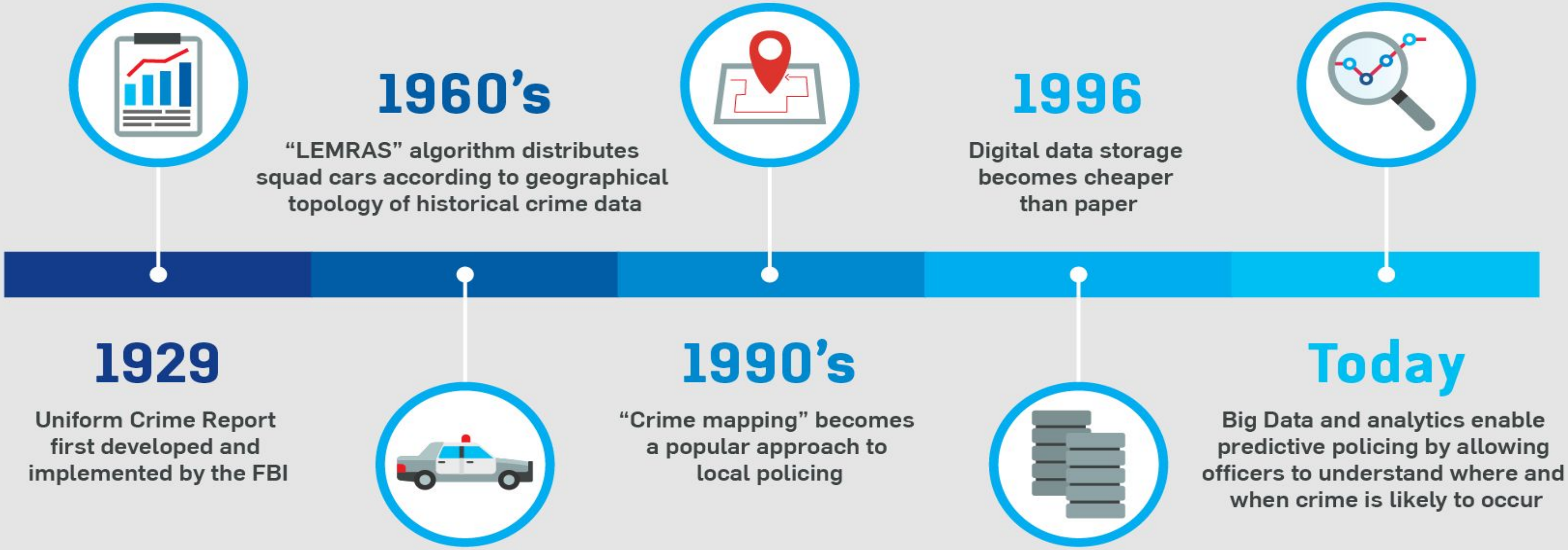
PREDICTIVE POLICING

PREDICTING CRIMES IN SAN FRANCISCO

WHAT IS PREDICTIVE POLICING?

Data plays an important role in allocating police resources and lowering crime rates:



1929

Uniform Crime Report first developed and implemented by the FBI

1960's

"LEMRAS" algorithm distributes squad cars according to geographical topology of historical crime data

1990's

"Crime mapping" becomes a popular approach to local policing

1996

Digital data storage becomes cheaper than paper

Today

Big Data and analytics enable predictive policing by allowing officers to understand where and when crime is likely to occur

Image from Coolfire Solutions

San Francisco is the cultural, commercial, and financial center of Northern California. San Francisco is the 15th most populous city in the United States, and the fourth most populous in California, with 881,549 residents as of 2019.

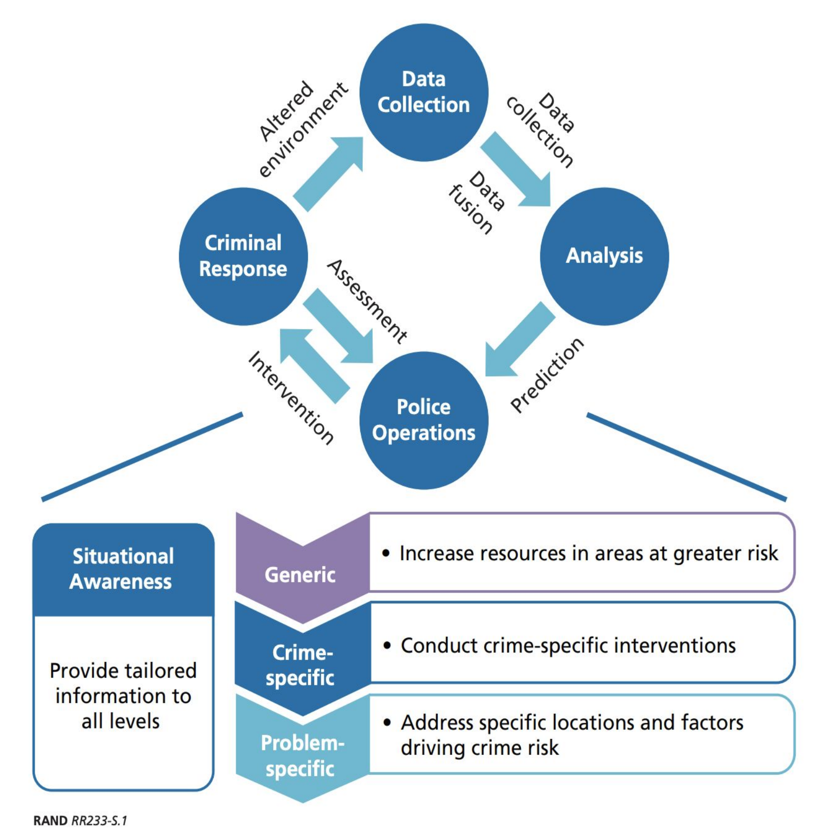
Golden Gate Bridge
Emeryville
24
80
580
Piedmont
880
Oakland
NORTH BEACH
FINANCIAL DISTRICT
SOMA
San Francisco
Alameda
61
MISSION DISTRICT
1
280
Bay Farm Island
Daly City
101
Brisbane
San Francisco Bay
Colma
1
Map data ©2020 Google

> “As data piles up, we have ourselves a genuine gold rush. But data isn't gold. I repeat, data in its raw form is boring crud. The gold is what's discovered therein.”
>
> – Eric Siegel, *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie or Die.*

Why is Predictive Policing important and how does it work ?

Law Enforcement is facing unprecedented new challenges. They are under increasing public scrutiny while expected to deal with a growing number of threats even as budgets continue to shrink.

Predictive policing allows law enforcement officers to make the most of their limited resources by deploying them more accurately in place and time.



- Increase resources in areas at greater risk
- Conduct crime-specific interventions
- Address specific locations and factors driving crime risk

RAND *RR233-S.1*

Our Goal

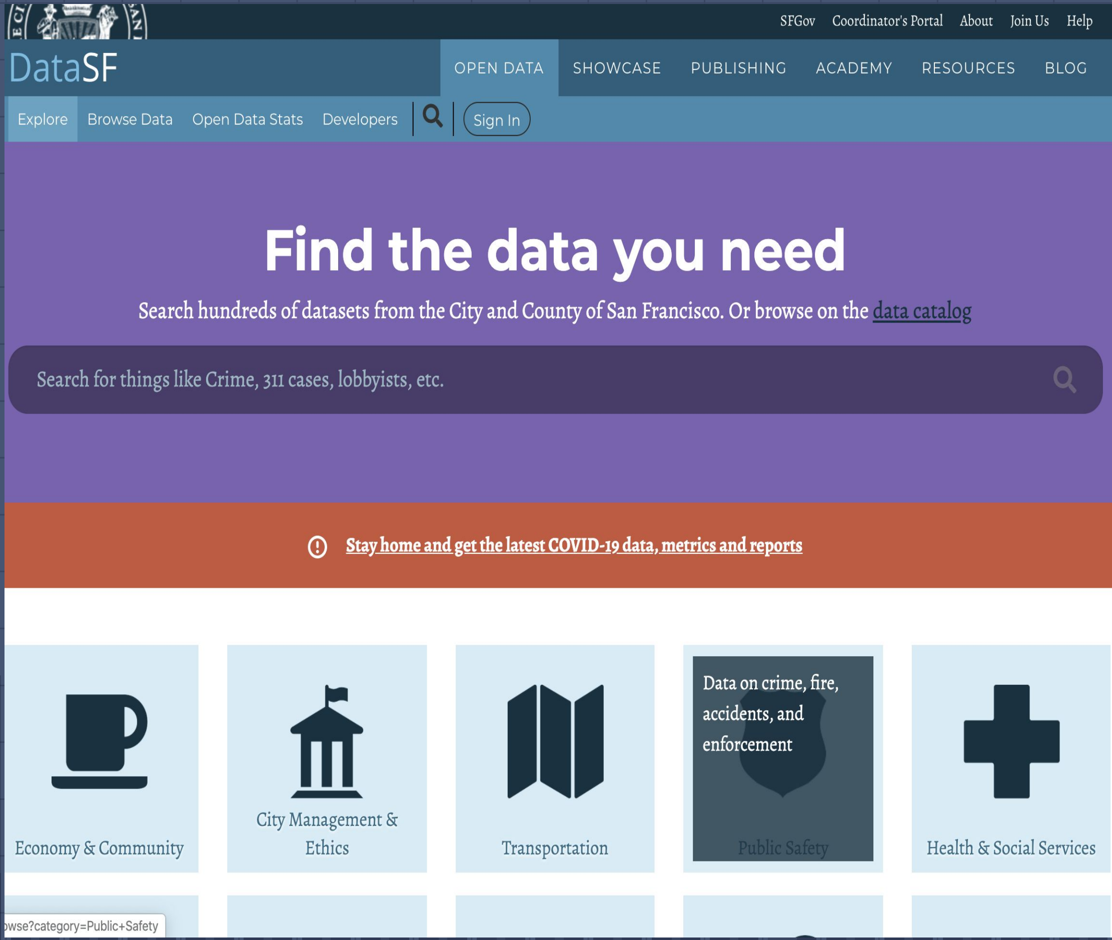
Study the crimes trend in the different districts of San Francisco.

Predict crimes and make a comparison between the predicted and the actual crimes that happened in 2019

Dataset

The *Data* used for this project is from the open data project by the city and county of San Francisco.

The Police Department Incident Reports from 2003 to present is available.

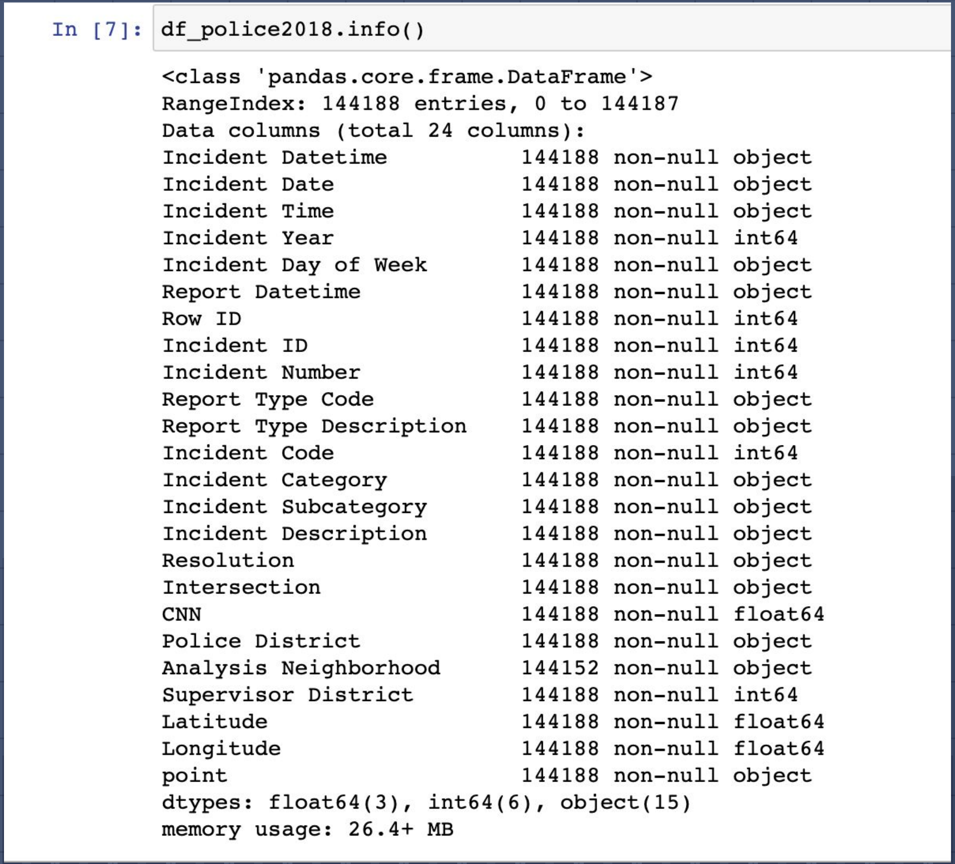


Find the data you need

Search hundreds of datasets from the City and County of San Francisco. Or browse on the data catalog

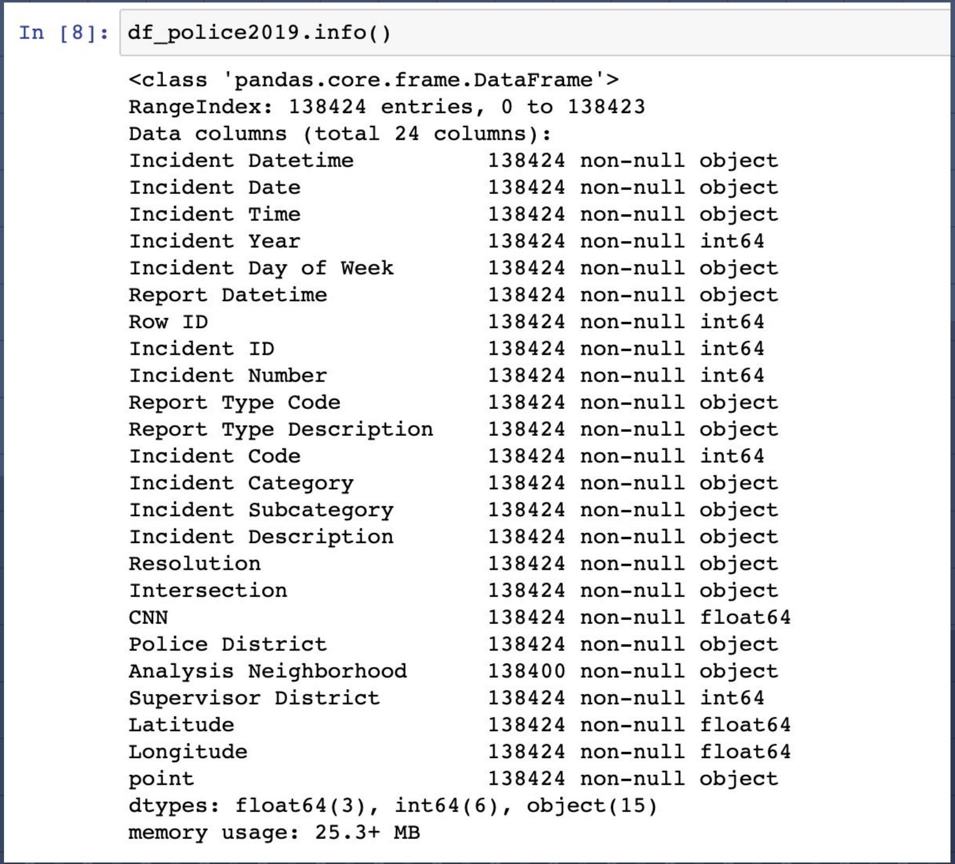
Stay home and get the latest COVID-19 data, metrics and reports

Example of the 2018 and 2019 datasets



```
In [7]: df_police2018.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144188 entries, 0 to 144187
Data columns (total 24 columns):
Incident Datetime          144188 non-null object
Incident Date              144188 non-null object
Incident Time              144188 non-null object
Incident Year              144188 non-null int64
Incident Day of Week       144188 non-null object
Report Datetime            144188 non-null object
Row ID                     144188 non-null int64
Incident ID                144188 non-null int64
Incident Number            144188 non-null int64
Report Type Code           144188 non-null object
Report Type Description    144188 non-null object
Incident Code              144188 non-null int64
Incident Category          144188 non-null object
Incident Subcategory       144188 non-null object
Incident Description       144188 non-null object
Resolution                 144188 non-null object
Intersection               144188 non-null object
CNN                        144188 non-null float64
Police District            144188 non-null object
Analysis Neighborhood      144152 non-null object
Supervisor District        144188 non-null int64
Latitude                   144188 non-null float64
Longitude                  144188 non-null float64
point                      144188 non-null object
dtypes: float64(3), int64(6), object(15)
memory usage: 26.4+ MB
```



```
In [8]: df_police2019.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 138424 entries, 0 to 138423
Data columns (total 24 columns):
Incident Datetime          138424 non-null object
Incident Date              138424 non-null object
Incident Time              138424 non-null object
Incident Year              138424 non-null int64
Incident Day of Week       138424 non-null object
Report Datetime            138424 non-null object
Row ID                     138424 non-null int64
Incident ID                138424 non-null int64
Incident Number            138424 non-null int64
Report Type Code           138424 non-null object
Report Type Description    138424 non-null object
Incident Code              138424 non-null int64
Incident Category          138424 non-null object
Incident Subcategory       138424 non-null object
Incident Description       138424 non-null object
Resolution                 138424 non-null object
Intersection               138424 non-null object
CNN                        138424 non-null float64
Police District            138424 non-null object
Analysis Neighborhood      138400 non-null object
Supervisor District        138424 non-null int64
Latitude                   138424 non-null float64
Longitude                  138424 non-null float64
point                      138424 non-null object
dtypes: float64(3), int64(6), object(15)
memory usage: 25.3+ MB
```

Process

Use the 2010 to 2017 Data for the graphical part to show the trends and the crime categories in the different districts of San Francisco.

Try to predict the number of crimes incidents in 2019 considering the zip code, the day of week and the time.

The 2018 dataset will be used to train the model and the 2019 to make the prediction

Tools

Open Refine

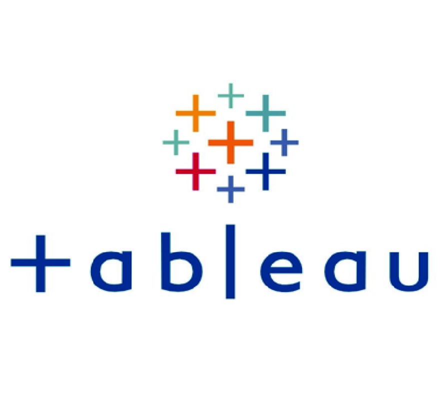
Cleaning Process

Anaconda/Jupyter Notebook

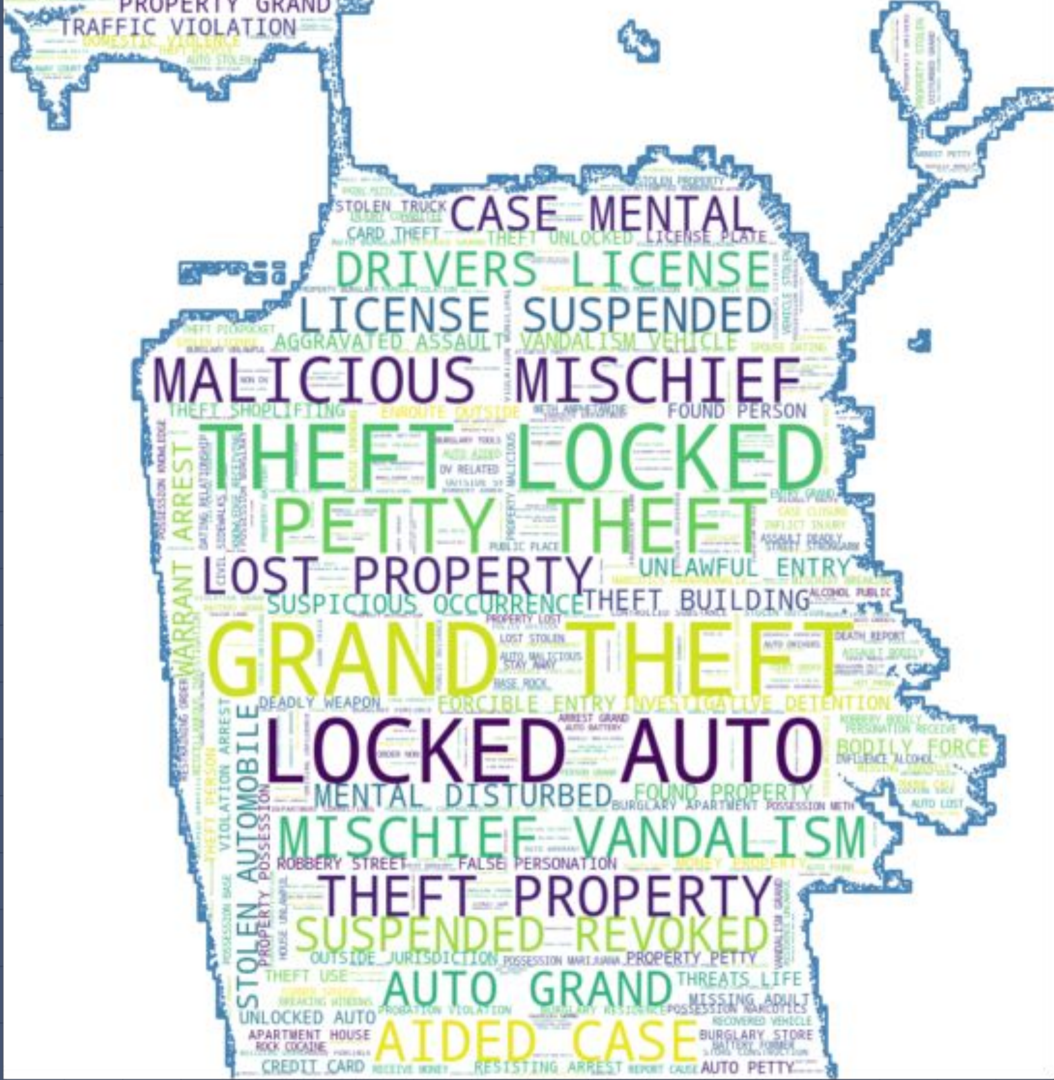
Train & Predict the Model

Tableau

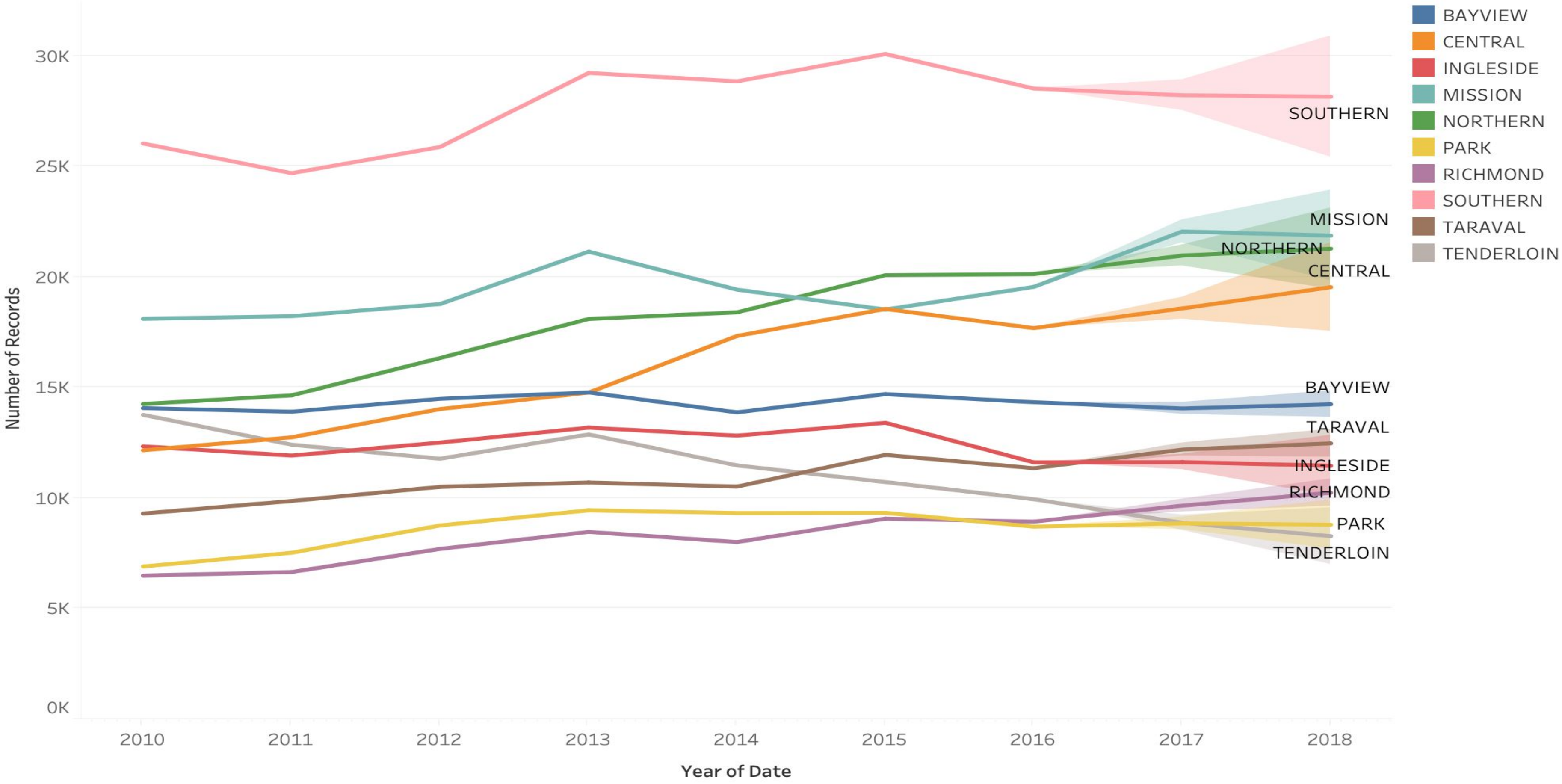
Visualization



Study of Crimes in San Francisco (2010 - 2017)



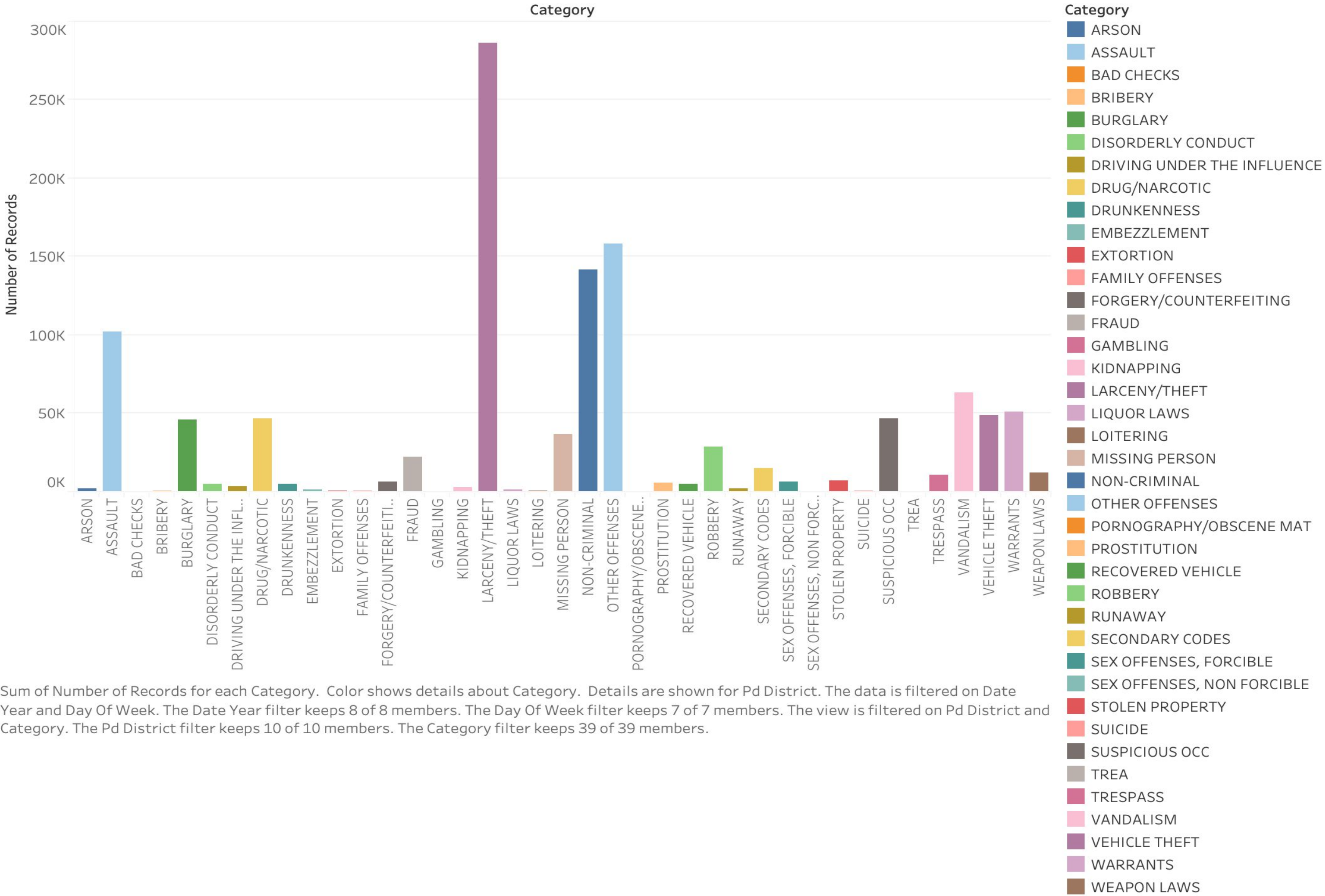
Crimes Records



The trend of sum of Number of Records (actual & forecast) for Date Year. Color shows details about Pd District. The marks are labeled by Pd District. The data is filtered on Date Year, Date Month and Day Of Week. The Date Year filter keeps 8 of 8 members. The Date Month filter keeps 12 of 12 members. The Day Of Week filter keeps 7 of 7 members. The view is filtered on Pd District, which keeps 10 of 10 members.

Interactive Graph here🔗

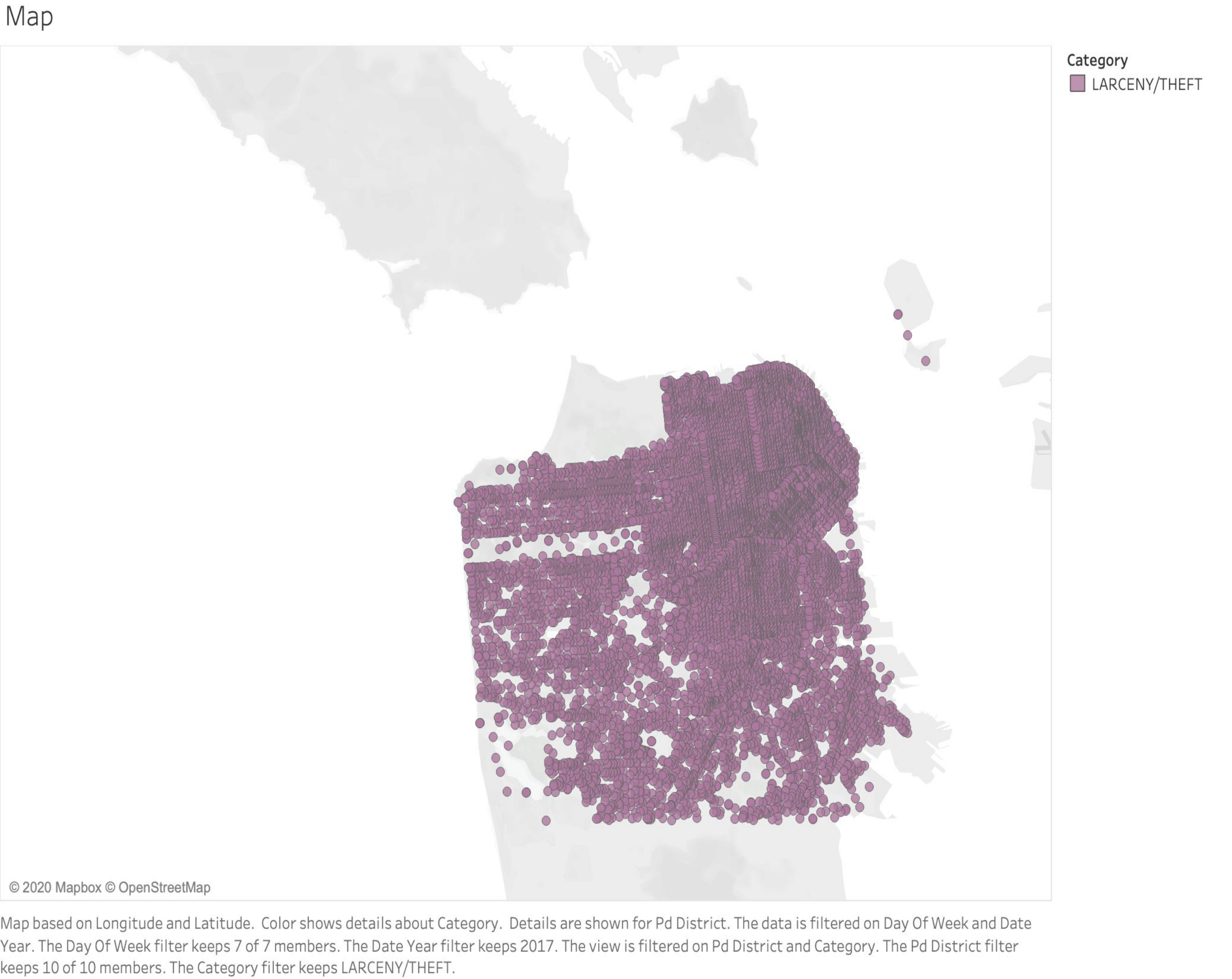
Crimes Category



Sum of Number of Records for each Category. Color shows details about Category. Details are shown for Pd District. The data is filtered on Date Year and Day Of Week. The Date Year filter keeps 8 of 8 members. The Day Of Week filter keeps 7 of 7 members. The view is filtered on Pd District and Category. The Pd District filter keeps 10 of 10 members. The Category filter keeps 39 of 39 members.

47826 Larceny incidents recorded in 2017

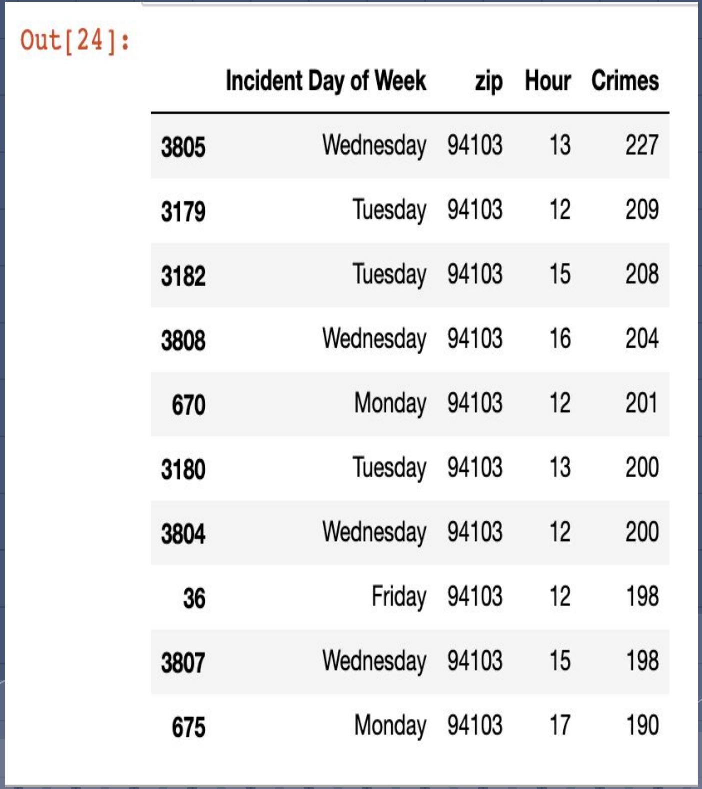
Interactive Map of categories records by district 🔗



Map

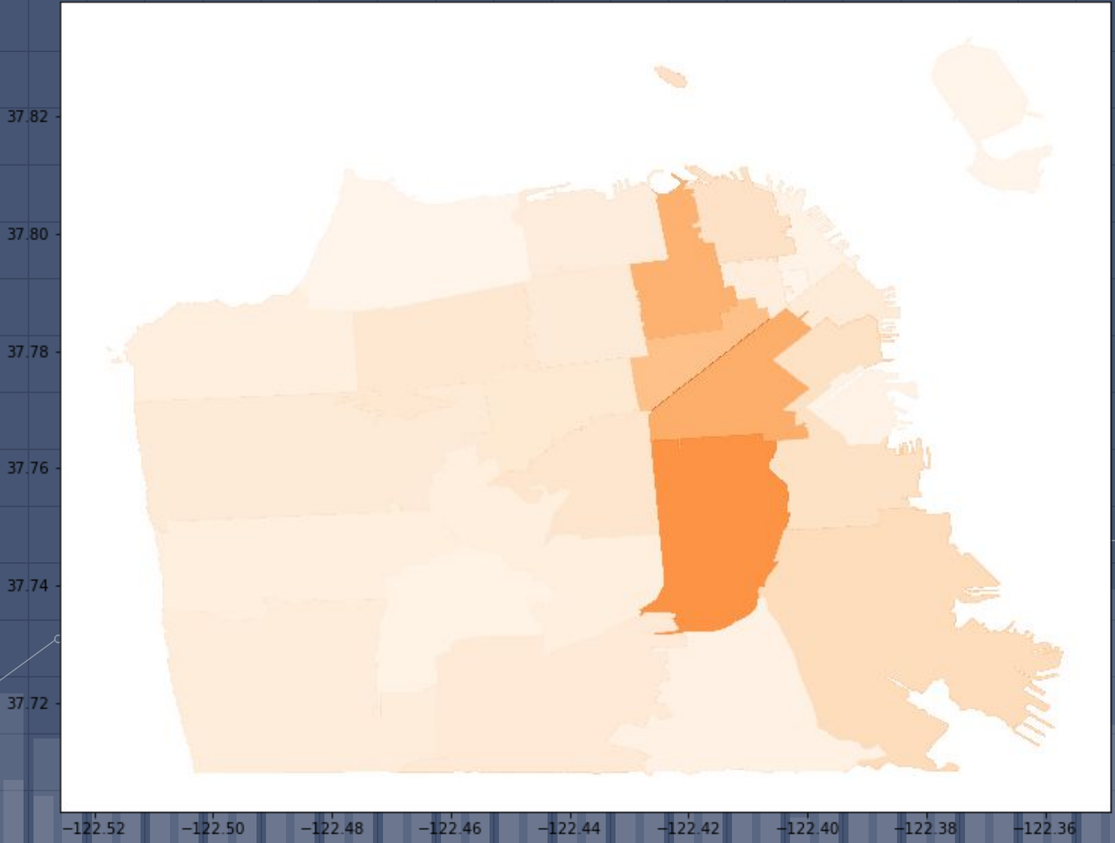
Map based on Longitude and Latitude. Color shows details about Category. Details are shown for Pd District. The data is filtered on Day Of Week and Date Year. The Day Of Week filter keeps 7 of 7 members. The Date Year filter keeps 2017. The view is filtered on Pd District and Category. The Pd District filter keeps 10 of 10 members. The Category filter keeps LARCENY/THEFT.

Crimes in 2018 (days and hours)



Out[24]:

	Incident Day of Week	zip	Hour	Crimes
3805	Wednesday	94103	13	227
3179	Tuesday	94103	12	209
3182	Tuesday	94103	15	208
3808	Wednesday	94103	16	204
670	Monday	94103	12	201
3180	Tuesday	94103	13	200
3804	Wednesday	94103	12	200
36	Friday	94103	12	198
3807	Wednesday	94103	15	198
675	Monday	94103	17	190



Prediction Process

Worked with the given latitude and longitude to get the zip code of each area.

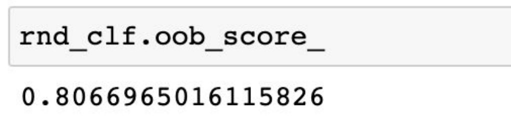
Select the day of the week, day and zip code to get the number of crimes recorded per day, hour and area

Run two different models and use the most efficient to predict the crimes in 2019 given the features selected

Models

A random forest regressor.

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

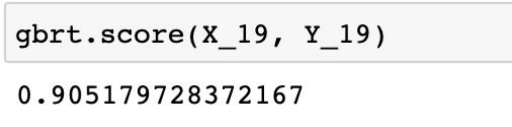


```
rnd_clf.oob_score_

0.8066965016115826
```

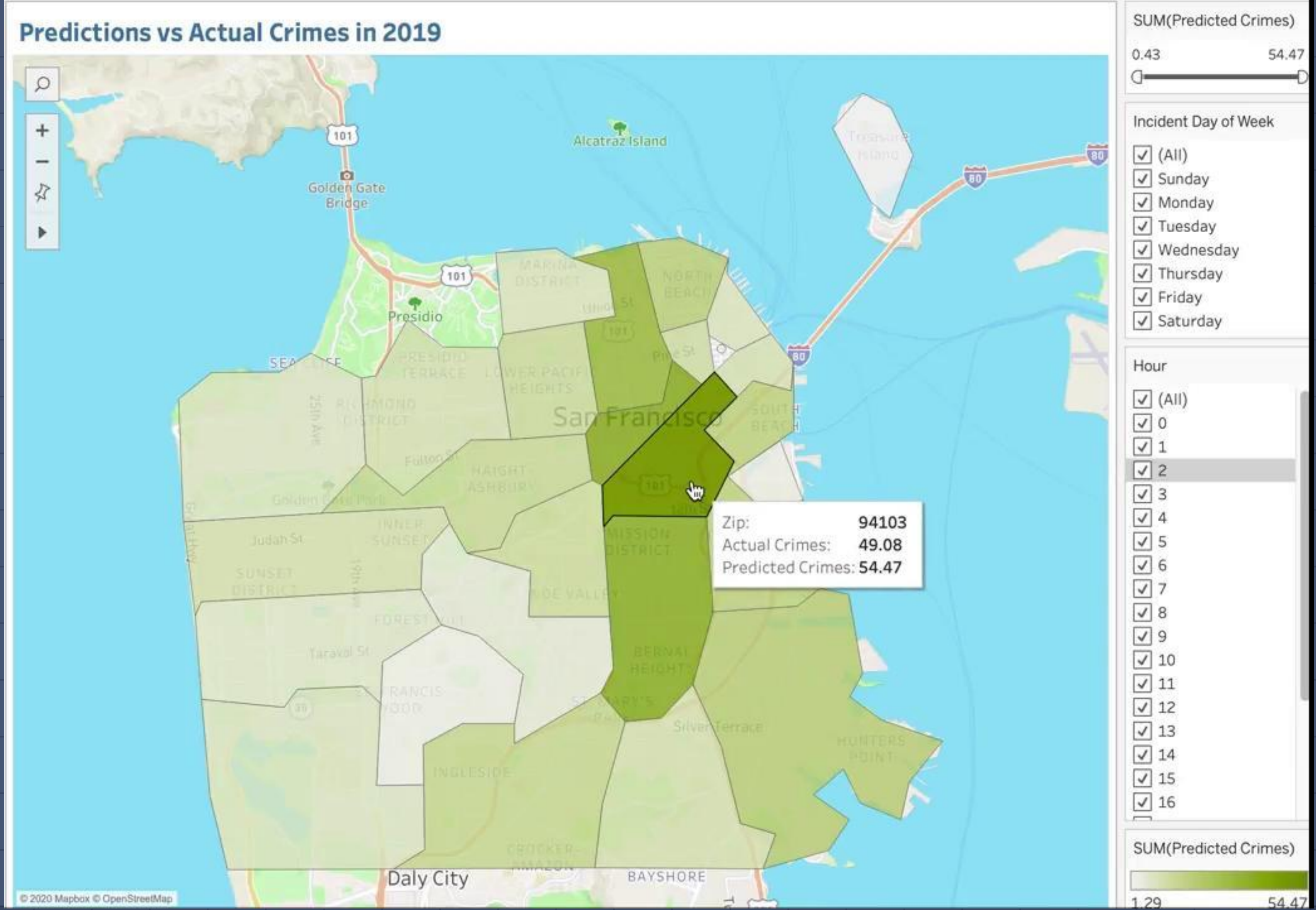
Gradient Boosting for regression.

GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.



```
gbrt.score(X_19, Y_19)

0.905179728372167
```



Predictions vs Actual Crimes in 2019

Zip: 94103
Actual Crimes: 49.08
Predicted Crimes: 54.47

SUM(Predicted Crimes)
0.43 54.47

Incident Day of Week
☑ (All)
☑ Sunday
☑ Monday
☑ Tuesday
☑ Wednesday
☑ Thursday
☑ Friday
☑ Saturday

Hour
☑ (All)
☑ 0
☑ 1
☑ 2
☑ 3
☑ 4
☑ 5
☑ 6
☑ 7
☑ 8
☑ 9
☑ 10
☑ 11
☑ 12
☑ 13
☑ 14
☑ 15
☑ 16

SUM(Predicted Crimes)
1.29 54.47

2019 Prediction Results vs Actual Crimes

Interactive Graph 🔗

Conclusion

The model shows that we can use past crimes patterns to predict current crime patterns with 90% accuracy and help send law enforcement where it is needed at the right time.

For future work, we might include stacking models to improve accuracy or testing other cities. Train police officers to add more input to the data during collection; the weather, distance to liquor stores and homeless shelters are also important