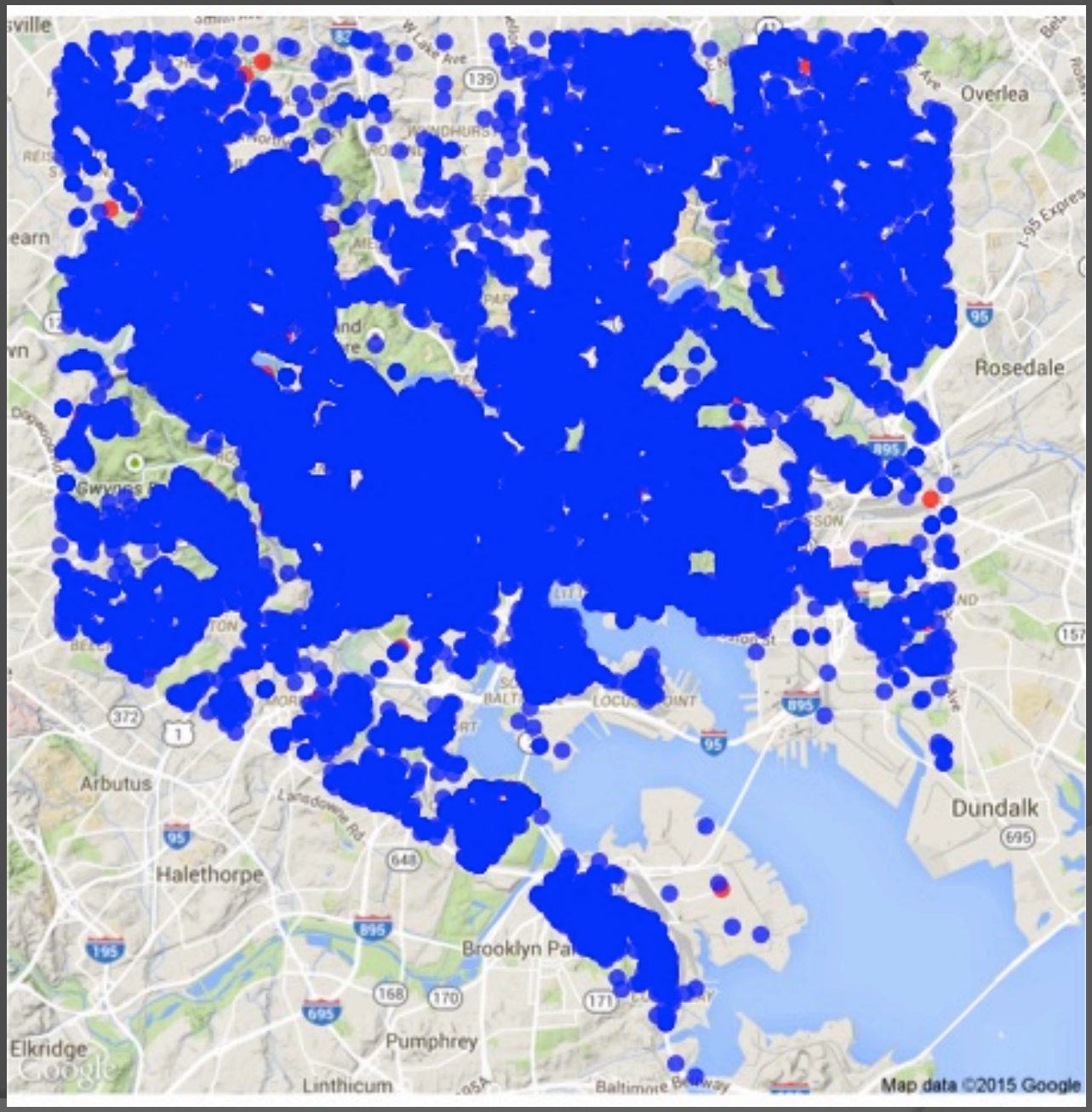
Arson


Baltimore
Rosedale
Overlea
Dundalk
Arbutus
Halethorpe
Brooklyn Park
Pumphrey
Elkridge
Linthicum
Gwynns Falls/ Leakin Park
The Maryland Zoo in Baltimore
Map data ©2015 Google

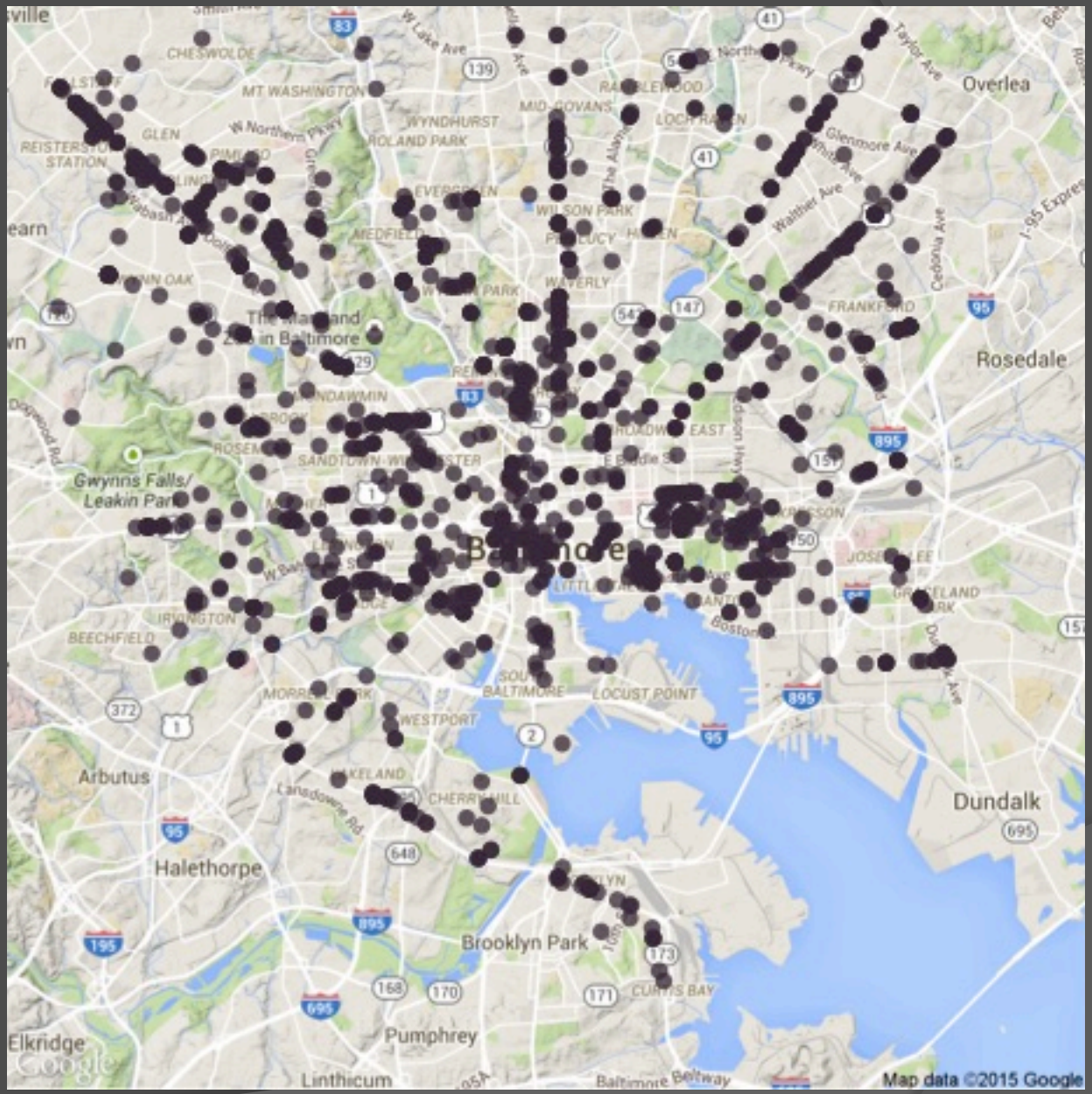
Beer & Murder


Overlea
Rosedale
Gwynns Falls/
Leakin Park
Arbutus
Halethorpe
Dundalk
Brooklyn Park
Pumphrey
Elkridge
Linthicum
Baltimore Beltway
Map data ©2015 Google

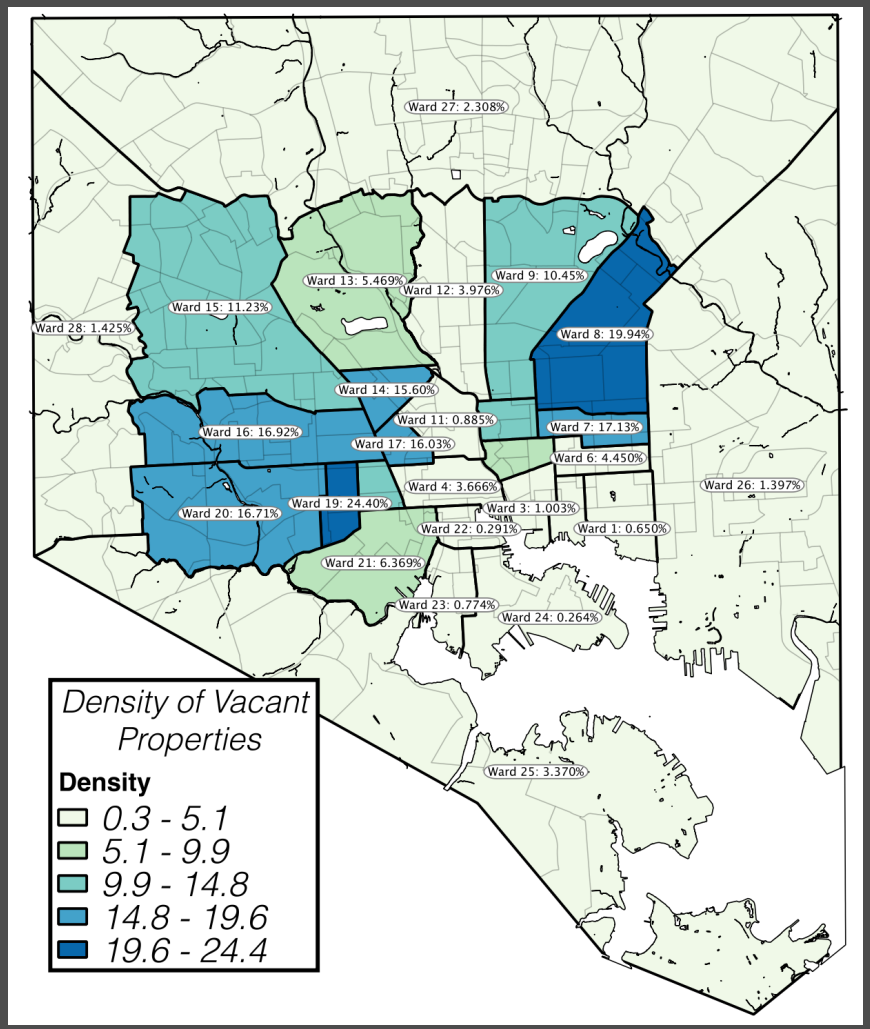
Auto Theft & GTA

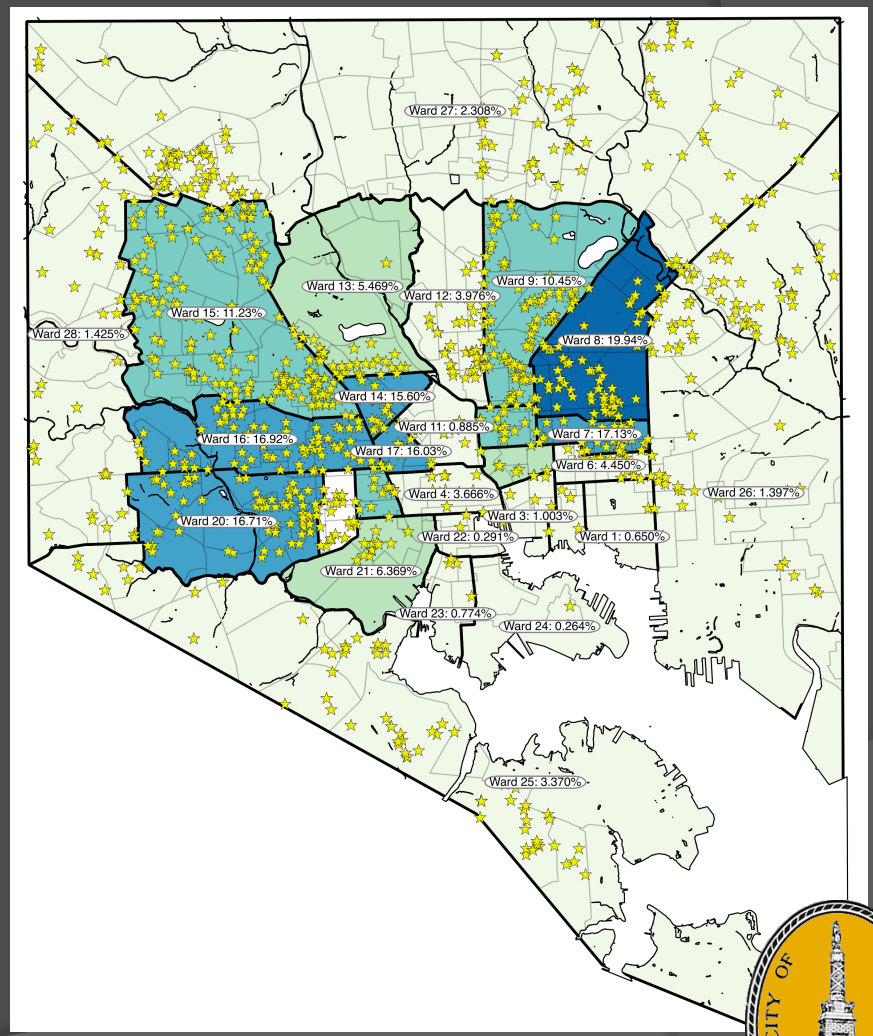


Store Robbery



Vacancies & Murders





Prediction– Poisson Regression

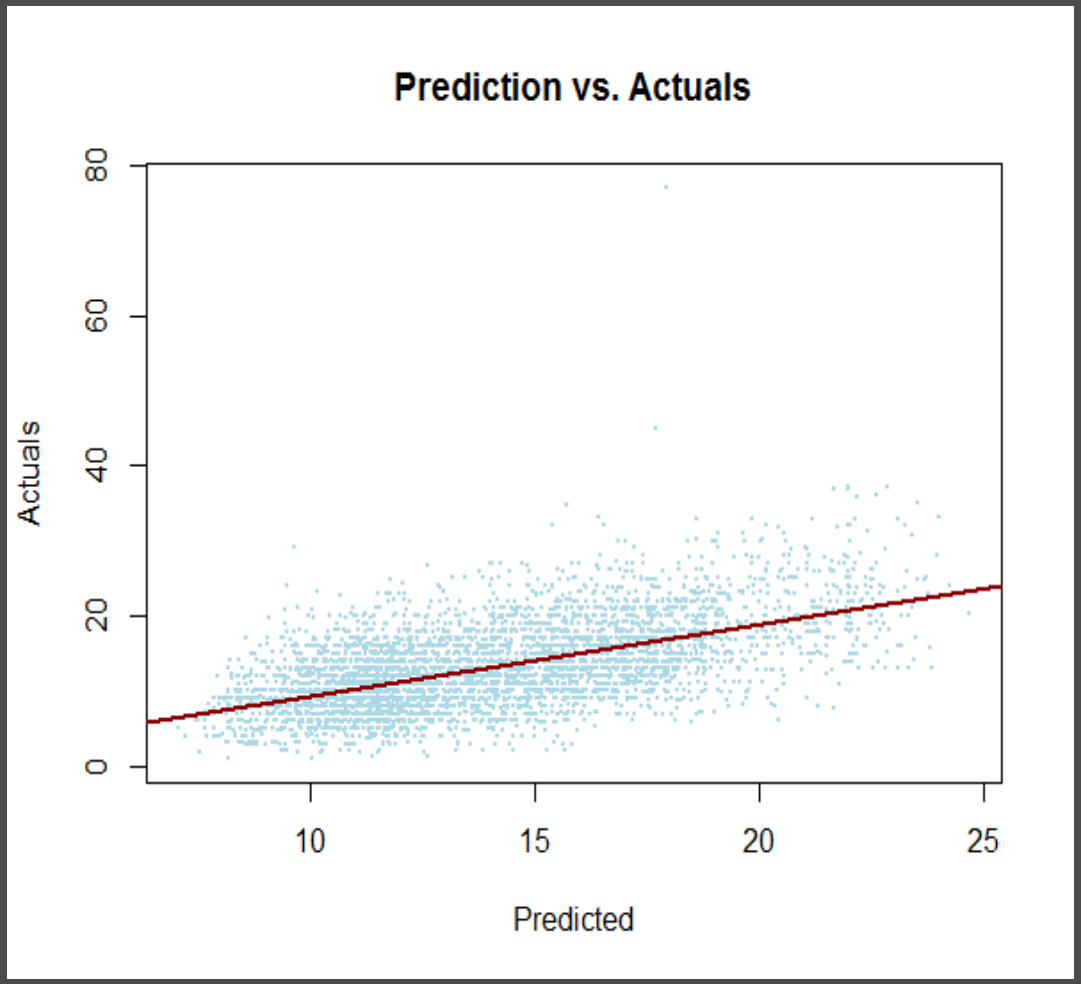
- Poisson Regression:
  - Crime rates are essentially counts
  - Frequency evaluated at the police district level
  - Grouped data by day, police district (9 total)
  - Aggregate data on Ethnicities, Age, Households, & Vacancies for all neighborhoods in the district
  - Randomly sample of obs. used for training/ testing datasets.

Poisson Regression: Results

SELECTED PARAMETER ESTIMATES:								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
tmin		1	0.0051	0.001	0.0031	0.0072	24.61	<.0001
tmax		1	0.0042	0.0008	0.0026	0.0057	28.15	<.0001
AGE65ovr		1	18.2919	0.8528	16.6204	19.9634	460.03	<.0001
Blk_AfAm		1	42.1374	0.9162	40.3417	43.933	2115.42	<.0001
PrecipMM		1	-0.0019	0.0002	-0.0023	-0.0014	59.65	<.0001
White		1	41.8469	0.9719	39.942	43.7518	1853.9	<.0001
AGE45_64		1	-18.6871	0.7962	-20.2476	-17.1266	550.89	<.0001
AGE35_44		1	121.0235	2.5006	116.1223	125.9246	2342.25	<.0001
AGE15_17		1	-116.458	2.3512	-121.066	-111.85	2453.3	<.0001

- 1% increase in proportion of people aged 35-44 in a police district increases the crime frequency by 3.
- Decrease in proportion of people aged 15-17 by 1% increases the crime count by 3.
- 2% increase in proportion of either African American or White ethnicities increases crime incidents by 3.
- A degree Celsius increase in minimum or maximum temperature increases crime rate by 1.

Poisson Regression: Results

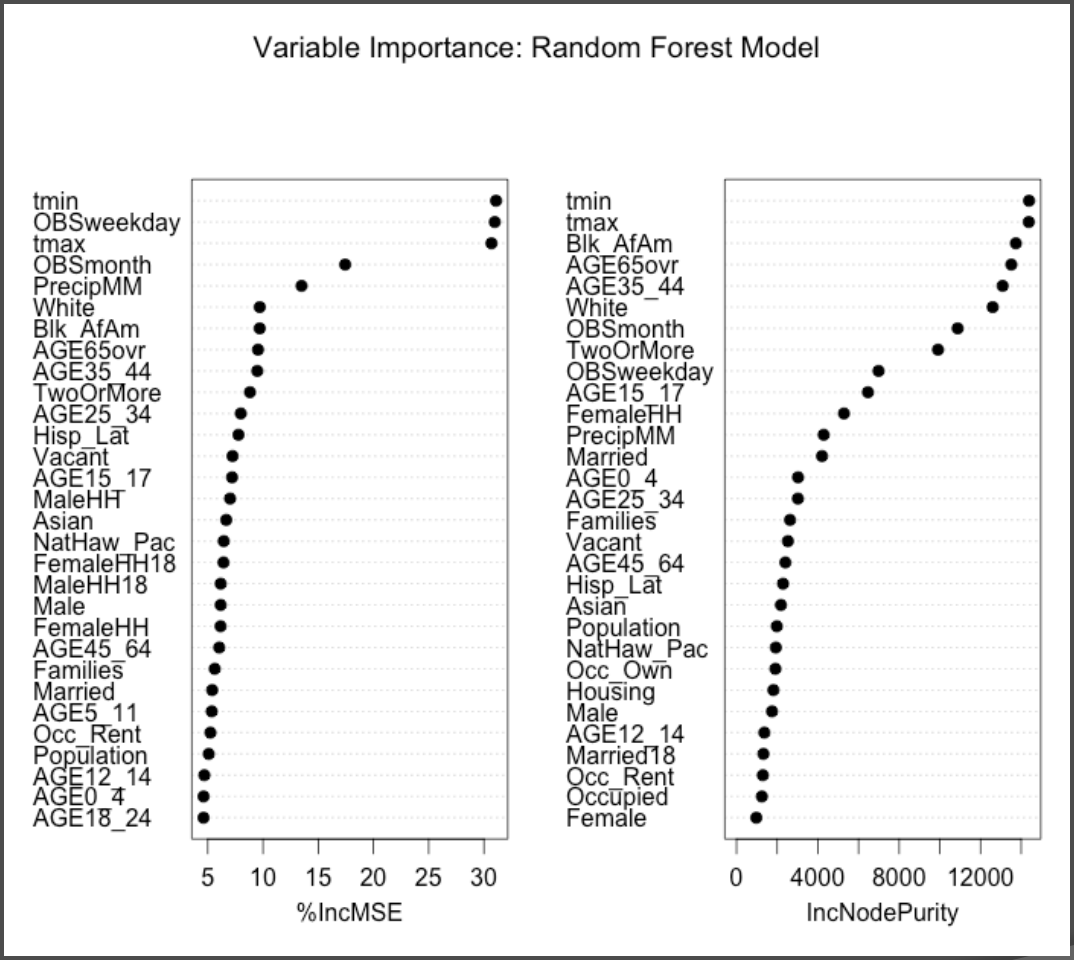


- Prediction vs. Actuals:
  - Pseudo R2 = 0.3563
  - Correlation(PRED vs. ACT) = 0.5981

Prediction – Random Forest

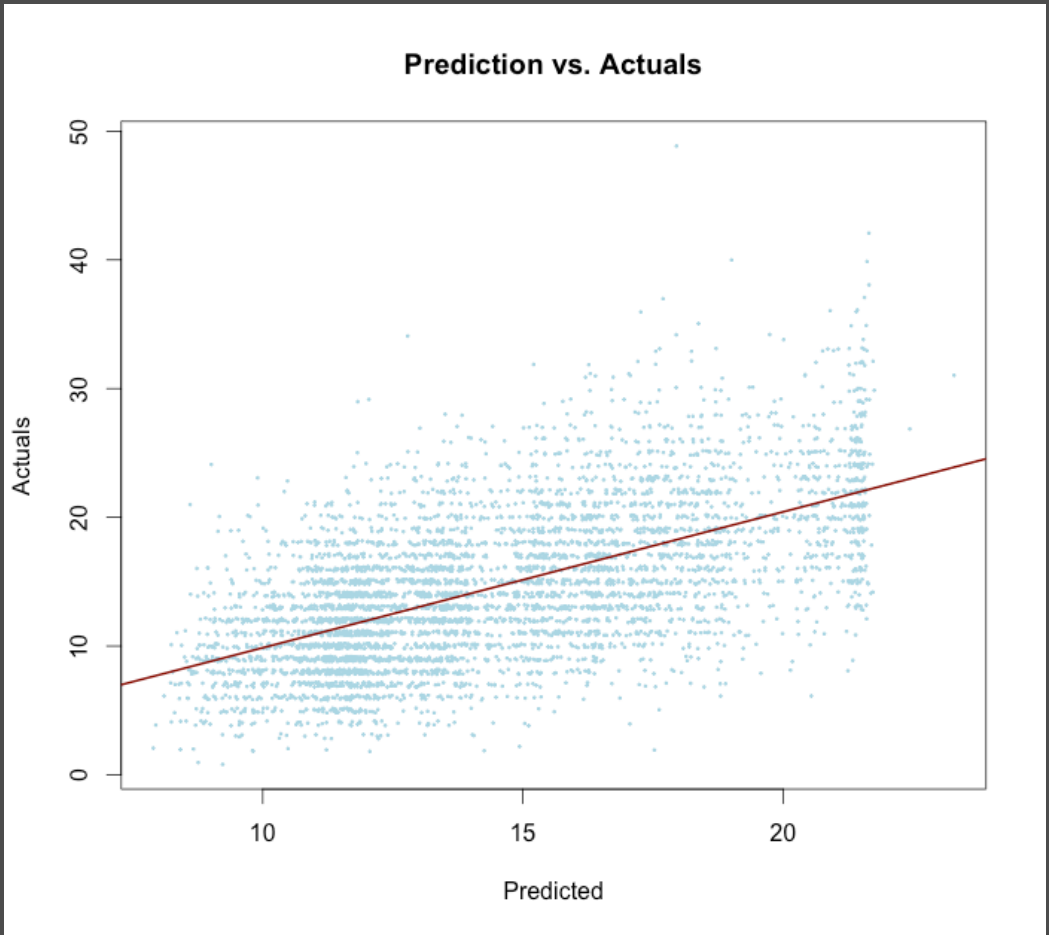
- Regression Trees:
  - Robust to noise, outlier, overfitting
  - Use all available independent variables from the data
  - 500 trees, 6 variables in each
  - 70% for training, 30% for testing
  - Generate Variable Importance Plot to assess variables with greatest influence.
  - Plot predictions against actuals

Random Forest: Results



- Var. Importance:
  - Weather is a major factor
  - Timeframes make a big difference
  - Ethnic boundaries are also important

Random Forest: Results



- Prediction vs. Actuals:
  - Pseudo R2 = 0.3561
  - Correlation(PRED vs. ACT) = 0.6554

Conclusions & Recommendations

- Criminal activity CAN be predicted using existing data about the region
- Random Forest Models scale easily
- Future models should include new data:
  - Income levels
  - Proximity to regional landmarks
  - Date specific information (Festivals, holidays, protests, etc.)

Thank you!

Questions?