Predicting and Analysis of Crime in San Francisco

Sameer Darekar
MS in Data Science
Indiana University
sdarekar@iu.edu

Rohit Dandona
MS in Data Science
Indiana University
rdandona@iu.edu

Vignesh Sureshbabu
MS in Data Science
Indiana University
vsureshb@iu.edu

1. OBJECTIVES AND SIGNIFICANCE

Data mining is an innovative, interdisciplinary, and a growing area of research and analysis to build paradigms and techniques across various fields for deducing useful information and hidden patterns from data. Criminology and crime analysis is one such area which deals with huge amount of data containing past criminal records. One particular reason which makes the application of data mining techniques in this field apt is the intricate nature of correlation of data sets containing the criminal records.

The primary goals of the project are

1. Analyze and visualize the spatial and temporal relationship of crime with various attributes.
2. Predict the category of crime in a particular location on analyzing the input variables such as latitude, longitude, date, day and location.
3. Suggesting the safest paths between two places.

Considering the increasing crime rate reducing the crime is a major challenge for the police department. To reduce the crime one must be aware of the patterns of crime which take place in a particular area and how they are related with time. We want to try and make life easier for the police department to identify the patterns in crime and place the personnel and patrol vehicles accordingly

2. BACKGROUND

2.1 Important Concepts and Background Information.

We have used following concepts in our project we will briefly describe them in this section.

2.1.1 Naïve Bayes Classifier:

We have used two kinds of Naïve Bayes in our project depending on the approaches which we will discuss in the subsequent sections

a. Gaussian Naïve Bayes Classifier:

The core assumption of the Gaussian Naive Bayes classifier is that, when the data is continuous, the attribute values of each class follows Gaussian or Normal distribution [2]. For a continuous attribute, the data is segmented based on the class variable. For each segment of data belonging to a particular class variable, the mean and variance of the data is computed. Considering the mean as m, and variance v, for the corresponding class variable c, the probability distribution of the observed value n can be computed by substituting n in the equation:

$$P(x = n|c) = \left(\frac{1}{\sqrt{2}\prod v}\right) e^{-(v-m)2/2v}$$

b. Multinomial Naïve Bayes Classifier:

The classifier is based on the multinomial event model where the frequencies of generation of certain events are represented by the sample of the model. The events are generated by $(p_1,\ldots,p_i)$ where p is given by the probability that an event occurs, and more than 1 event in case of multiclass problems [2]. The samples of the model are feature vectors, where it can be rendered as histograms, representing the count of number of times an event occurs in a particular instance [2]. The likelihood of observing a histogram x is given by:

$$p(X|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i pki^{x_i}$$

Where $x = (x_1,\ldots,x_n)$ is the feature vector (in this case, histogram counting the number of times an event occurs) [2]

The estimate of frequency-based probability of the classifier can be zero, if the given class and the corresponding feature vector is not covered together in the training data, resulting in zero for the overall probability. The situation discussed is handled by pseudocount, where an amount is added to the number of observed cases for changing the expected probability [3]. It is done to ensure that the resulting probability is not zero in any case.

Generally, the multinomial Naïve Bayes classifier is considered as the special case of naïve based classifier, using multinomial distribution for all the features [4].

2.1.2 Decision Tree Classifier:

Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve a classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached

2.1.3 Random Forest Classifier:

The Random Forest algorithm is one of the important methods of ensemble classifiers. The methods of classification by aggregating the predictions of multiple classifiers for improving the classification accuracy are known as ensemble methods. A set of base classifiers is constructed from the training data, and the classification of the test records is declared by taking a vote on the combining predictions made by the individual classifiers [5]

2.1.4 Support Vector Machines (SVM) Classifier:

Support Vector Machines are supervised learning models, that can be used for both classification and regression data analysis [6]. There are 2 types of SVMs

a. Linear Support Vector Machine

b. Non Linear Support Vector Machine

In this project, we have used Linear SVM only as the RBF kernel of the classifier takes very long time to train and doesn't give much accuracy.

The core concept of Linear SVM is to search for hyperplane with the maximum margin, hence also known as maximum margin classifier [6].

2.2 Previous Work

A lot of work has been performed in crime classification in general. Predicting surges and hotspots of crime is a major sector in this field pursued for study. Important contributions in this sector has been made by Bogomolov in [7], where specific regions in London are predicted as crime hotspots using demographic data and behavioral data from mobile networks. In [8], Chung-Hsien Yu deploy classification algorithms such as Support Vector Machines, Naive

Bayes, and Neural Networks, to classify city neighborhoods as crime hotspots. Toole, in [9], demonstrated the spatio-temporal significance of correlation in crime data by analyzing crime records for the city of Philadelphia. They were also able to identify clusters of neighborhoods affected by external forces.

Significant contributions have also been made in understanding patterns of criminal behavior to facilitate criminal investigations. Tong Wang in [10], presents a procedure to find patterns in criminal activity and identify individuals or groups of individuals who might have committed particular crimes. The different types of crimes comprising traffic violations, theft, sexual crimes, arson, fraud, violent crimes and aggravated battery and drug offenses are analyzed using various data mining techniques such as association mining, prediction and pattern visualization. [11]. Similar kind of crime analysis on a data set from UCI machine learning repository using various data mining techniques has been carried out in [12]. In [13] different social factors such as population density, median household income, percentage population below poverty line and unemployment rate are used to classify different states of the U.S into three classes of crime low, medium, and high. Naïve Bayes classification and decision trees are used for the mining crime data.

In various Kaggle competitions, python libraries were used for implementing the common data mining techniques. Few of the notable libraries were Keras for neural networks [14], Lasagna for neural networks [15]. One another interesting approach used for similar analysis was model assembling [16].

2.3 What makes this Interesting?

Finding interesting patterns in crime which were previously unknown which can help the police department to mitigate crime in San Francisco and also help to arrest the guilty. If we know beforehand which crimes occur in which particular area, at what time then efficiency of police patrolling can be enhanced by multifold. All these results are embedded in the result section which is discussed later in the report.

3. METHODOLOGY

3.1 Dataset Information

The data set is obtained from the "San Francisco Crime Classification" Kaggle competition [1]. It has been derived from the San Francisco Police Department Crime Incident Reporting system. It includes crime data captured over 12 years. The data set is classified into train and test data sets which is available on the competition website [1].

The training data consists of nine columns – Dates (timestamp of the crime incident), Descript (description of the crime), DayOfWeek (Day of the crime incident), Pd District (name of the police department district), Resolution (state of the action against the crime, ex ARRESTED, BOOKED, etc.), Address (exact address of crime), X (longitude of the location of crime), Y (latitude of location of crime), and, Category (category of the crime incident). The category is the class variable for prediction. The total number of records in the training data set are 8,78,049.

The test data comprises of six columns, excluding Descript, Resolution, and Category the class variable from the training data set. The total number of records in the test data set are 8,84,261.

The majority class of the dataset is LARCENY/THEFT which consists of around 19% of the total training data so if we assign all values to this class we must get an accuracy of 19%, so the accuracy of the classifier should at least be more than 19% and this will be our baseline.

3.2 Evaluation Strategy Information

There are 2 data sets the training data which is labeled and the testing data which is unlabeled, we need to submit our prediction to the kaggle team to get our score and rank. For evaluating the accuracy of the classifier on the training dataset we used the 5 fold cross validation and we get the results which are shown in subsequent sections.

3.3 Methodology:

The project consists of three parts firstly the visualization of data secondly the classification of 39 class variables and predicting the safest distance, but after visualization of the

data on a map we found out that the crime is mostly concentrated on 2 regions of San Francisco and hence the other regions will be comparatively safe hence it is no used predicting the safest path hence we drop the idea.

For the enhanced understanding we have combined results with methodologies in this section. We used two approaches as follows

Approach 1: converting continuous variables to categorical

For testing these classifiers on the data we transformed the continuous variables from the dataset like Address, time, X (Longitude) and Y (Latitude) and further map these categorical variables with a particular number as the scikit learn library works only on numbers.

The dates column which was like *"5/13/2015 11:53:00 AM "* was further split into year, month, day and hour. X and Y were discretized into 5 bins.

Address which was of the form

1. 1500 Block of LOMBARD ST
2. OAK ST / LAGUNA ST

Was processed and only the street was extracted. The first one is an actual location whereas the second is the intersection of two streets. Thus the number of unique addresses was drastically reduced and later each unique address was mapped with a unique number.

The first 5 rows of the transformed dataset is as follows

DayOfWeek	PdDistrict	Address	X	Y	time	day	month	year
2	4	7677	3	0	23	13	5	2015
2	4	7677	3	0	23	13	5	2015
2	4	10556	3	0	23	13	5	2015
2	4	1641	3	0	23	13	5	2015
2	5	762	4	0	23	13	5	2015

Table 1: Initial Dataset

Initially, we selected all the variables and fed the data to the classifier, all of the classifiers gave the accuracy below par that is it was below 19% for all the classifiers as shown in table 2 left.

Later on we decided we find Pearson's correlation between all variables and category as follows and then select the top correlated terms as shown in table 2 right.

Classifier	Accuracy(%)
Gaussian Naïve Bayes	18.92
Decision Tree	3.35
Random Forest(estimators=2)	6.62
Support Vector Machines	4.21

	Category
DayOfWeek	0.001078
PdDistrict	-0.040674
Address	0.050874
X	0.02951
Y	0.00287
time	0.023524
day	0.000805
month	0.000008
year	-0.021803

Table 2: Initial Accuracy and Pearson Correlation

Then we dropped the columns having lower correlation coefficient that is nearer to 0 so we eliminated month, day and DayOfWeek. After running all the classifiers we got the accuracy as follows in table 3.

Classifier	Accuracy (%)
Gaussian Naïve Bayes	18.68
Decision Tree	10.04
Random Forest(estimators=2)	10.38
Support Vector Machines	5.32

Table 3: Accuracy after removal of least correlated columns

In Table 3 we can see that the accuracy of all classifiers have increased except that of Naïve Bayes. This clearly indicates that there is a nonlinear relationship between the class and other columns hence the accuracy is below the baseline that is 19%. So we need to combine this attributes such that they give us a good nonlinear correlation and hence increase the accuracy.

After this we also tried normalizing the values of each rows and columns using Z-Score Normalization and Min Max normalization but still the result did not change.

After many trial and errors and considering the combination of various attributes we got the best accuracy for the following combinations:

['DayOfWeek', 'PdDistrict', 'Address', 'day', 'month']

Classifier	Accuracy (%)
Gaussian Naïve Bayes	20.20
Decision Tree	13.46
Random Forest(estimators=2)	12.67
Support Vector Machines	7.37

Table 4: best accuracy after trial and error

The highest accuracy we got was that for Naïve Bayes that was just somewhat above 20% which was better than the benchmark but still needed some improvement, whereas the other classifiers were still performing badly. After all these results we thought of some other approaches to move on then we moved to the next approach

Approach 2: Binarization of categorical variables

In this approach the data points consist of only binary variables 1 and 0 depending if the attribute is present in the data point. As shown in figure 1.

pD_BAYVIEW	pD_CENTRAL	pD_INGLESIDE	pD_MISSION	pD_NORTHERN	pD_PARK	\
0	0	0	0	1	0	
0	0	0	0	1	0	
0	0	0	0	1	0	
0	0	0	0	1	0	
0	0	0	0	0	1	

pD_RICHMOND	pD_SOUTHERN	pD_TARAVAL	pD_TENDERLOIN	...	day_Tuesday	\
0	0	0	0	...	0	
0	0	0	0	...	0	
0	0	0	0	...	0	
0	0	0	0	...	0	
0	0	0	0	...	0	

day_Wednesday	X_0	X_1	X_2	X_3	X_4	X_5	Y_0	Y_1
1	0	0	0	1	0	0	1	0
1	0	0	0	1	0	0	1	0
1	0	0	0	1	0	0	1	0
1	0	0	0	1	0	0	1	0
1	0	0	0	1	0	0	1	0

Figure 1: first 5 rows of Initial Dataset after Binarization

Here we see that the first row of the dataset is in district Northern hence its value 1 and 0 for other districts and similarly day is Wednesday. We have made 5 bins for continuous variable X and 2 bins for 2 bins for Y after trial and error for optimacy.

Here instead of Gaussian we used multinomial Naïve Bayes which is best suited for such kind of data.

After running the classifiers over this transformation we got the following accuracies

Classifier	**Accuracy (%)**
Multinomial Naïve Bayes	21.96
Decision Tree	22.08
Random Forest(estimators=2)	22.06
Support Vector Machines	20.33

Table 5: initial accuracy for Binarization

Here we see that the accuracy has significantly increased and all classifiers give accuracy above the baseline that is 19%.

After this we created many extra features from the existing dataset which worked for increasing the accuracy of the classifiers.

Dimensionality reduction

a. Generating feature based on time:

   If we see from above visualizations (Insert figure numbers) we see that the state of crime changes with time firstly we created the feature awake which is 1 when the people are awake and 0 otherwise in night hours, the accuracy further increased after grouping them into 4 categories as early morning, morning, afternoon, evening, and night as we see in figure (). We further generalized and divided them into a group of 6 categories as per hours as follows

   [4,5,6,7], [8,9,10,11], [12,13,14,15], [16,17,18,19], [20,21,22,23], [0,1,2,3] as just arbitrary alphabets a,b,c,d,e,f respectively.

b. Generating feature based on address:

   There are around 9000 unique addresses in the previous approach after just taking the streets replicating them in this approach won't work as it would increase the dimensionality and in turn may suffer from the curse of dimensionality. Hence we decided to make the feature as intersection if the address contains "/" as it indicates intersection of two streets. It indicates 1 if it is an intersection and 0 otherwise

c. Generating features based on month:

   We also tried to group months into seasons viz. summer, winter, spring and fall but it had an adverse effect on accuracy hence later dropped it.

d. Generating features based on days of month:

   We also tried to generate feature first half indicating 1 if day is less than 15 and 0 otherwise and other combinations by grouping but this too had an adverse effect on accuracy hence dropped it.

After trial and error method with all the methods we got best accuracy till now considering the following feature vector.

['pD_BAYVIEW', 'pD_CENTRAL', 'pD_INGLESIDE', 'pD_MISSION', 'pD_NORTHERN', 'pD_PARK', 'pD_RICHMOND', 'pD_SOUTHERN', 'pD_TARAVAL', 'pD_TENDERLOIN', 'day_Friday', 'day_Monday', 'day_Saturday', 'day_Sunday', 'day_Thursday', 'day_Tuesday', 'day_Wednesday', 'a', 'b', 'c', 'd', 'e', 'f', 'intersection', 'X_0', 'X_1', 'X_2', 'X_3', 'X_4', 'X_5', 'Y_0', 'Y_1']

The accuracy of various classifiers we got is as follows:

Classifier	Accuracy (%)
Multinomial Naïve Bayes	22.5
Decision Tree	23.16
Random Forest(estimators=2)	23.06
Support Vector Machines	21.27

Table 6: Best accuracy till date

After this we submitted our results to Kaggle.

4. RESULTS

The results of the project is divided into two parts:

4.1 Visualization of Crime

The dataset has been intrinsically studies to portray embedded correlations. **Tableau and R** have been used to preprocess the data and create/present the visualizations.

The date field in the dataset is a timestamp present in the “MM/dd/yyyy hh:mm:ss” format and is not of much utility if used as such.

Tableau provides the facility of parsing through a timestamp through a utility called “DATEPART” and extract specific segments of the timestamp.

A part of the analysis required representing the data by five separate parts of the day. The following construct was used to derive a column in Tableau.

4.1.1 Volumetric Analysis:

There are 39 categories of crime report incidents available in the dataset including "Other offenses" and "Non-criminal".

Larceny and theft was the most common crime in San Francisco between January 1st 2003 and May 13th 2015, with a total of 174,900 reported incidents. The next two highest crime categories, "Other offenses" (98,281 reported incidents) and "Non-criminal" (98,172 reported incidents), are omitted from discussion in the analyses in Figure 2. Instead, the key focus is on the following high crime categories: larceny and theft, assault (76,876 reported incidents), drugs (53,971 reported incidents), vehicle theft (53,781 reported incidents), vandalism (44,725 reported incidents), drugs (21,910 reported incidents) and burglary (19,226 reported incidents).

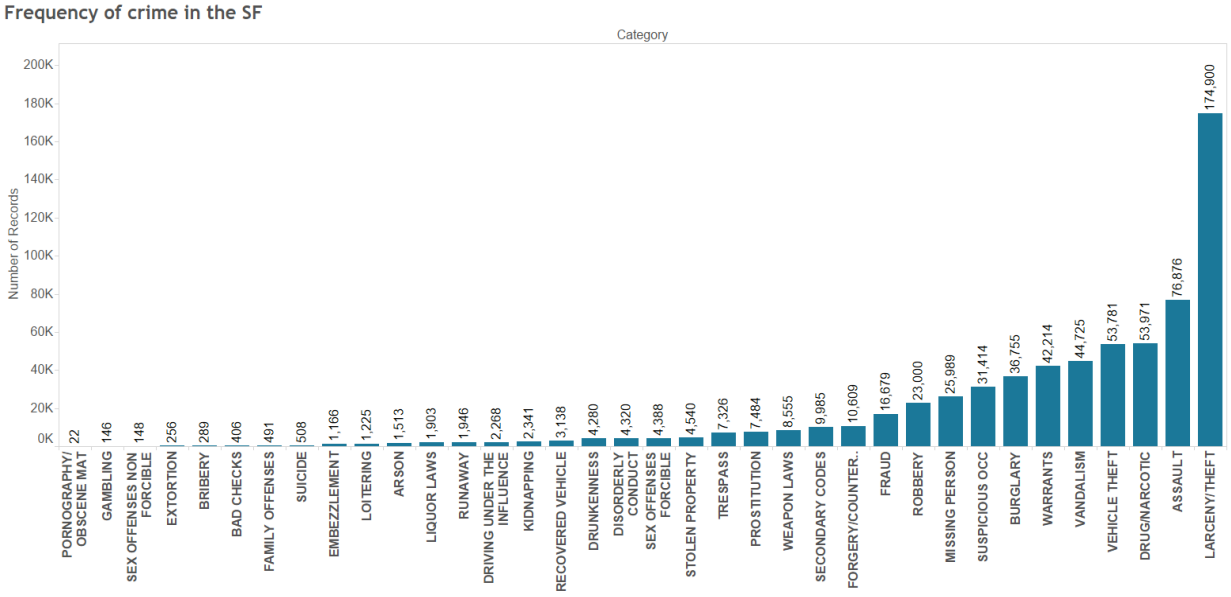


Figure 2: Crime trends in San Francisco

4.1.2 Month wise distribution:

Since the crime incidents recorded in 2015 are only till May, those records have been removed for this analysis. The month of February is the safest and is observed to have the least crime incidents reported. October is the least safe with the maximum number of reported crime incidents (Figure 3).

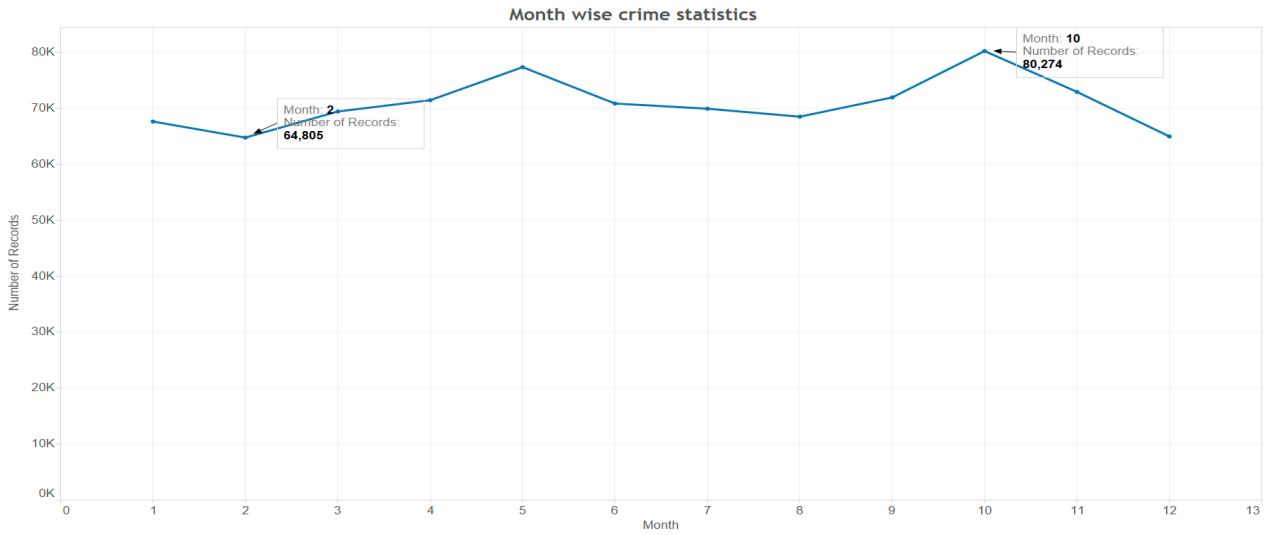


Figure 3: Month wise crime statistics

4.1.3 Region wise distribution:

A contour plot to represent the region wise density of crime leveraging the available latitude/longitude information of the reported crime, is as follows:

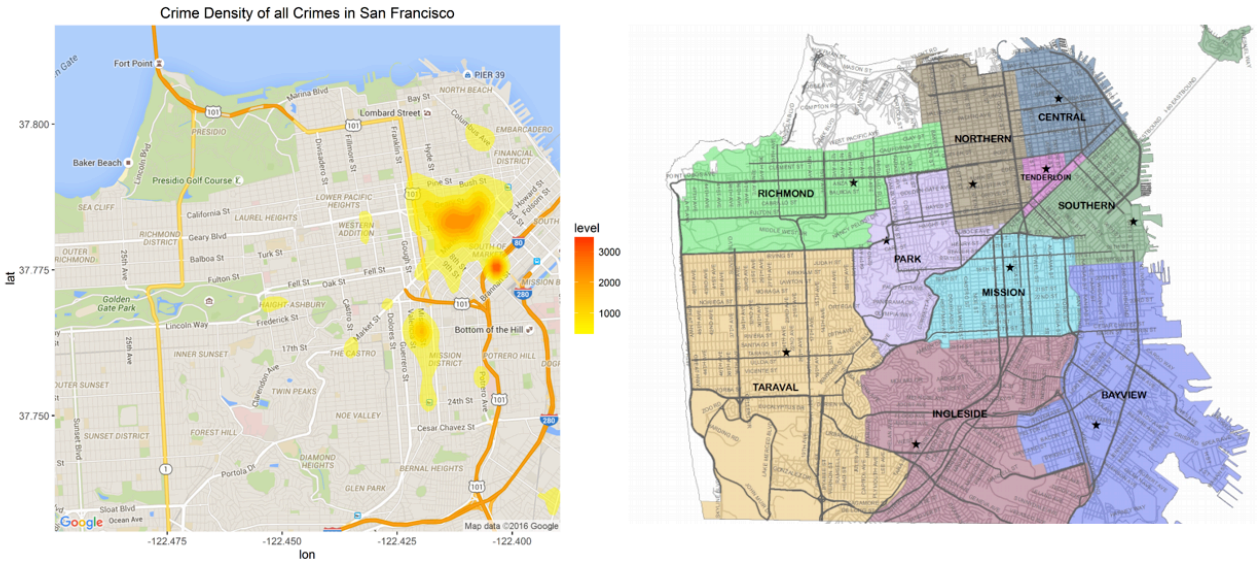


Figure 4: Regional crime density

A giant hotspot can be observed in the Southern and Tenderloin region (Figure 4) with relatively less dense plots in the surrounding neighborhoods. Majority of the crime seems to be concentrated in these and surrounding regions.

Plotting a line graph to represent crime in each region on each day of a week uncovers some interesting trends as shown in Figure 5:

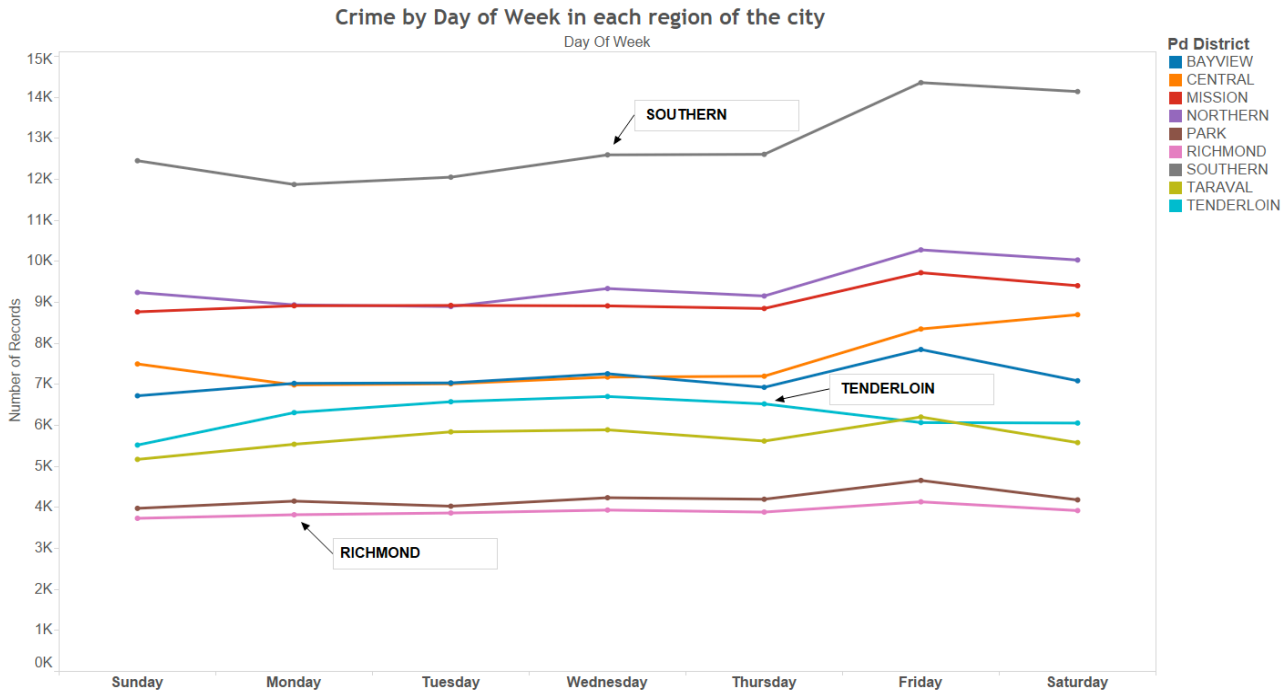


Figure 5: Day wise crime trends per region

Clearly, the Southern region is the most notorious with the maximum number of crimes reported followed by the Northern region. Richmond is the safest place to be in San Francisco.

4.1.4 Crime rate by day of the week:

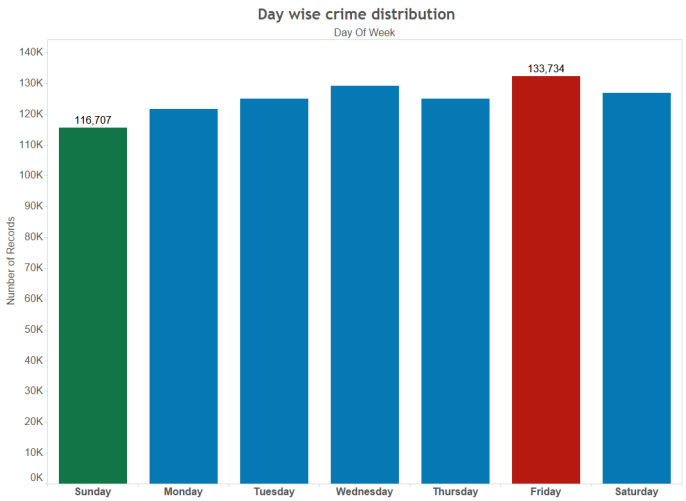


Figure 6: Day wise crime trends

The rate of crime is observed to increase over the weekend (Figure 6). The number of reported incidents seems to be the maximum on Fridays with a total number of 133,743 reported incidents. For almost all regions, the same trend can be observed. Tenderloin, however, is an exception as the crime rate in this region decreases over the weekend and is the highest on Wednesdays.

Wednesday, Friday and Saturday are the most crime prominent days. Interestingly, the least amount of incidents were reported on Mondays (116,707 reported incidents) followed by Sundays. In the Southern, Northern, Central and Mission districts, the number of crimes increased sharply on Friday and Saturday, then declined for rest of the week.

4.1.5 Specific crime trends

We examine the crime trends of a hand full of crimes (Kidnapping, Fraud, Robbery, Missing Person, Burglary, Vandalism, Vehicle Theft, Drug/Narcotics, Assault and Larceny/Theft). The following trends are observed (Figure 7):

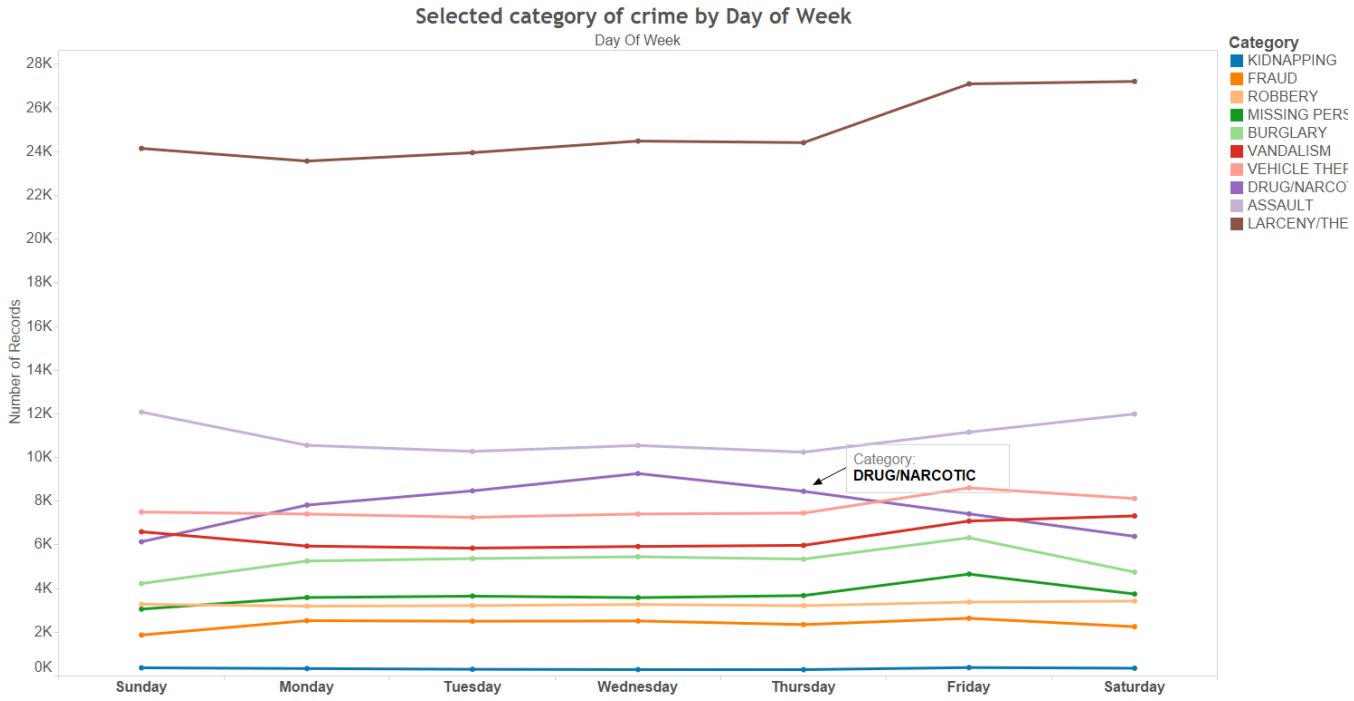


Figure 7: Specific crime study

Interestingly, Drug/Narcotic related crimes decrease over the weekend (unlike other categories). The related reported incidents are the maximum on a Friday.

Among the high crime categories, larceny and theft tend to occur on Friday and Saturday. On the other hand, the occurrences of assault steadily increased from Thursday to Sunday, while burglary generally occurred more often during the weekdays than the weekends. Drug crimes were most commonly reported on Wednesday and least reported during the weekends, and robberies happened by approximately the same frequency every day.

4.1.6 Hourly trend analysis for specific crimes:

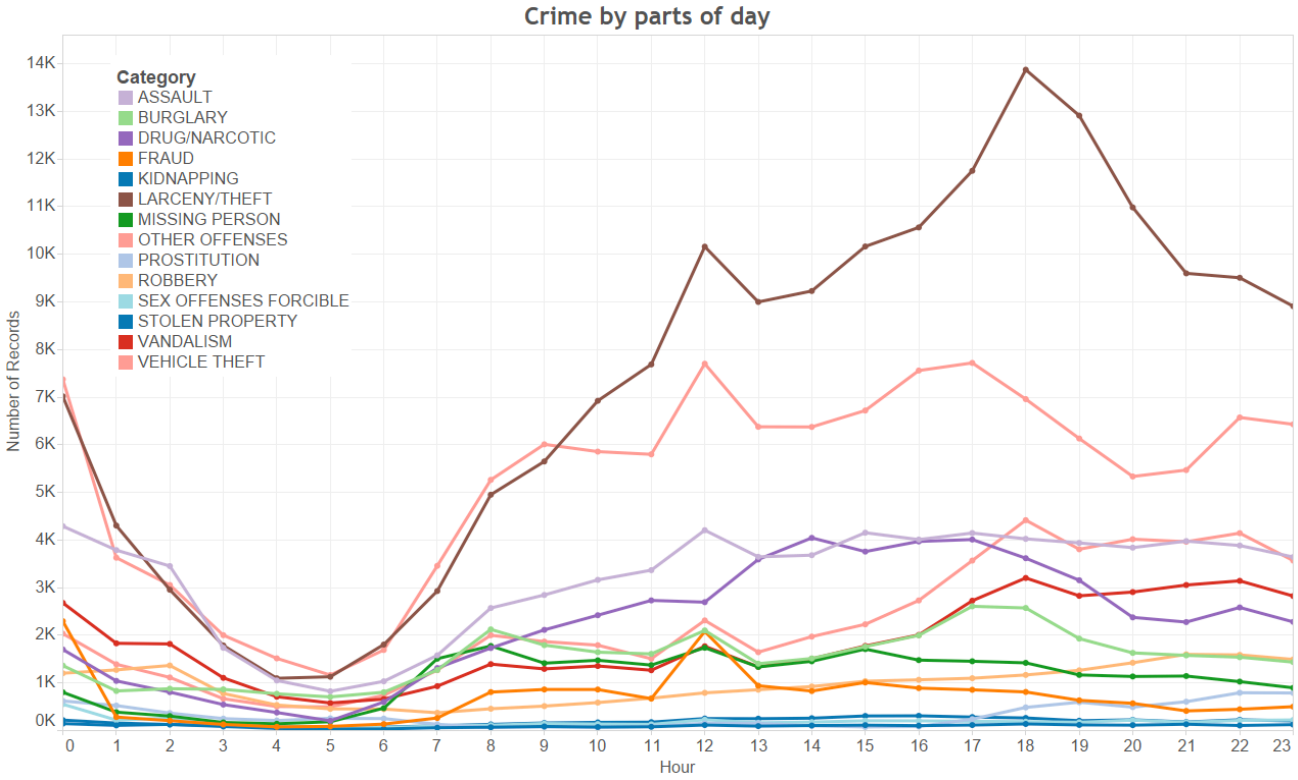


Figure 8: Hourly trend of each crime (selected)

The hourly trends unravel some interesting crime facts:

5 am is the safest part of a day with 8,637 reported incidents and 6 pm is the most dangerous hour with 55,104 reported incidents.

Surprisingly, 12 pm is the second most dangerous hour during a day and in fact is the hour where incidents reported under some crime categories is maximum.

Kidnapping and stolen property incidents take place uniformly throughout the day

4.1.7 Violent Crimes in San Francisco: Assault, Robbery and Sex Offences (forcible)

Distribution:

Figure 9: Violent crime geographical implementation

4.1.8 Violent crime density (each) by parts of day:

Discretization of time into following groups for visualization.

Early Morning:	4am to 8am	Afternoon:	12pm to 5 pm
Morning:	8am to 12pm	Evening:	5pm to 9pm
Night:	9pm to 4am		

4.1.9 Robbery:

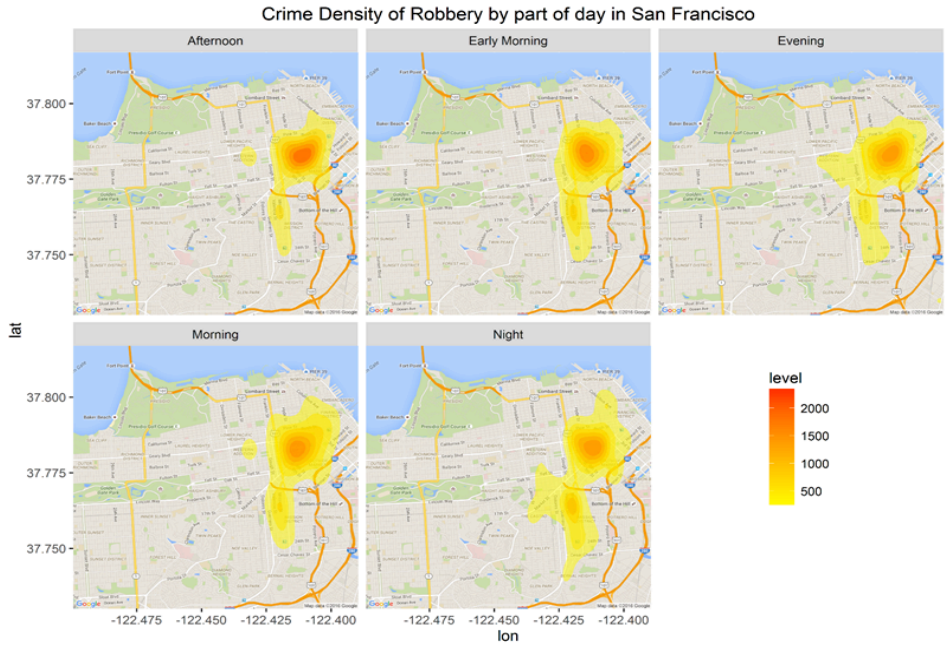


Figure 10: Geographical representation for Robbery

A slight shift in epicenter of the crime can be observed over different parts of the day.

4.1.10 Sex Offences (forcible):

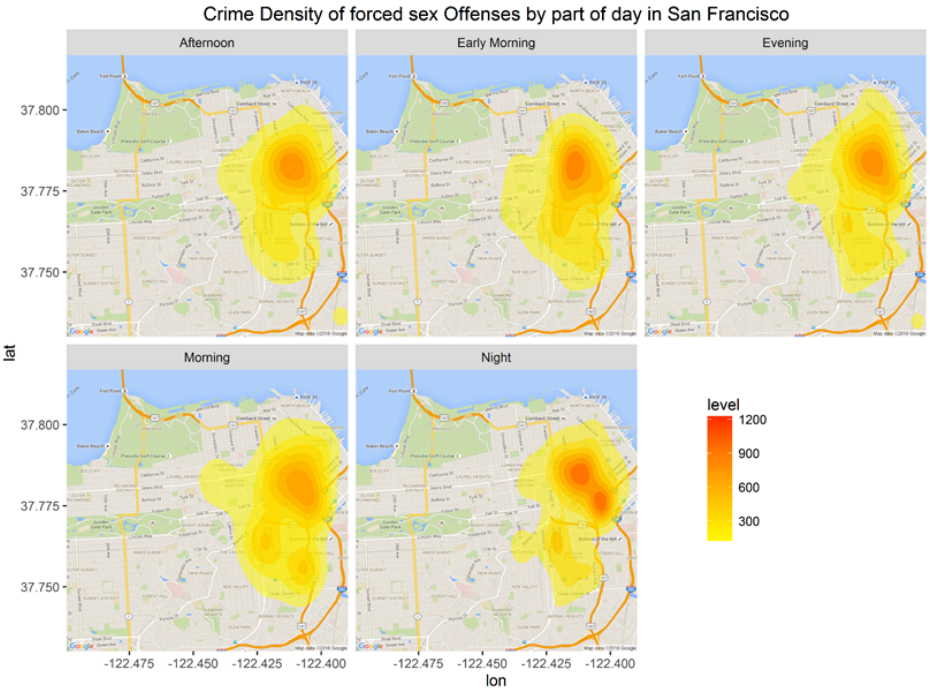


Figure 11: Geographical representation for Sex Offences

Different Patterns can be observed at different time with increase in density at night in figure 11.

4.1.11 Larceny and Theft:

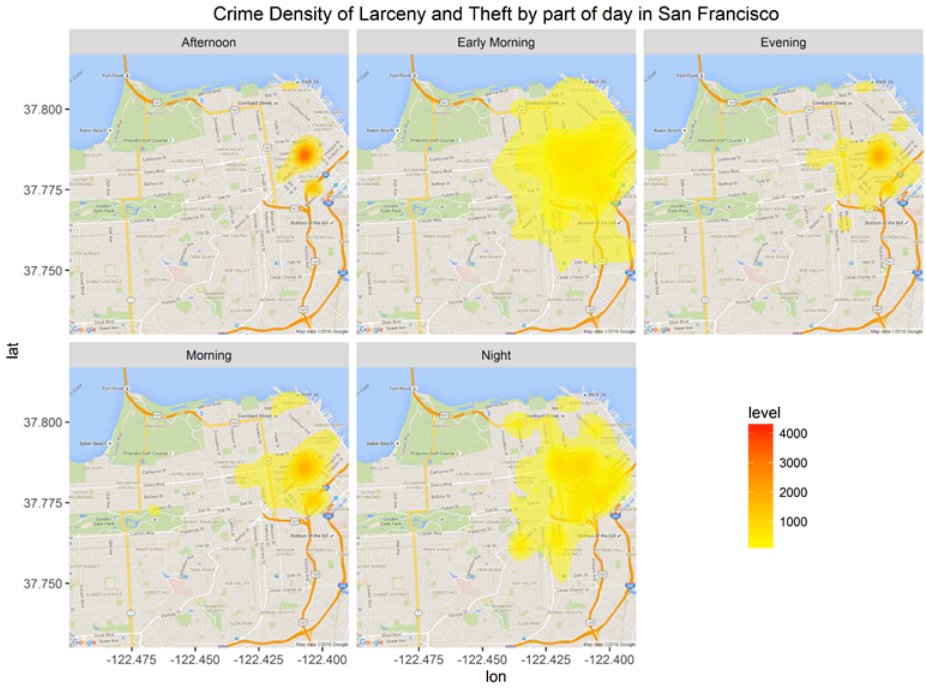


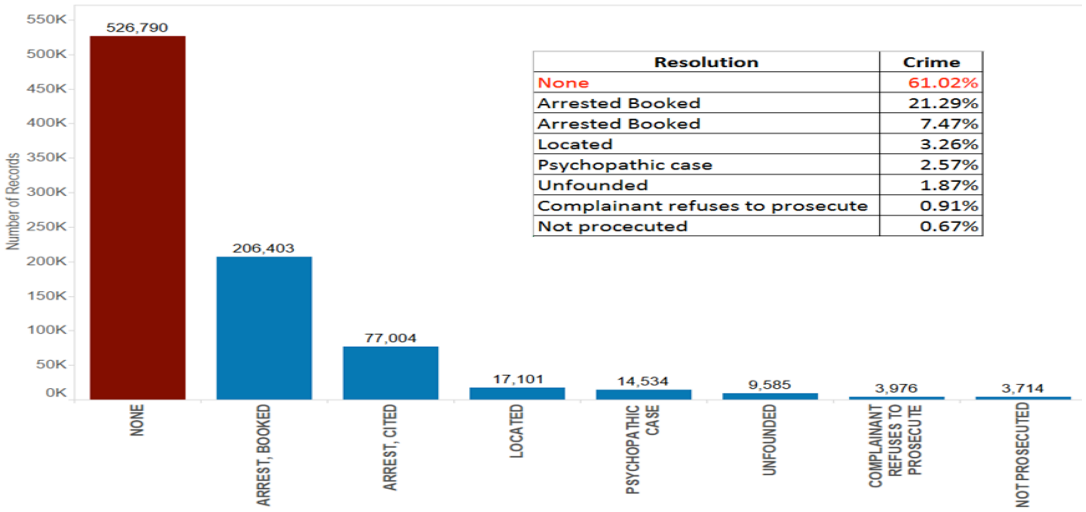
Figure 12: Geographical representation for Sex Offences

Crime is spread across multiple regions early in the mornings and at night. It is concentrated on a single region in the afternoon.

4.1.12 Food for thought

Figure 13 shows a bar graph representing the percentage of incidents which have been investigated by law enforcement agencies vs the ones that haven't been.

Surprisingly, for 61.02% of the reported incidents no conclusive action has been taken.



Resolution	Crime
None	61.02%
Arrested Booked	21.29%
Arrested Booked	7.47%
Located	3.26%
Psychopathic case	2.57%
Unfounded	1.87%
Complainant refuses to prosecute	0.91%
Not procecuted	0.67%

Figure 13: Incident resolution percentage

The following graph (Figure 14) shows the top three streets where the maximum crime incidents were reported [Bryant Street, Market Street, Mission Street]:

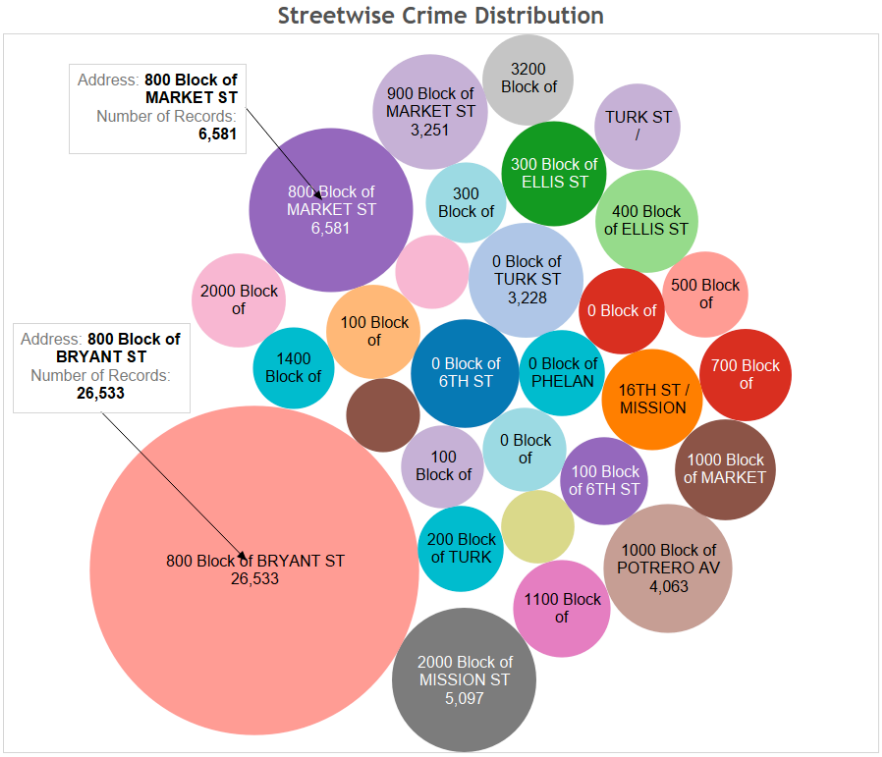


Figure 14: Street crime analysis

The police commission office is located on 850 Bryant Street which is just near the 800 Bryant Street which has the most crime rate in terms of streets (Figure 15)

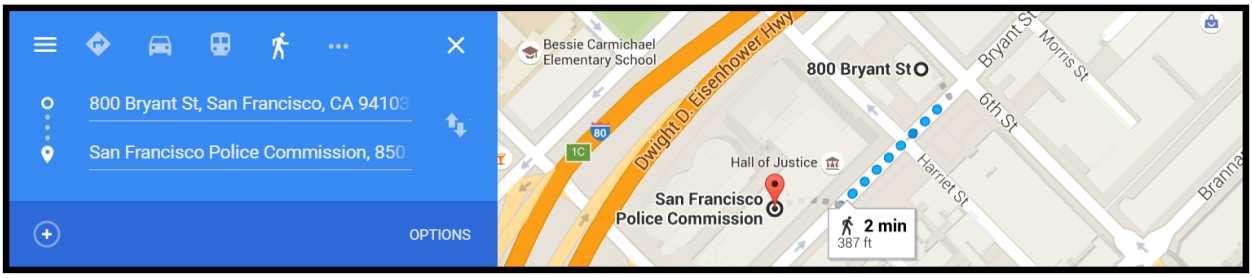


Figure 15

Market Street, with the second highest number of crime incidents reported, has a police station 0.5 miles away from it (Figure 16):

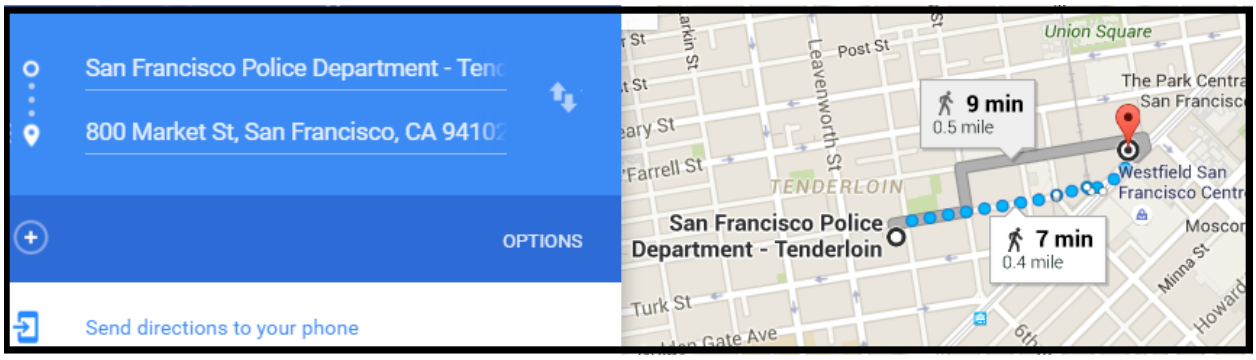


Figure 16

4.2 Predictive Analysis

Final predictive results of the analysis (approach 1 vs approach 2 as described in the methodologies section of this report) in table 7:

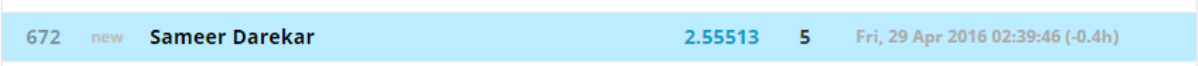
Classifier	Approach 1	Approach 2
	Accuracy (%)	Accuracy (%)
Gaussian Naïve Bayes	20.2	0.6
Multinomial Naïve Bayes	9.18	22.5
Decision Tree	13.46	23.16
Random Forest(estimators=2)	12.67	23.06
Support Vector Machines	7.37	21.27

Table 7: best accuracy after trial and error

The highest accuracy we got via approach 1 was that for Naïve Bayes that was just somewhat above 20% which was better than the benchmark but still needed some improvement, whereas the other classifiers were still performing badly. After all these results we thought of some other approaches to move on then we moved to the next approach

In approach 2, the results obtained after binerization and dimensionality reduction are much more superior to the once obtained in approach 1.

We submitted our results to Kaggle with the decision tree classifier which had the best accuracy on training dataset. We got a **rank of 1149 with a score of 2.64 out of around 1886 competitors**, and later when we sent our submission for Naïve Bayes, we got a rank of **672 with a score of 2.55.**



From the link *https://www.kaggle.com/c/sf-crime/leaderboard*

The test dataset seemed to more favorable to Naïve Bayes classifier than the decision tree.

5. INDIVIDUAL TASKS

The core ideas of the project were formed by the team after a lot of ground work and brain storming sessions. Equal efforts were put in by the team on the exploratory data analysis part.

Rohit: Performed data visualizations on Tableau, developed the SVM classifier using scikit learn in python. Initially he tried implementing it without pre installed packages in python. The drawback was poor accuracy and time consumption during training the classifier. Finding patterns from visualizations.

Sameer: Worked on data preprocessing. Developed Naïve Bayes, Decision Tree and Random Forest classifiers using scikit learn in python. Contributed by helping me in spatial data analysis in R. Formulated approach to improve the accuracy. Worked on creating new features from existing features.

Vignesh: Contributed in data pre processing. Developed spatial visualizations using the ggmap2 library with suggestions from Sameer. Worked on developing naïve bayes classifier without libraries, poor accuracy and took around several hours for training the classifier. Contributed on developing naïve bayes using scikit learn. Worked on decision trees using R. Finding patterns from visualizations.

6. CONCLUSIONS AND FUTURE WORK

Based on the visual analysis, the following pointers can be presented to the law enforcement agencies of San Francisco, to enforce an enhanced crime check:

- Tenderloin and Southern are identified as the most notorious districts and need special attention at any part of the day or month. The crime rate is directly proportional to how famous a neighborhood since Tenderloin and Southern are downtown regions with the maximum number of social hotspots.
- The crime rate has been observed to be the maximum at 6 pm and minimum at 5 pm, suggesting direct proportionality to the density of population on the streets.

- SFPD needs to work more proactively as the 61% of cases are still not resolved. Amazingly, there are major crime hotspots within a mile radius of some police establishments. The patterns discovered could be instrumental in deploying patrol and planning the man power of the police department.
- Finding out why February has the lowest and October has the highest reported crime, may be an added incentive towards efficient crime check.
- Decision tree works best on training data but Multinomial Naïve Bayes gave us a better rank on Kaggle on test dataset.
- Making use of Principal Component Analysis (PCA) for dimensionality reduction and then checking the accuracy of the classifiers may improve performance.
- In the future we can analyze the twitter feeds of a particular area and combine the idea of sentimental analysis to observe any relations or interesting patterns relating the general sentiment of the local people and crime rate. Planning to implement the same with Prof. Muhammad Abdul-Mageed during summer.
- Deep Learning techniques like Theano and Keras libraries can be used to improve the accuracy.

7. REFERENCES

[1] San Francisco crime data set. DOI= https://www.kaggle.com/c/sf-crime/data

[2] Naïve Bayes Classifier DOI=https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[3] Pseudocount DOI=https://en.wikipedia.org/wiki/Pseudocount

[4] Difference between Multinomial and Gaussian Naïve Bayes DOI=http://stats.stackexchange.com/questions/33185/difference-between-naive-bayes-multinomial-naive-bayes

[5] Introduction to Data Mining - by P.-N. Tan et al., Pearson 2006 (Section Random Forests Pg. 293

[6] Support Vector Machines DOI= https://en.wikipedia.org/wiki/Support_vector_machine

[7] Bogomolov, Andrey and Lepri, Bruno and Staiano, Jacopo and Oliver, Nuria and Pianesi, Fabio and Pentland, Alex.2014. Once upon a crime: Towards crime prediction from demographics and mobile data, Proceedings of the 16th International Conference on Multimodal Interaction.

[8] Yu, Chung-Hsien and Ward, Max W and Morabito, Melissa and Ding, Wei.2011. Crime forecasting using data mining techniques, pages 779-786, IEEE 11th International Conference on Data Mining Workshops (ICDMW)

[9] Toole, Jameson L and Eagle, Nathan and Plotkin, Joshua B. 2011 (TIST), volume 2, number 4, pages 38, ACM Transactions on Intelligent Systems and Technology

[10] Wang, Tong and Rudin, Cynthia and Wagner, Daniel and Sevieri, Rich. 2013. pages 515-530, Machine Learning and Knowledge Discovery in Databases

[11] Chen, Hsinchun, et al. ”Crime data mining: a general framework and some examples.” Computer 37.4 (2004): 50-56.

[12] Iqbal, Rizwan, et al. ”An experimental study of classification algorithms for crime prediction.” Indian Journal of Science and Technology 6.3 (2013): 4219-4225.

[13] UCI Machine Learning Repository (2012). Available DOI= http://archive.ics.uci.edu/ml/datasets.html

[14] Keras: Theano-based Deep Learning library (2015). Available DOI= http://keras.io

[15] Lasagne: a lightweight library to build and train neural networks in Theano (2015). Available DOI= https://github.com/Lasagne/Lasagne

[16] Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738.