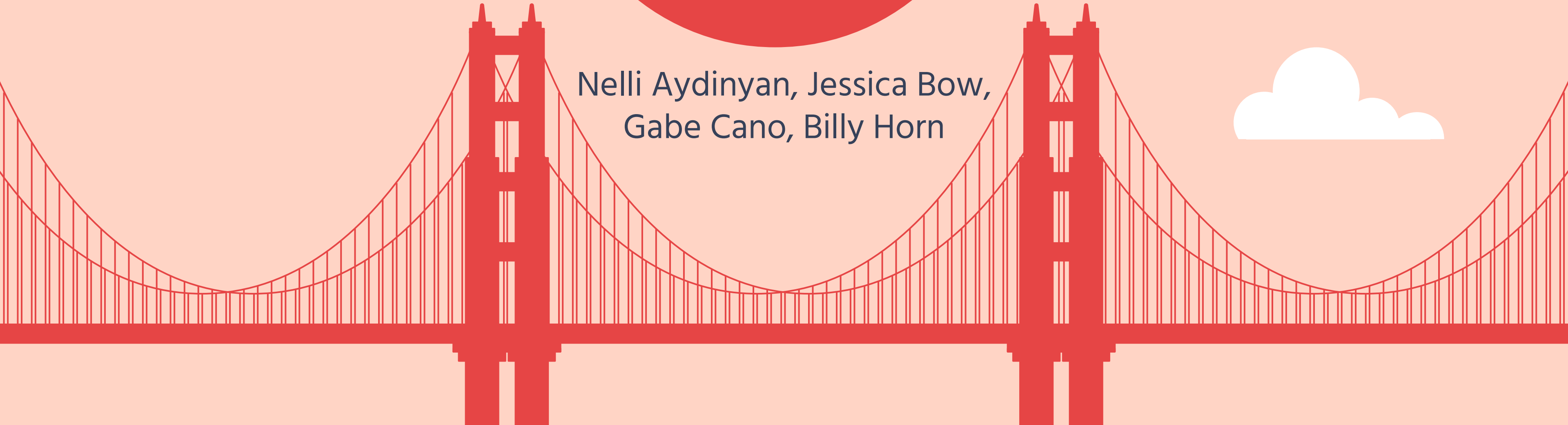
SAN FRANCISCO Crime Prediction

Nelli Aydinyan, Jessica Bow,
Gabe Cano, Billy Horn



Overview

Problem Statement

Data Background

Data Visualizations

Folium Mapping

Models

Conclusions

Problem Statement

As part of the public outcry to defund the police and dismantle and restructure law enforcement as a whole, there has been a call to provide funding for new and existing services to step in where applicable.

Based on the specific features of a crime, we endeavor to predict which type of intervention is needed for a particular crime.

Methodology for Crime Categorization

Police (Emergency)

Any crime that can be considered an immediate danger to human life or public safety

Police (Non-Urgent)

Serious crimes that don’t pose an immediate threat to human lives

Specialty

Crimes where other services may be a more qualified first response, including mental illness, substance abuse, sex crimes, domestic crimes, etc.

Data Background

300K+ crimes in San Francisco

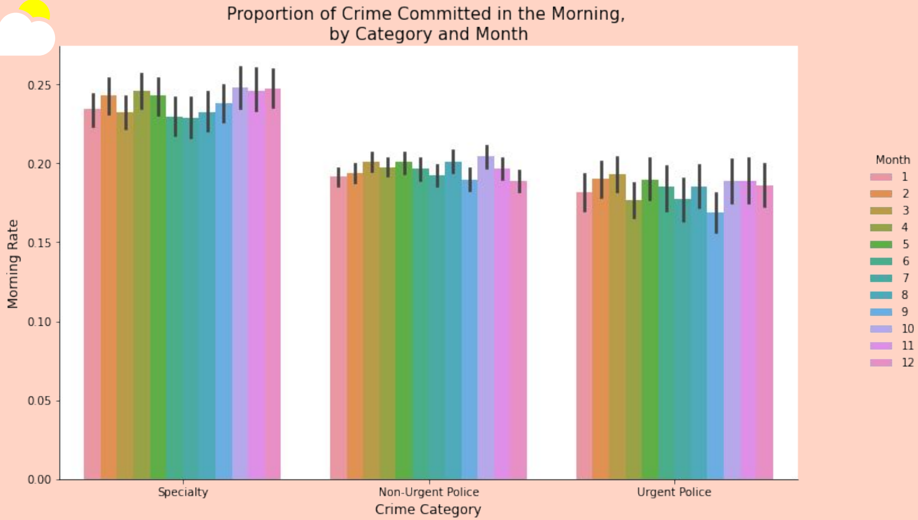
Ranges from 2016 - 2018

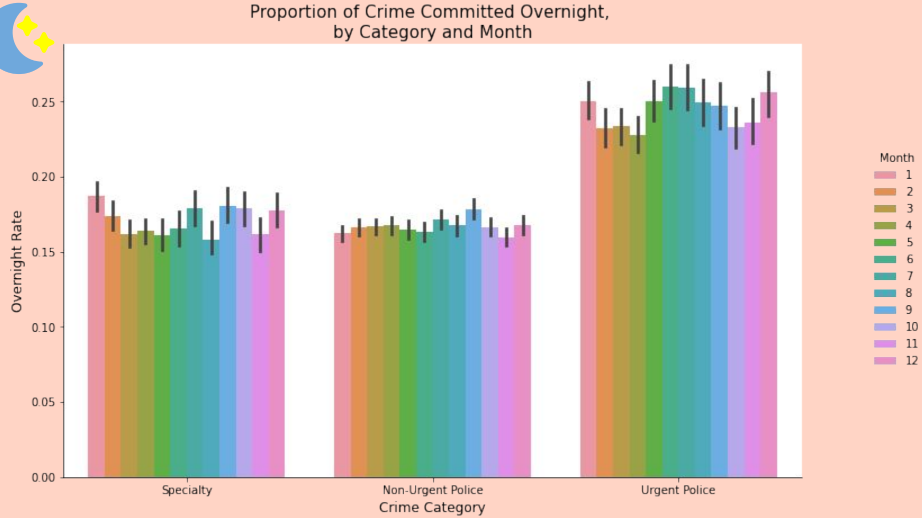
Includes:

- Geospatial data
- Crime category
- Crime description

Crimes: Morning vs. Night

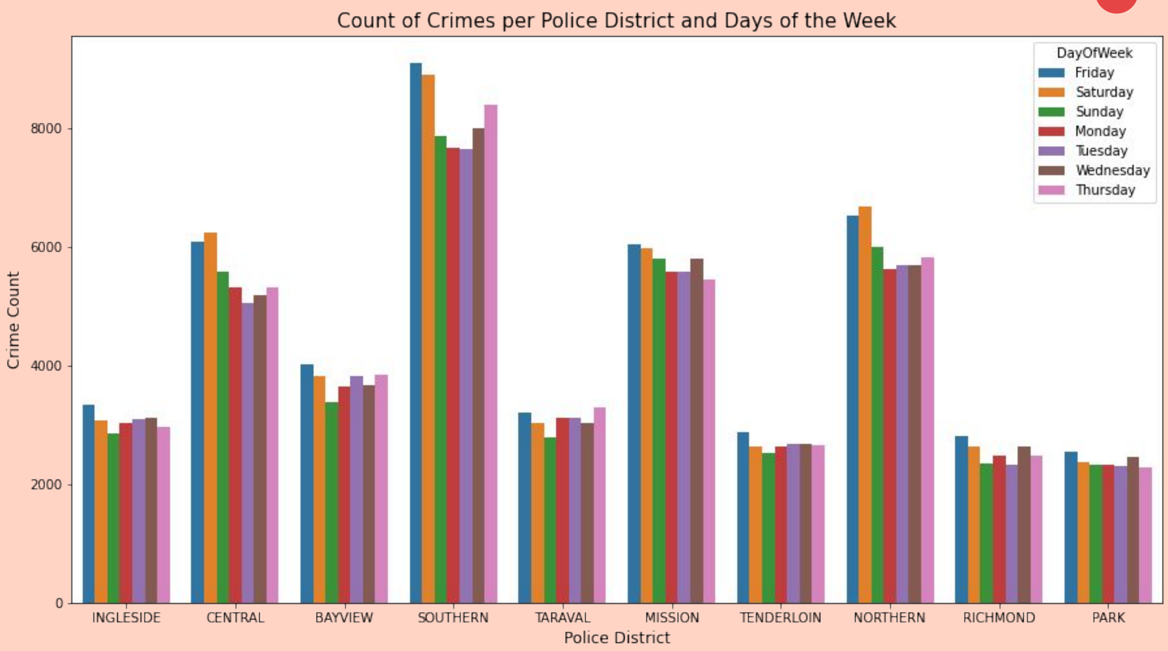
- Specialty crimes are highest in the morning
- Morning crimes are much higher in winter months
- Crimes requiring police intervention are highest at night
- Urgent crimes are highest in summer





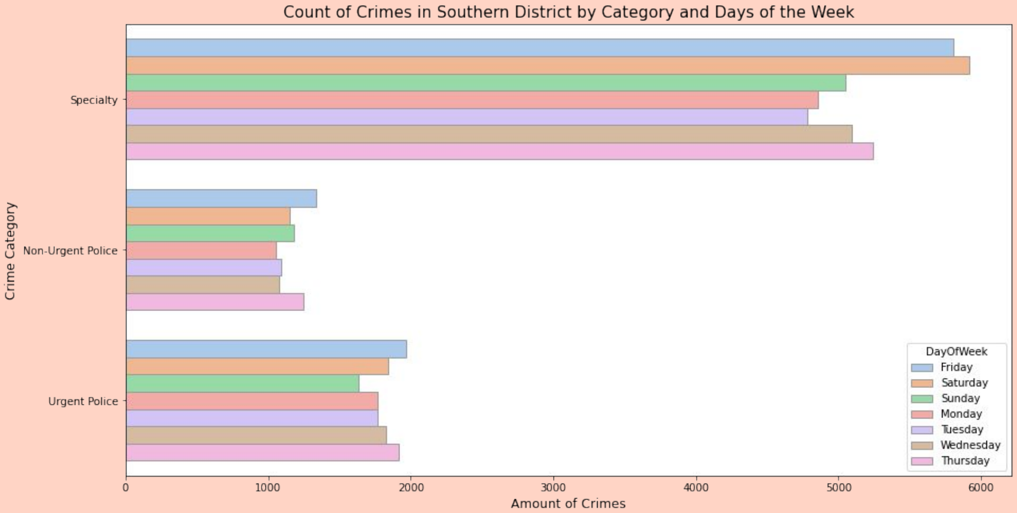
Crime Count: Police District and Days of the Week

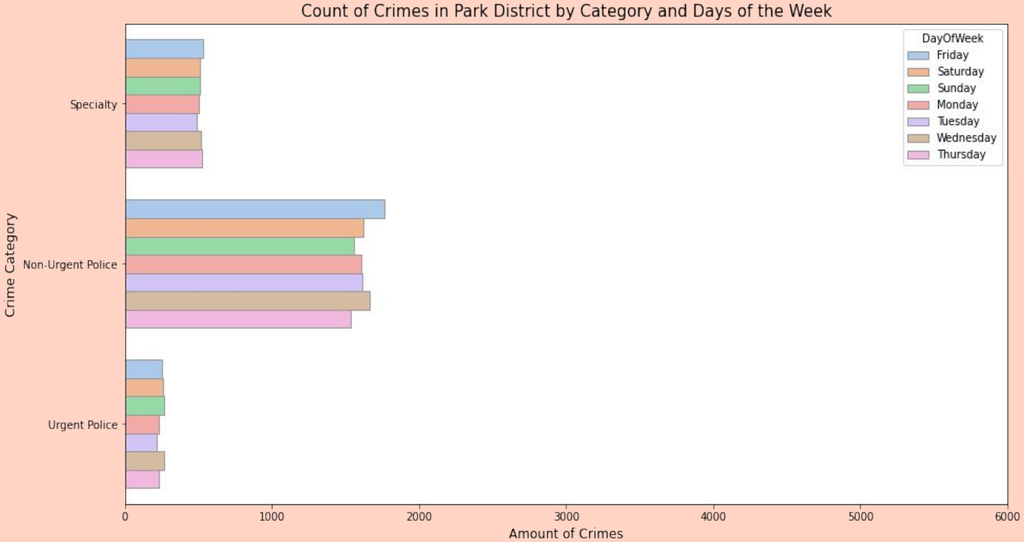
- Across districts, crimes tend to happen most often on Fridays and Saturdays
- Southern is the police district that's had the most reported crime
- Tuesdays and Sundays are safest days of the week.

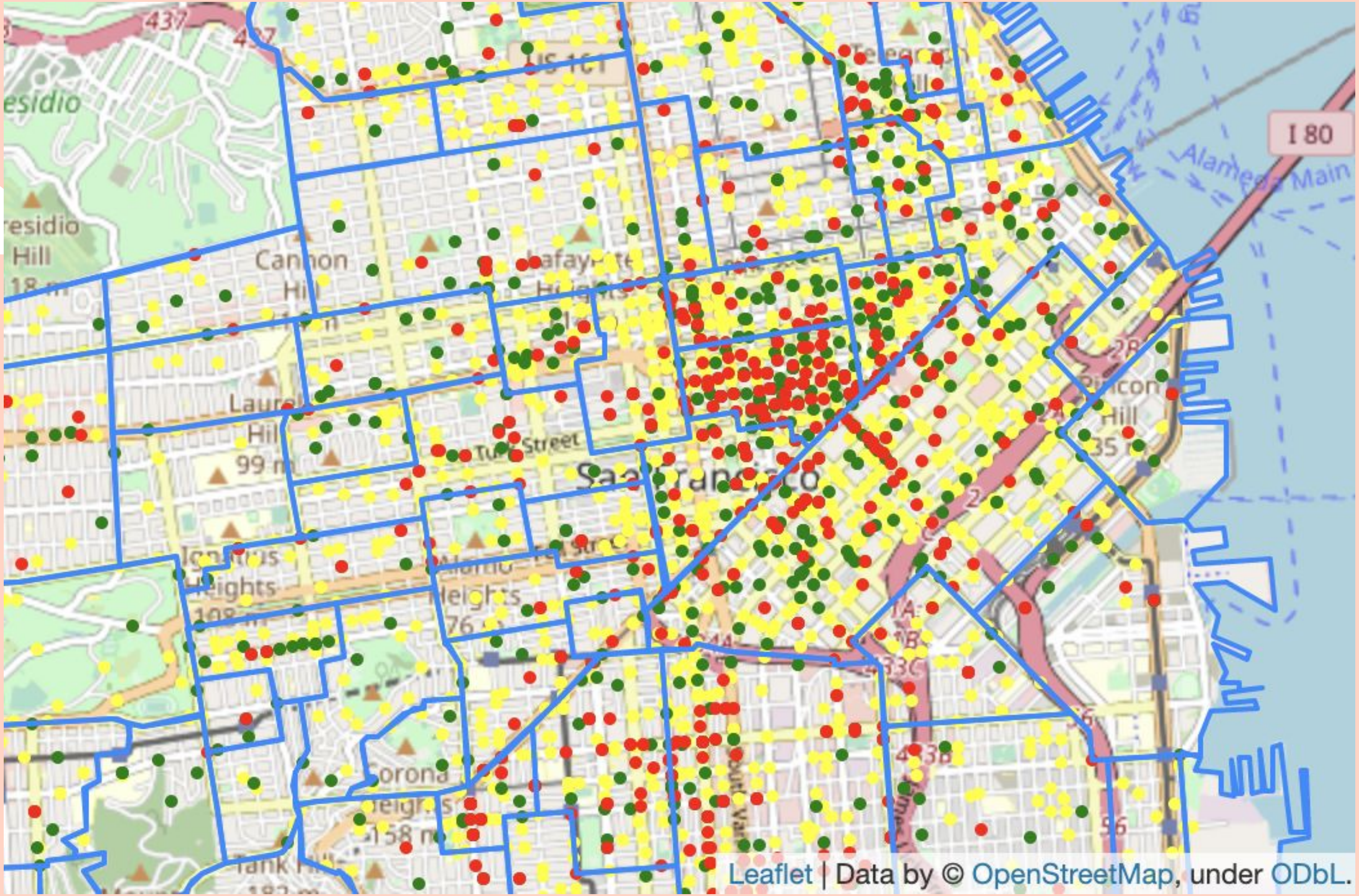


Highest vs. Lowest Crime Districts

- Most crimes in Park District are non-urgent
- Southern has around 3x amount of crime overall and 10x amount of urgent crimes
- Almost 10x the amount of urgent crimes

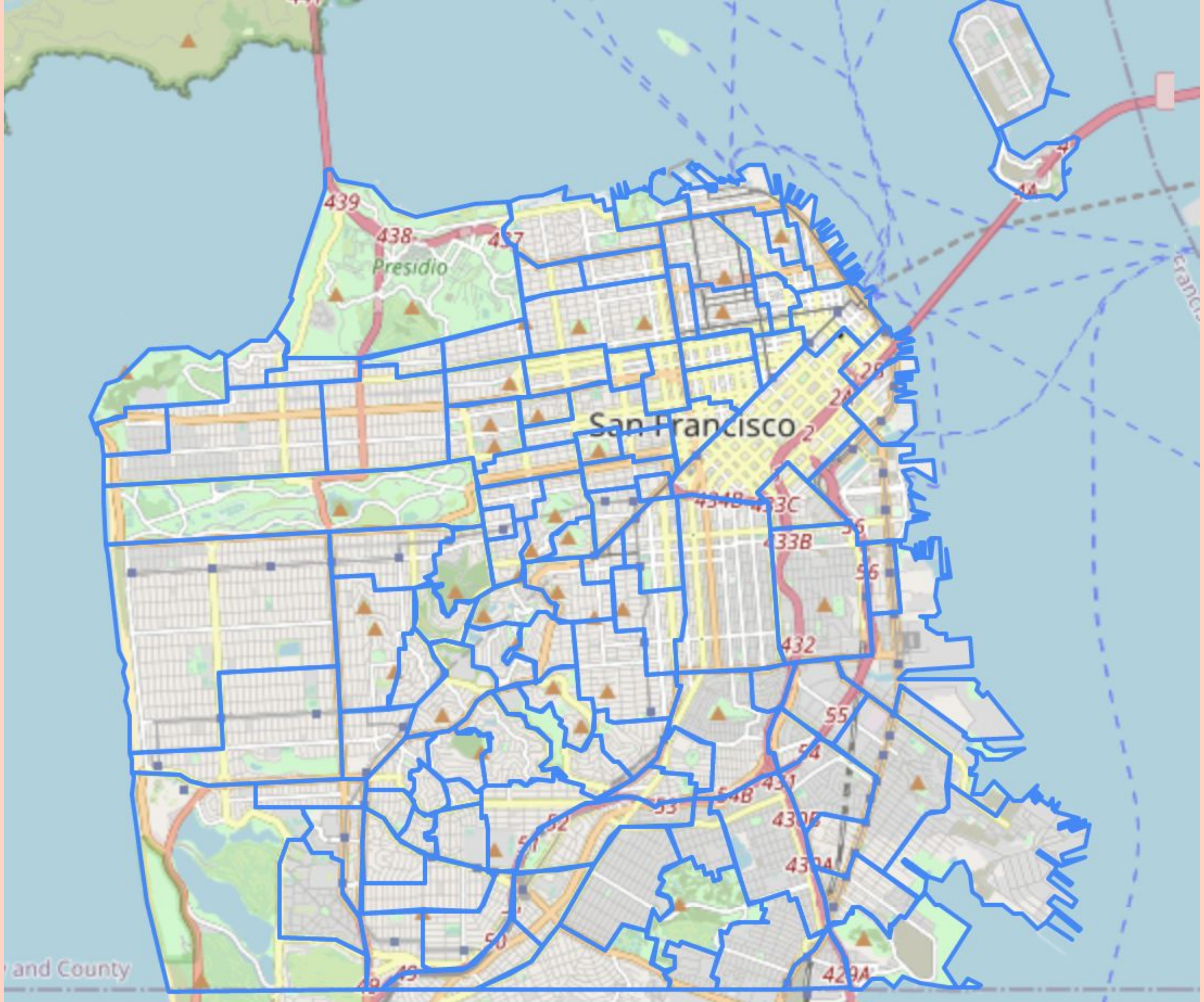


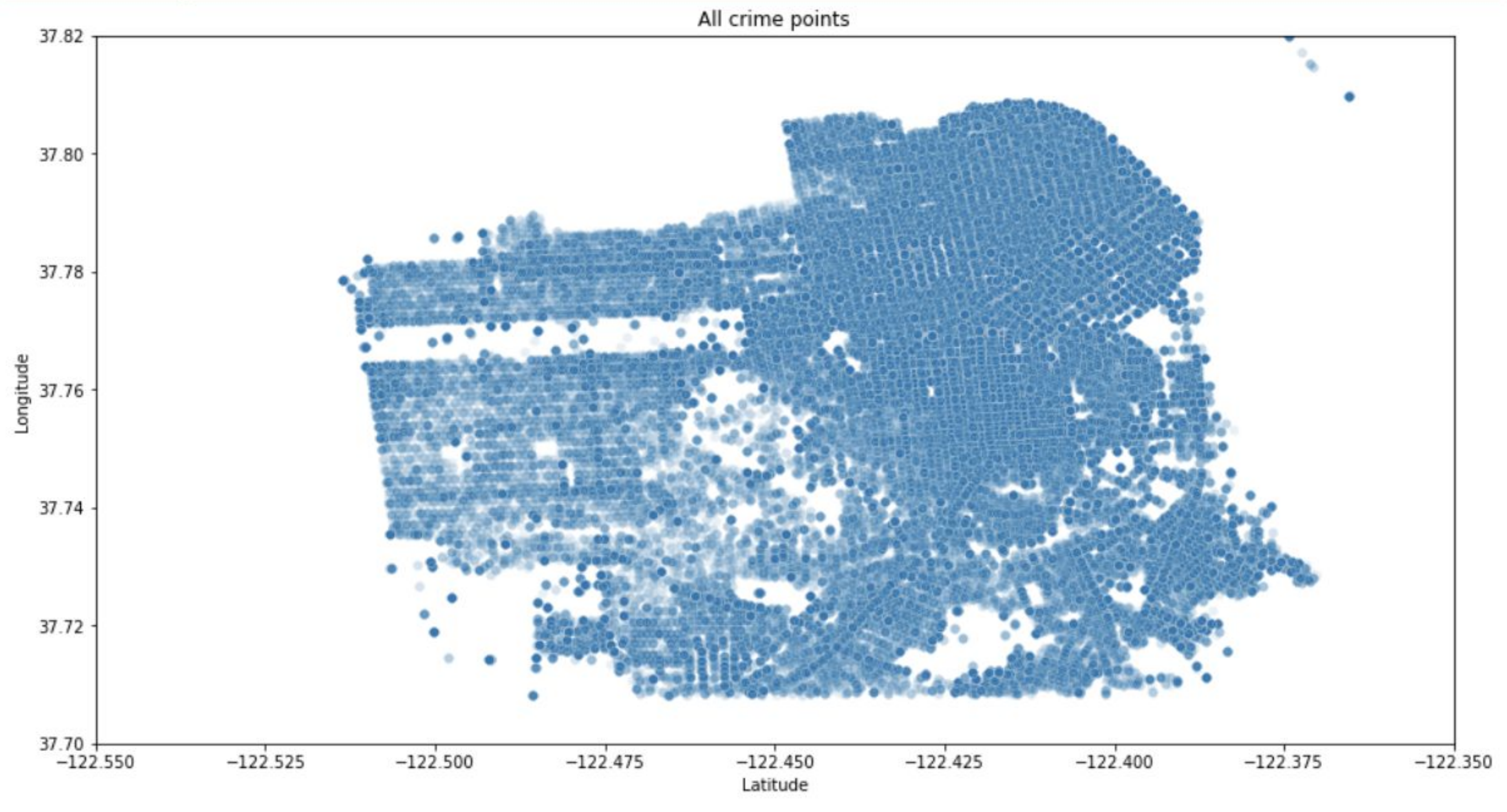




Crime between 11pm-6am for Saturdays in 2018

- Police
- Non Urgent PD
- Specialty Services


San Francisco
Presidio
and County


All crime points
Longitude
Latitude
37.82
37.80
37.78
37.76
37.74
37.72
37.70
−122.550
−122.525
−122.500
−122.475
−122.450
−122.425
−122.400
−122.375
−122.350

Shapely Example

```
geom = [Point(xy) for xy in zip(crime_df['X'], crime_df['Y'])]
```

```
for i in range(len(neighb_geo)):
    crime_geo.loc[crime_geo.intersects(neighb_geo.loc[i, 'geometry']), 'neighbourhood'] = neighb_geo['name'].loc[i]
```

Folium Example

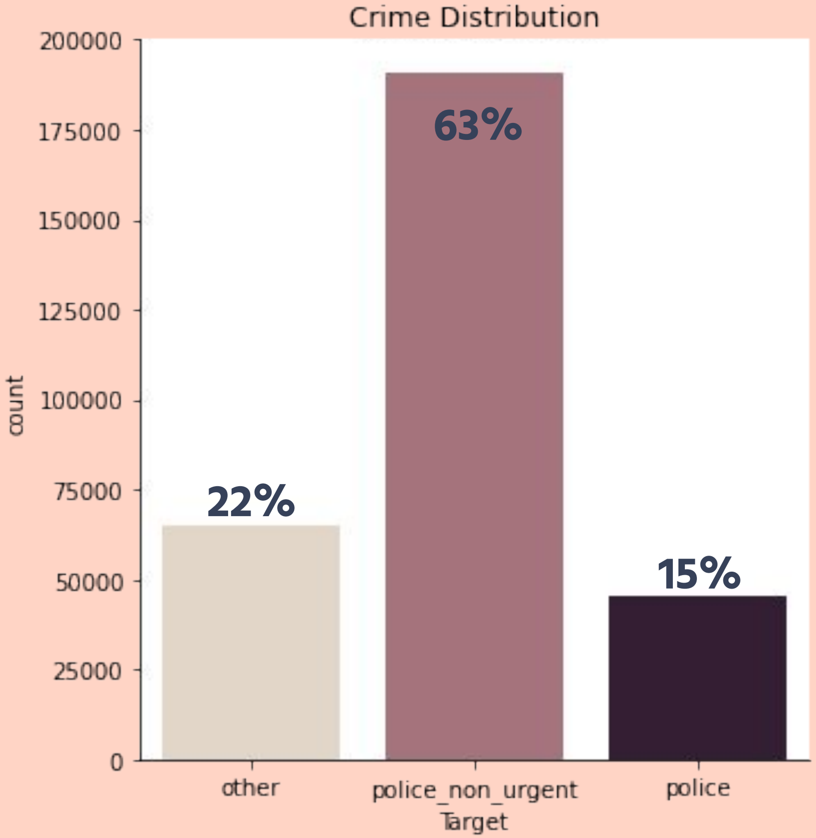
```
sf_map_1 = folium.Map(location=[37.8, -122.5])

pnu.apply(lambda row:folium.CircleMarker(location=[row['Y'], row['X']], radius=1, color='yellow').add_to(sf_map_1), axis=1)
sp_sup.apply(lambda row:folium.CircleMarker(location=[row['Y'], row['X']], radius=1, color='green').add_to(sf_map_1), axis=1)
pol.apply(lambda row:folium.CircleMarker(location=[row['Y'], row['X']], radius=1, color='red').add_to(sf_map_1), axis=1)

neighb_geo.apply(lambda row:folium.Polygon(locations=[row['lon_lat']], weight=2).add_to(sf_map_1), axis=1)

sf_map_1
```

Greatly Unbalanced Target Class



Model Results Summary

	TRAIN	TEST
Logistic	53%	53%
Bagging	88%	68%
Random Forest	89%	69%
AdaBoost	48%	47%
XGBoost	55%	54%

Modeling

- Over 10,000 model combinations
- Baseline Accuracy: 33%
- Best Model Accuracy: 55%,

  +24% over baseline
- XGBoost

Model Feature Weights

Weight	Feature
0.0163	neighbourhood_South of Market
0.0138	neighbourhood_Downtown / Union Square
0.0129	neighbourhood_Fishermans Wharf
0.0127	neighbourhood_Eureka Valley
0.0122	neighbourhood_Central Waterfront
0.0122	neighbourhood_Chinatown
0.0122	neighbourhood_Sunnyside
0.0120	neighbourhood_Mint Hill
0.0119	neighbourhood_Upper Market
0.0003	DayOfWeek_Saturday
0.0003	DayOfWeek_Tuesday
0.0003	DayOfWeek_Wednesday
0.0003	DayOfWeek_Monday
0.0002	Month
0.0002	Day
0.0002	DayOfWeek_Sunday
0.0002	Year_2018
0	neighbourhood_Candlestick Point SRA
0	neighbourhood_Sherwood Forest
0	neighbourhood_Treasure Island
0	neighbourhood_Westwood Highlands

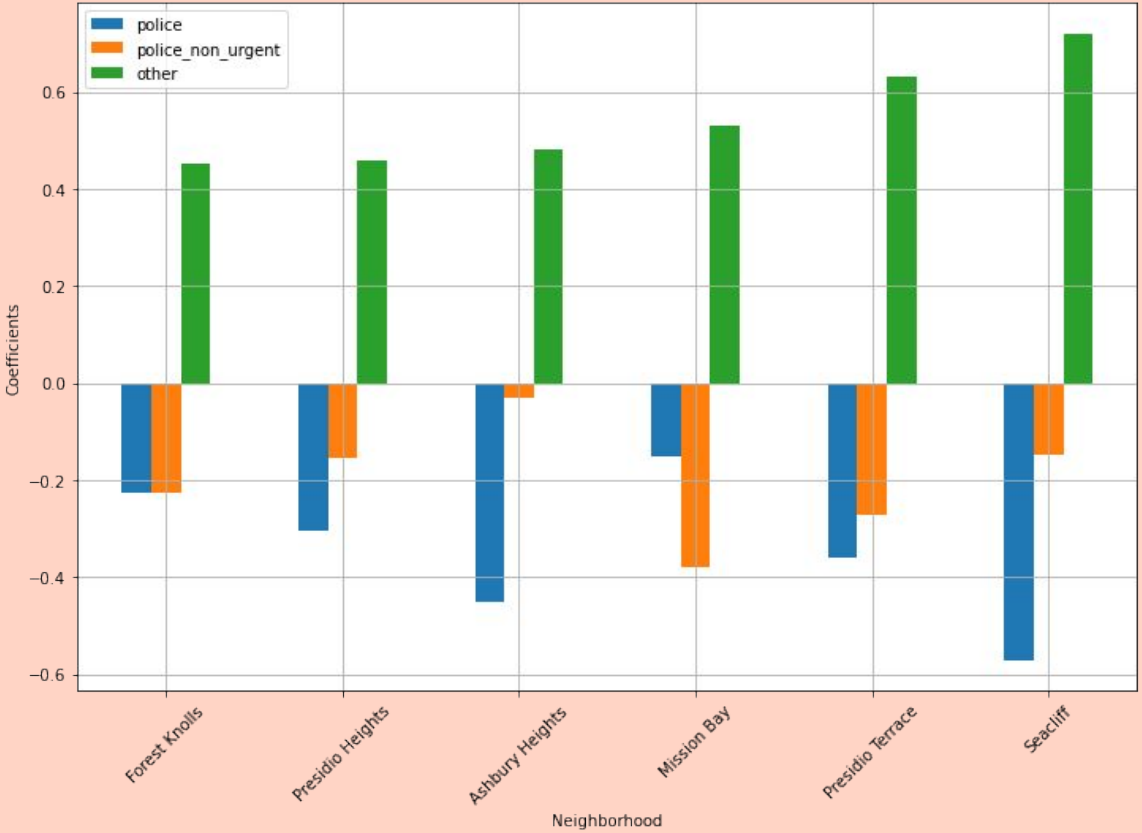
S. of Market Crimes:

- Total: 43,016
- Police: 33,832
- Specialty: 9,184

Treasure Island Crimes:

- Total: 25
- Police: 21
- Specialty: 4

Model Coefficients



Predictive Power:

- Can provide log odds of crime category by neighborhood

Conclusions/Future Considerations

- More robust feature engineering
- Refining target categories
- More NLP on crime description column
- Investigate crime resolution vs neighborhood demographics

Thank You!
Questions?

Nelli Aydinyan, Jessica Bow,
Gabe Cano, Billy Horn