Crime in Chicago

Zihao Guo zg866
Shuting Gu sg5686
Anshan He ah4734
Ziyu Lei zl2350

Problem Statement

- Background Introduction
  - Crimes in Chicago
- Goal
  - Estimate when or where crime is committed
  - Predict crime type

Data Understanding

- Data Set

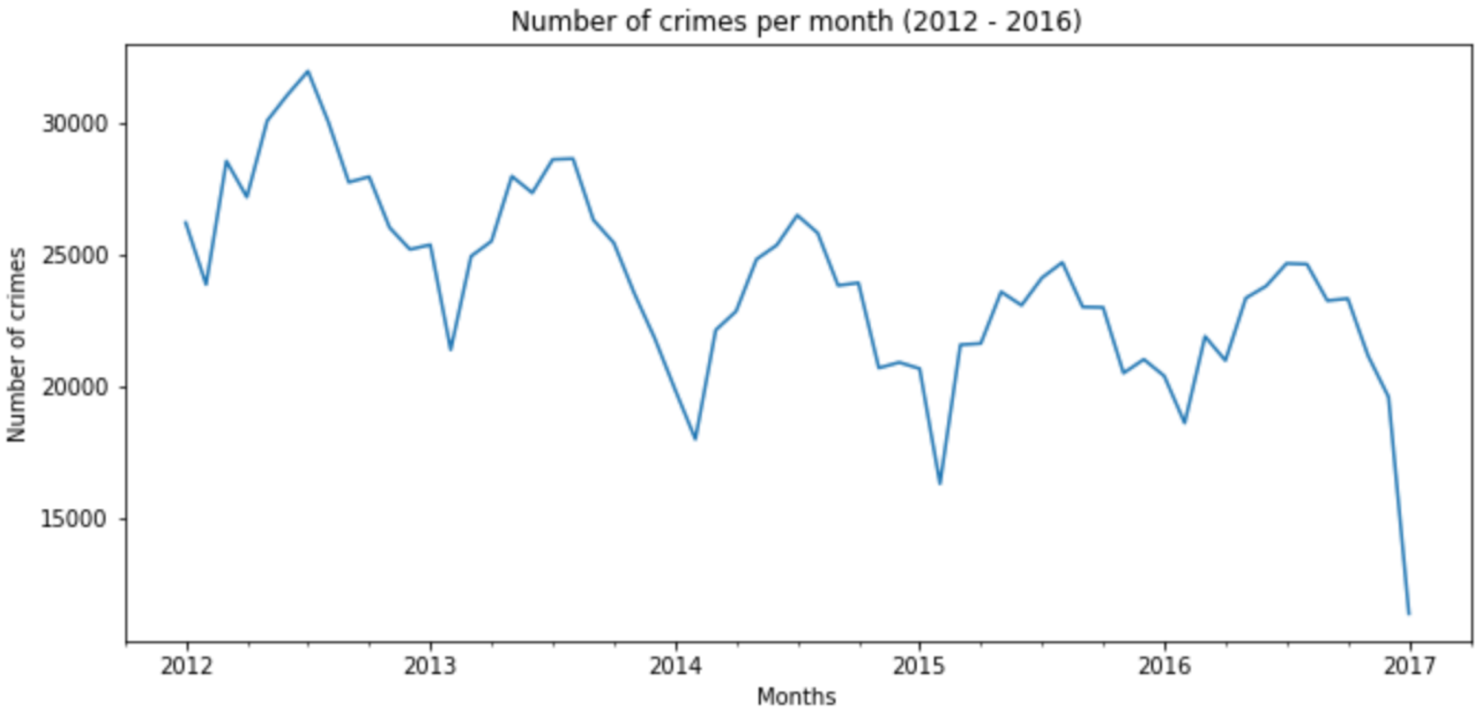
Data Summary

	number_nan	number_distinct	distinct vals
ID	0	1456714	[10508693, 10508695, 10508697, 10508698, 10508...
Date	0	582146	[05/03/2016 11:40:00 PM, 05/03/2016 09:40:00 P...
Block	0	32774	[013XX S SAWYER AVE, 061XX S DREXEL AVE, 053XX...
Primary Type	0	33	[BATTERY, PUBLIC PEACE VIOLATION, THEFT, WEAPO...
Description	0	342	[DOMESTIC BATTERY SIMPLE, RECKLESS CONDUCT, SI...
Location Description	1658	142	[APARTMENT, RESIDENCE, STREET, SIDEWALK, CHA H...
Arrest	0	2	[True, False]
Domestic	0	2	[True, False]
District	1	24	[10.0, 3.0, 15.0, 6.0, 1.0, 2.0, 24.0, 7.0, 18...
Community Area	40	78	[29.0, 42.0, 25.0, 44.0, 35.0, 38.0, 1.0, 67.0...
Latitude	37083	368076	[41.864073157, 41.782921527, 41.894908283, 41....
Longitude	37083	367942	[-87.70681860799999, -87.60436317, -87.7583719...

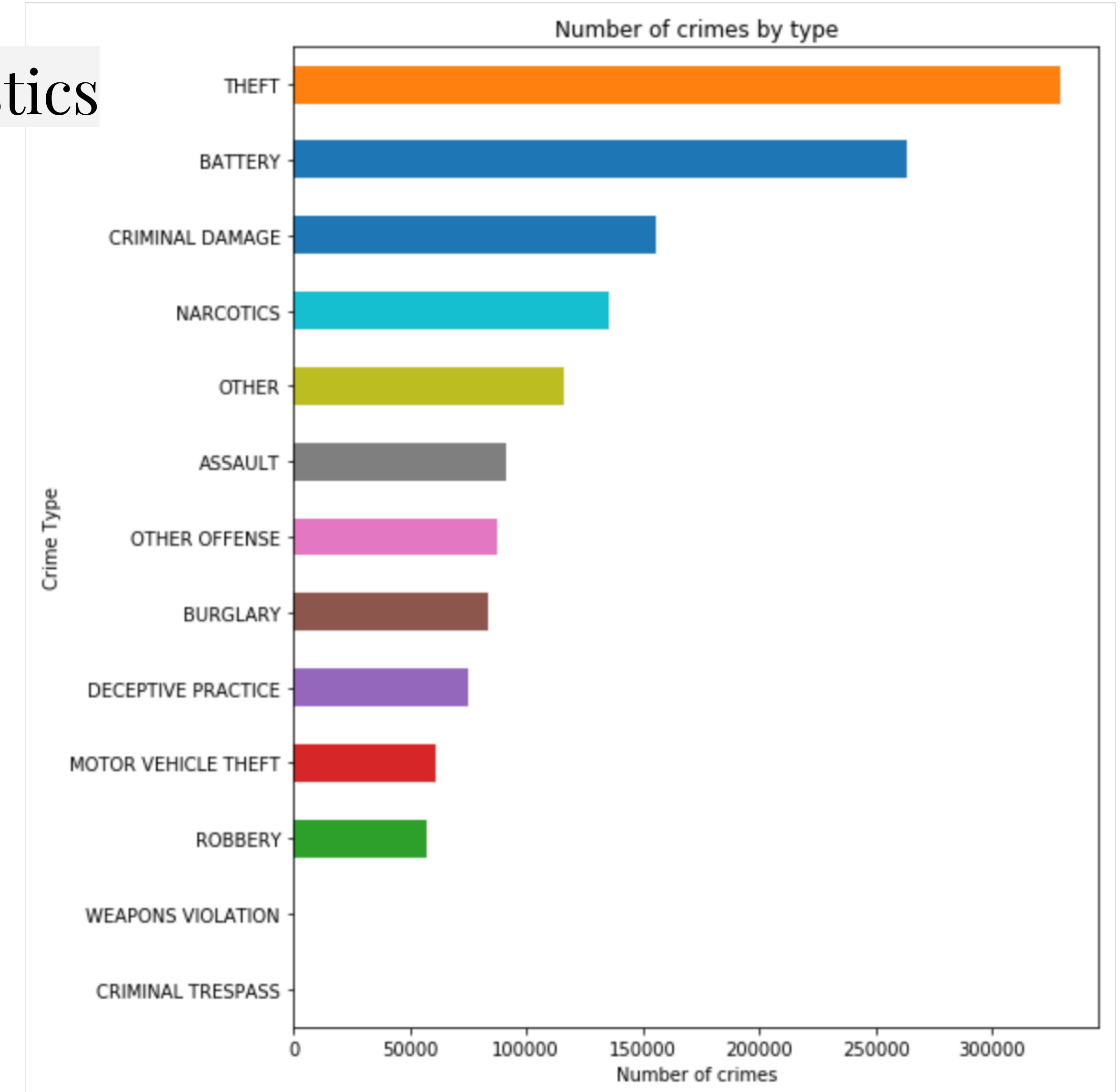
Data Preparation

- Data Cleaning
  - Missing Values: drop null
- Feature Engineering
  - Transformation
  - Normalization

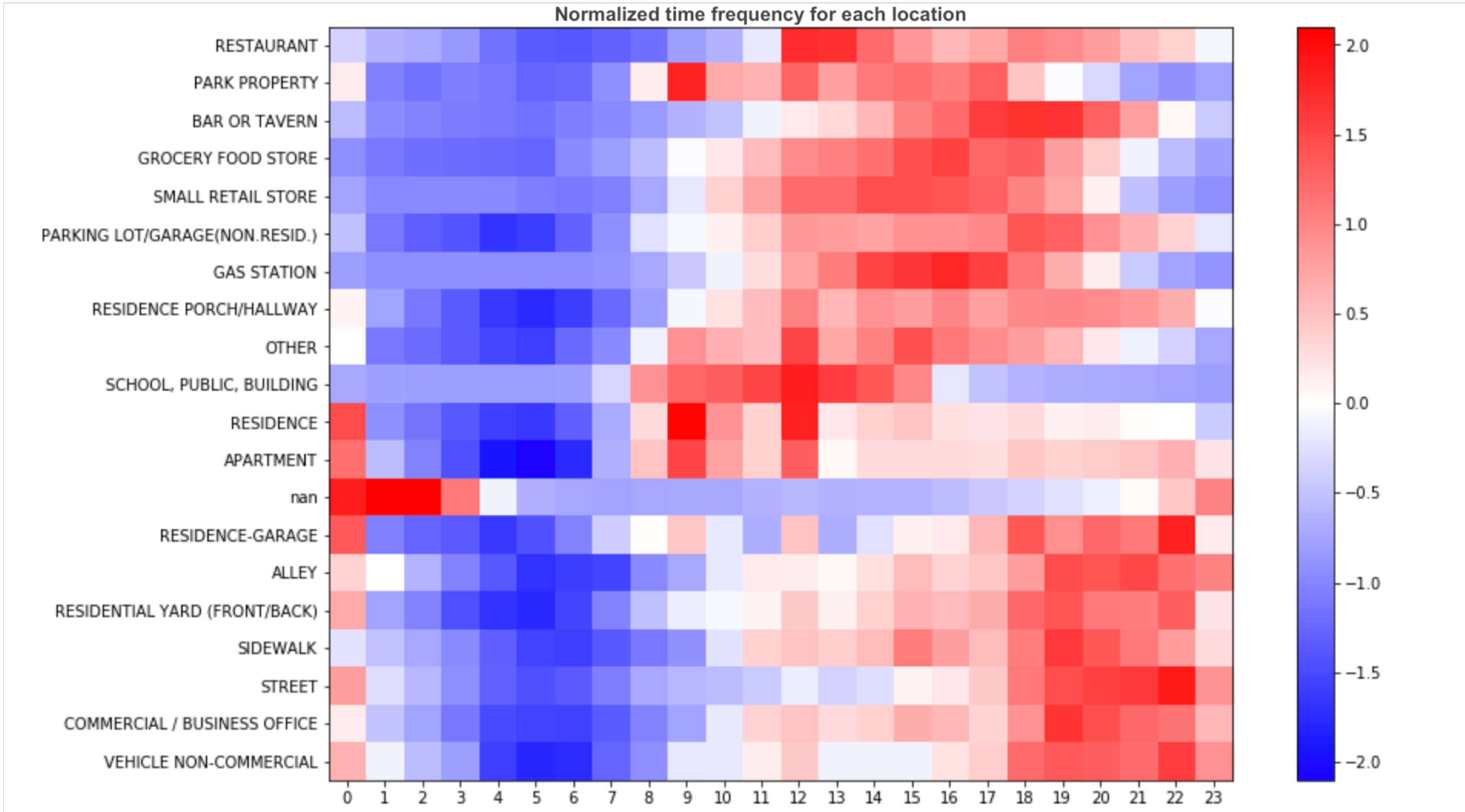
Methodology
Descriptive Statistics



Methodology
Descriptive Statistics

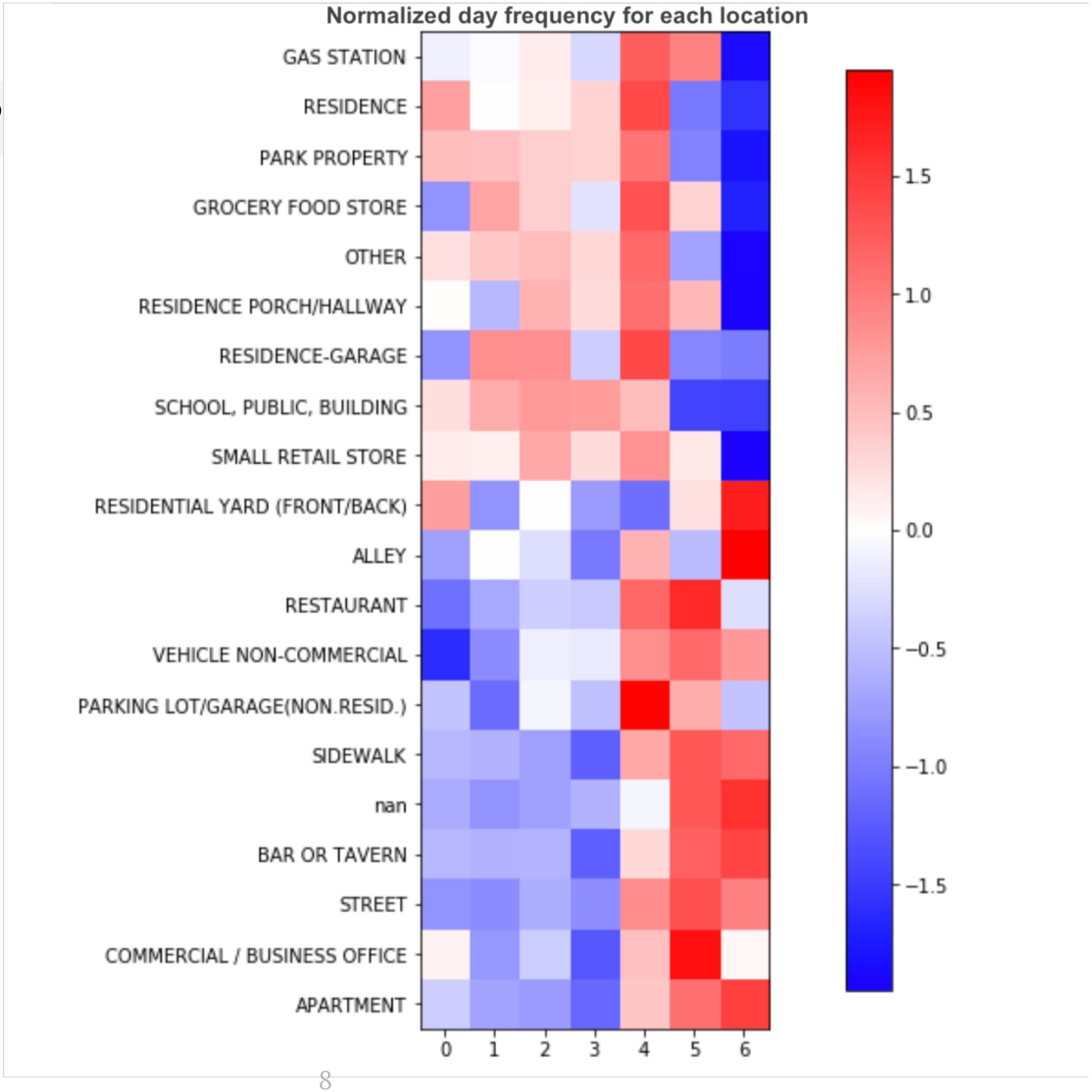


Methodology
When or Where?

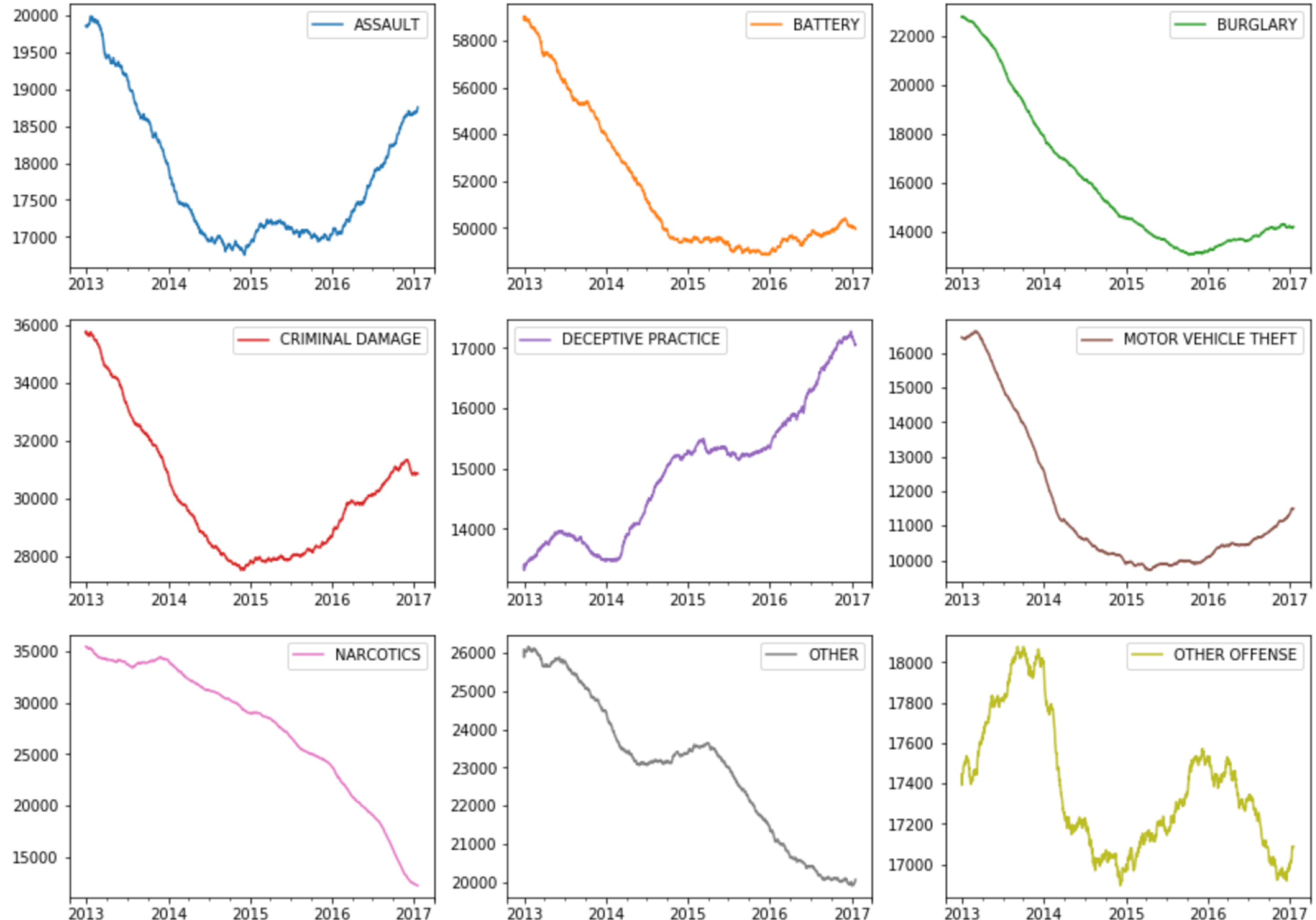


Methodology

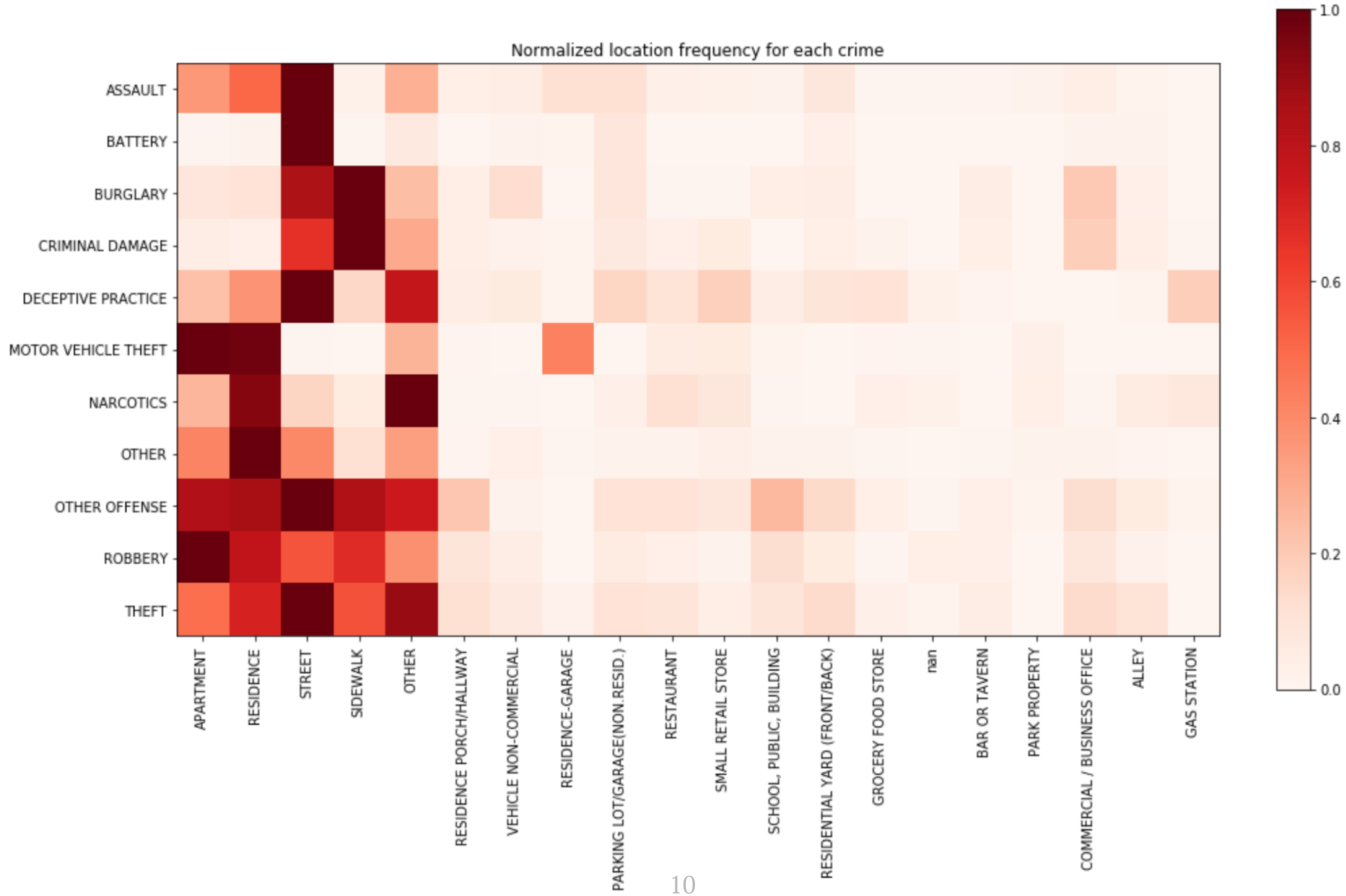
When or Where?



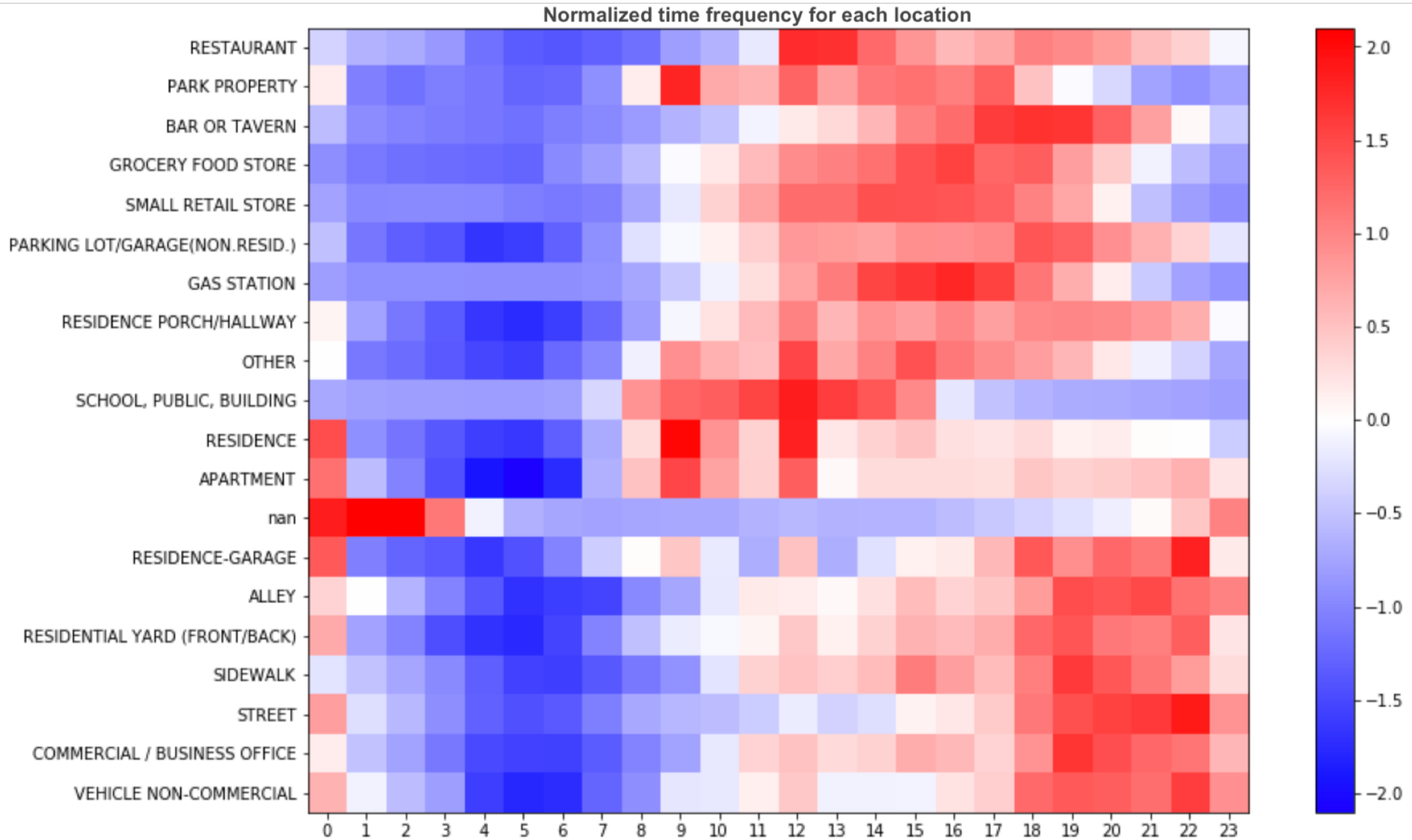
Methodology
Crime Type



Methodology
Crime Type



Methodology
Crime Type



Methodology
Crime Type

- Data Mining Techniques
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - KNN
  - XGBoost
- Evaluation Metric
  - F1 Score

Results

Models	F1 Score
Logistic Regression	0.226
K-Nearest Neighbors	0.361
Random Forests	0.391
Decision Tree	0.397
XGBoost	0.418

Conclusion

- Relationships between crime frequency, time and location
- Insights from crime type prediction
- Future work

Thank You!