# SF CRIME CLASSIFICATION

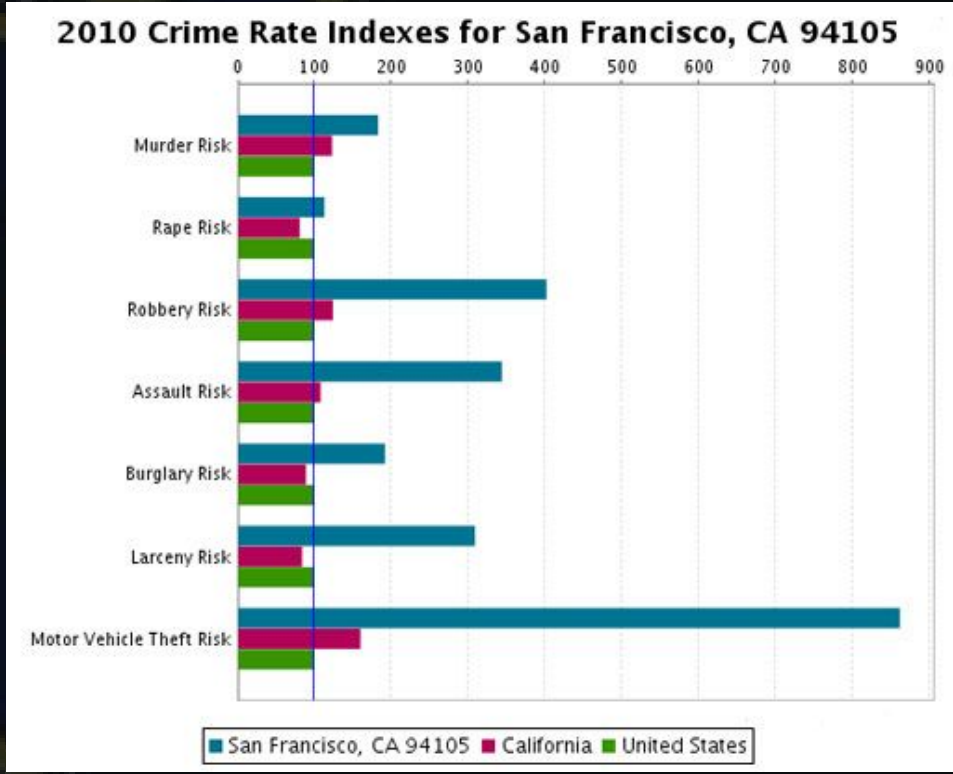
'The Outliers'
Khyati Pawde : 704590958
Mansi Shah : 504591572
Pallavi A. Kotkar : 504589593
Sravani Kamisetty : 304414410

# PROBLEM

- The project uses the SF crime dataset (nearly 12 years of crime reports)
- Given a time and location -> predict the category of crime
- Contains incidents derived from SFPD Crimes
- The data ranges from 01/01/2003 to 05/13/2015.
- The training set and test set rotate every week, meaning week 1,3,5 belong to test set, week 2,4,6,8 belong to training set.
- Kaggle project

# MOTIVATION

- Between 1934 and 1963 Alcatraz Island near San Francisco housed some of the most notorious criminals
- SF is known for the tech scene
- But with rising wealth inequality, housing shortages and proliferation of expensive digital devices there has been a rise in the crime rate
- The city saw more the 20% jumps in both the rate of property crime such as thefts and burglary and rate of violent crimes such as robbery and assault between 2012 and 2013



# MOTIVATION

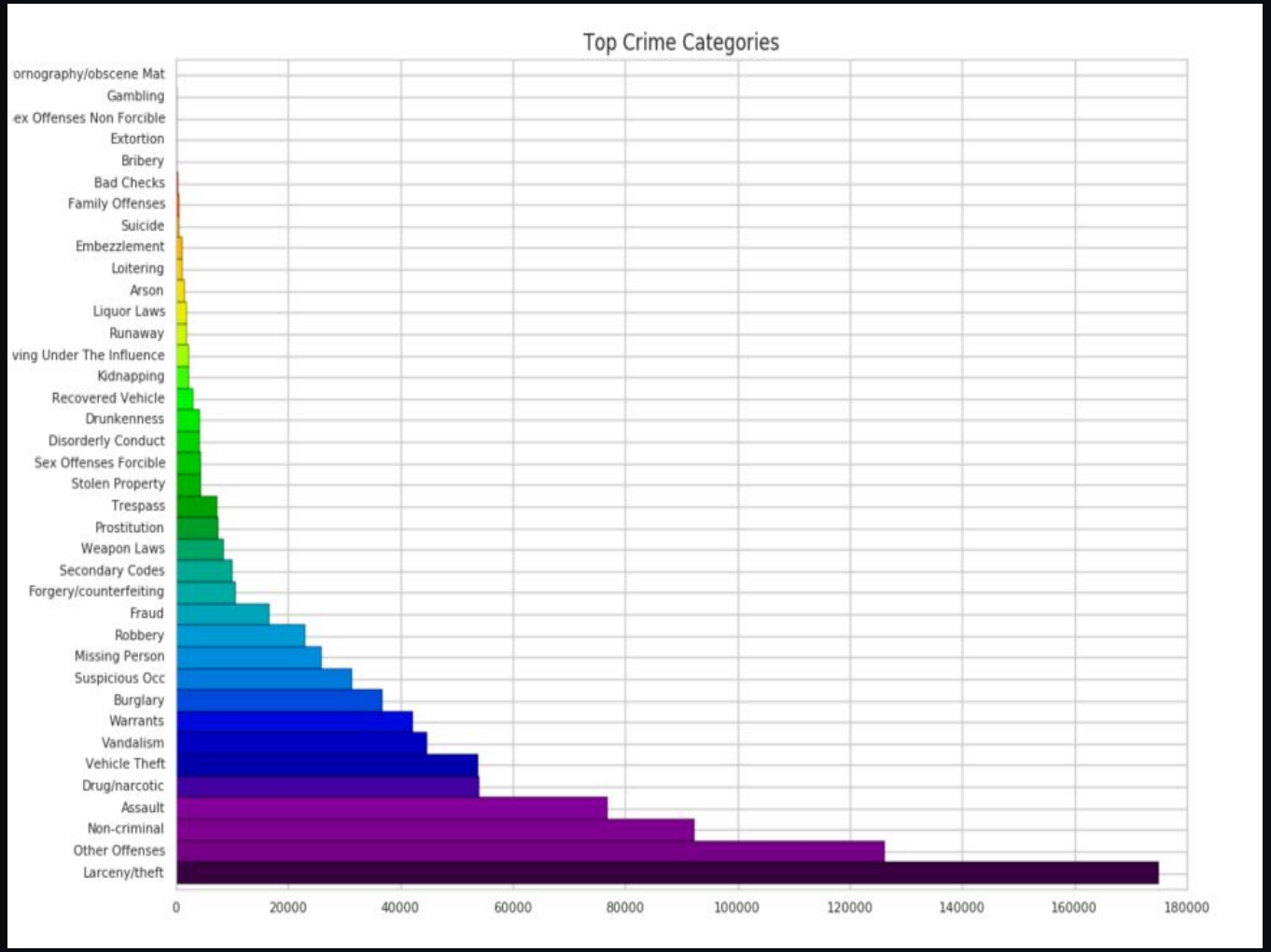
- The primary motivation of this project is to make use of the huge amount of data available to help gauge the current crime scenario in San Francisco.
- Find underlying patterns in the data which can reveal certain criminal psychologies.
- Patterns when uncovered can help the public as well as the police department in abatement of such crimes.

# DATA

| Type | Training | Testing |
|---|---|---|
| Size | 121 MB | 87 MB |
| Number of Observations: | 878049 | 884262 |
| Data Fields: | Date | Date |
| | Descript | DayOfWeek |
| | DayOfWeek | PdDistrict |
| | PdDistrict | Address |
| | Resolution | X |
| | Address | Y |
| | X | |
| | Y | |
| | Category | |


# CATEGORIES

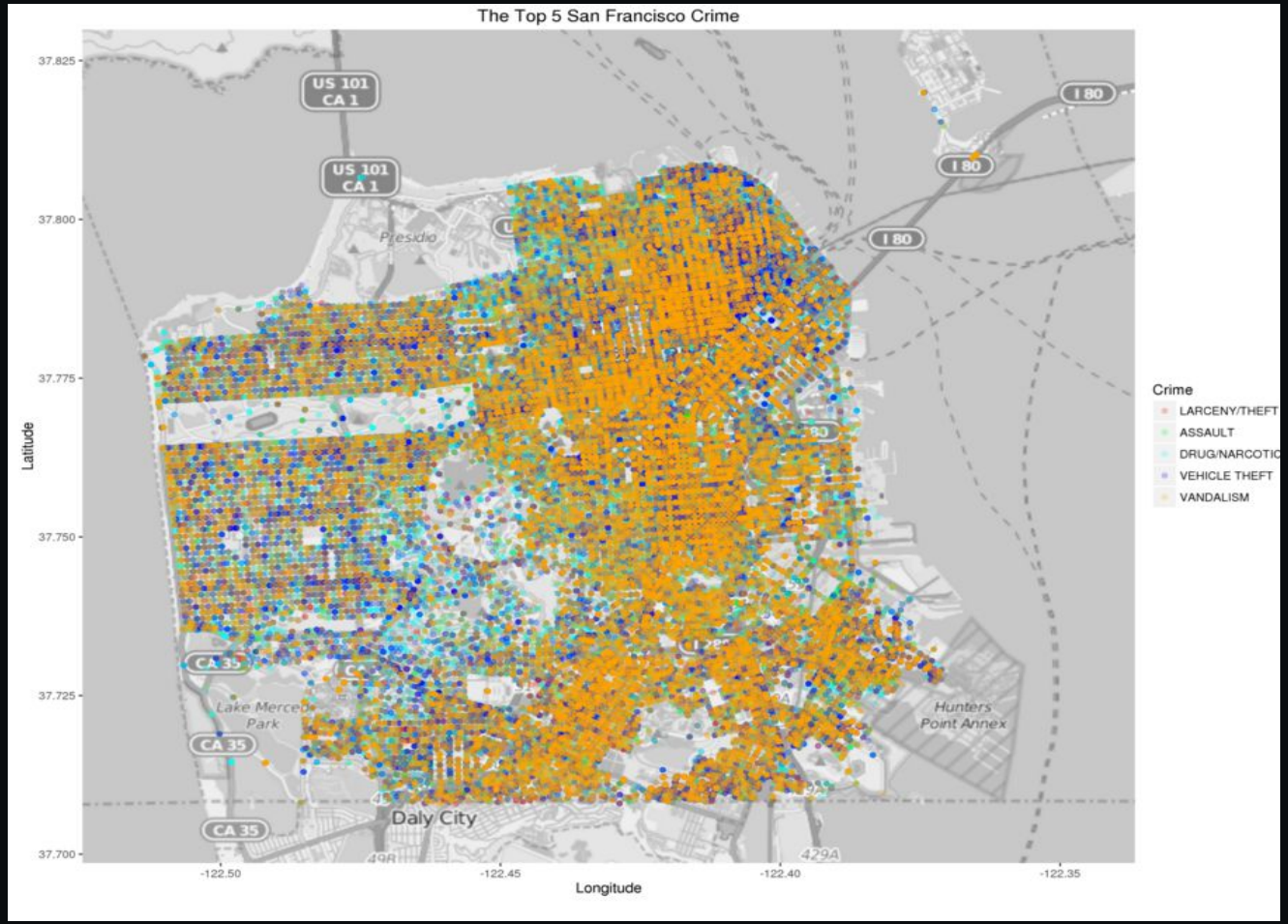

# EXPLORATORY ANALYSIS



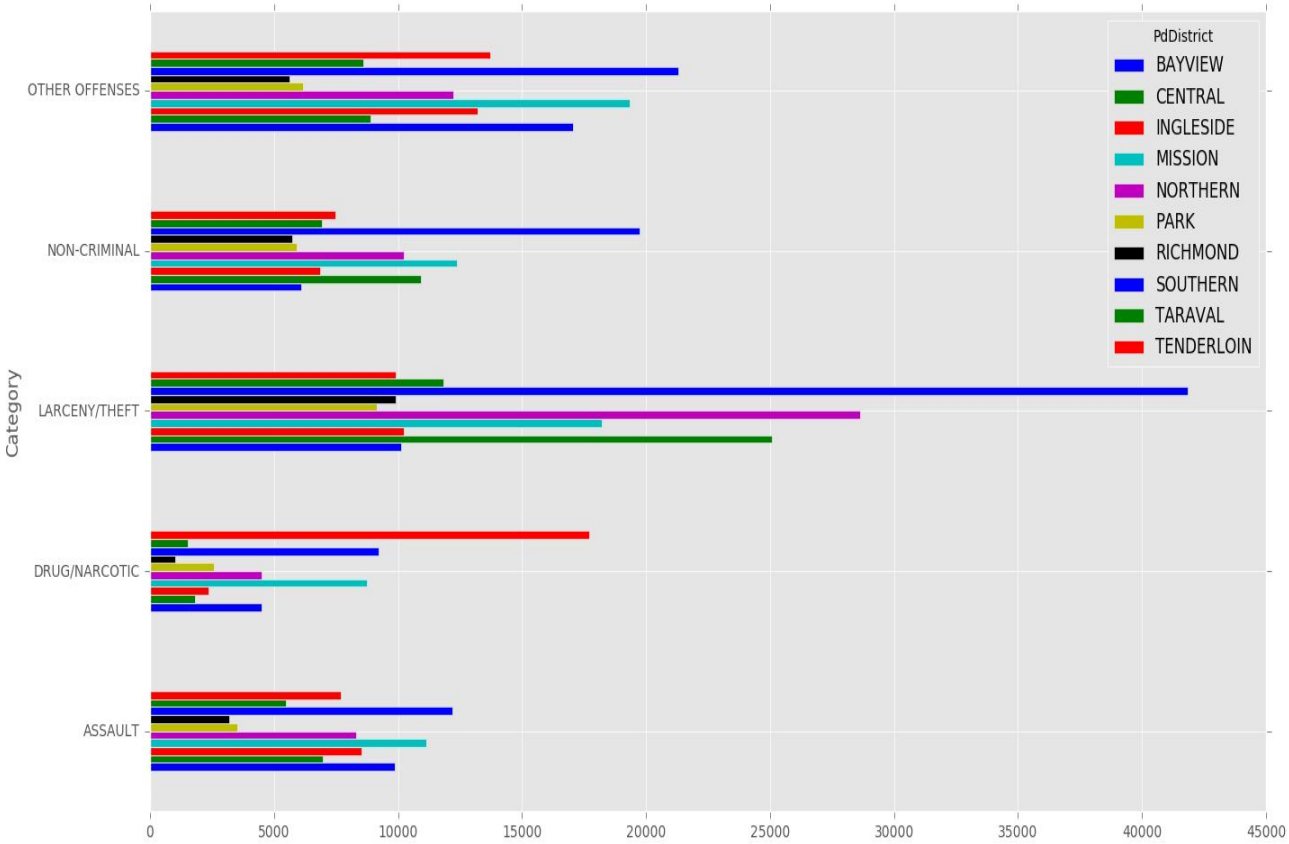

90% of the data

# EXPLORATORY ANALYSIS CONTINUED

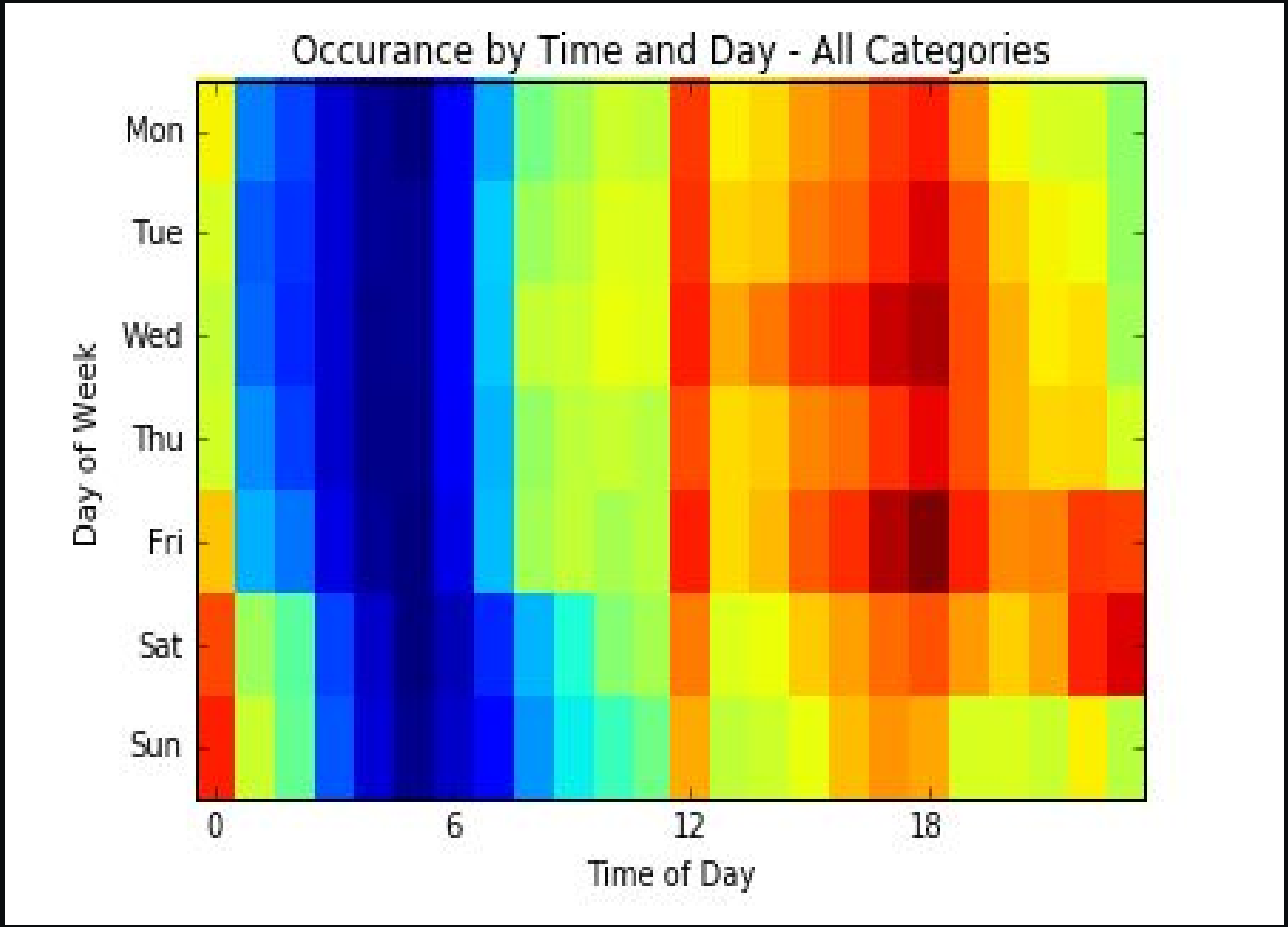
SF map with 5 crimes



# Districts v/s Crime



# Day v/s Time


Occurance by Time and Day - All Categories
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Day of Week
0
6
12
18
Time of Day

# OUTPUT LABELS

| ID | Larceny | Drug | ... |
|---|---|---|---|
| 1 | [0-1] | [0-1] | ... |
| 2 | ... | ... | ... |
| ... | ... | ... | ... |

# EVALUATION TECHNIQUE

Multiclass Logarithmic Loss Function :

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

Number of class labels

Number of cases in the test set

1 : if observation i is in class j
0 : otherwise

Predicted probability of observation i that belongs to class j

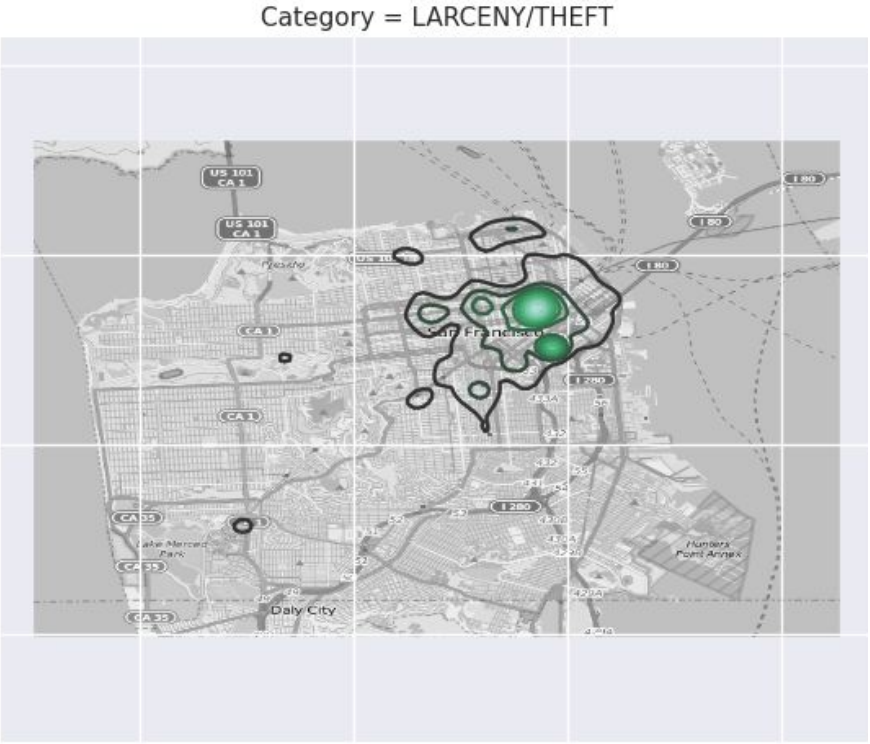
# METHODS

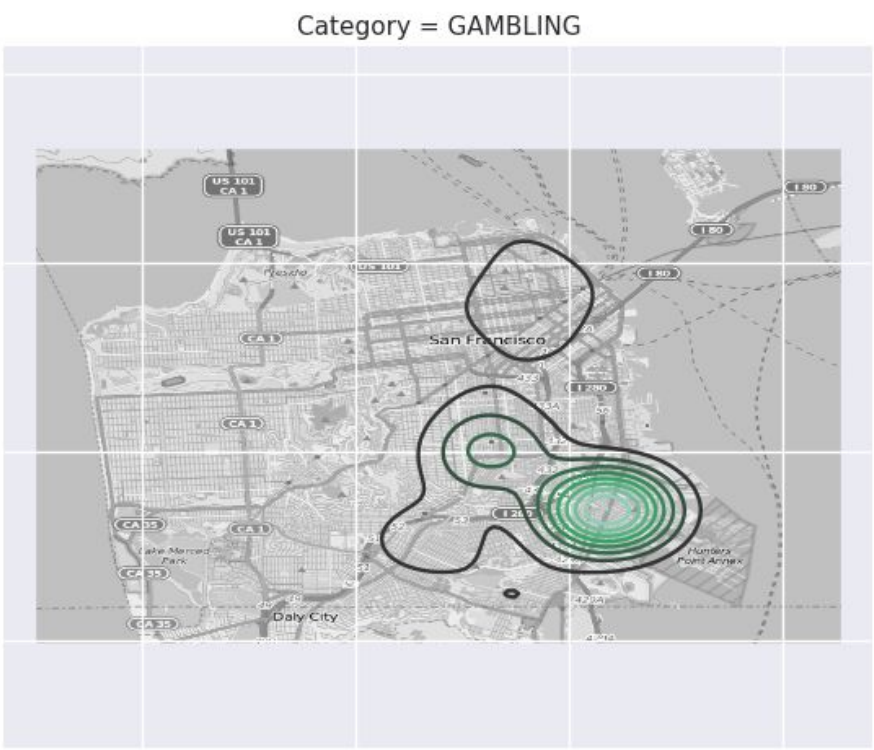
- Classification without clustering
- Clustering
- Classification with clustering
- Binomial Regression

# CLASSIFICATION WITHOUT CLUSTERING

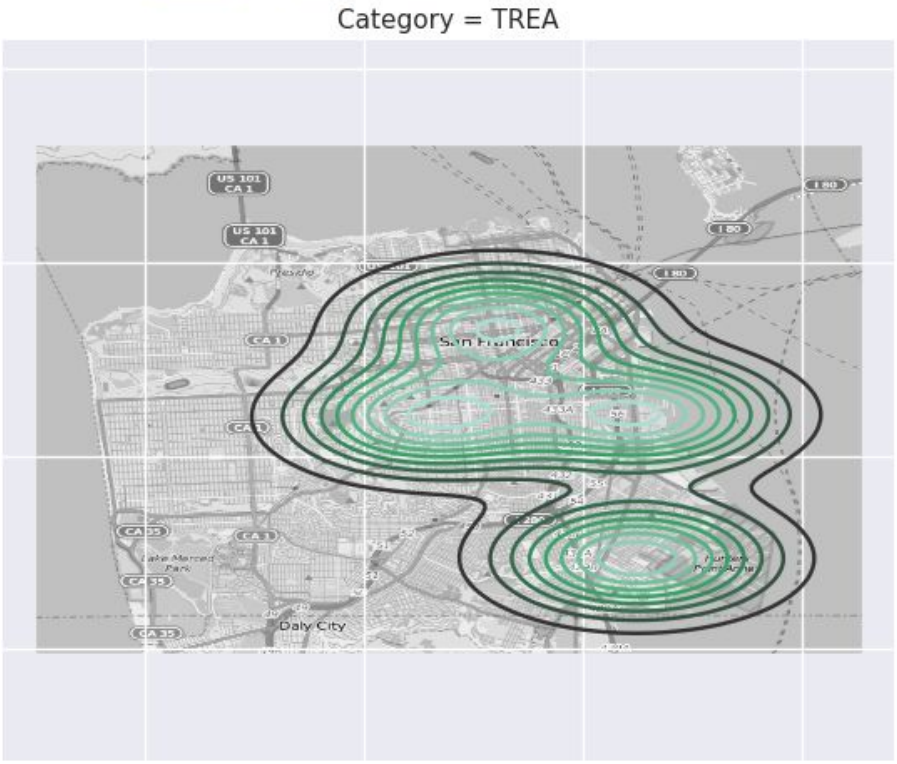
| METHOD | LOG LOSS | |
|---|---|---|
| | EVALUATION | TESTING |
| KNN | 25.5457 | N/A |
| SVM | 34.3241 | N/A |
| NAIVE BAYES | 26.8343 | N/A |
| DECISION TREES | 24.7645 | N/A |
| RANDOM FORESTS | 23.0341 | 24.02 |

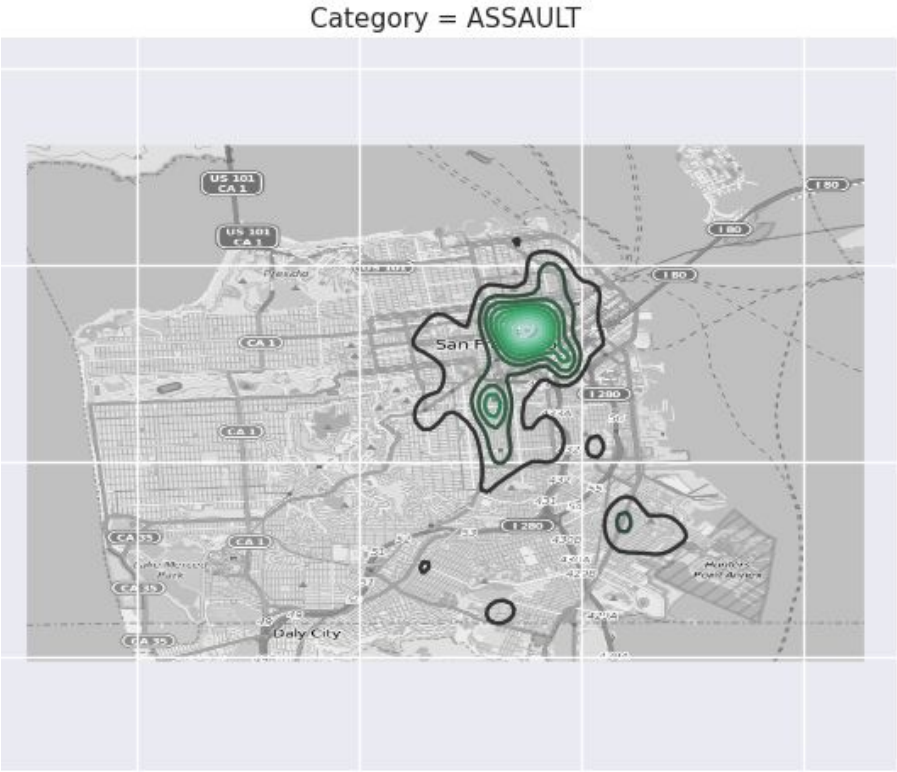
# Category Density



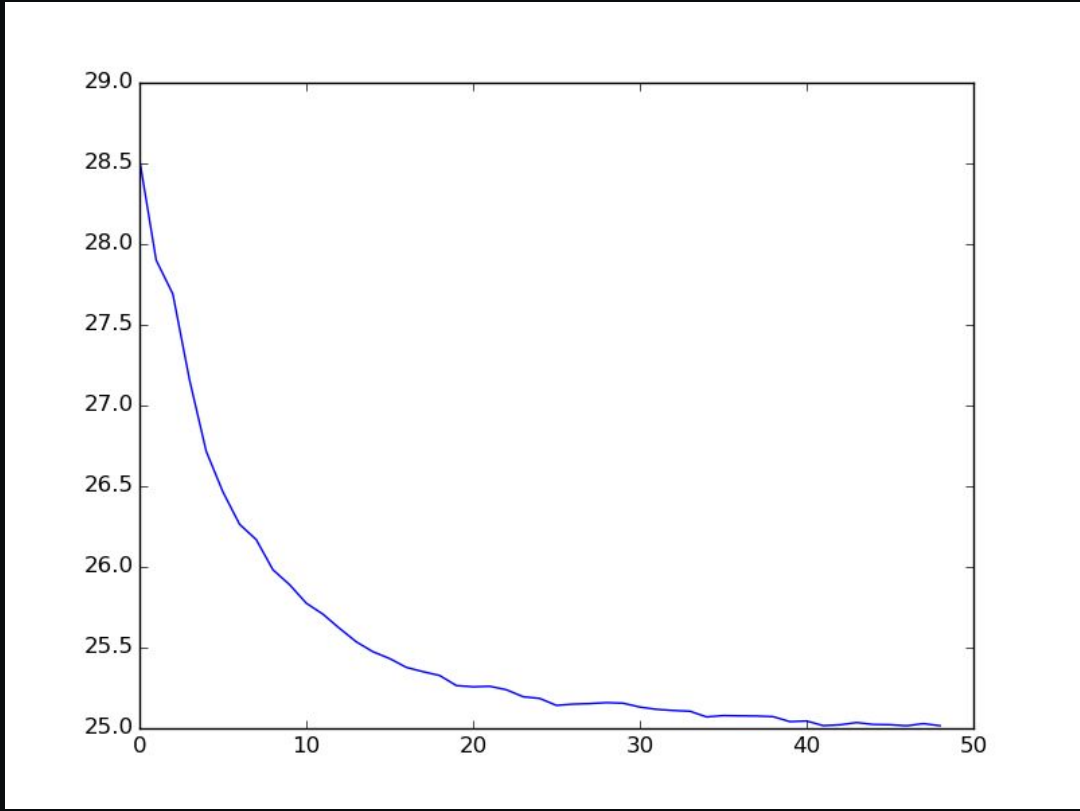
# Category Density



# CLUSTERING

- ➔ K clusters were created using the latitude and longitude features.
- ➔ Further classification was done within these clusters using the rest of the features.
- ➔ With KNN the log-loss was minimum with K = 40.
- ➔ K - Means was used for clustering



# CLASSIFICATION WITH CLUSTERING

| METHOD | LOG LOSS | |
|---|---|---|
| | EVALUATION | TESTING |
| SVM | 13.8976 | 13.9878 |
| NAIVE BAYES | 12.551 | 12.421 |
| DECISION TREES | 10.93 | 10.89 |
| RANDOM FORESTS | 6.1498 | 6.26 |
| LOGISTIC REGRESSION | 2.6523 | 2.60025 |

# LOGISTIC REGRESSION

- Best Performance with the lowest log loss compared to other methods
- Used all features
- Used all categories
- Used binomial variance instead of logit

The link function provides the relationship between the linear predictor and the mean of the distribution function.

| Family | Variance | Link |
|---|---|---|
| gaussian | gaussian | identity |
| binomial | binomial | logit, probit or cloglog |
| poisson | poisson | log, identity or sqrt |
| Gamma | Gamma | inverse, identity or log |
| inverse.gaussian | inverse.gaussian | 1/mu^2 |
| quasi | user-defined | user-defined |

# ACCURACY

- The evaluation technique used is log loss.
- The Kaggle Leaderboard #1 | log loss: 2.05079
- Our log loss: **2.60025**

# THANK YOU!