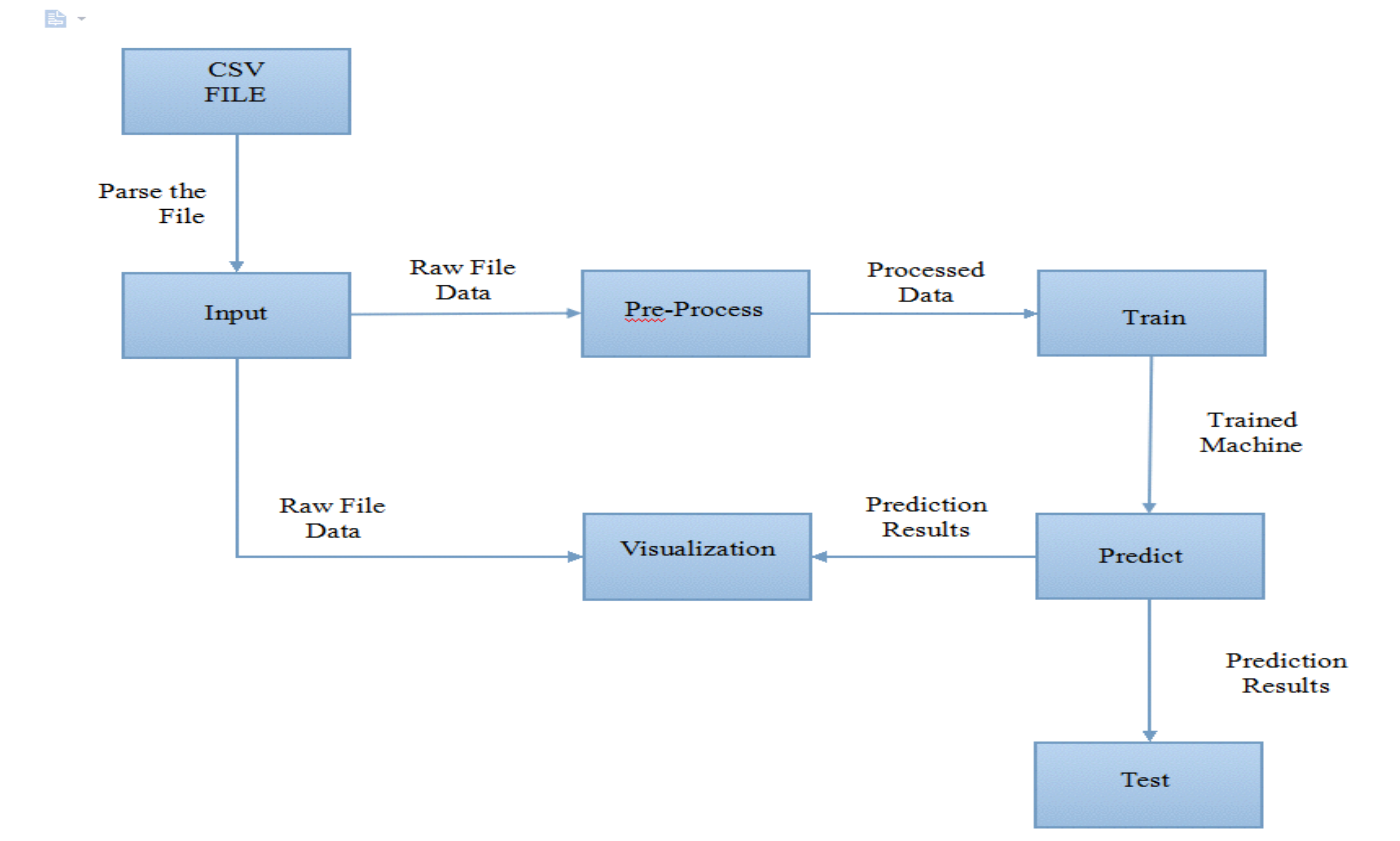
TIM245 Presentation

San Fransisco Crime Classification

Members:
Borui Wang, Han Chen

Introduction

Design Document: System Overview



CSV FILE

Parse the File

Input

Raw File Data

Pre-Process

Processed Data

Train

Trained Machine

Predict

Prediction Results

Visualization

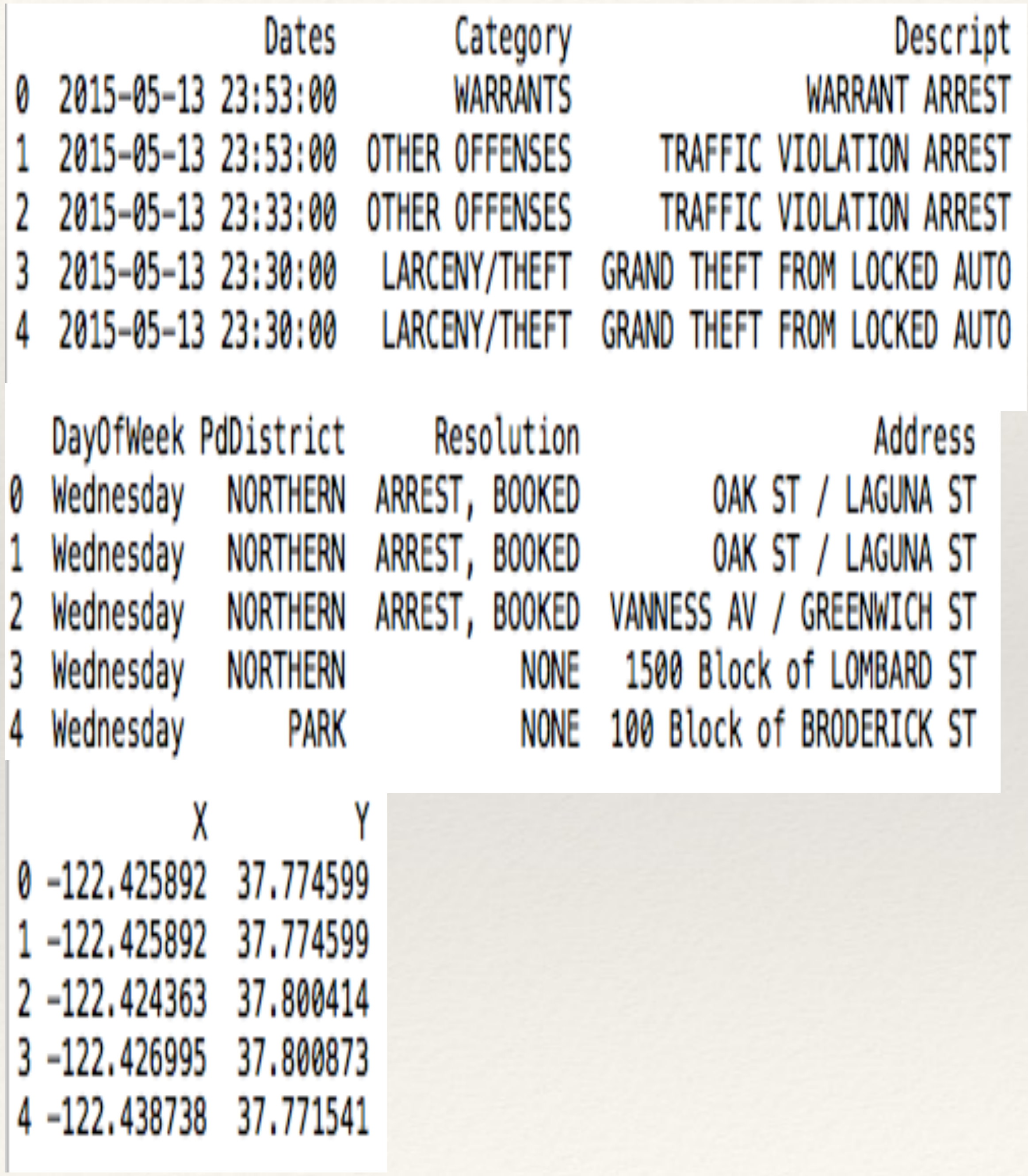
Raw File Data

Prediction Results

Test

About the Dataset

1. There’re 7 attributes in the dataset. More than 1,600k records from 2003-01-01 to 2015-05-13.
2. There’re 39 kinds of Category,
   879 kinds of Descript,
   17 kinds of Resolution.
3. Wrong coordinates.(< -122)

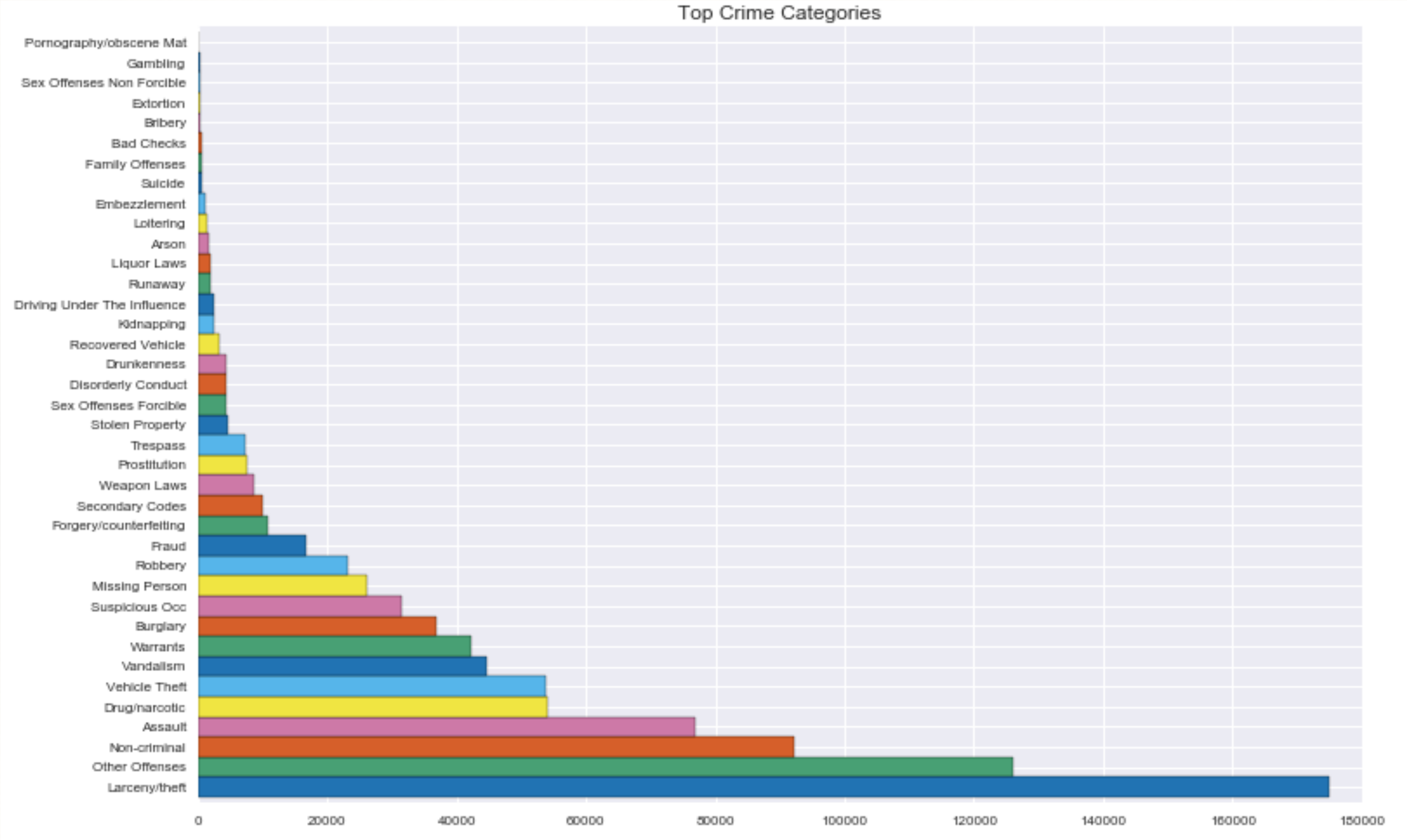


```
                 Dates        Category                 Descript
0  2015-05-13 23:53:00        WARRANTS           WARRANT ARREST
1  2015-05-13 23:53:00  OTHER OFFENSES  TRAFFIC VIOLATION ARREST
2  2015-05-13 23:33:00  OTHER OFFENSES  TRAFFIC VIOLATION ARREST
3  2015-05-13 23:30:00   LARCENY/THEFT  GRAND THEFT FROM LOCKED AUTO
4  2015-05-13 23:30:00   LARCENY/THEFT  GRAND THEFT FROM LOCKED AUTO

   DayOfWeek PdDistrict      Resolution                    Address
0  Wednesday   NORTHERN  ARREST, BOOKED         OAK ST / LAGUNA ST
1  Wednesday   NORTHERN  ARREST, BOOKED         OAK ST / LAGUNA ST
2  Wednesday   NORTHERN  ARREST, BOOKED  VANNESS AV / GREENWICH ST
3  Wednesday   NORTHERN            NONE   1500 Block of LOMBARD ST
4  Wednesday       PARK            NONE  100 Block of BRODERICK ST

            X          Y
0 -122.425892  37.774599
1 -122.425892  37.774599
2 -122.424363  37.800414
3 -122.426995  37.800873
4 -122.438738  37.771541
```

About the Dataset



Hotspot

1. What is a 'HotSpot'?

   A 'hotspot' is a geographic zone on the map with a greater probability that a crime will occur

2. How do we use HotSpot?

   plot all records of the same crime

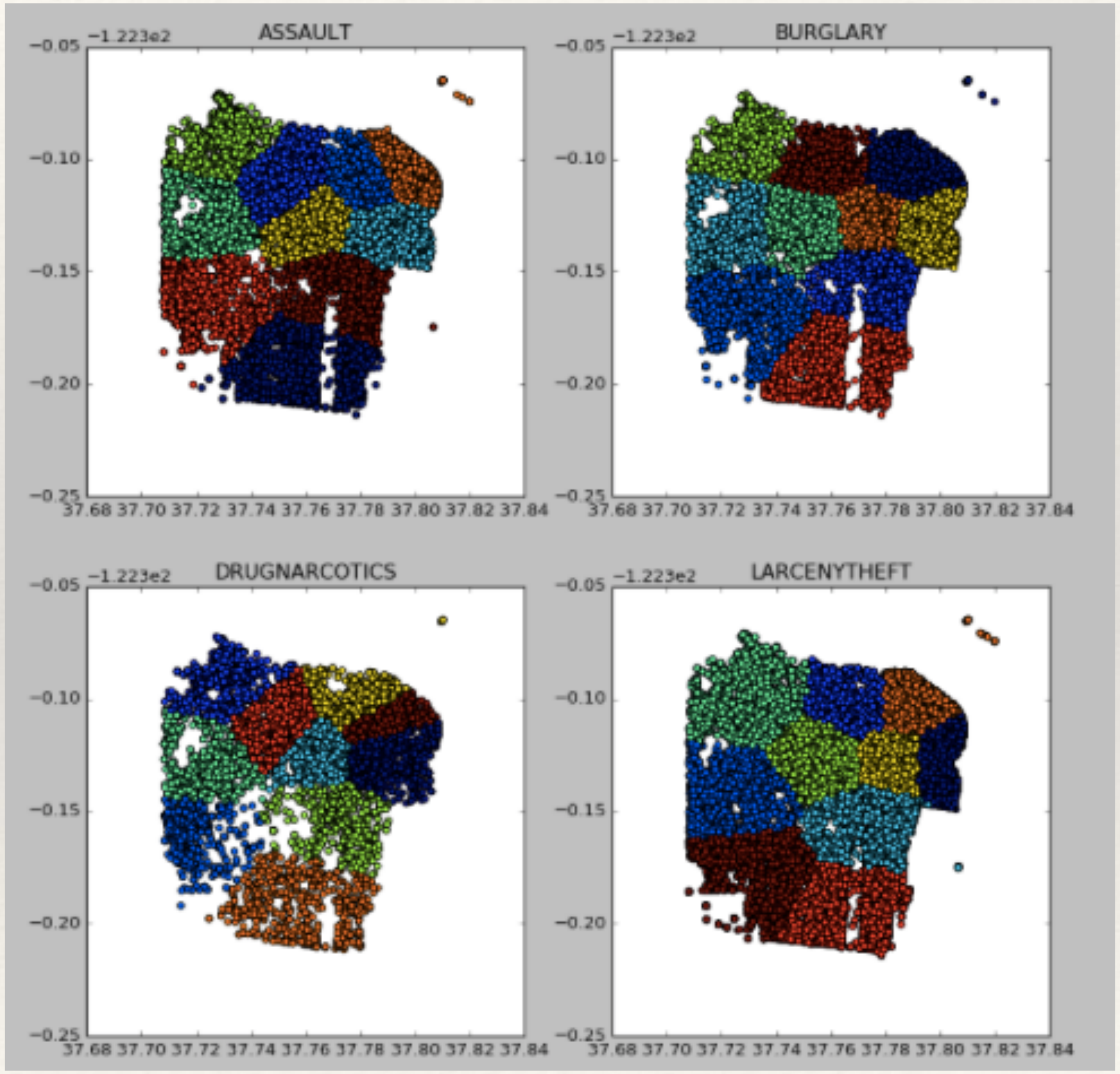
3. Algorithm

   run a clustering algorithm, currently k-means

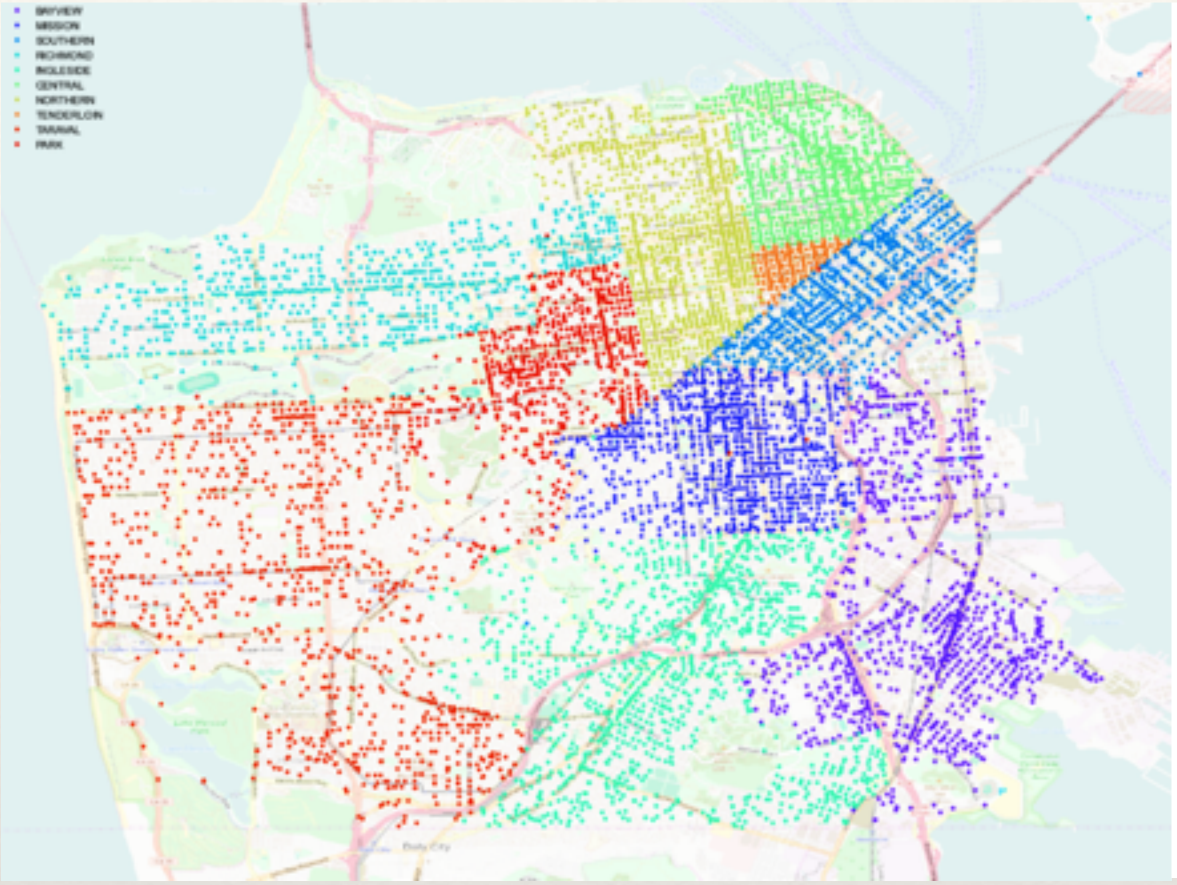
   assign the most dense clusters as hotspots

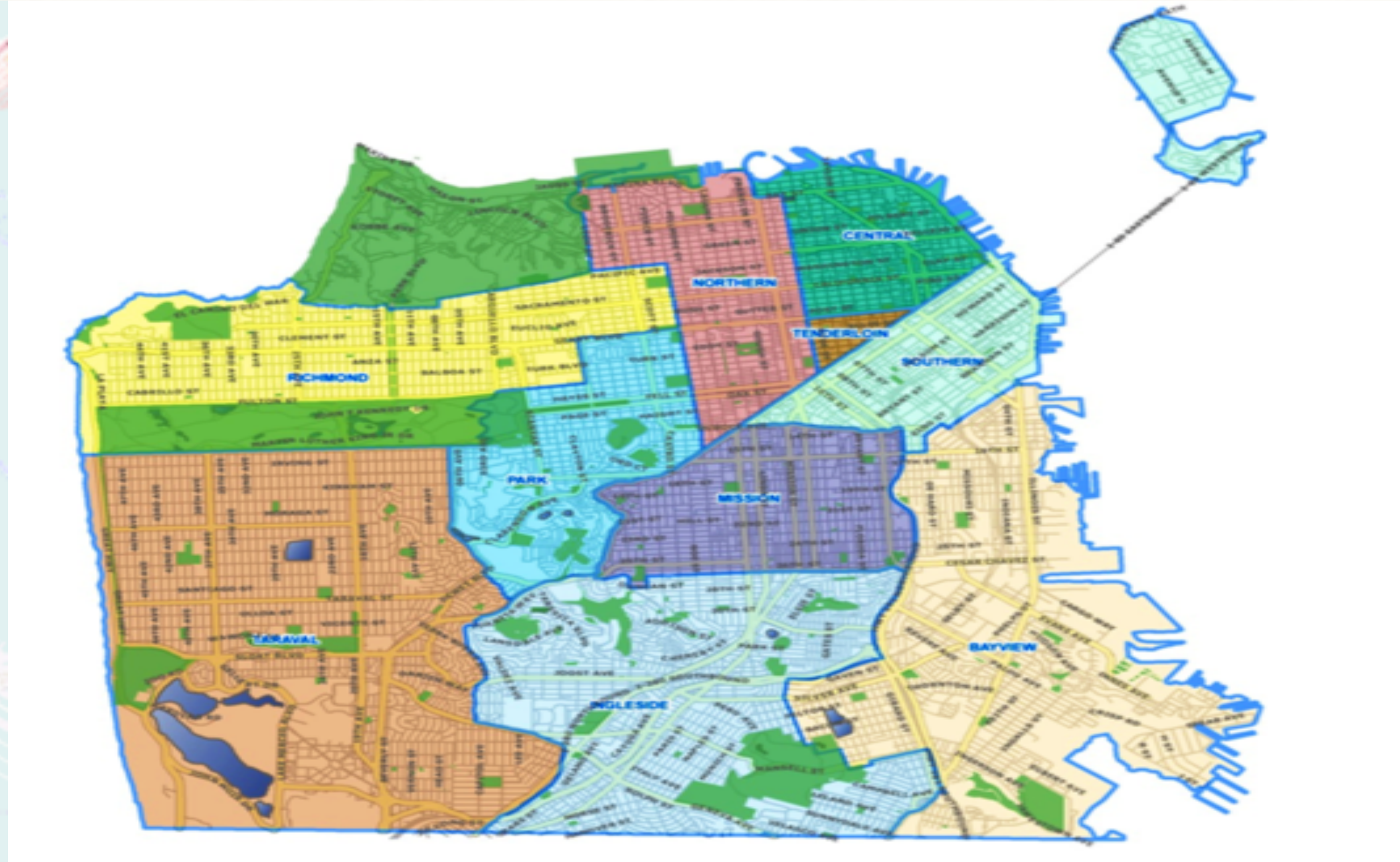
4. How to calculate Density?

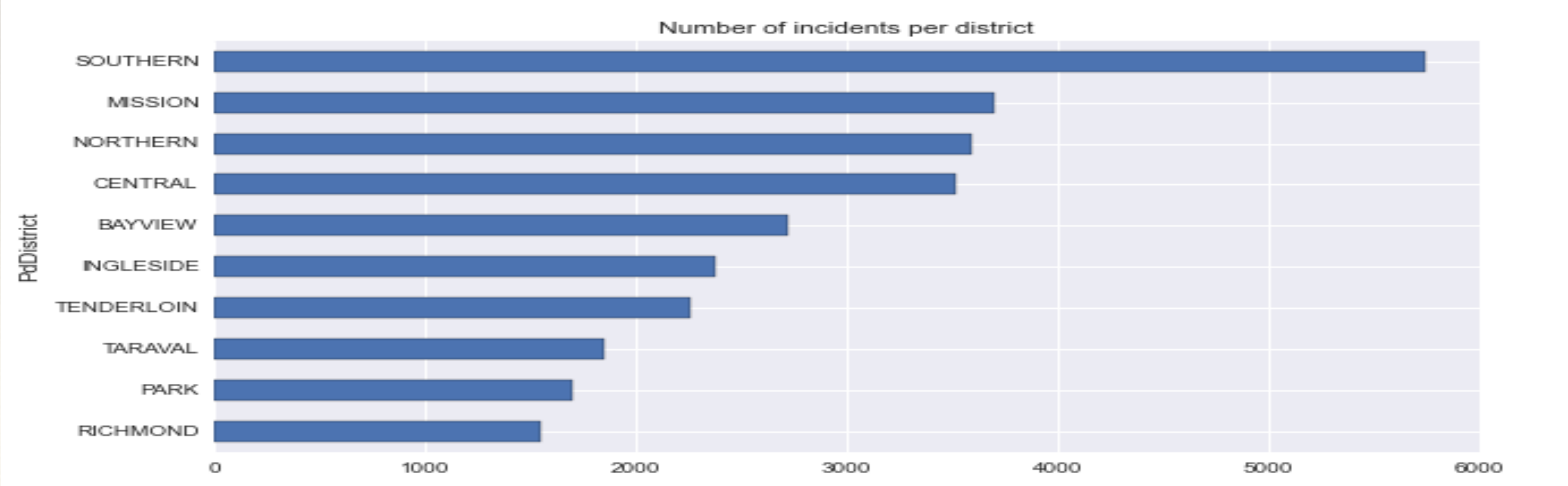
   density = number of points / [(Xmax - Xmin) * (Ymax - Ymin)]



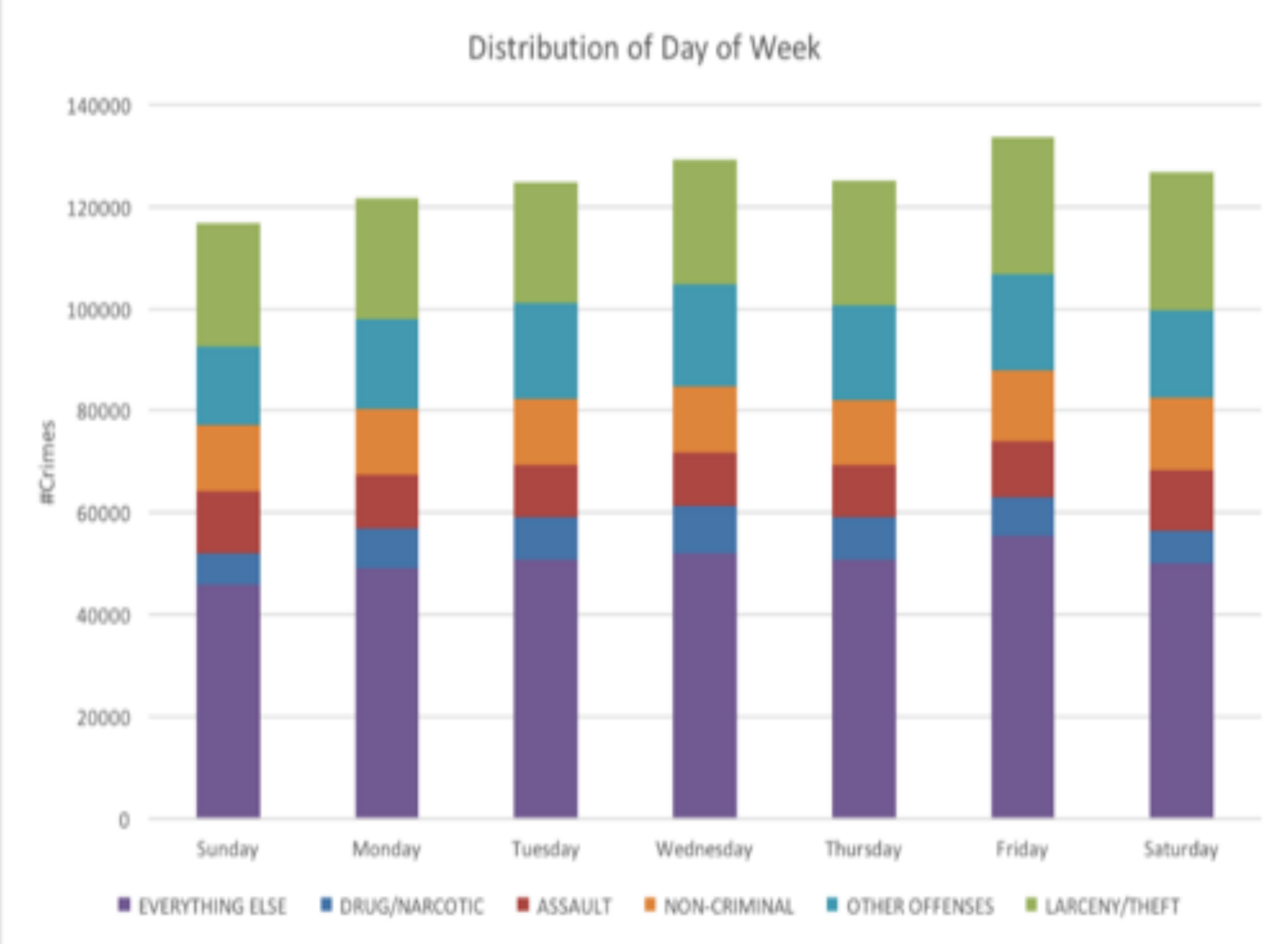
Crime in Different Areas

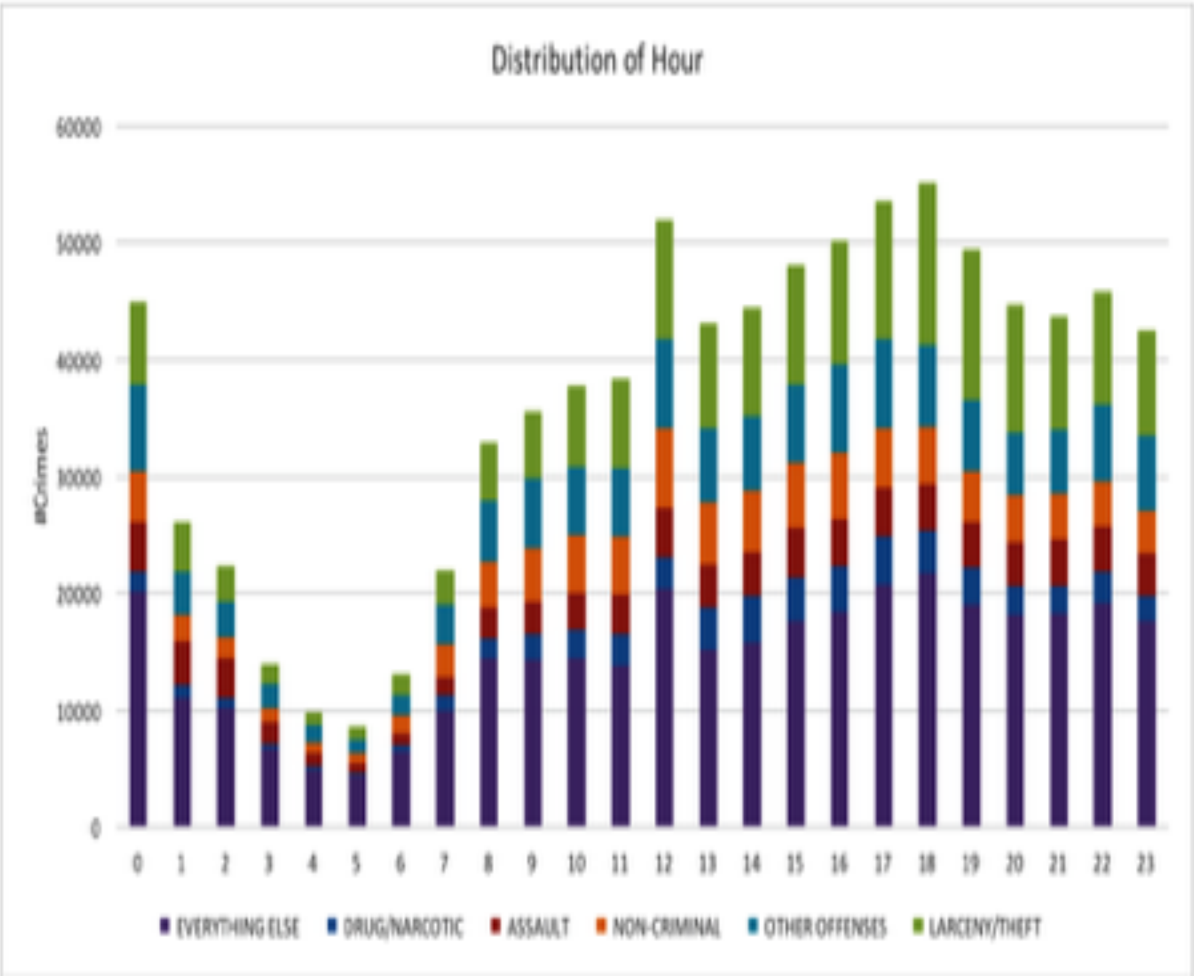


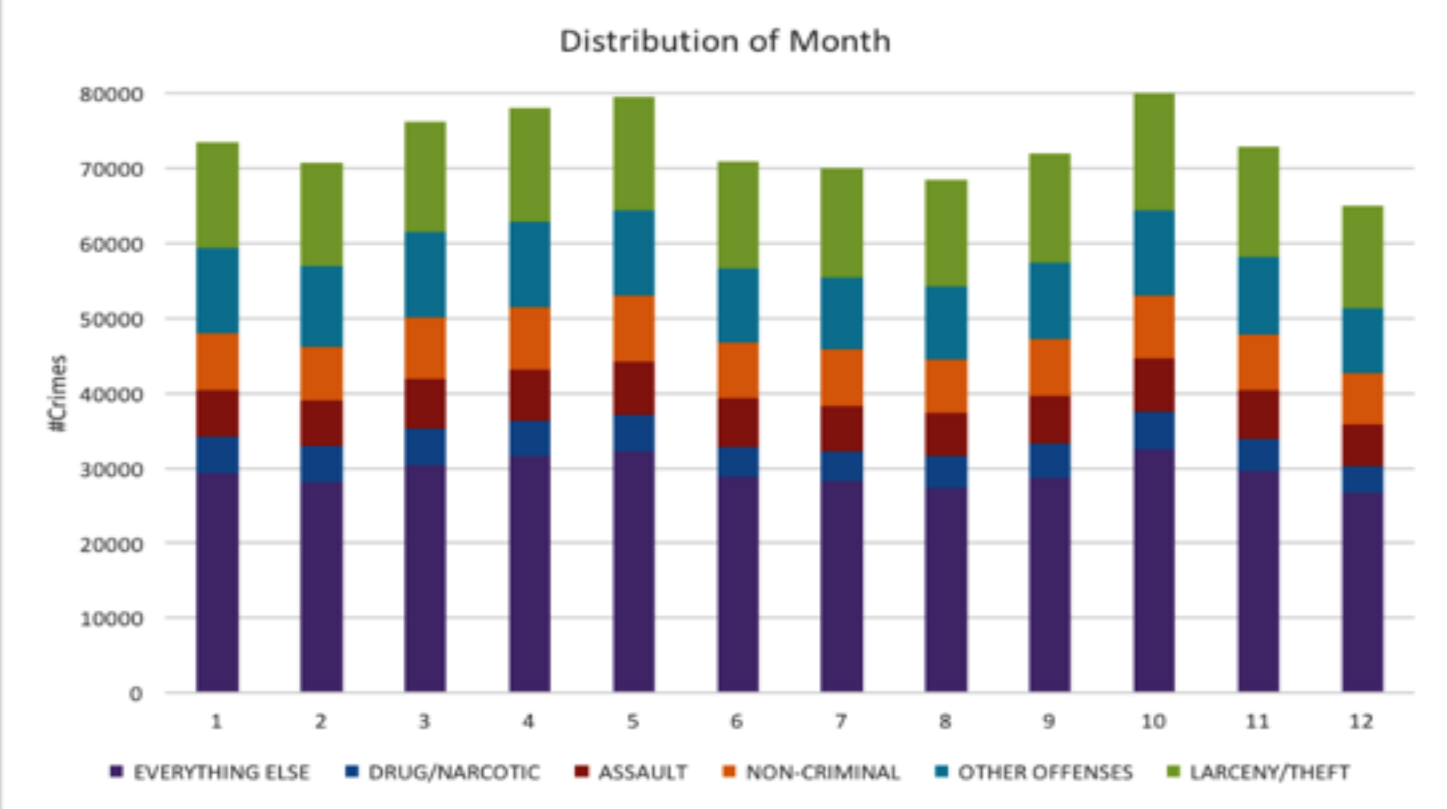




Distribution of Data







Feature Selection

```
Dates -> 2015-05-13 23:53:00
Category -> VEHICLE THEFT, labels(39)
Descript -> detailed description of the crime incident (only in train.csv)
DayOfWeek -> Wednesday
PdDistrict -> CENTRAL, name of the Police Department District
Resolution -> how the crime incident was resolved (only in train.csv)
Address -> the approximate street address of the crime incident
           23,228 different addresses
X - Longitude
Y - Latitude
```

Data Preprocessing

Numerical attributes

Dates -> 2015-05-13 23:53:00 -> Year, Month, Day and Hour

X - Longitude

Y - Latitude

Nonnumerical

Category -> VEHICLE THEFT, labels(39) -> LabelEncoder

DayOfWeek -> Wednesday

PdDistrict -> CENTRAL, name of the Police Department District

pd.get_dummies() to cover text to binary array

Training model

Cross-validation

We used a single train-test split for our train data set and we split 30% train data for test set.

Model selection

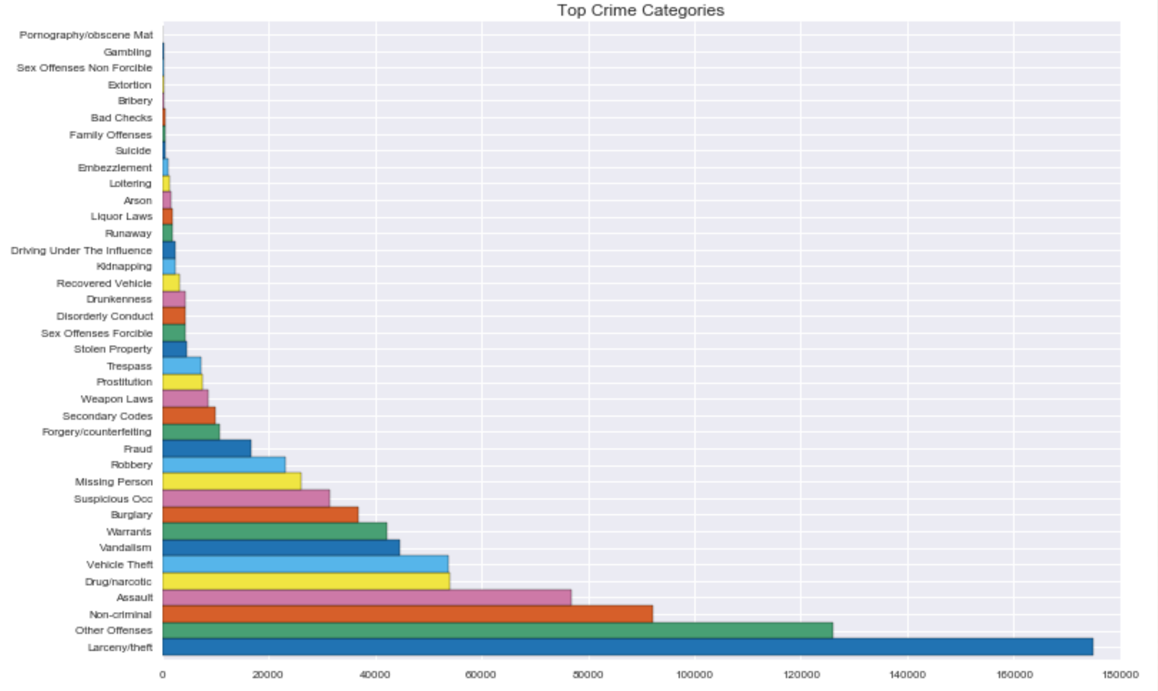
Logistic Regression, Naive Bayes, SVM, Random Forest

Result

	Logistic Regression	Naive Bayes	Random Forest
Score	0.22	0.22	0.27
Log-loss	2.61	2.61	2.1

Conclusion

1. Random forest is much better for tangled feature
2. Reason for low accuracy
   1. Too many labels and less features
   2. Features are decentralized
3. Focus on specific crimes (top 4)



Questions?