Understanding and Predicting Crime in San Francisco

Capstone Project

Elton Yeo, DSI 13, General Assembly

The City of San Francisco and San Francisco Police Dept have limited resources…

- Where should they focus their efforts to reduce crime in the city and increase safety?
- How can they modify their policy or operational approaches to reduce crime?

This project has two objectives:

1. **Develop insights into the areas with the most crimes**, and consider alternative policy and operational measures to reduce crime

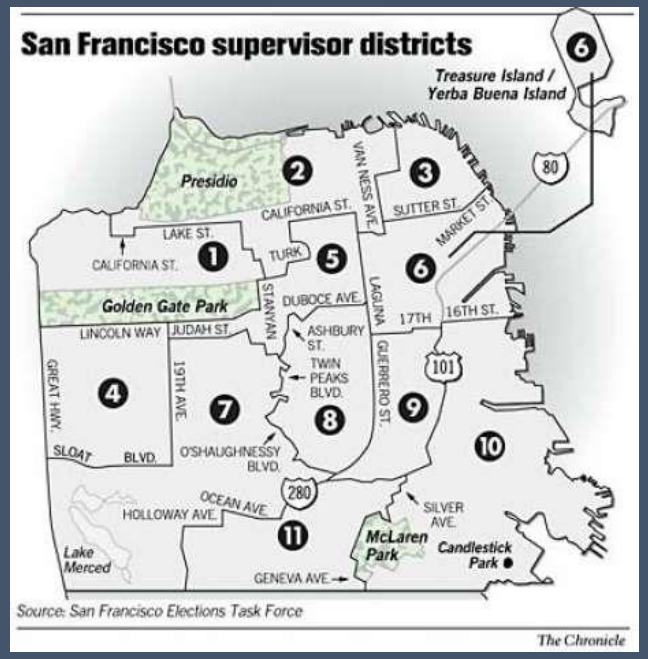
2. **Predict the number of preventable crime** given a specific zipcode, day of the week, and hour, which will help the police plan their patrol schedules and resources to reduce crime

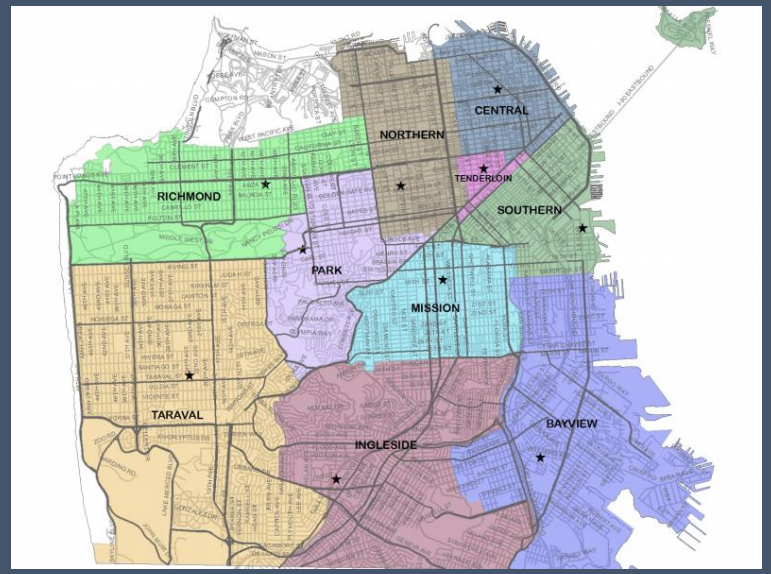
Data


SFGov
Coordinator's Portal
About
Help
DataSF
OPEN DATA
SHOWCASE
PUBLISHING
ACADEMY
RESOURCES
BLOG
Explore
Browse Data
Open Data Stats
Developers
Sign In
Police Department Incident Reports: 2018 to Present
Public Safety
View Data
Visualize
Export
API
For a detailed overview, see: https://support.datasf.org/help/police-department-incident-reports-2018-to-present-overview
Updated
April 22, 2020
This dataset includes police incident reports filed by officers and by individuals through self-service online reporting for non-emergency cases. Reports included are those for incidents that occurred starting January 1, 2018 onward and have been approved by a supervising officer.

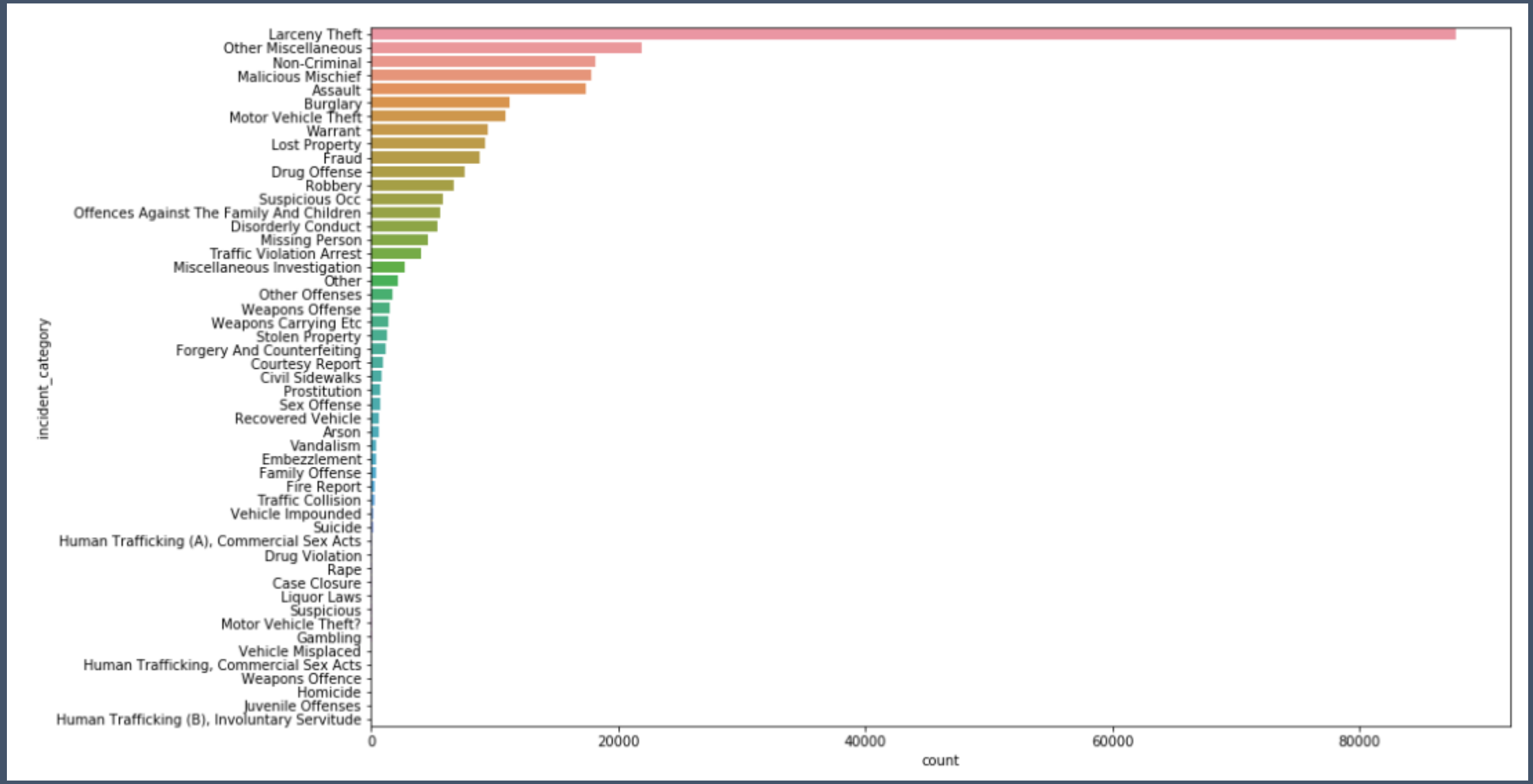
Making sense of our data e.g. locations

- Latitude, longitude, intersection, sf find neighbourhood, analysis neighbourhood, analysis neighbourhood**s**, current police district, current supervisor districts, supervisor district – what do these mean and are they important for our analysis and prediction?

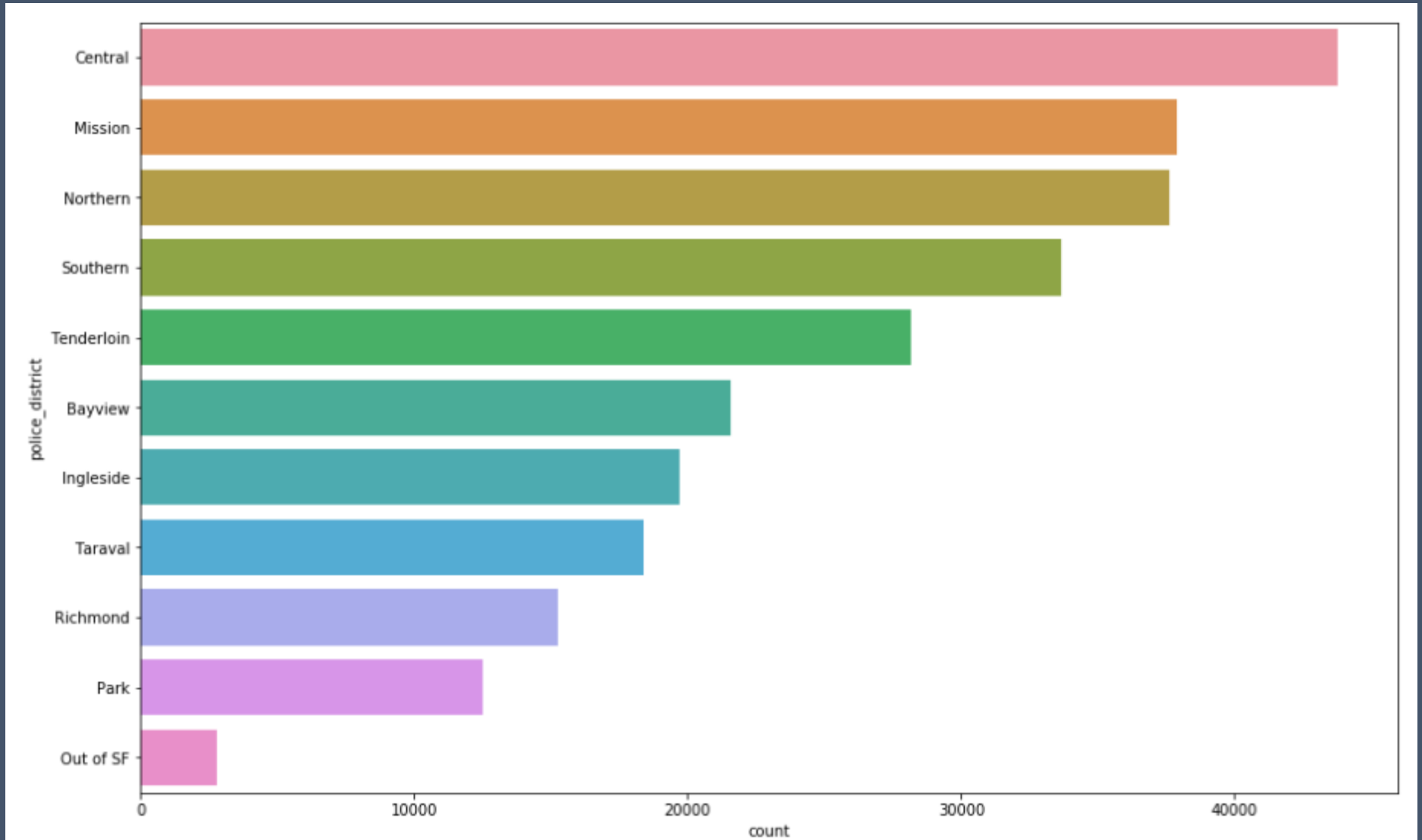




Theft occurred the most frequently – but how do we treat these different categories of crime?



Central police district had the highest number of crimes

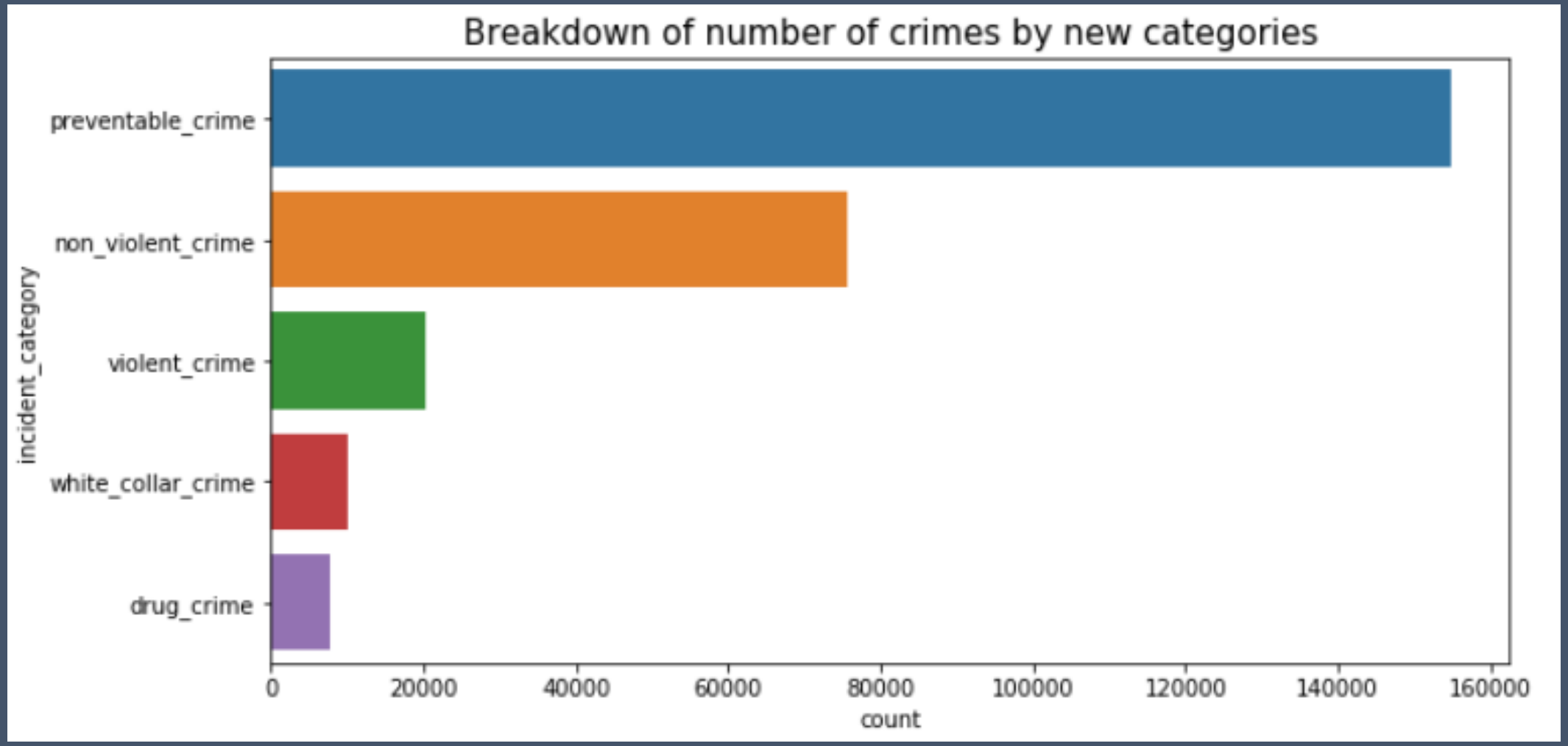


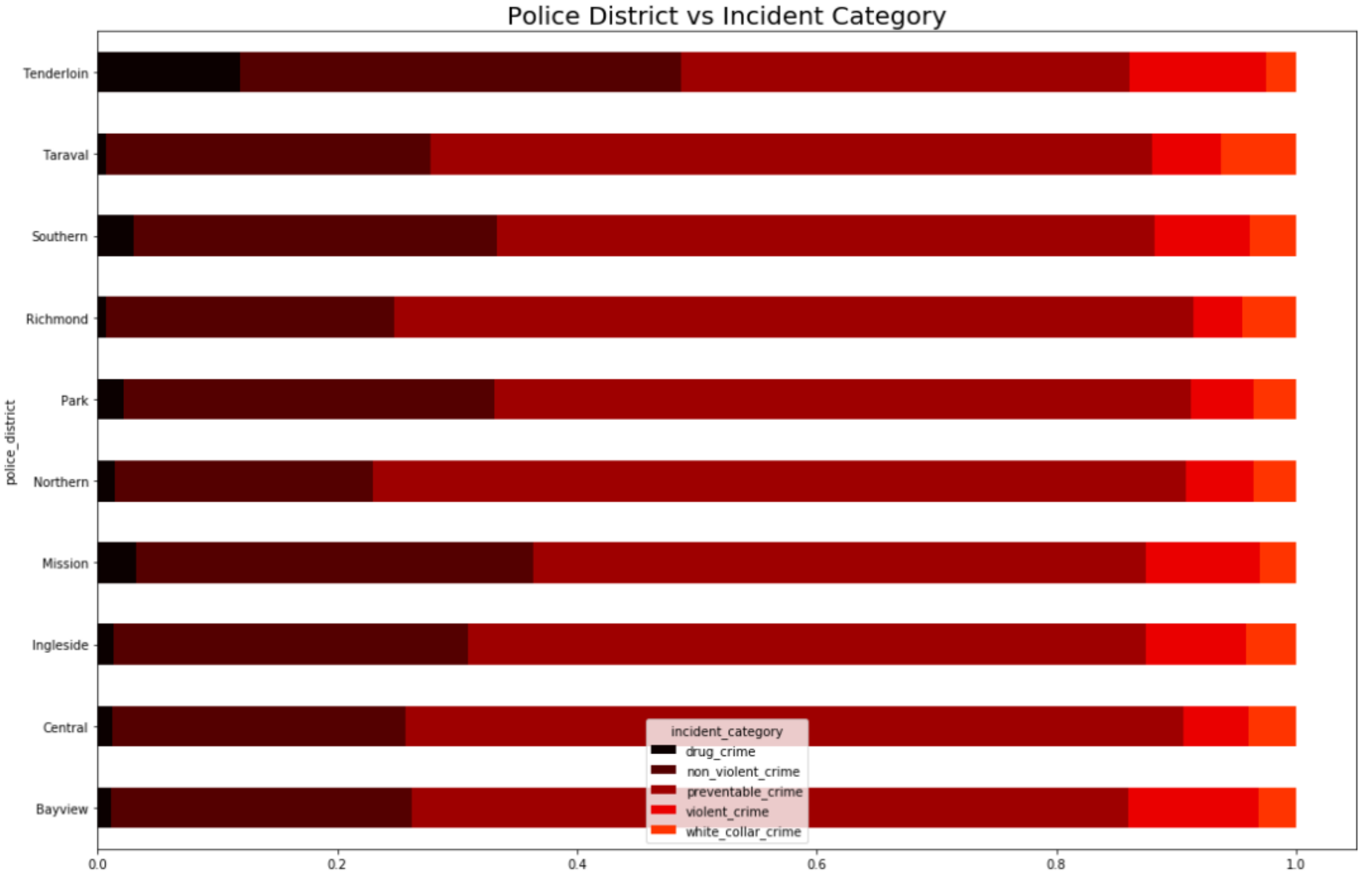

Police District vs Incident Category
incident_category
Weapons Offense
Weapons Offence
Weapons Carrying Etc
Warrant
Vehicle Misplaced
Vehicle Impounded
Vandalism
Traffic Violation Arrest
Traffic Collision
Suspicious Occ
Suspicious
Suicide
Stolen Property
Sex Offense
Robbery
Recovered Vehicle
Rape
Prostitution
Other Offenses
Other Miscellaneous
Other
Offences Against The Family And Children
Non-Criminal
Motor Vehicle Theft?
Motor Vehicle Theft
Missing Person
Miscellaneous Investigation
Malicious Mischief
Lost Property
Liquor Laws
Larceny Theft
Juvenile Offenses
Human Trafficking, Commercial Sex Acts
Human Trafficking (B), Involuntary Servitude
Human Trafficking (A), Commercial Sex Acts
Homicide
Gambling
Fraud
Forgery And Counterfeiting
Fire Report
Family Offense
Embezzlement
Drug Violation
Drug Offense
Disorderly Conduct
Courtesy Report
Civil Sidewalks
Case Closure
Burglary
Assault
Arson
0.0
0.2
0.4
0.6
0.8
1.0
police_district
Bayview
Central
Ingleside
Mission
Northern
Out of SF
Park
Richmond
Southern
Taraval
Tenderloin

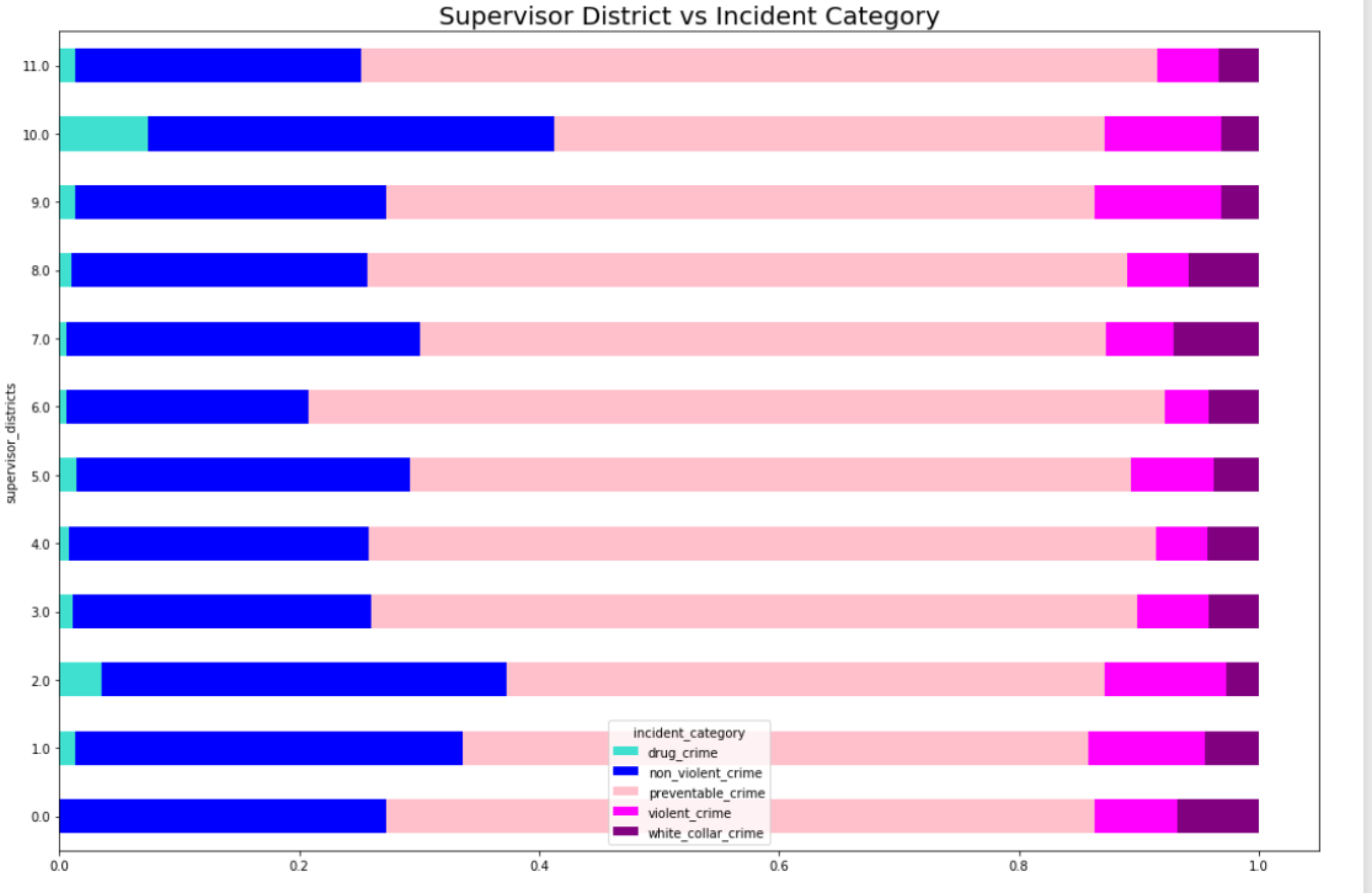
Can we group crime into more meaningful categories for analysis and prediction?

Preventable Crime	Violent Crime	Drug Crime	White-Collar Crime	Non-Violent Crime
• 'Larceny Theft', 'Malicious Mischief', 'Burglary', 'Motor Vehicle Theft', 'Robbery', 'Stolen Property', 'Disorderly Conduct', 'Vandalism', 'Arson', 'Motor Vehicle Theft', 'Suspicious Occ', 'Offences Against The Family And Children', 'Suspicious', 'Family Offense', 'Prostitution', 'Civil Sidewalks'	• 'Assault', 'Weapons Offense', 'Weapons Carrying Etc', 'Rape', 'Weapons Offence', 'Homicide'	• 'Drug Offense', 'Drug Violation'	• 'Fraud', 'Forgery And Counterfeiting', 'Embezzlement'	• 'Lost Property', 'Missing Person', 'Recovered Vehicle', 'Traffic Collision', 'Vehicle Impounded', 'Suicide', 'Vehicle Misplaced', 'Traffic Violation Arrest', 'Non-Criminal', 'Other Miscellaneous', 'Warrant', 'Gambling', 'Human Trafficking, Commercial Sex Acts', 'Human Trafficking (B), Involuntary Servitude', 'Juvenile Offenses', 'Liquor Laws', 'Case Closure', 'Human Trafficking (A), Commercial Sex Acts', 'Fire Report', 'Sex Offense', 'Courtesy Report', 'Other Offenses', 'Other', 'Miscellaneous Investigation'

Preventable crimes were the most frequently committed category of crime

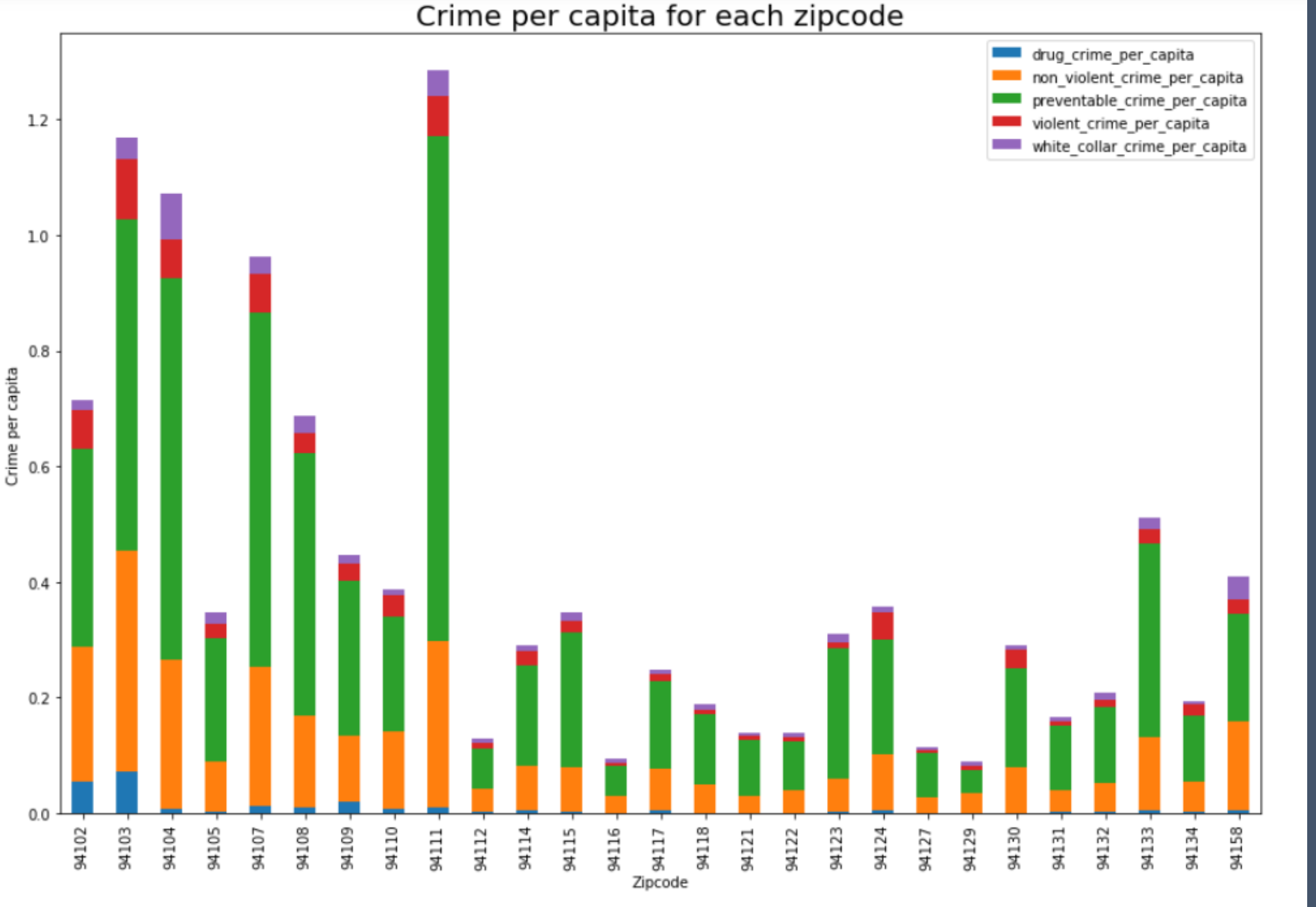


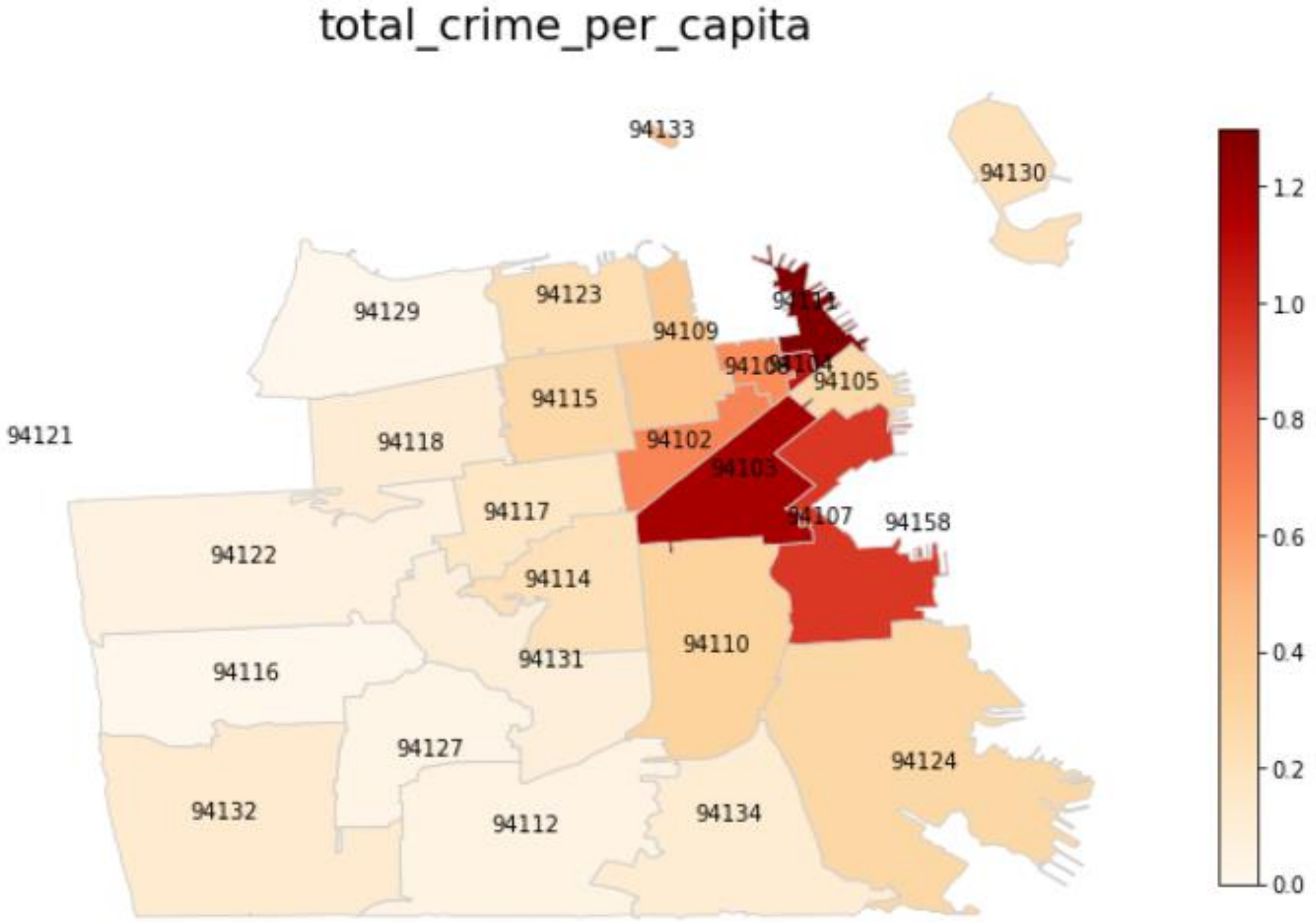

Police District vs Incident Category
police_district
Tenderloin
Taraval
Southern
Richmond
Park
Northern
Mission
Ingleside
Central
Bayview
0.0
0.2
0.4
0.6
0.8
1.0
incident_category
drug_crime
non_violent_crime
preventable_crime
violent_crime
white_collar_crime

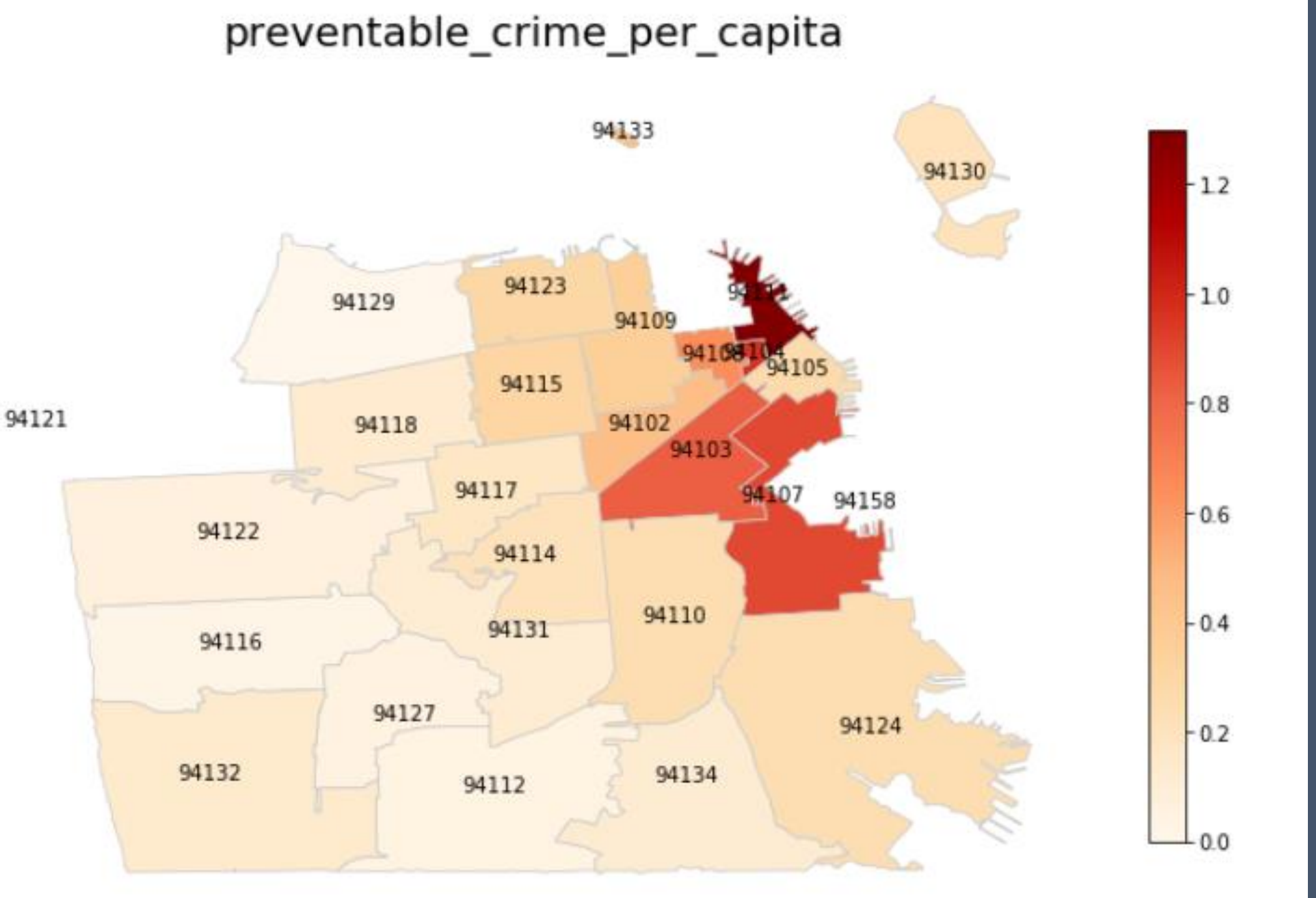

Supervisor District vs Incident Category
supervisor_districts
11.0
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0.0
0.2
0.4
0.6
0.8
1.0
incident_category
drug_crime
non_violent_crime
preventable_crime
violent_crime
white_collar_crime

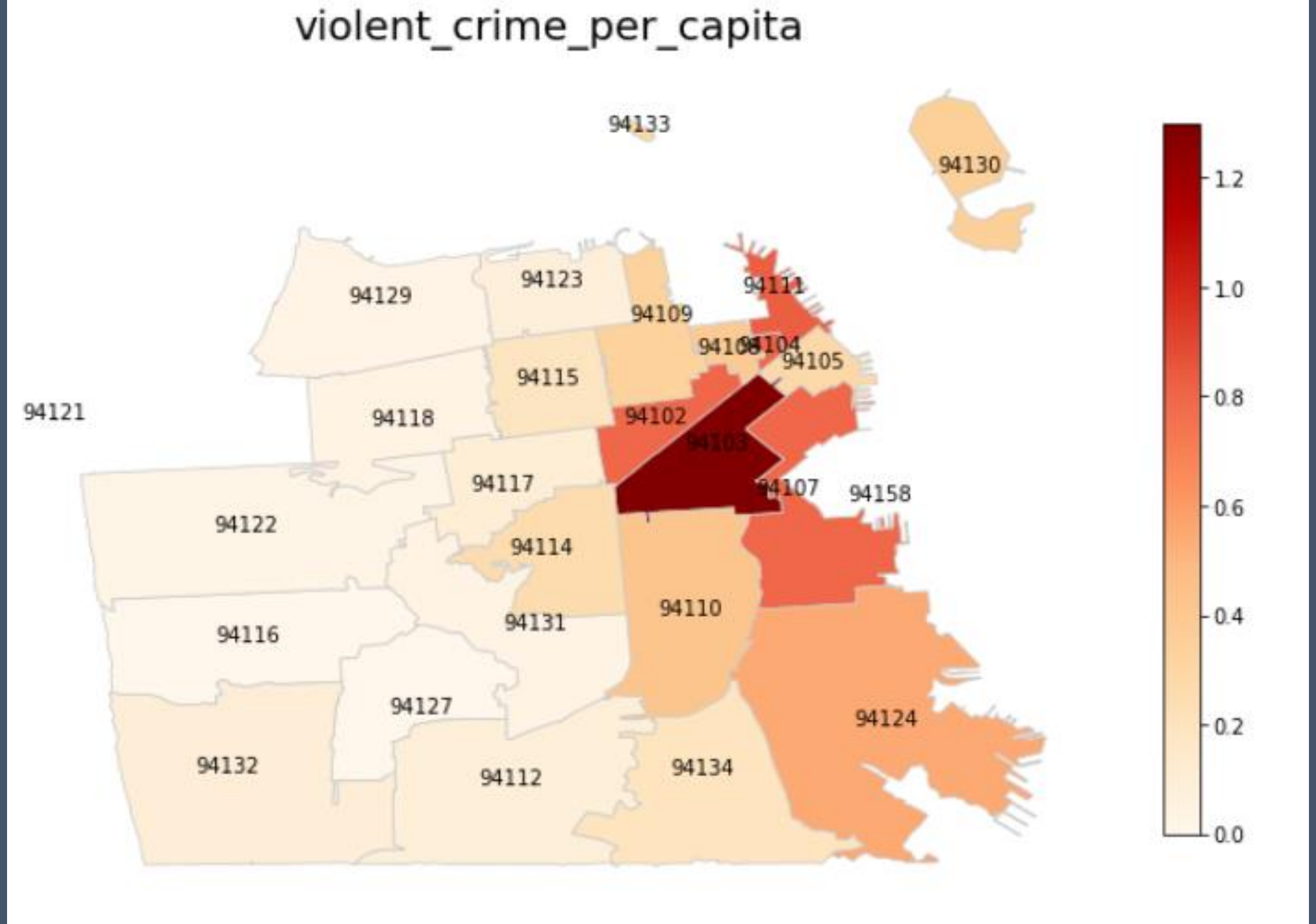
But what about the number of crimes?

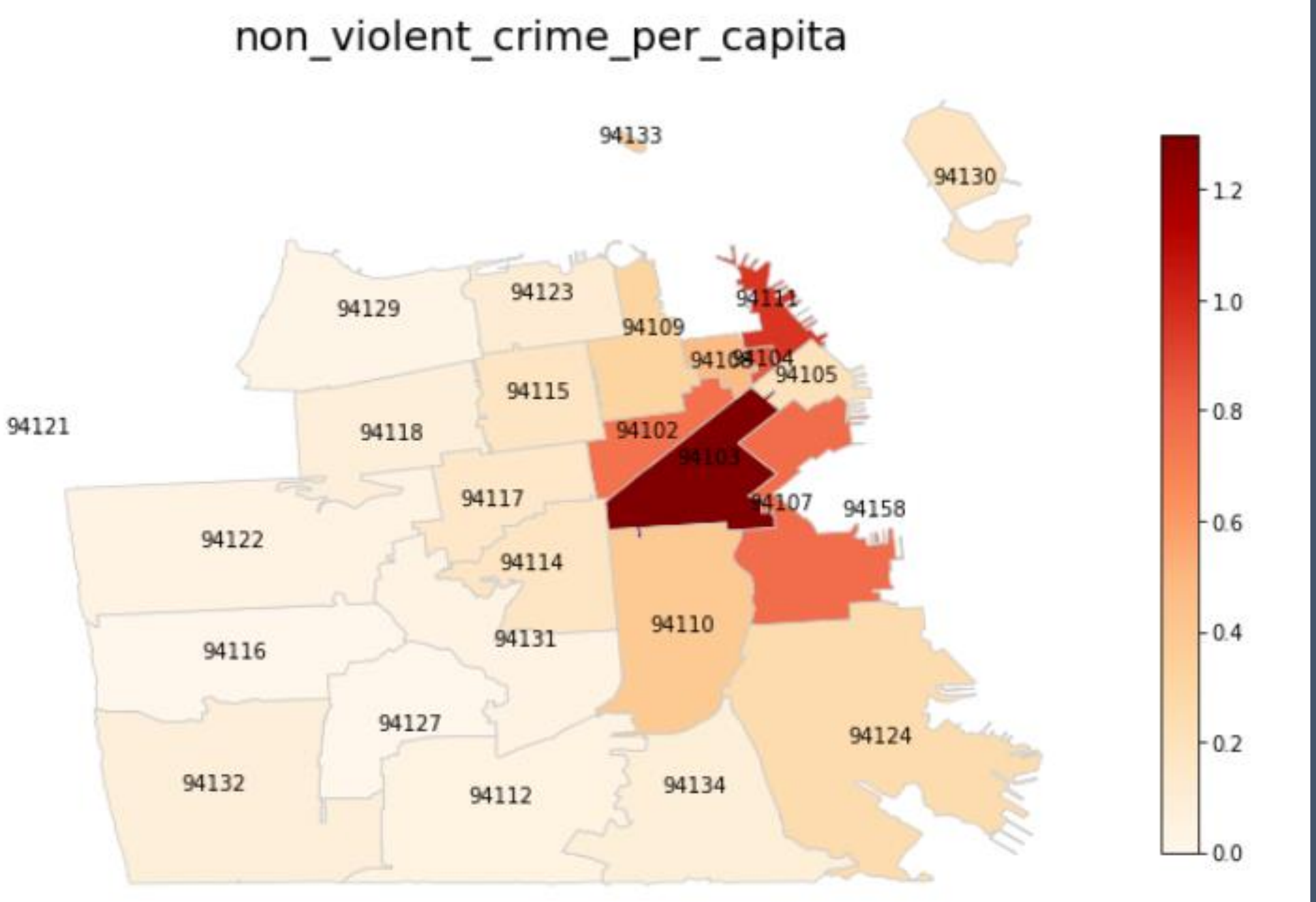
- How can we meaningfully compare the number, and not just the proportion, of crimes across different districts?

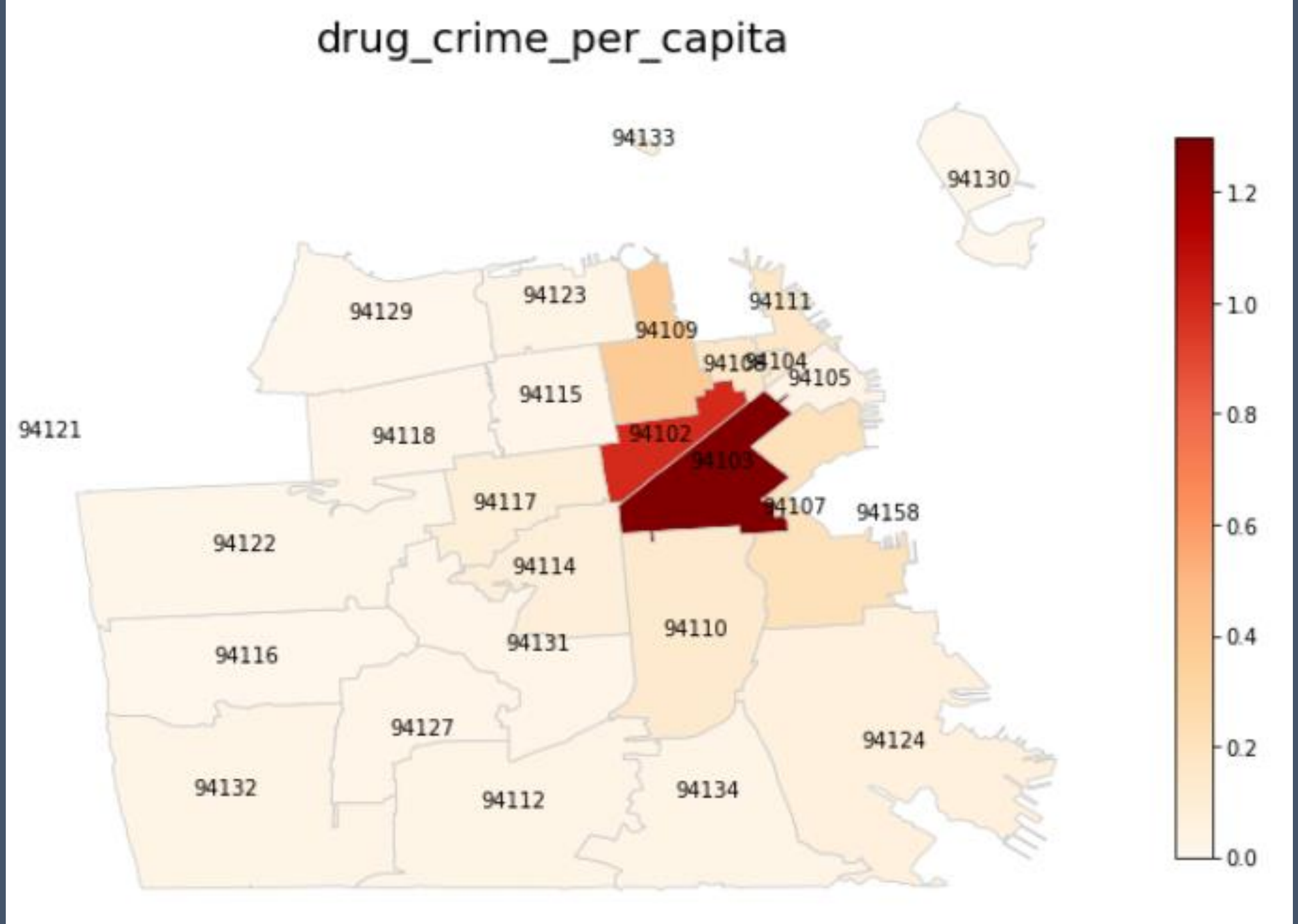

Crime per capita for each zipcode
drug_crime_per_capita
non_violent_crime_per_capita
preventable_crime_per_capita
violent_crime_per_capita
white_collar_crime_per_capita
Crime per capita
0.0
0.2
0.4
0.6
0.8
1.0
1.2
94102
94103
94104
94105
94107
94108
94109
94110
94111
94112
94114
94115
94116
94117
94118
94121
94122
94123
94124
94127
94129
94130
94131
94132
94133
94134
94158
Zipcode

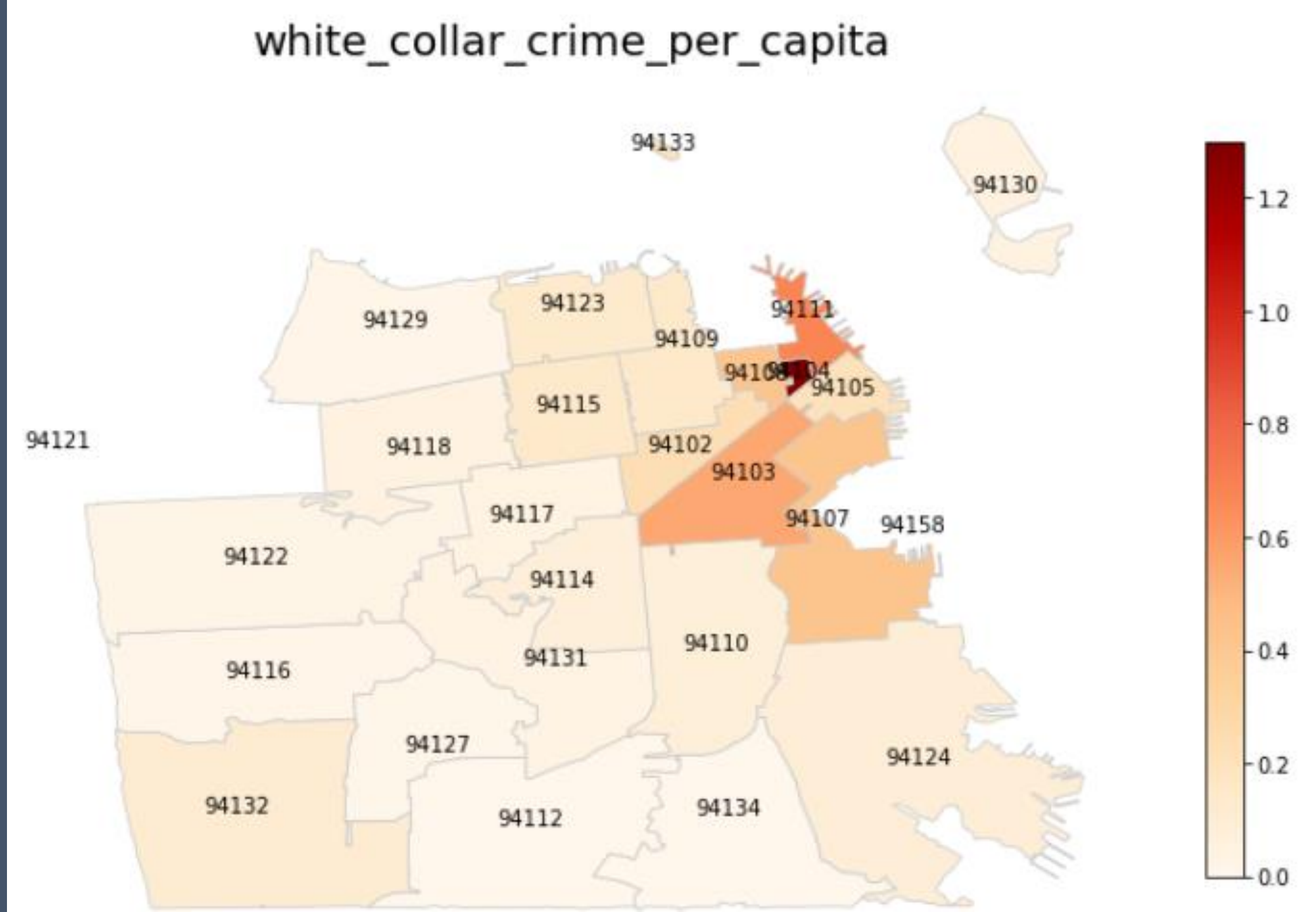

total_crime_per_capita
94133
94130
94129
94123
94109
94115
94118
94121
94102
94103
94117
94107
94158
94122
94114
94110
94131
94116
94127
94124
94132
94112
94134
94105
1.2
1.0
0.8
0.6
0.4
0.2
0.0


preventable_crime_per_capita
94133
94130
94129
94123
94109
94108
94104
94105
94115
94121
94118
94102
94103
94117
94107
94158
94122
94114
94110
94131
94116
94127
94124
94132
94112
94134
1.2
1.0
0.8
0.6
0.4
0.2
0.0


violent_crime_per_capita
94133
94130
94129
94123
94111
94109
94108
94104
94105
94115
94121
94118
94102
94103
94117
94107
94158
94122
94114
94110
94131
94116
94127
94124
94132
94112
94134
1.2
1.0
0.8
0.6
0.4
0.2
0.0


non_violent_crime_per_capita
94133
94130
94129
94123
94111
94109
94108
94104
94105
94115
94121
94118
94102
94103
94117
94107
94158
94122
94114
94110
94131
94116
94127
94124
94132
94112
94134
1.2
1.0
0.8
0.6
0.4
0.2
0.0

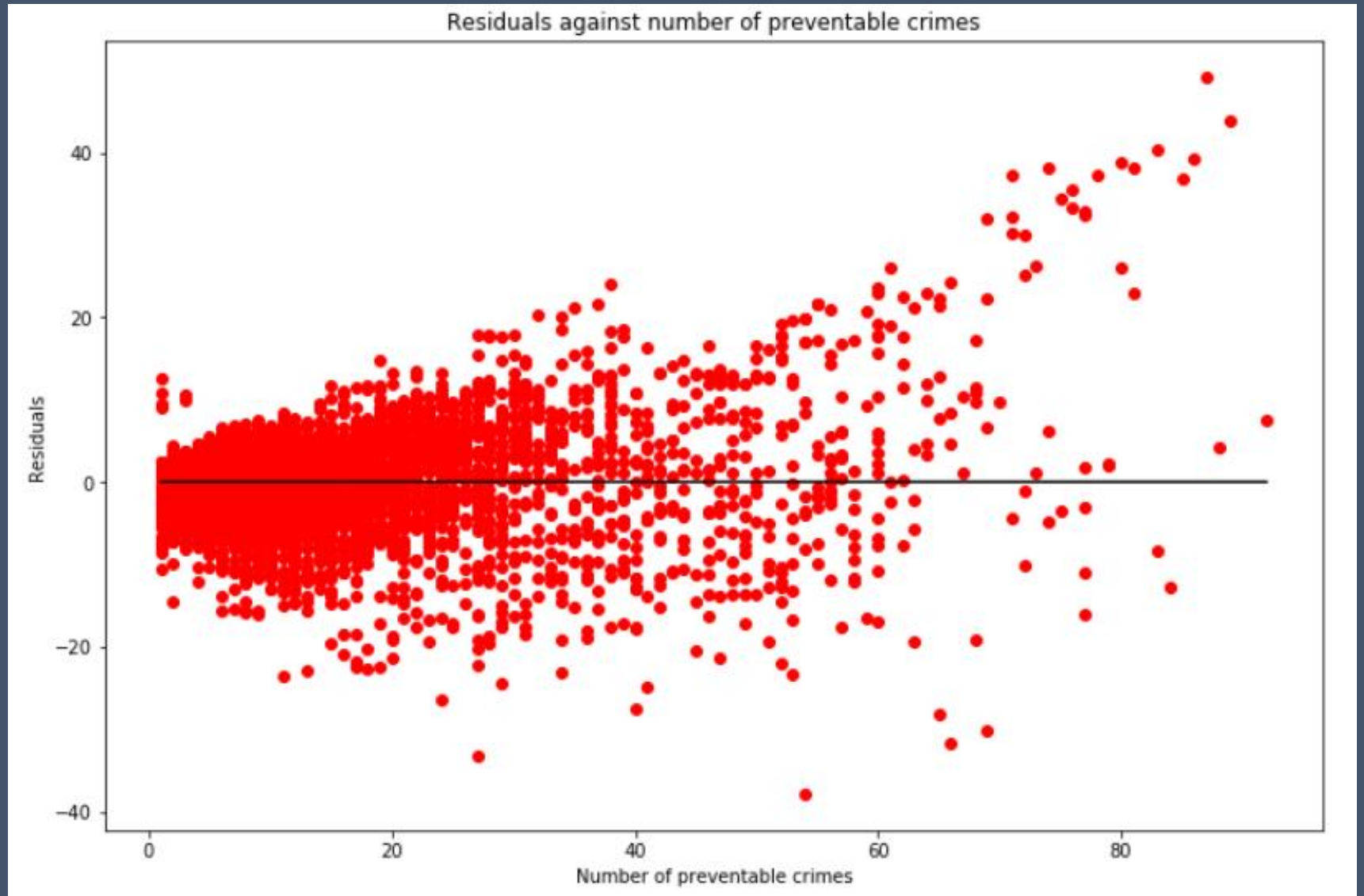

drug_crime_per_capita
94133
94130
94129
94123
94111
94109
94108
94104
94105
94115
94121
94118
94102
94103
94117
94107
94158
94122
94114
94110
94131
94116
94127
94124
94132
94112
94134
1.2
1.0
0.8
0.6
0.4
0.2
0.0


white_collar_crime_per_capita
94133
94130
94129
94123
94111
94109
94105
94115
94121
94118
94102
94103
94117
94107
94158
94122
94114
94110
94116
94131
94127
94124
94132
94112
94134
1.2
1.0
0.8
0.6
0.4
0.2
0.0

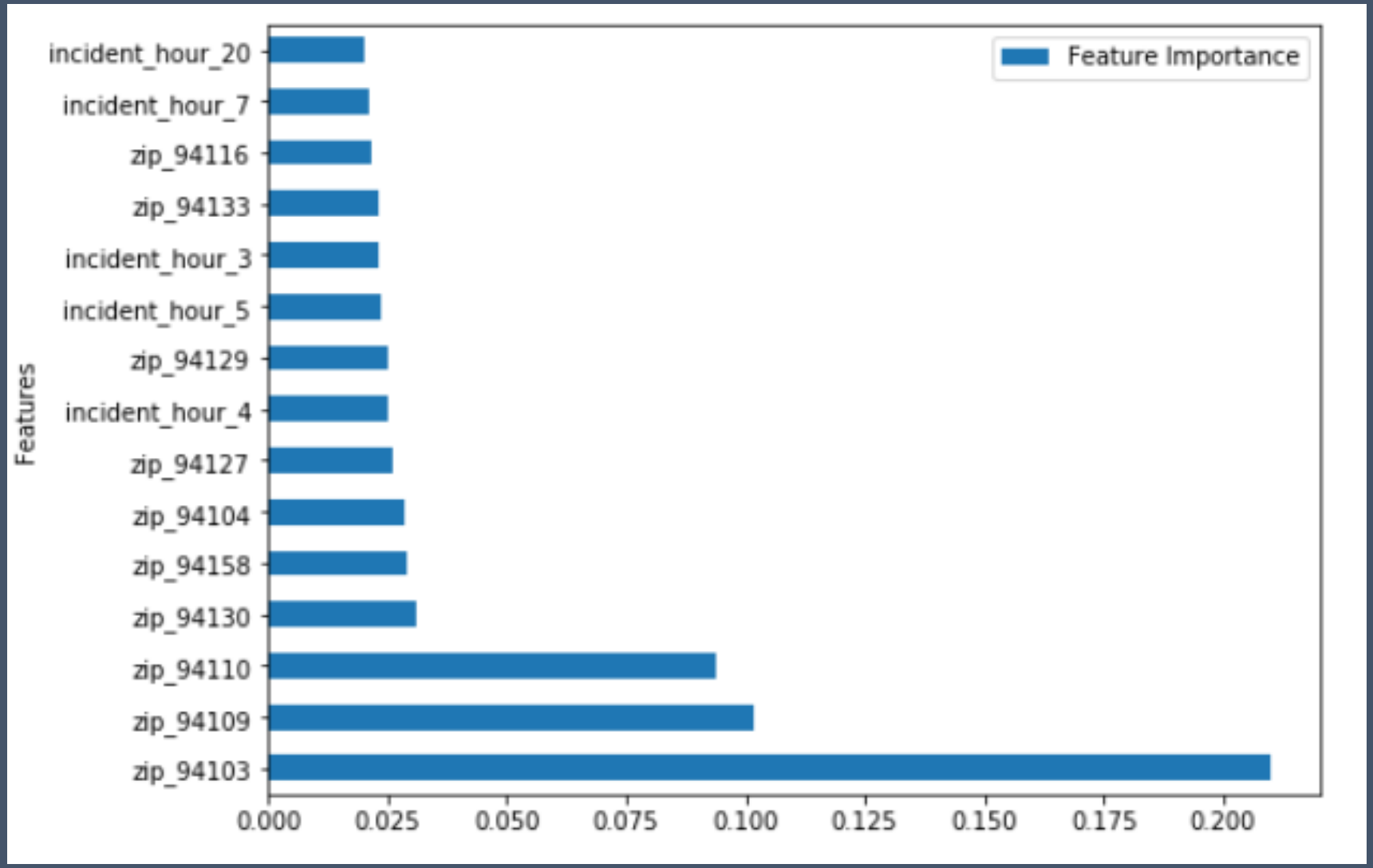
Predicting preventable crimes based on zipcode, day of week, and hour of day

Model trained on 2018 data	Mean Squared Error for X_test, y_test	R^2 score for X_test, y_test
Baseline	247	-
Linear Regression	65.9	0.733
LassoCV	65.7	0.734
RidgeCV	65.8	0.734
Random Forest Regressor with GridsearchCV	47.1	0.809
XGBoost with GridsearchCV	38.3	0.845

Success with R^2 score of 0.8 when XGBoost model is used to predict from test data



Zipcodes were the most important features in predicting the number of preventable crimes



Practical Steps for SF Police to take

	incident_day_of_week	zip	incident_hour	preventable_crime	xgb_predict
2470	Thursday	94103	19	1.326923	1.907094
1260	Saturday	94103	19	1.269231	1.876870
43	Friday	94103	19	1.615385	1.858716
3079	Tuesday	94103	19	1.250000	1.791903
1259	Saturday	94103	18	1.480769	1.787161
1871	Sunday	94103	19	1.038462	1.766315
3694	Wednesday	94103	19	1.596154	1.757343
1261	Saturday	94103	20	1.480769	1.690898
648	Monday	94103	19	1.307692	1.676706
3078	Tuesday	94103	18	1.769231	1.623973

Next Steps

- Include further demographic data
- Include more years of data
- Corroborate quantitative insights with qualitative experiences of Police officers ad policymakers

Any questions?