Title
An Informative and Predictive Analysis of the San Francisco Police Department Crime Data

Author
Wu, Xiaoxu

Publication Date
2016

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

An Informative and Predictive Analysis of the San Francisco Police Department Crime Data

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Xiaoxu Wu

2016

Abstract of the Thesis

An Informative and Predictive Analysis of the San Francisco Police Department Crime Data

by

Xiaoxu Wu

Master of Science in Statistics

University of California, Los Angeles, 2016

Professor Yingnian Wu, Chair

It is the responsibility of the San Francisco Police Department to protect the local community from various crimes and to improve the local security environment. With the development of modern statistics tools, we can learn from the past data and give suggestions for future strategy.

In this thesis, we study the San Francisco Police Department crime dataset from 01/01/2013 through 05/13/2015. Informative analysis regarding timing and location for different crimes are examined. Visualization methods are proposed for related features. We also discuss possibilities of predicting the crime categories given time and location data using the k-nearest-neighbor model and the logistic regression model.

The thesis of Xiaoxu Wu is approved.

Nicolas Christou

Robert L. Gould

Yingnian Wu, Committee Chair

University of California, Los Angeles

2016

To my mother . . .

who—among so many other things—

saw to it that I learned to touch-type

while I was still in elementary school

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

San Francisco Police Department has been working hard to make a better living for the local residents as well as tourists. With in depth analysis of the historical dataset provided by SF open data, we would be able to discover regular patterns for the crimes, which will help improve the safety and help police officers to better guard the city.

In addition, the methodology described in this these can be applied to future data or data in other cities with minimal modifications.

The organization of this thesis is as follows. We will start with data preprocessing in Chapter 2, where we will discuss the data overview, data cleaning and feature extraction. Chapter 3 will focus on informative analysis of the data, where we can see when and where do crimes occur more frequently. In Chapter 4, we will be discussing classification problems, where we can see the possibility of predicting the crime category using location and time information. Different models will be used and classification results will be further analyzed and compared.

CHAPTER 2

The SFPD Crime Data

2.1 Dataset Description

The dataset used in this thesis comes from SF OpenData [CS15]. It provides total of 878049 incidents from 1/1/2003 through 5/13/2015 across all San Francisco's 10 PD districts. Nine data fields are provided for each incidents.

- Dates: date and time when the incident happened. This field has the format of "yyyy-mm-dd HH:MM:SS". One example would be "2015-05-13 23:53:00". The date ranges from 2003-01-01 through 2015-05-13 and the time ranges from 00:00:00 through 23:59:59.

- Category: category of the incident. Total of 39 categories included in the dataset are listed in Table 2.1.

- Description: detailed description of the incident. This data field provides more detailed information regarding the incident. For example, the type of the property lost or damaged during a theft or vandalism.

- DayOfWeek: the day of week when the incident happened. There are seven possible levels for this field: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday.

- PdDistrict: the PD district where the incident happened. There are total of 10 PD district in San Francisco: Bayview, Central, Ingleside, Mission,

warrant	other offenses	larceny/theft
vehicle theft	vandalism	non-criminal
robbery	assault	weapon laws
burglary	suspicious occ	drunkeness
forgery/counterfeiting	drug/narcotic	stolen property
secondary codes	trespass	missing person
fraud	kidnapping	runaway
driving under the influence	sex offenses forcible	prostitution
disorderly conduct	arson	family offenses
liquor laws	bribery	embezzlement
suicide	loitering	sex offenses non-forcible
extortion	gambling	bad checks
trea	recovered vehicle	pornography/obscene mat

Table 2.1: list of all 39 crime categories

Northern, Park, Richmond, Southern Taraval, Tenderloin. The map of these 10 PD district are shown in Figure 2.1.

- Resolution: how the incident was resolved .
- Address: the street address when the incident happened.
- X: longitude of the incident location. San Francisco city longitude ranges from -122.5136 to -122.3649.
- Y: latitude of the incident location. San Francisco city latitude ranges from 37.70788 to 37.81998. Notice that there are 67 rows in the dataset where the location recording was not correct. An erroneous value of (-120.5, 90) was recorded for these incidents.

The first six rows of the dataset are shown in Figure 2.2.

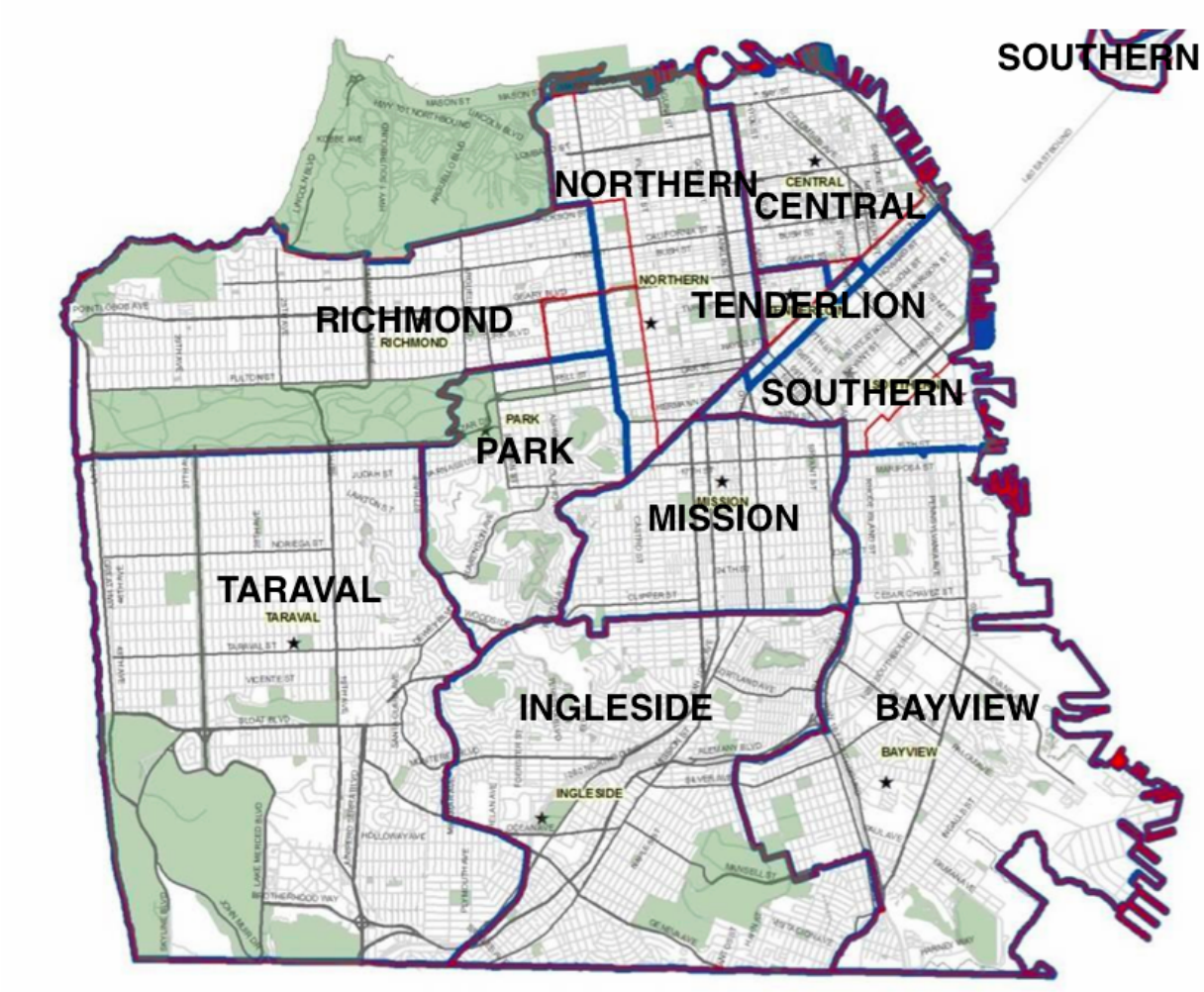


Figure 2.1: San Francisco PD District Map (figure modified from [Gro14])

	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution
1	2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED
2	2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED
3	2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED
4	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE
5	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	PARK	NONE
6	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM UNLOCKED AUTO	Wednesday	INGLESIDE	NONE

	Address	X	Y
1	OAK ST / LAGUNA ST	-122.4259	37.77460
2	OAK ST / LAGUNA ST	-122.4259	37.77460
3	VANNESS AV / GREENWICH ST	-122.4244	37.80041
4	1500 Block of LOMBARD ST	-122.4270	37.80087
5	100 Block of BRODERICK ST	-122.4387	37.77154
6	0 Block of TEDDY AV	-122.4033	37.71343

Figure 2.2: First six samples of the dataset

2.2 Data Cleaning

There are errors in the datasets that need to be fixed before proceeding to the data analysis.

2.2.1 Data Labeling

This large dataset is manually collected and the labelings are very noisy. We would like to fix incorrect or improper labelings with our best effort.

As is mentioned in the previous section, among the 39 crime categories, there is one type called "NON-CRIMINAL". Looking into the descriptions for these incidents, we can see that these incidents report non-criminal reports, e.g. death report, found property, aided cases etc. Since we are only interested in criminal incidents that can affect location resident's security environment, these incidents are deleted from the dataset.

In addition, we would like to eliminate the crime category "OTHER OFFENSES" because this category includes various crime incidents that are hard to label with the existing crime categories. Understanding the patterns of these crime incidents does not help predict the possible crime incidents.

We also noticed that incidents with a label "TREA" share the description of "trespassing or loitering near posted industrial property". Thus we can merge these incidents with the trespassing category. For the rest of this thesis, we will be studying the behavior of the rest 36 crime categories.

2.2.2 Longitude-latitude Errors

We've also mentioned that there are 67 events that have incorrect location information. While the San Francisco city has the range of -122.5136 to -122.3649 in longitude and 37.70788 to 37.81998 in latitude, those 67 events without cor-

rect location recorded was filled with the value of -120.5 for longitude and 90 for latitude. Fortunately, the PdDistrict field is still available for these events. Observing this, the mid-location of the corresponding PdDistrict was assigned to these events with erroneous location.

After these data cleaning steps, we are ready for some preliminary feature extraction in the next section.

2.3 Preliminary Feature Extraction

In order to present the most meaningful information from the data fields, the dataset was preprocessed with the following features extracted.

The first temporal feature is the "Year" extracted from the "Dates" data field. We are interested in seeing the crime trend during the past 12 years. Examples of interesting questions include whether the number of crimes decreases significantly over time, whether proportion of different crime categories change over time.

The second temporal feature is the "Month" extracted from the "Dates" data field. We are interested in seeing the crime trend over different seasons. Examples of interesting questions include what type of crime dominates for summer season, what type of crime is most significant during holiday season (i.e. November and December).

The third temporal feature is the "Hour" extracted from the "Dates" data field. We also expect to see hour of the day influencing number of crimes. Examples of interesting questions are what hours do we see largest number of crimes happening, what are the most frequent crimes during working hours, what are for night hours.

The fourth temporal feature is the "DayOfWeek". The days of the a week may have a significant influence over the type of crimes. For example, we may expect more drunkenness on Fridays and Saturdays, while home burglary may happen

more frequently during weekdays.

In addition, location of the crime are also of great importance. The spatial features that we are interested in are the "PdDistrict" and longitude-latitude. Crime spatial density distribution plotted using longitude-latitude information can be of great interest. The trends of different crimes across PD districts can also advice PD management policy.

CHAPTER 3

Informative Analysis

In this chapter, we will analyze the timing of different types of crime as well as the location. Visualization methods for related features will be proposed. These informative analysis can be of guiding value for local PD to better guard the community.

3.1 Temporal Analysis – crime categories and time

3.1.1 Interpreting the Year and Month Fields

The dataset includes all crime incidents from 2013-01-01 to 2015-05-13. We would like to start our temporal analysis for the general trends of the crimes on the scale of years and months. Figure 3.1 shows the total counts of crimes in each month for the top 5 city wide most common crime categories. Plots for all 36 crime categories can be found in Appendix A. Note that we have only 13 days for the May of 2015 and we need to normalize the total counts to 31 days for easier comparison. In order to eliminate the noise from month to month, the moving average over 12 months are also calculated and plotted in the red line. We can see a constant drop in drug crime after the year 2009. This can be possibly related to the drug lab scandal in 2009 according to NBC news [Gre10]. Figure 3.1 also shows a drastic drop of vehicle theft after 2006. This is due to the fact that before 2006, the vehicle recovery cases were mislabeled as vehicle thefts [Jen15].

3.1.2 Interpreting the DayOfWeek and Hour Fields

While the trend on the year and month scale can provide valuable information about the overall changing of the security environment, we are also interested in more detailed temporal analysis that can possibly imply patterns that we can learn from. Day of a week and hour of a day are two important periodic features that have potential impact on the crime categories. We can plot the pattern for each crimes changing with the day of the week and the time of the day. Figure 3.2 shows the overall crime rate heat map and five representative examples.

The overall heat map indicates that the 1am to 7am period has the lowest crime counts. This can be related to the fact that people are sleeping. Moreover, we can see the low crime rate patterns have a 2 hour shift on Saturdays and Sundays, which is matching the fact that people go to bed later and wake up later on weekends. In addition, the highest crime rates occurs around evening time from 6pm to 8pm on weekdays. And this peak extends through later of the night for Fridays.

While lots of crimes share similar patterns with the overall heat map with higher frequency near evening time and lower frequency after midnight, there are interesting facts for some crime categories. Figure 3.2d shows that prostitutions are rarely seen during daytime. Figure 3.2f shows that arsons usually occurred at late night through after midnight. Figure 3.2e shows that drunkenness happens more frequently Friday night and Saturday night, especially after midnight. The heat maps for all 36 crime categories are listed in Appendix B.

3.2 Spacial Analysis – crime categories and location

3.2.1 Interpreting the PD District Field

We've mentioned in Chapter 2 that there are 10 PD districts in San Francisco city. The total number of crimes within each district are summarized in Figure 3.3a. Noticing the difference in the size of the PD districts, we normalize the total number of crime counts by the local population using the demographic data reported in 2000 by US Census Bureau [Gro08]. The ranking among the PD districts was changed due to normalization with the population statistics (Figure 3.3b). One example is the Tenderloin district, where the total number of crime incidents ranks No. 6 among the 10 PD district of San Francisco city. But after normalizing with the local population, the Tenderloin becomes the top 2 PD district with the highest crime incidents counts.

We can take a closer look at the top crime categories in each PD district. Since we have a total of 36 crime categories, plotting all of them can be overwhelming and not informative. Instead, we plot only the five of the citywide most common crime categories and study their distributions among the 10 PD districts (Figure 3.4a). Larceny/Theft has the leading counts of incidents in most district. But the leading crime in the Tenderloin district is drug. This is matching the fact that Tenderloin is the most popular place for drug dealers. A similar bar plot showing the distribution of the top five crimes among the 10 districts is also informative (Figure 3.4b). Most Larceny/Theft happened in Southern district while most drug crimes happened in Tenderloin. We can also see that vehicle thefts and vandalisms are evenly distributed among the 10 districts.

3.2.2 Interpreting the Longitude-latitude Field

Crime distributions can be created with ggmap package in R [KW13]. Examples of crime distributions are presented in Figure 3.5. And the distributions of all 36 crime categories are listed in Appendix C. Plotting the crime distribution density over a map is a very informative visualization helping people to detect the hotspot for each crime category. For example, we can see that the hotspots of drug and prostitution are located near the area of the Tenderloin, the famous red light district and drug dealers' place of the San Francisco city. On the other hand, vehicle theft has an even distribution across the whole city. This observation is matching with the claim we made in section 3.2.1.

3.2.3 Interpreting the Address Field

In addition to the PD district and longitude-latitude information provided by the dataset, there is also valuable information in the address column. Observing that the addresses can be categorized into two types: 1) number of blocks for certain street and 2) the cross of two streets. Fortunately, the type of an address can be easily extracted by parsing the "/" punctuation. Overall, there are 26.8% of crimes that happened at a street corner. Figure 3.6 shows the percentage of incidents that happened at a street corner for each crime category. We can see a large amount of driving under influences happened at a street corner because traffic related crimes can occur more frequently at a street corner. And it is reasonable that crimes like suicide or sex offenses seldom happened at a street corner. It is possible that for some incidents without known exact address, an approximate street corner address was assigned. But the fact that the overall street corner incidents takes 26.8% of the total number tells us that such cases are not too frequent. And the large variation of this feature among different crime categories can still be helpful for the classification purposes.

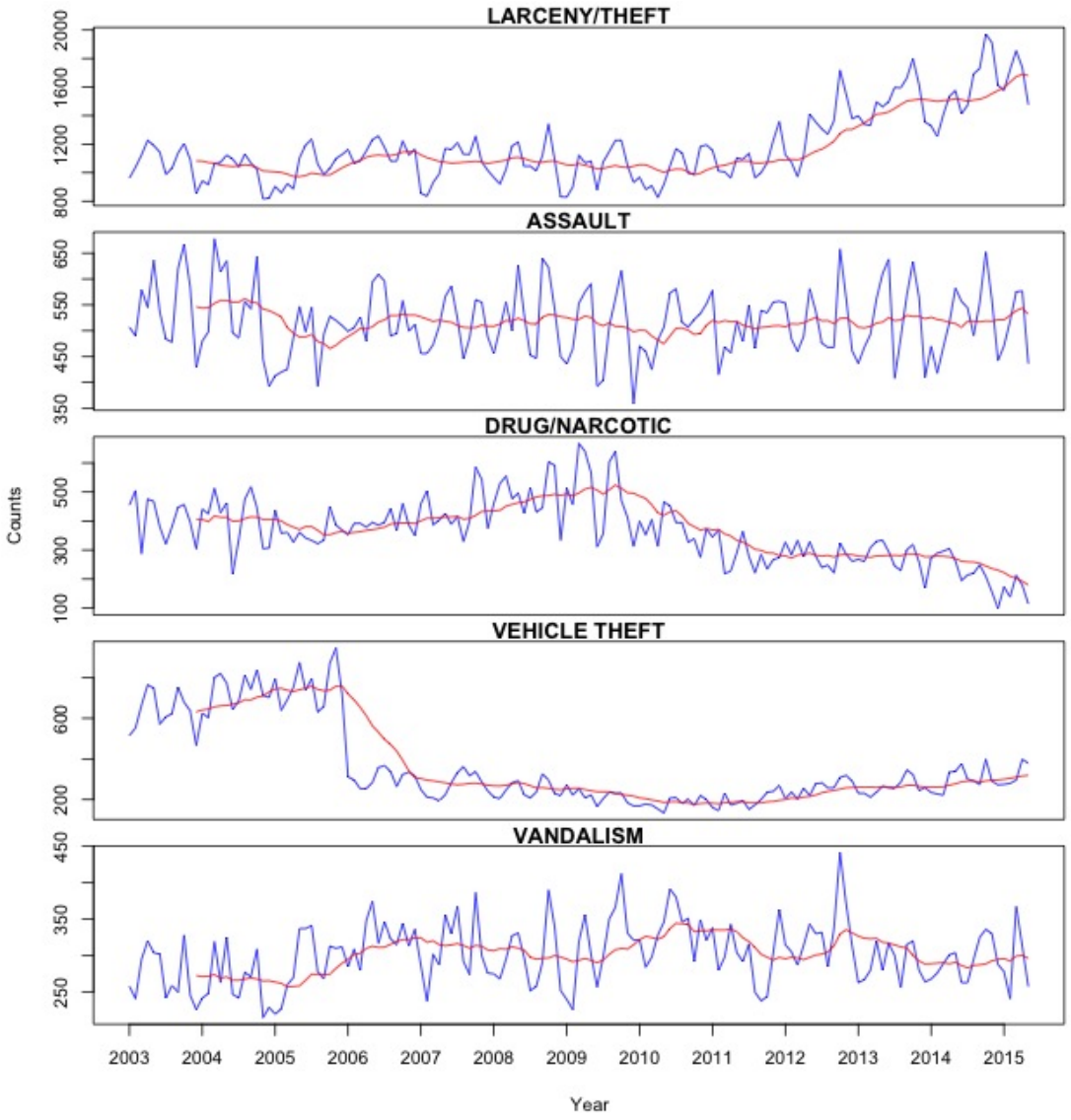


Figure 3.1: Total crime counts in each month (blue line) and the 12 month moving average (red line) from 2003 through 2015 for the top 5 city wide most common crimes.

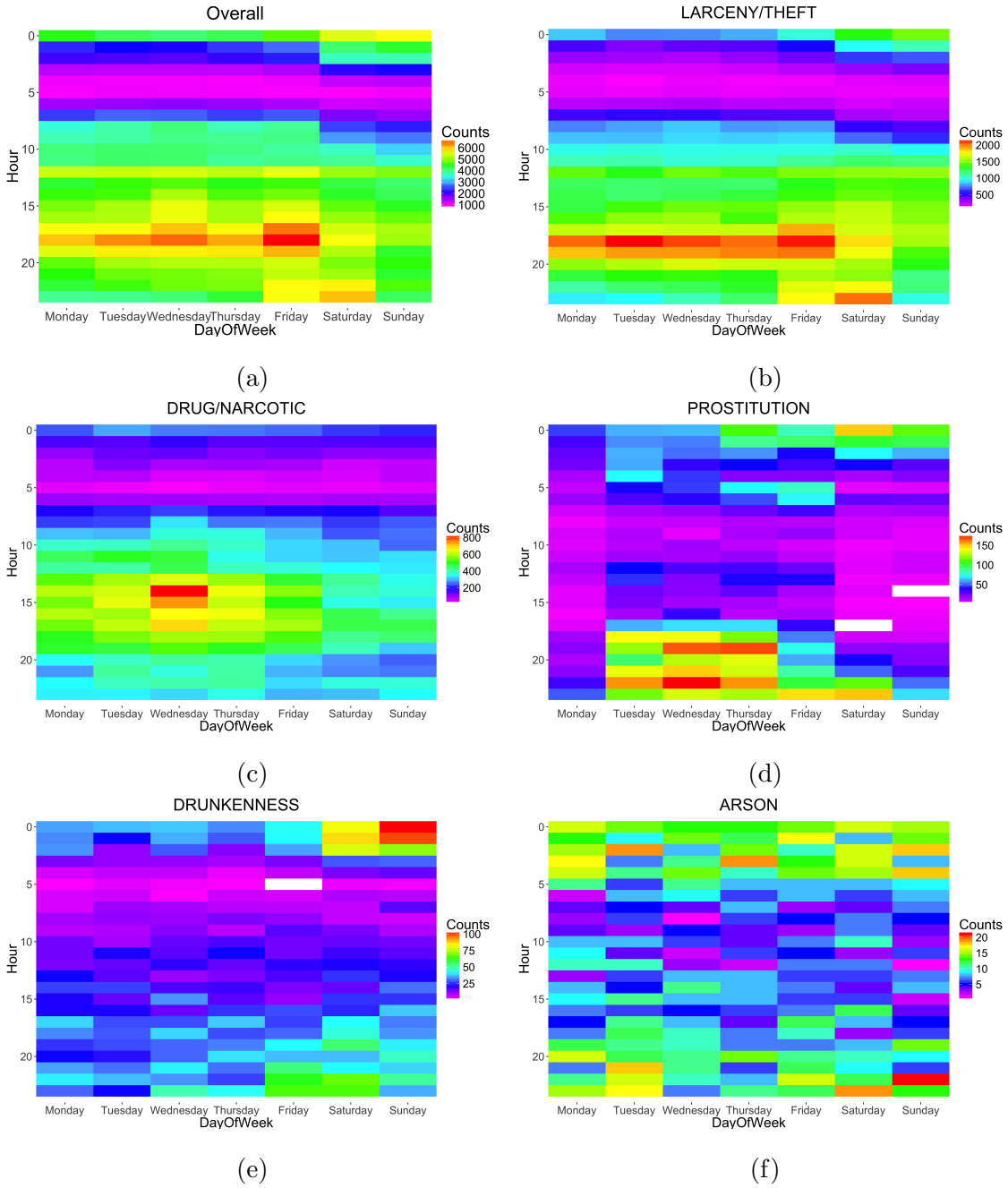
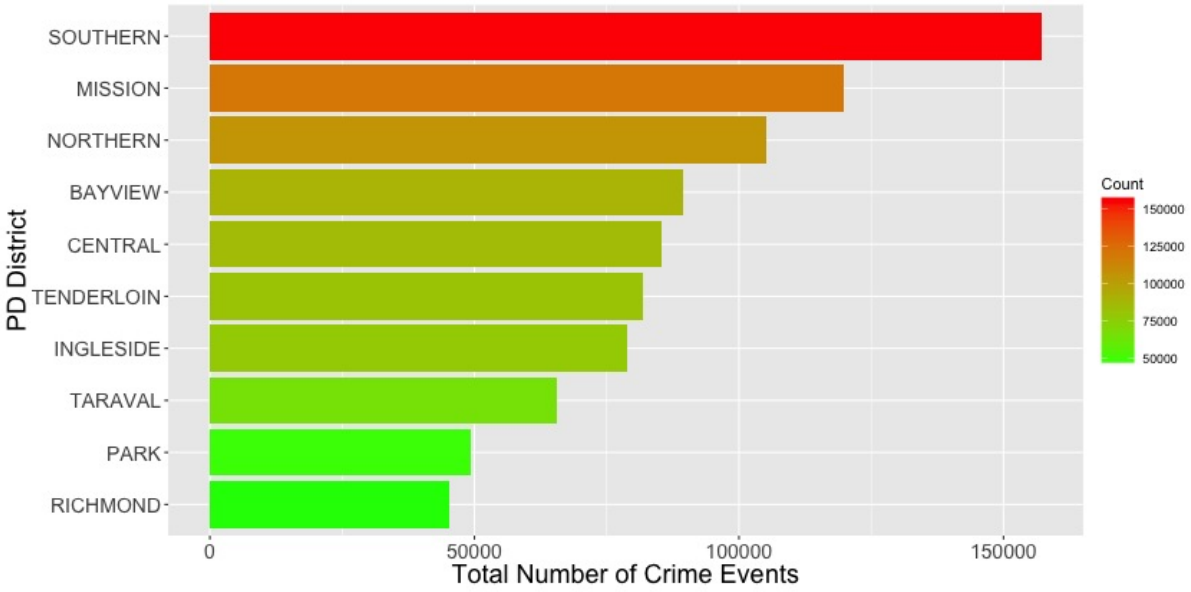
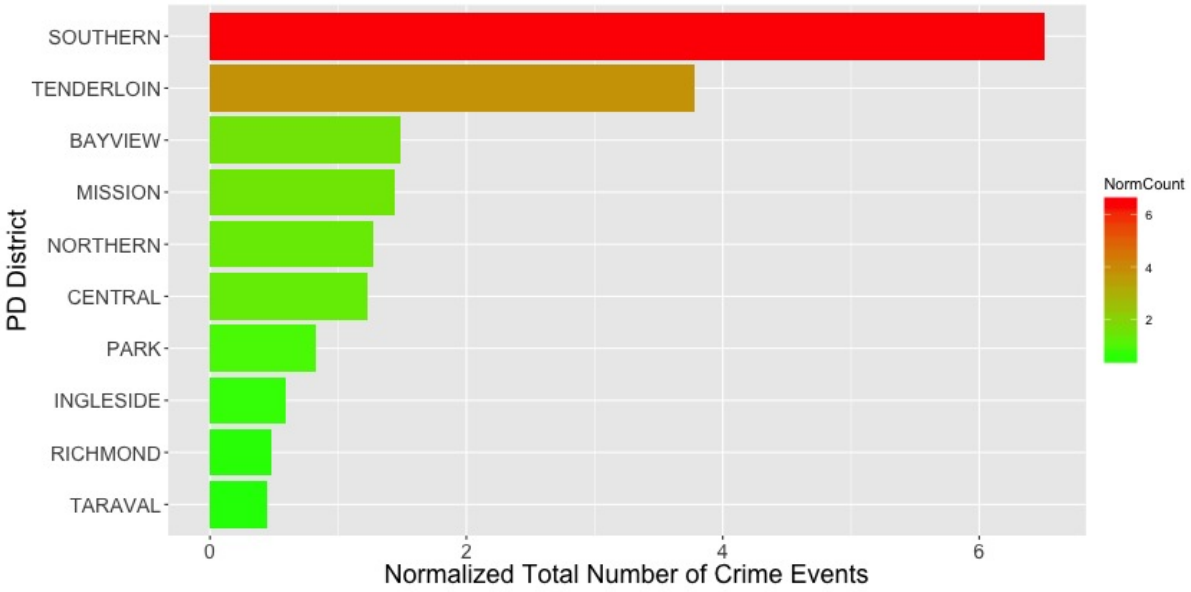


Figure 3.2: Crime Distribution over Day of Week and Hour of the Day

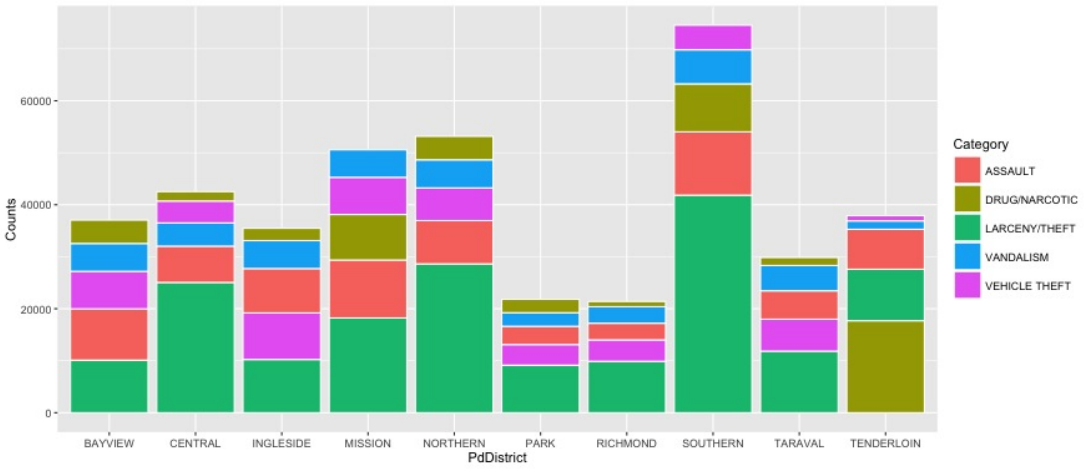


(a) Crime Incidents Counts

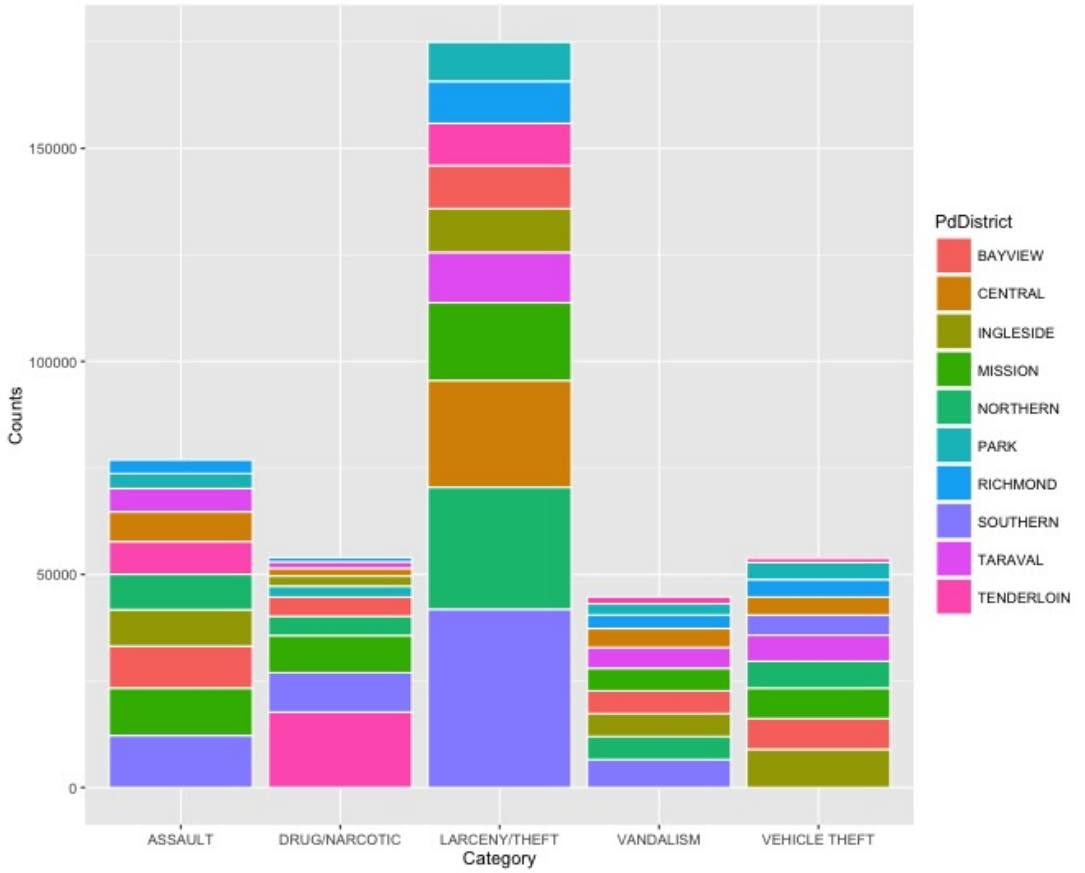


(b) Crime Incidents Counts Normalized by Population

Figure 3.3: Crime Incidents by PD District



(a)



(b)

Figure 3.4: Top 5 citywide crime categories in the 10 PD districts

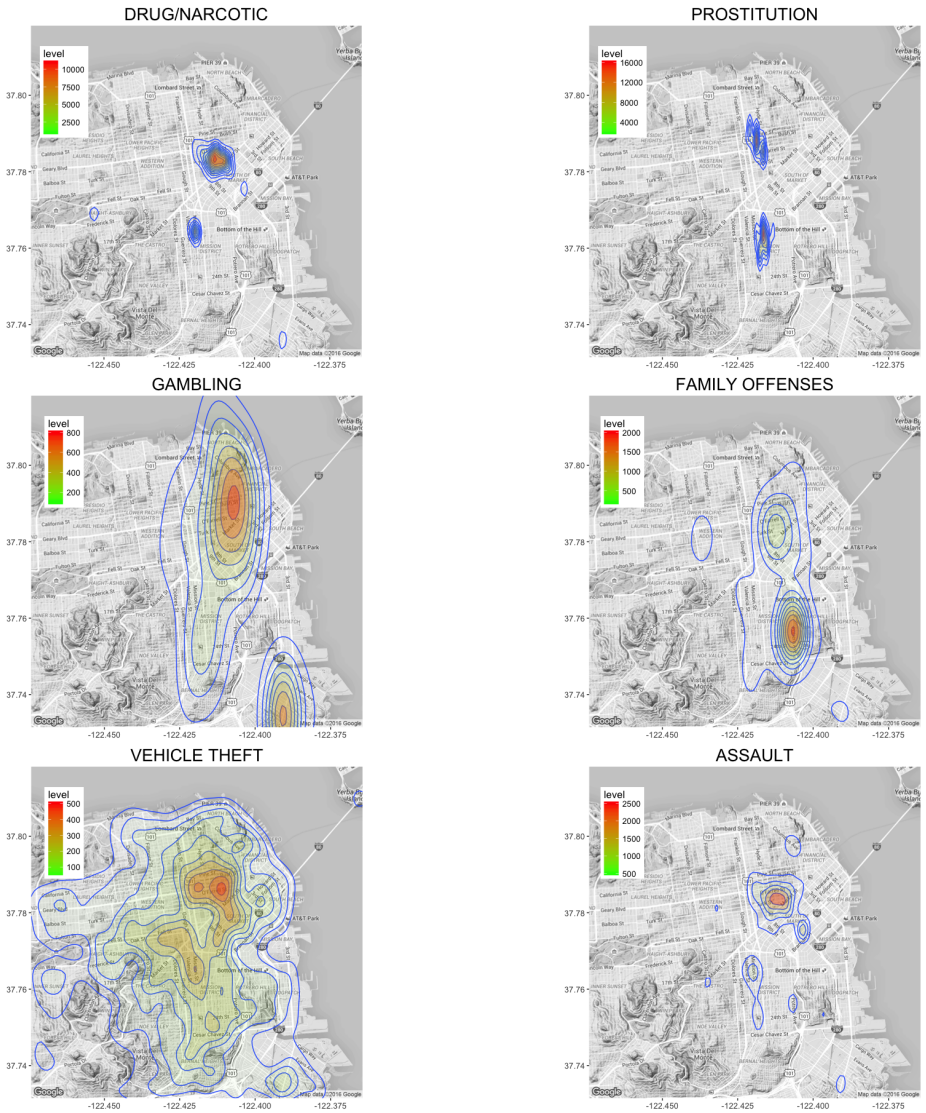


Figure 3.5: Mapping the Distribution for Each Crime Categories

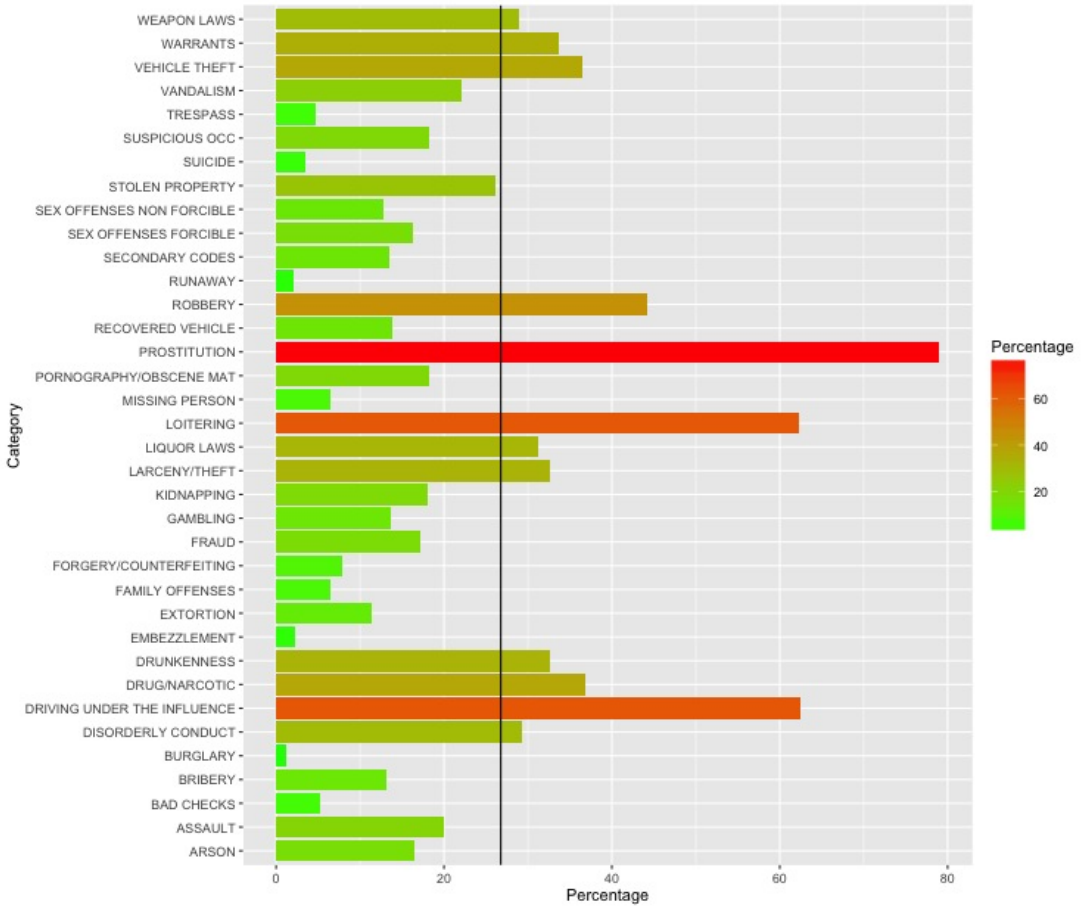


Figure 3.6: Percentage of Crime Counts that Happens on a Street Corner for All 36 Crime Categories. Overall Street Corner Crime Percentage Marked by the Black Line

CHAPTER 4

Predictive Analysis

In this chapter, the k-nearest-neighbor model and the logistic regression model will be applied to the existing dataset for the prediction of the crime categories. Classification results will be reported. We will see how machine learning algorithms can help predict crime category and aid police to protect our communities better.

The first model that we tried is the k-nearest-neighbor using location features from the dataset. The second model that we tried is one-vs-all logistic regression model, where we fit one logistic regression model for each category and then combine them to make our final decision.

4.1 Evaluation Metrics

4.1.1 Classification Accuracy

Classification models are usually evaluated with the accuracy metric where the prediction labels are compared with the ground truth and the percentage of correct predictions can be reported as classification accuracy. In this thesis, we are studying a classification problem with 36 crime categories. So the accuracy benchmark for uniform blind guess is 2.78% (1 divided by 36).

4.1.2 Multi-class Log-loss

In addition, we would like to introduce multi-class log-loss as a second evaluation metric. Equation 4.1 shows how to compute the multi-class log-loss metric where $p_{i,j}$ is the prediction probability for the i^{th} sample in the j^{th} class; $y_{i,j}$ is 1 if the ground truth label for the i^{th} sample is class j and 0 otherwise; N is the total number of data samples; M is the total number of classes.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j} log(p_{i,j}) \tag{4.1}$$

If we could achieve perfect matching between the prediction and the ground truth label, the multi-class log-loss would be equal to zero. For any other imperfect classification results, the multi-class log-loss would be a positive number. We are working on a 36-class problem, and the multi-class log-loss benchmark can be computed with Equation 4.2 if uniform random labels are applied for the samples. For uniform random labels, correct predictions are achieved with a probability of $\frac{1}{36}$, and log loss of zeros are applied for these correctly classified samples. For incorrect predictions with a probability of $\frac{35}{36}$, log loss should be equal to $-\log 0$, which is positive infinity. In order to make this number numerically computable, we use 10^{-15} for the log.

$$logloss_{benchmark} = -(\frac{1}{36}\log 1 + \frac{35}{36}\log 10^{-15}) = 33.58 \tag{4.2}$$

We are also interested in this metric because this dataset is hand labeled and very noisy. In the previous chapters, we have found lots of data labeling errors and there could be many more uncovered. Thus, this log-loss metric, measuring the cross-entropy between the prediction probabilities and the target labels, is more fair.

4.2 K-nearest-neighbour Model

We first try using k-nearest neighbor model with the longitude-latitude feature for crime category classification problem. Since the longitude-latitude feature represents the geological location of an incident, no feature normalization is needed and we can directly use the Euclidean distance. We partition the dataset into 80% training and 20% testing. Varying the number of neighbors for the k-nearest-neighbor model, we can evaluate both the classification accuracy and the log-loss (Figure 4.1).

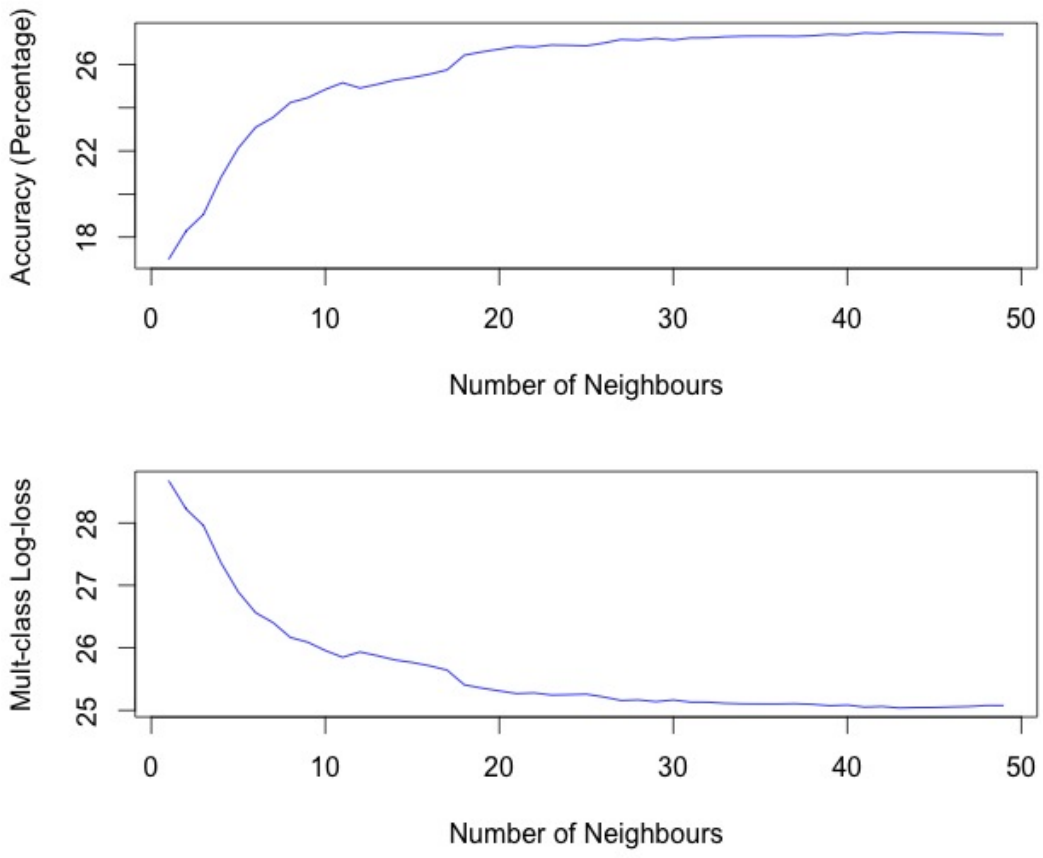


Figure 4.1: k-nearest-neighbor model performance with respect to the number of neighbors k

For general knn models, the classification accuracy increases as the number

Number of Neighbors	10	20	30	40	50	100	150	200
Accuracy (%)	24.09	25.61	26.06	26.25	26.31	26.29	26.06	25.88
Log-loss	26.22	25.69	25.54	25.47	25.45	25.46	25.54	25.60

Table 4.1: k-nearest neighbor model 5-fold cross validation performance with respect to the number of neighbors

of neighbors k increase by reducing the effect of noise. This behavior is clearly shown in Figure 4.1. But as k approaches the total number of training samples, the accuracy should drop significantly. Finding the elbow point will yield a best parameter k for the model. Unfortunately, the dataset that we are studying in this thesis has more than 800,000 samples, and it is not practical to examine the performance with respect to k up to such a large number.

We selectively varied the k and examined the model performance by 5-fold cross validation. The results are shown in Table 4.1. We can see that the accuracy is increasing as k increases. Moreover, a small amount of performance degradation can be seen as k exceeds 50.

Observing the fact that the performance difference for k equals 30 versus k equals 50 are minimal but the computation time increases significantly with larger k, we will choose k equals 30 for this problem.For k equals 30, an accuracy of 27.14% and a multi-class log-loss of 25.17 can be achieved.

4.3 Logistic Regression Model

In order to classify 36 crime categories, we split the data set into 80% training data and 20% testing data. We then fit one logistic regression model for each one crime categories returning the possibilities for each sample belonging to the current crime category.

The prediction variables that we are using in our model are listed in Table 4.2.

Feature name	Description
X:Y	Longitude latitude
PdDistrict	The PD district
DayOfWeek	The day of the week
Hour	The hour of the day
Year	The year
Month	The month of the year
StreetCorner	Happened at a street corner or not

Table 4.2: List of Features for the Logistic Regression Model

And a total accuracy of 28.51% and multi-class log-loss of 2.45 can be achieved with this model.

CHAPTER 5

Conclusion

In this thesis, we discussed the San Francisco PD crime data from 01/01/2003 to 05/13/2015. We started with data preprocessing and data cleaning in Chapter 2. We then studied the informative analysis on both the timing and the location of the incidents in Chapter 3. From these studies, we were able to extract frequency patterns for different crime categories. Understanding when and where certain types of crimes are more likely to occur can increase the efficiency of police officers and suggest policy changes. In addition, in Chapter 4, we discussed the possibility of classifying the crime categories given the incidents time and location information. For a large dataset with noisy labelings, we were able to achieve an accuracy of 28.51% for a 36-class classification problem using only seven features. Seeing the fact that the accuracy benchmark for uniform blind guess is 2.78% (1 divided by 36), this is a reasonably good performance.

With this thesis, we achieved more understandings of the local security environment of the San Francisco city from informative visualizations and predictive classifications. And these understandings can hopeful aid SFPD to better serve the city. More importantly, similar methods can be applied to future data or crime data from other cities. We hope that a larger population can benefit from a better security environment with the help of data analysis.

For future work, more interesting questions can be further studied. One example of such questions is that given the type of a crime, when and where are they most likely to occur.

APPENDIX A

Monthly Counts

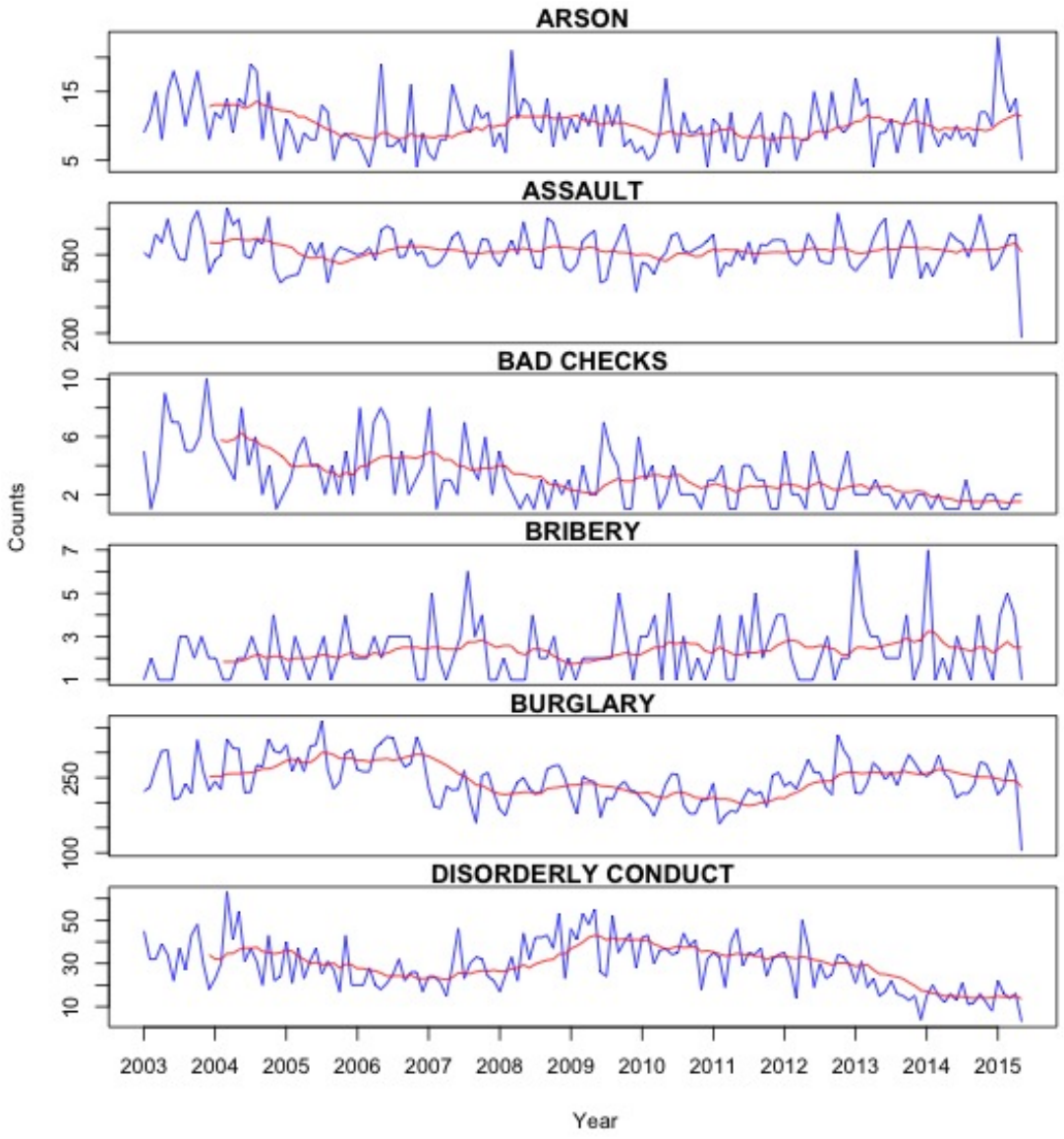


Figure A.1: Crime Counts Trend Over the 12 years (PART I)

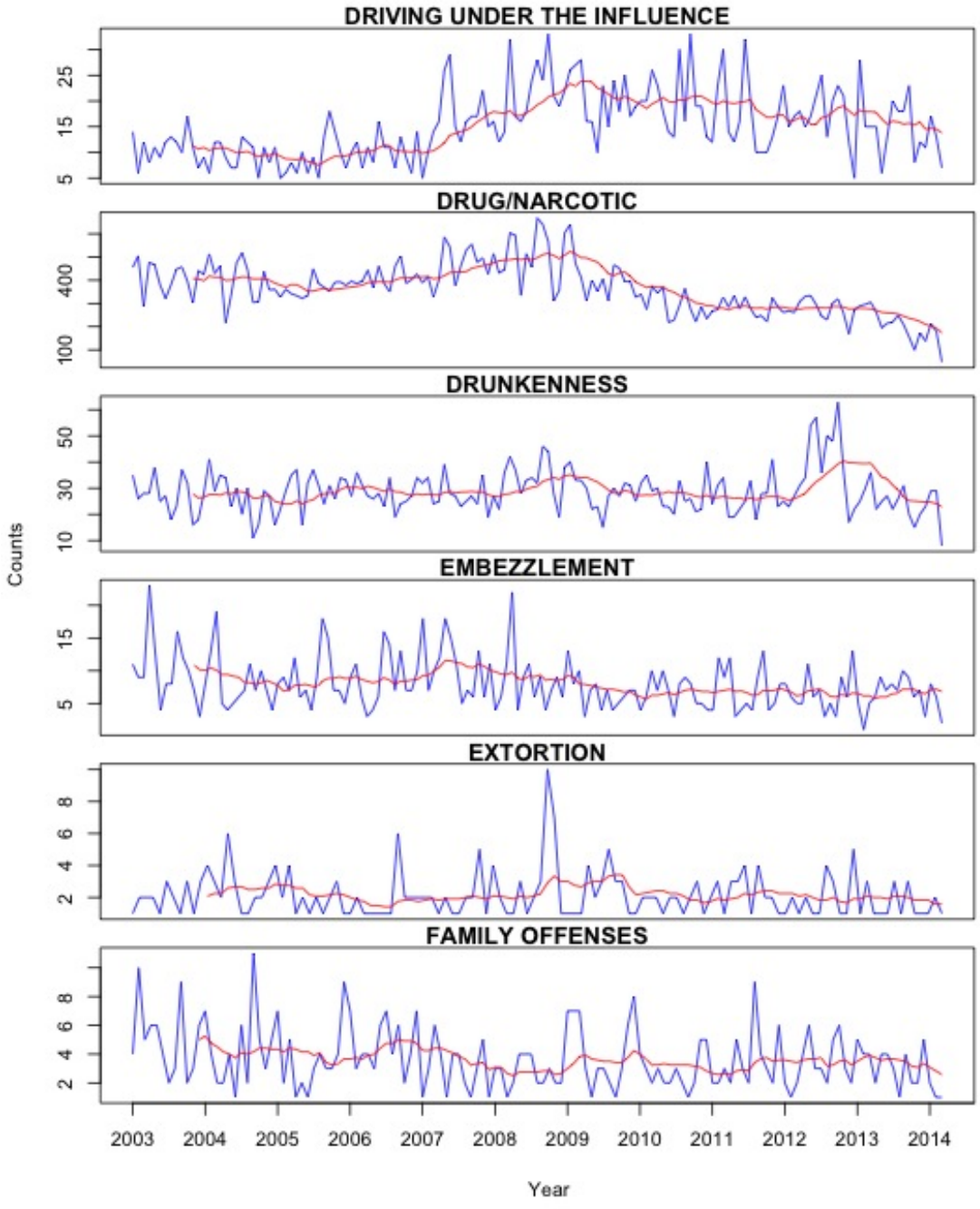


Figure A.2: Crime Counts Trend Over the 12 years (PART II)

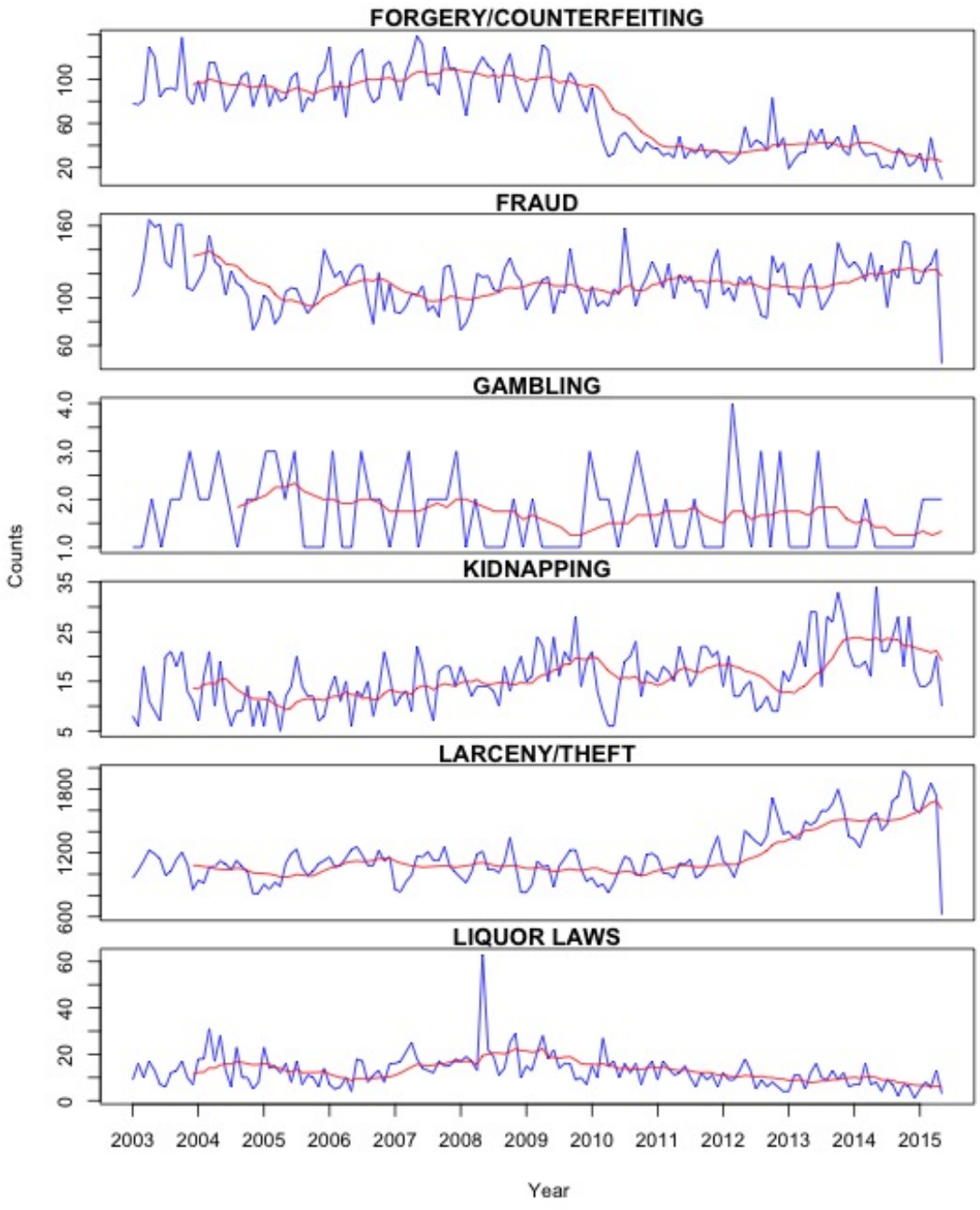


Figure A.3: Crime Counts Trend Over the 12 years (PART III)

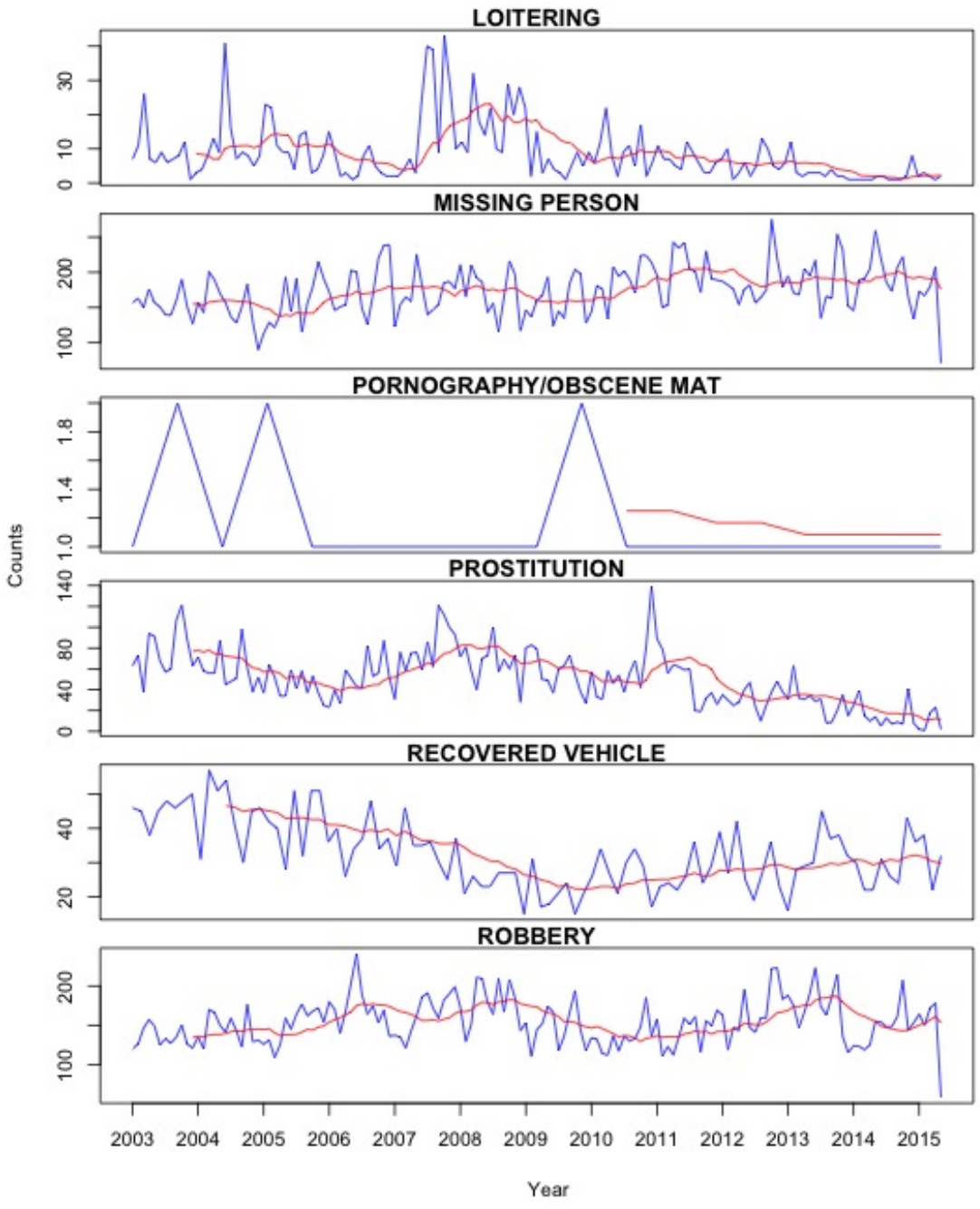


Figure A.4: Crime Counts Trend Over the 12 years (PART IV)

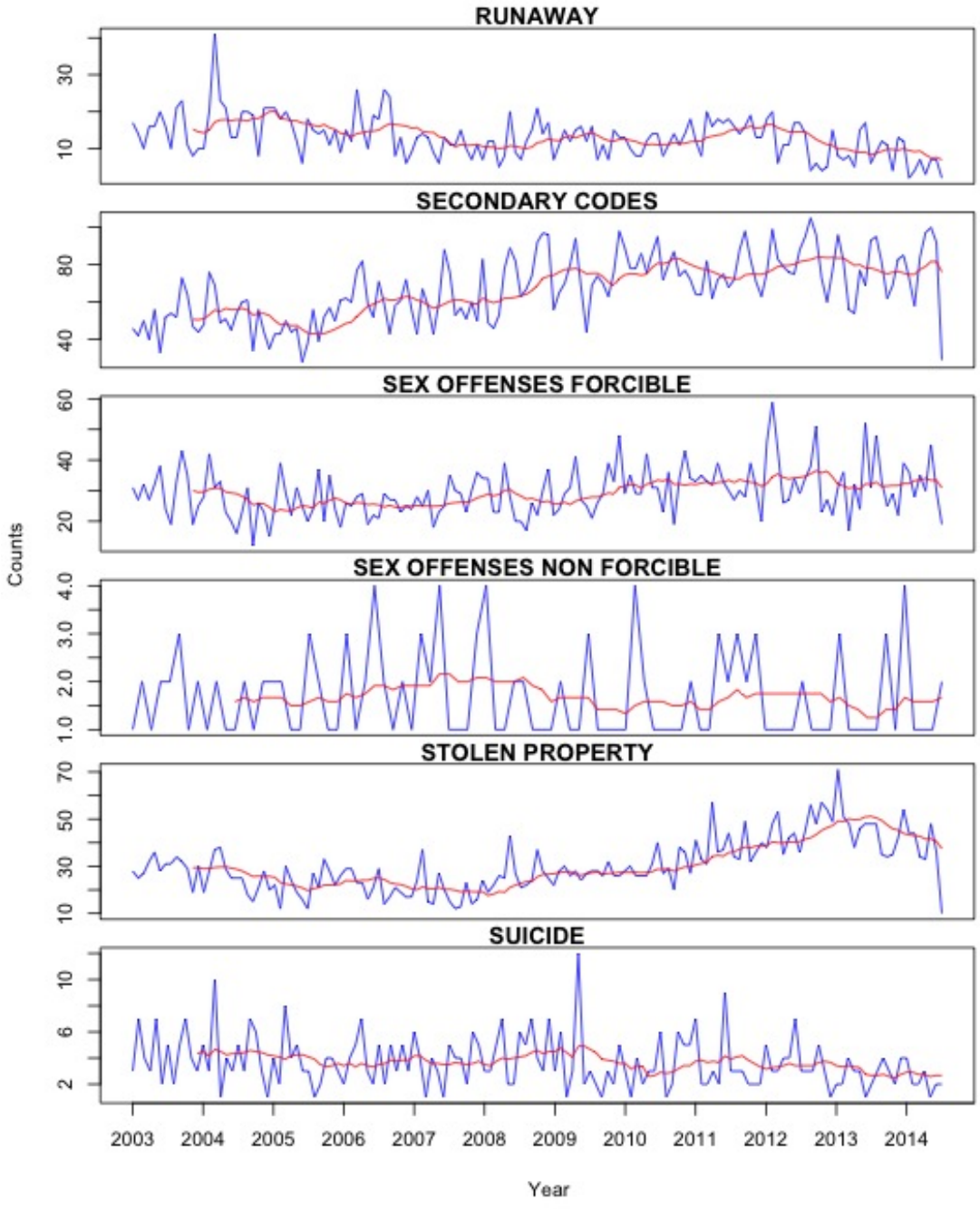


Figure A.5: Crime Counts Trend Over the 12 years (PART V)

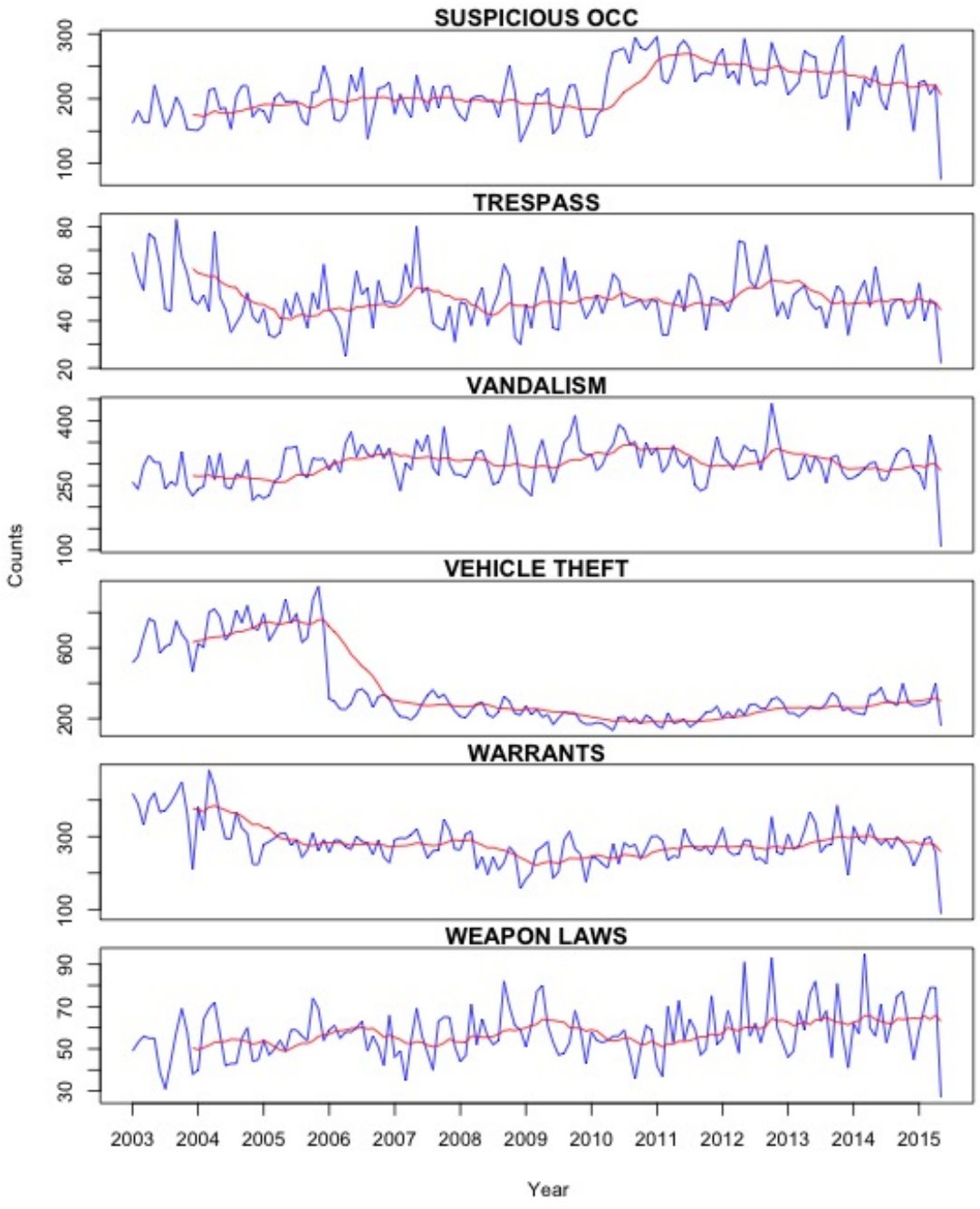


Figure A.6: Crime Counts Trend Over the 12 years (PART VI)

APPENDIX B

Weekdays and Hours Counts Heat Maps

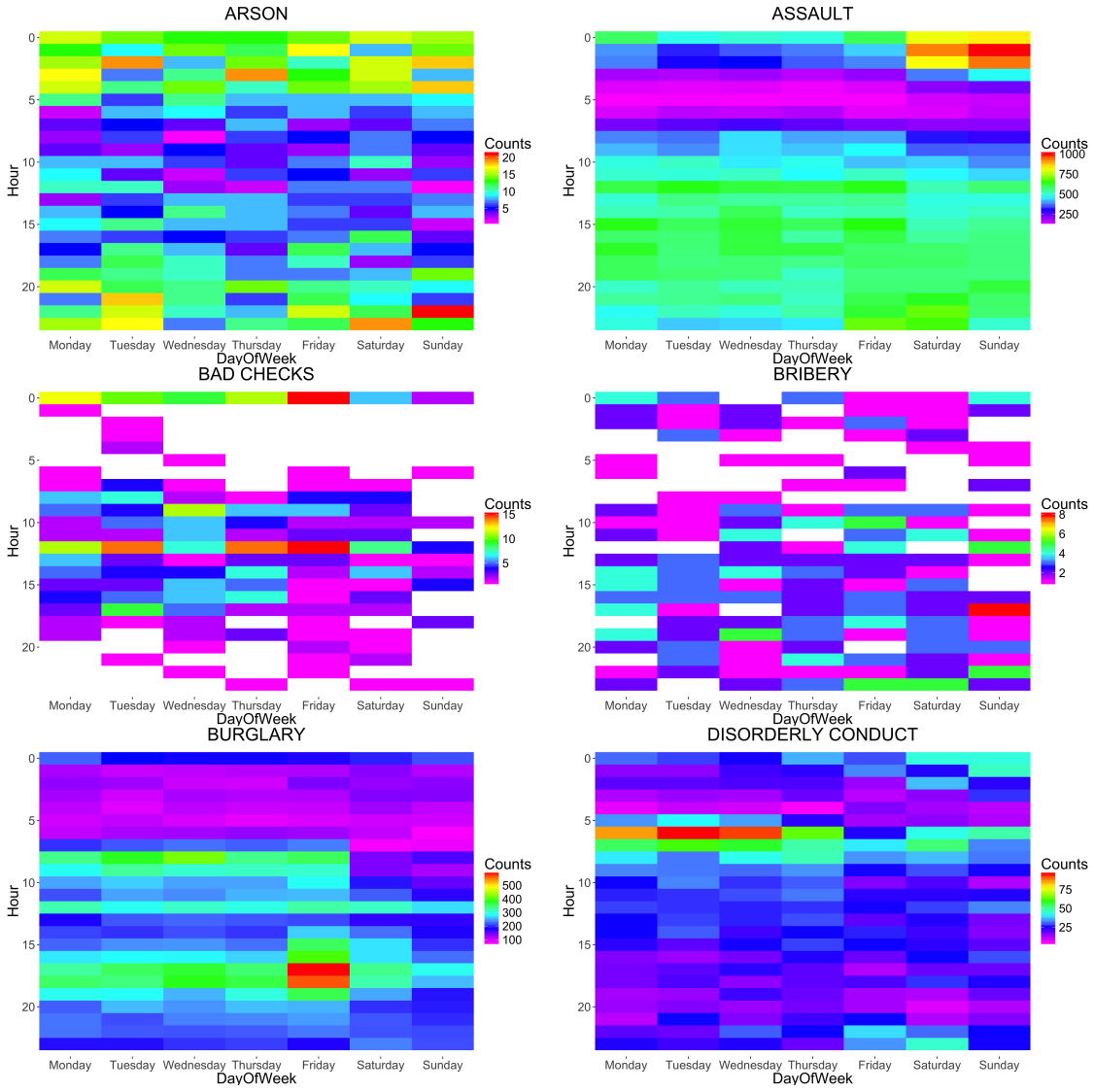


Figure B.1: Crime Distribution over Day of Week and Hour of Day (PART I)

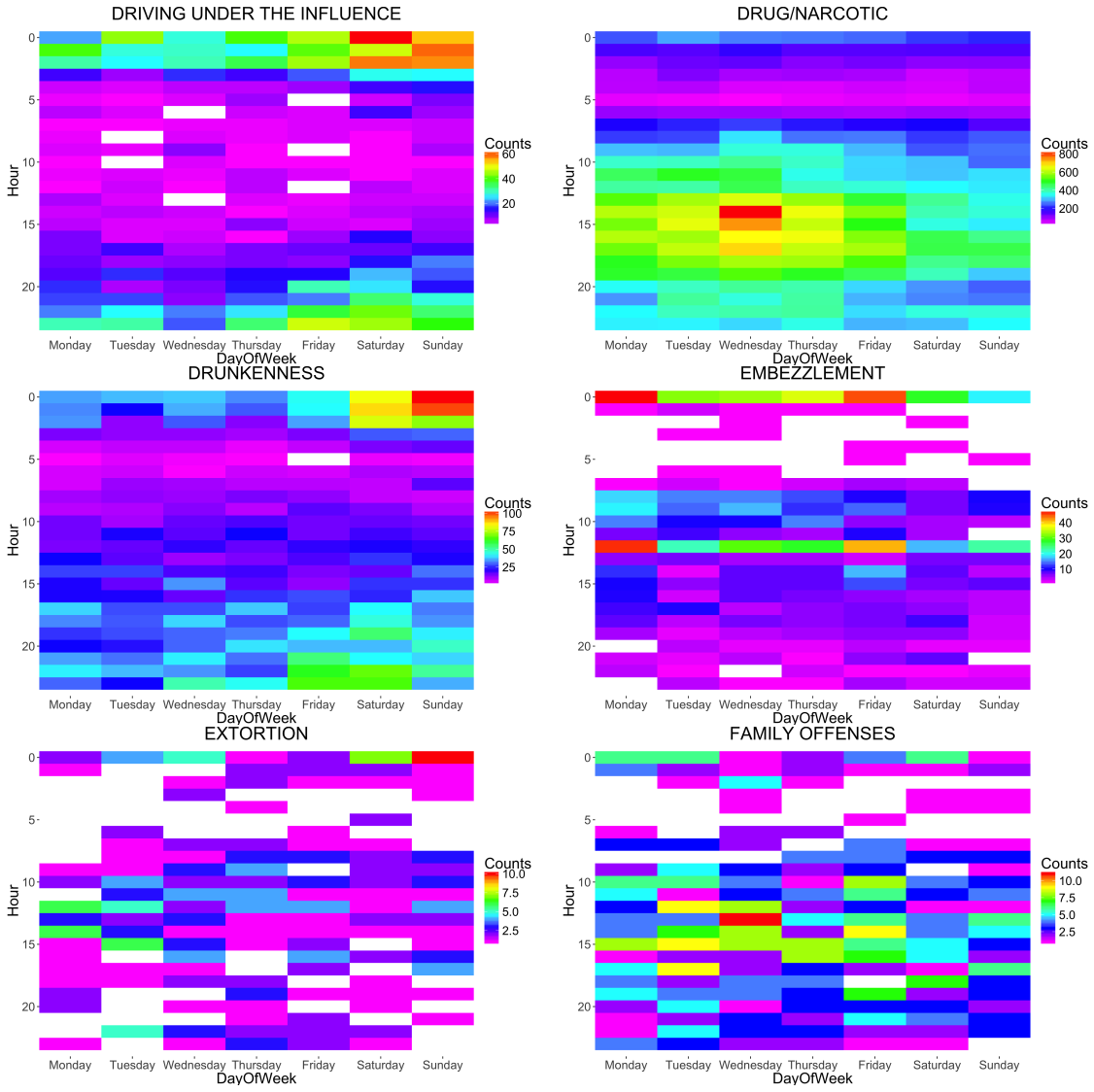


Figure B.2: Crime Distribution over Day of Week and Hour of Day (PART II)

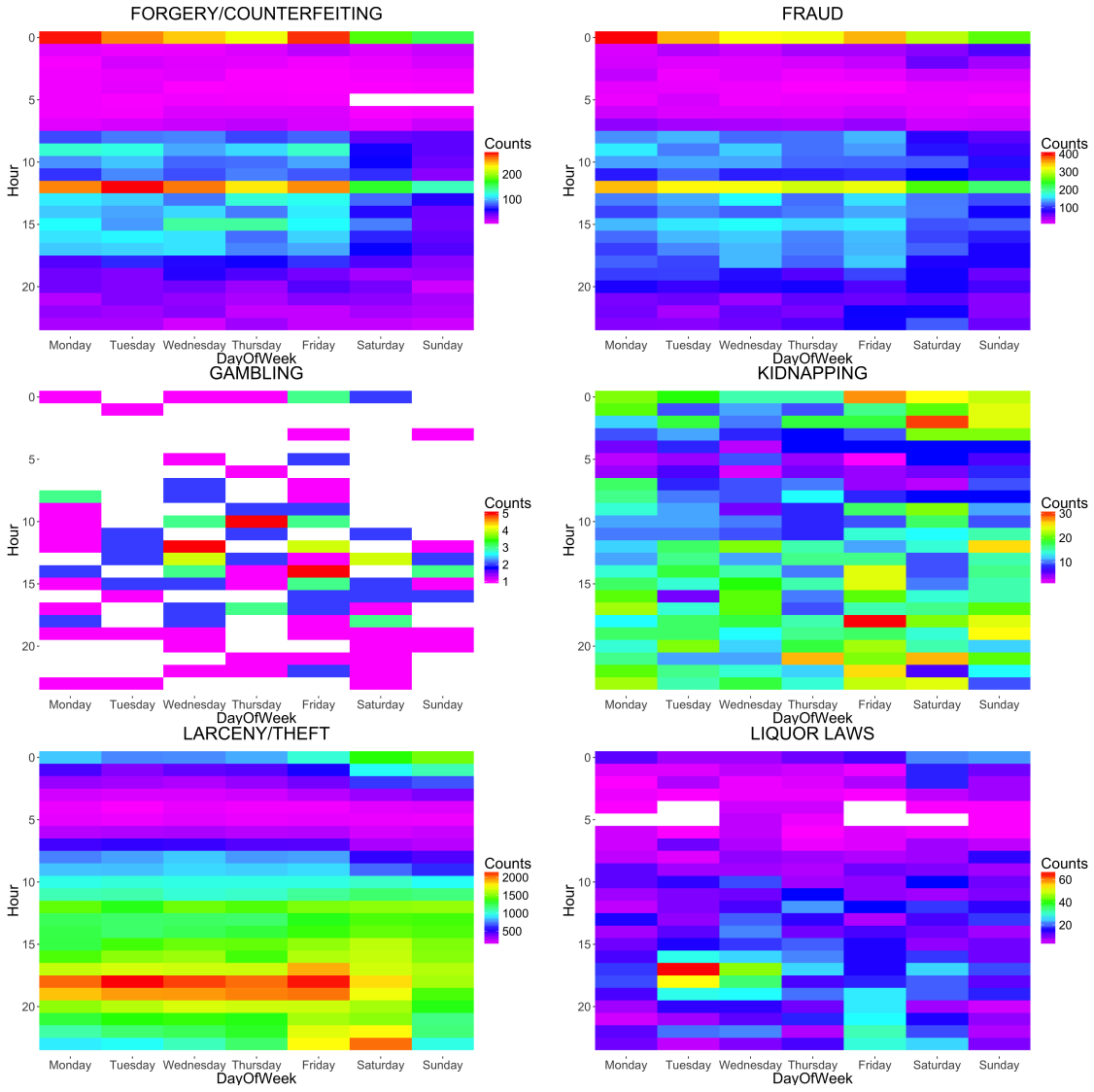


Figure B.3: Crime Distribution over Day of Week and Hour of Day (PART III)

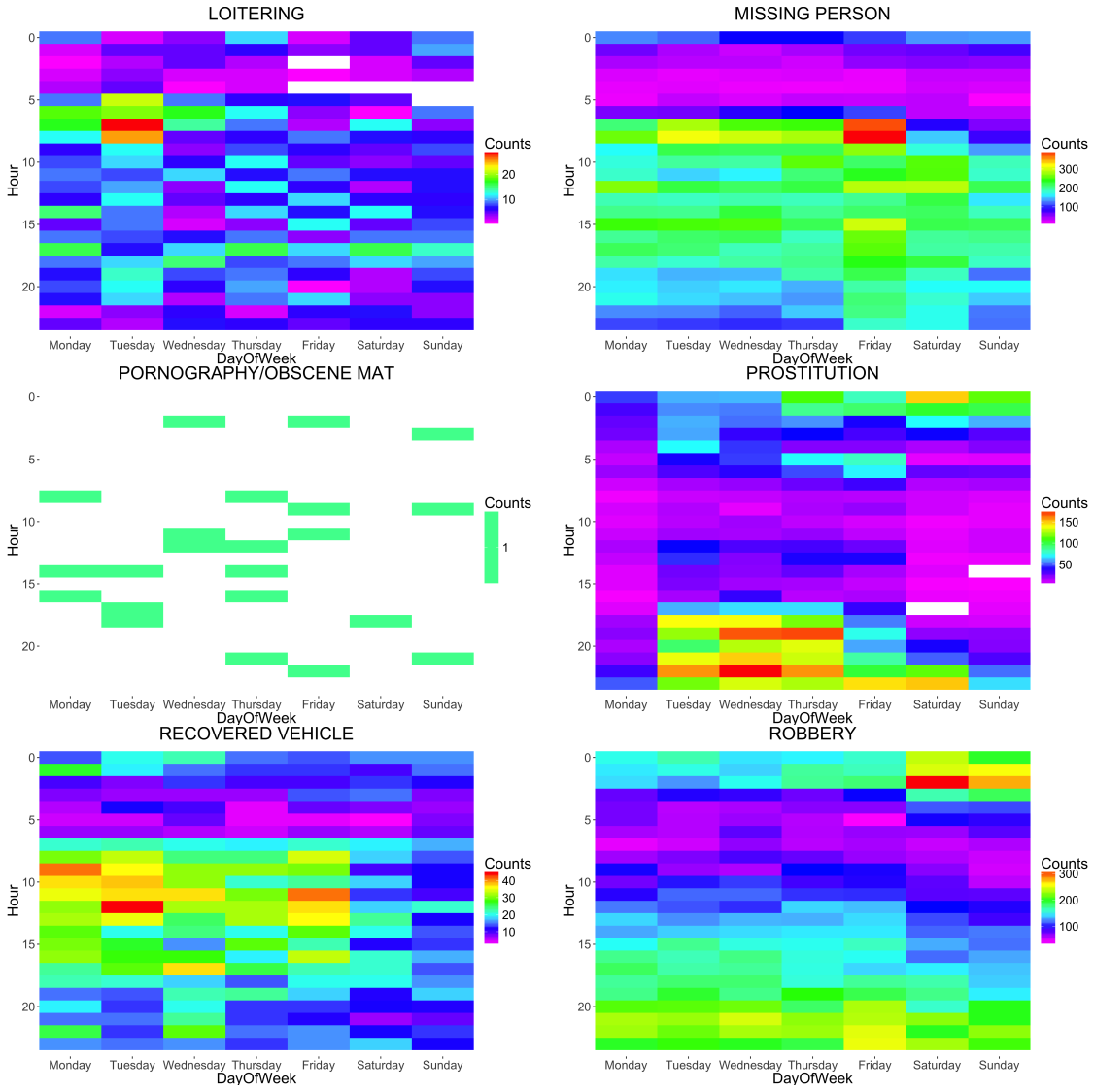


Figure B.4: Crime Distribution over Day of Week and Hour of Day (PART IV)

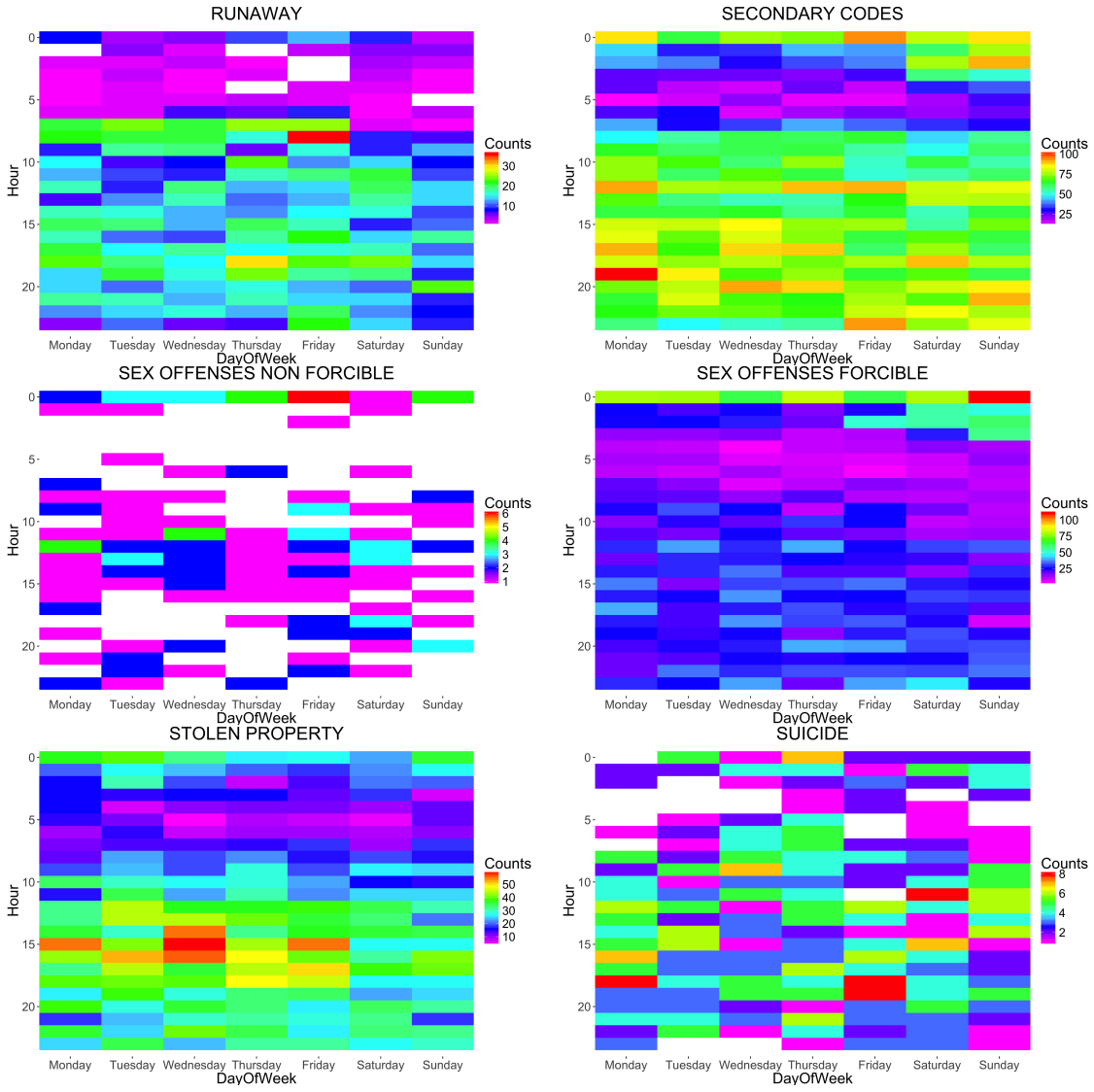


Figure B.5: Crime Distribution over Day of Week and Hour of Day (PART V)

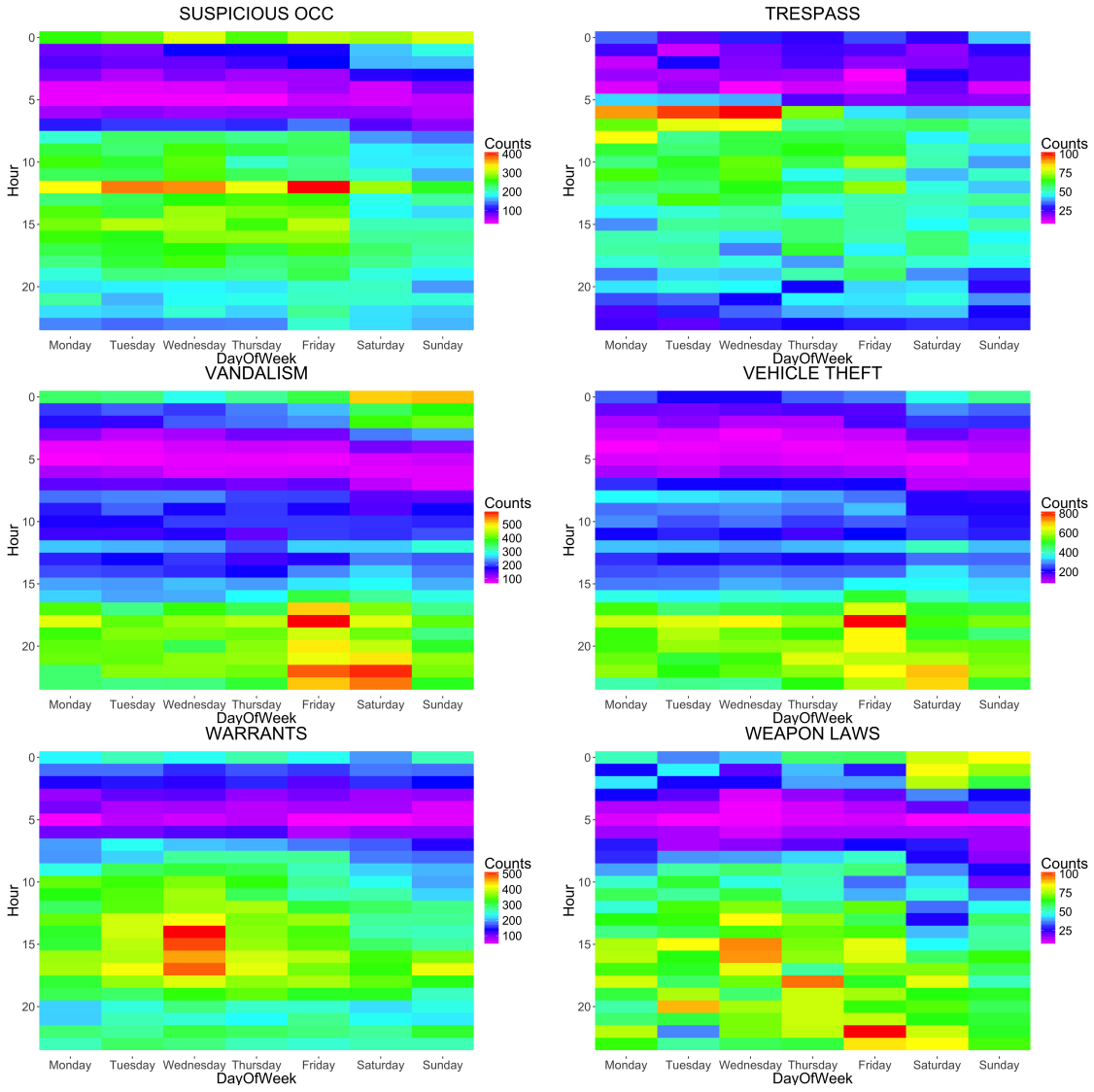


Figure B.6: Crime Distribution over Day of Week and Hour of Day (PART VI)

APPENDIX C

Crime Density Distributions Over the City Map

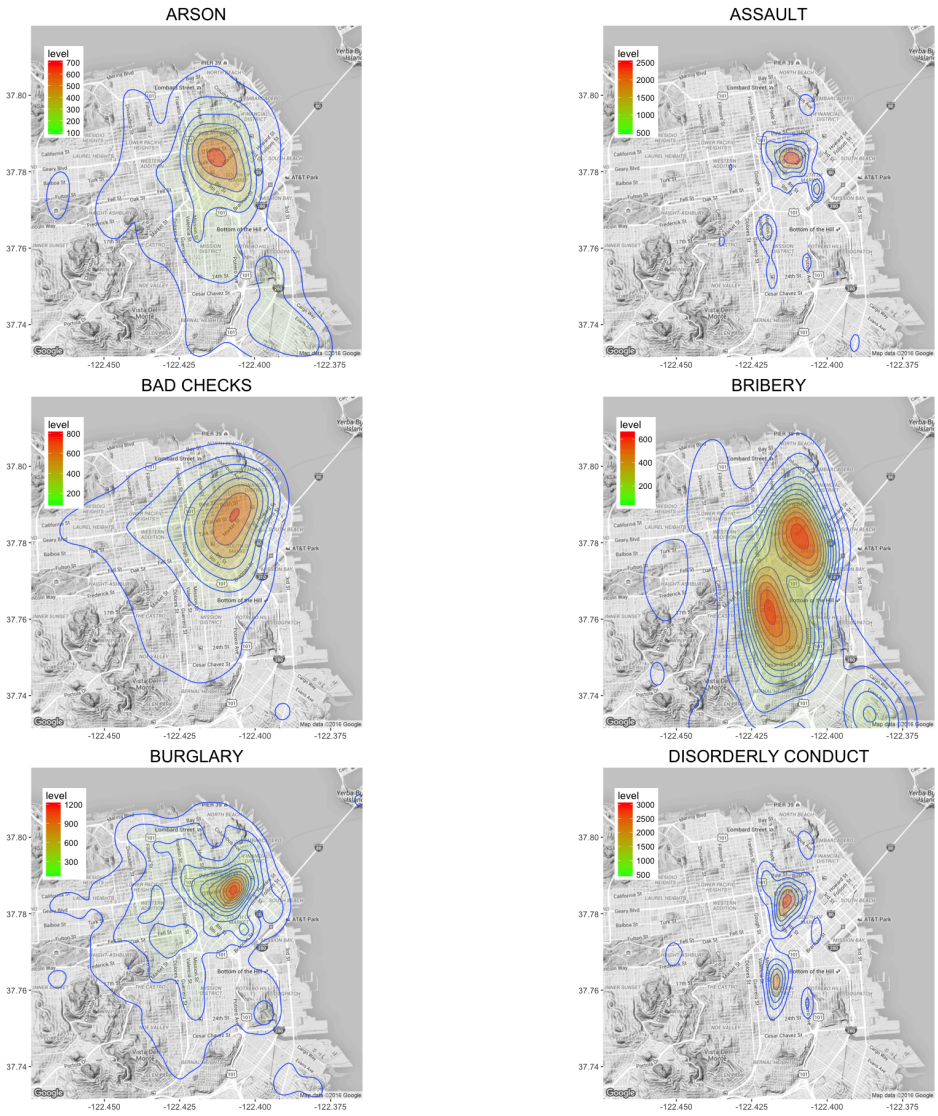


Figure C.1: Crime Distribution Density Over a Map (PART I)

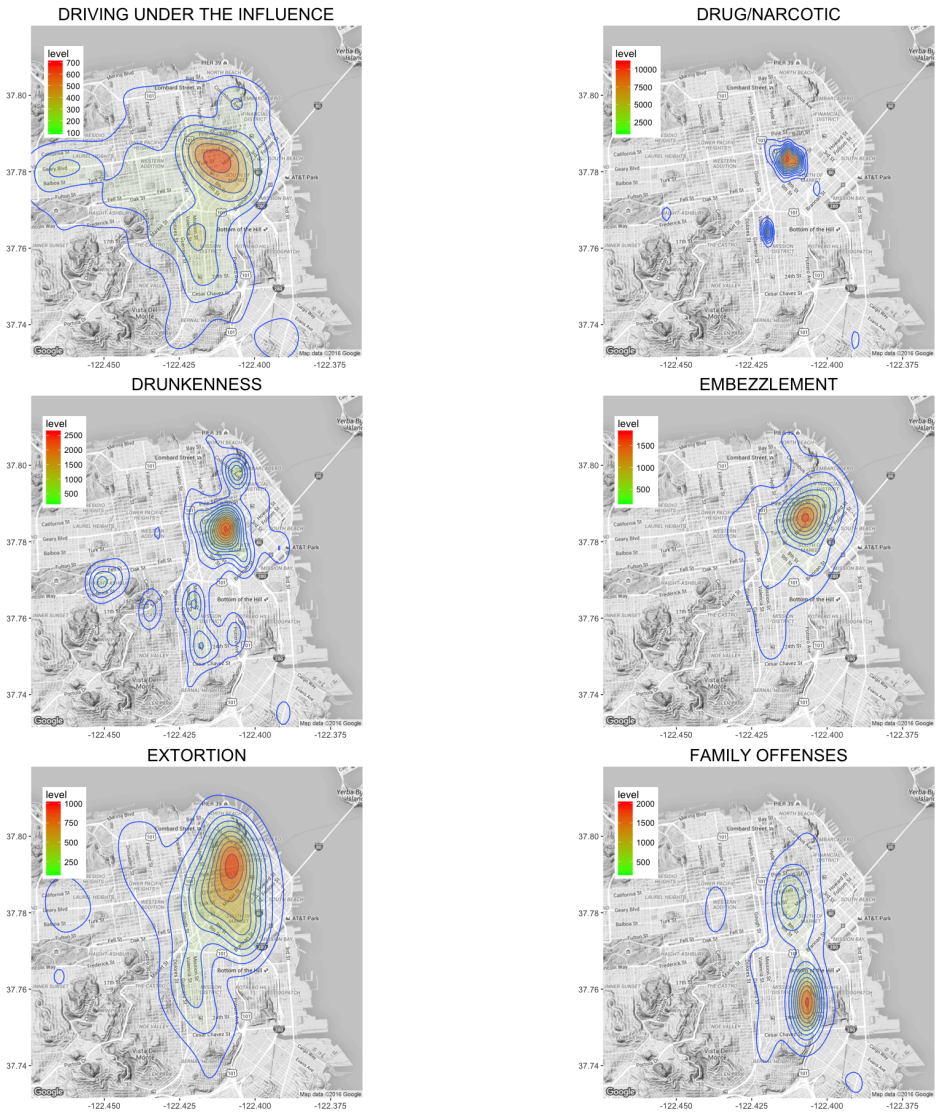


Figure C.2: Crime Distribution Density Over a Map (PART II)

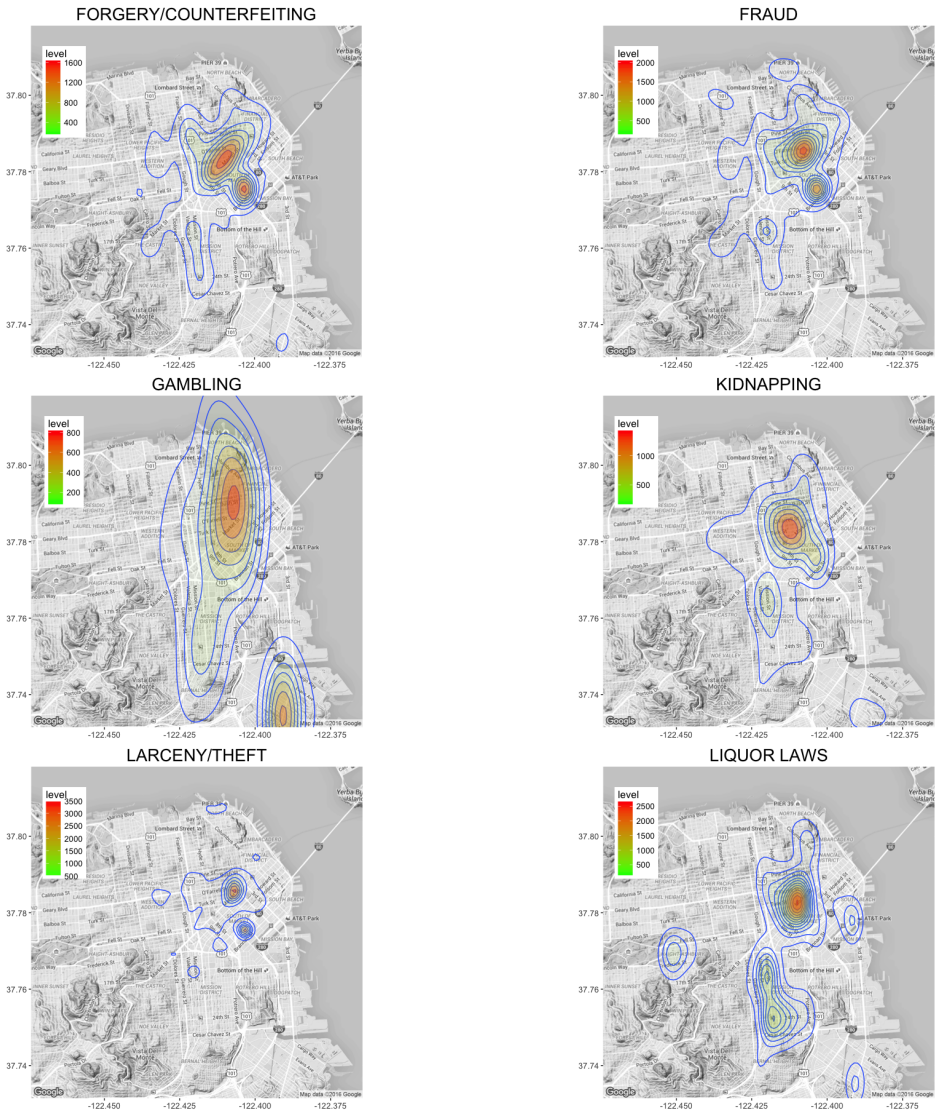


Figure C.3: Crime Distribution Density Over a Map (PART III)

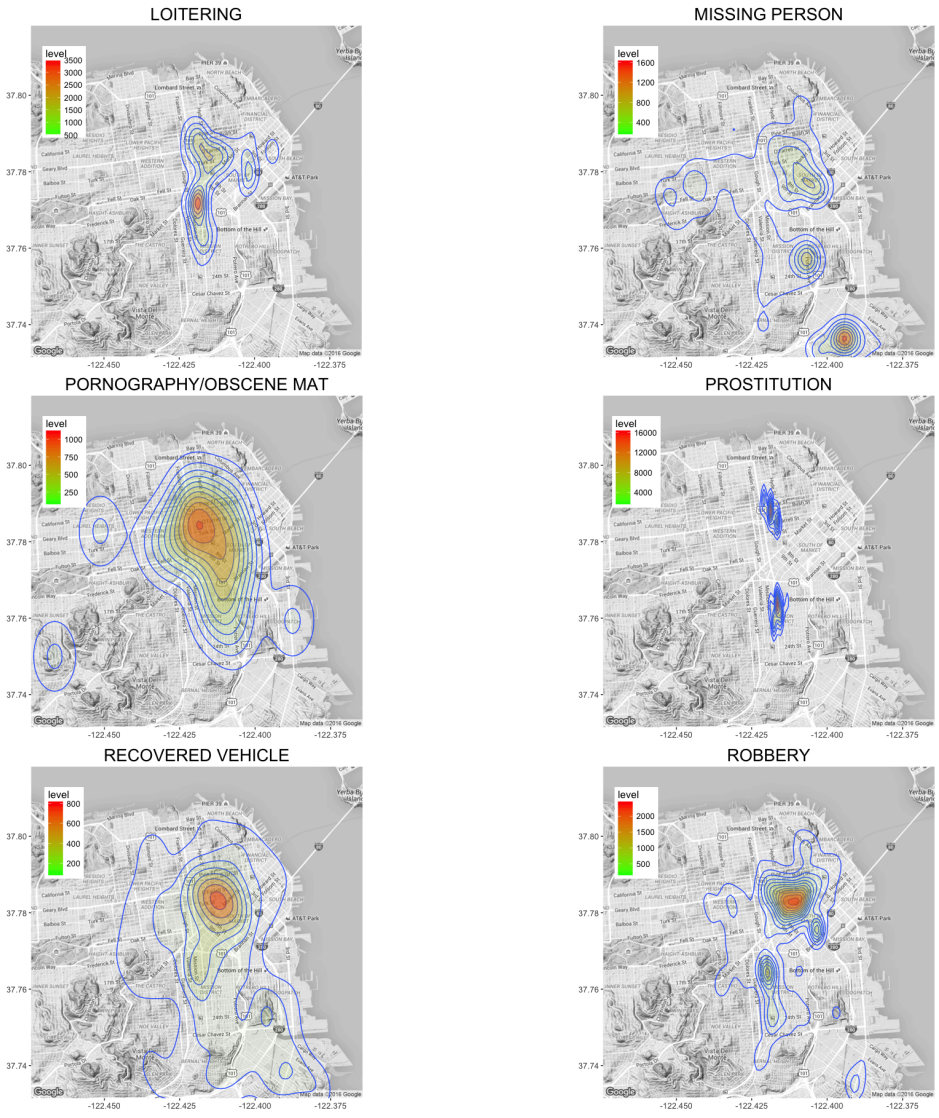


Figure C.4: Crime Distribution Density Over a Map (PART IV)

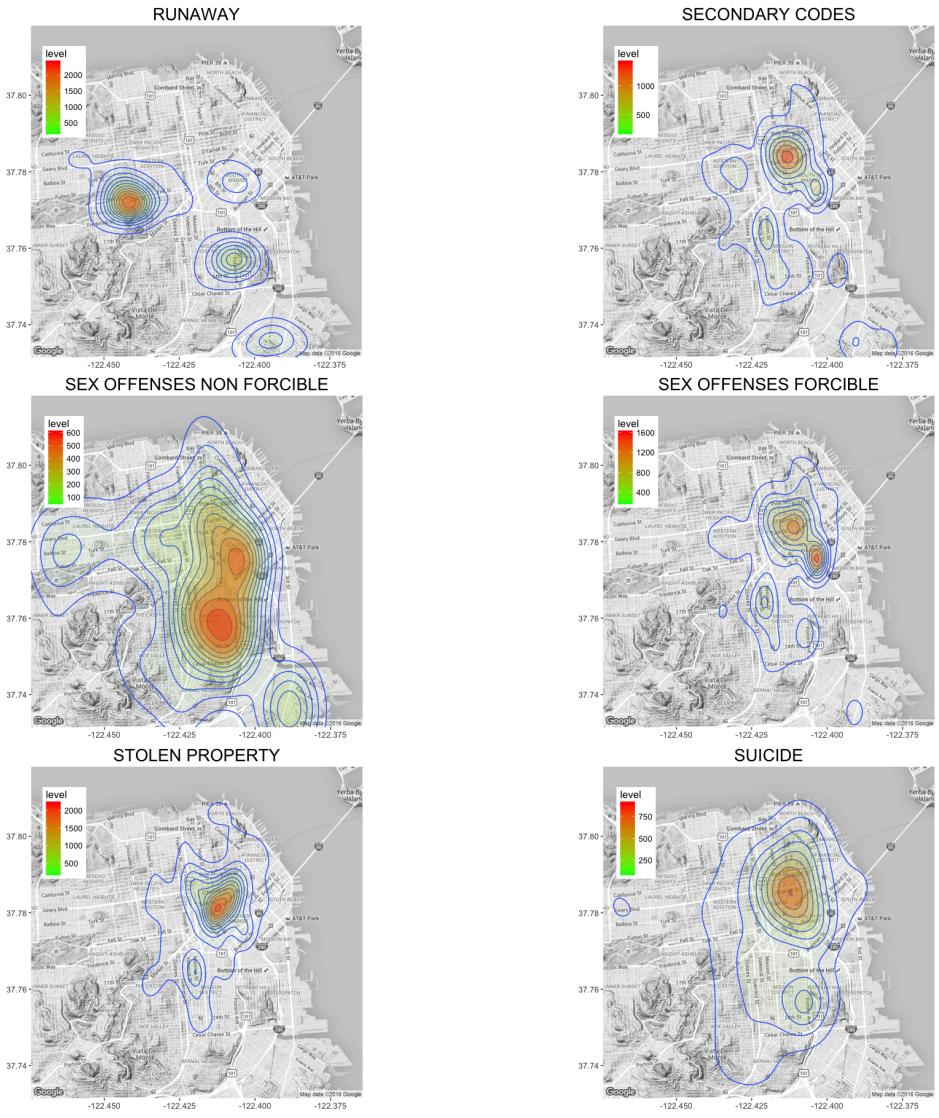


Figure C.5: Crime Distribution Density Over a Map (PART V)

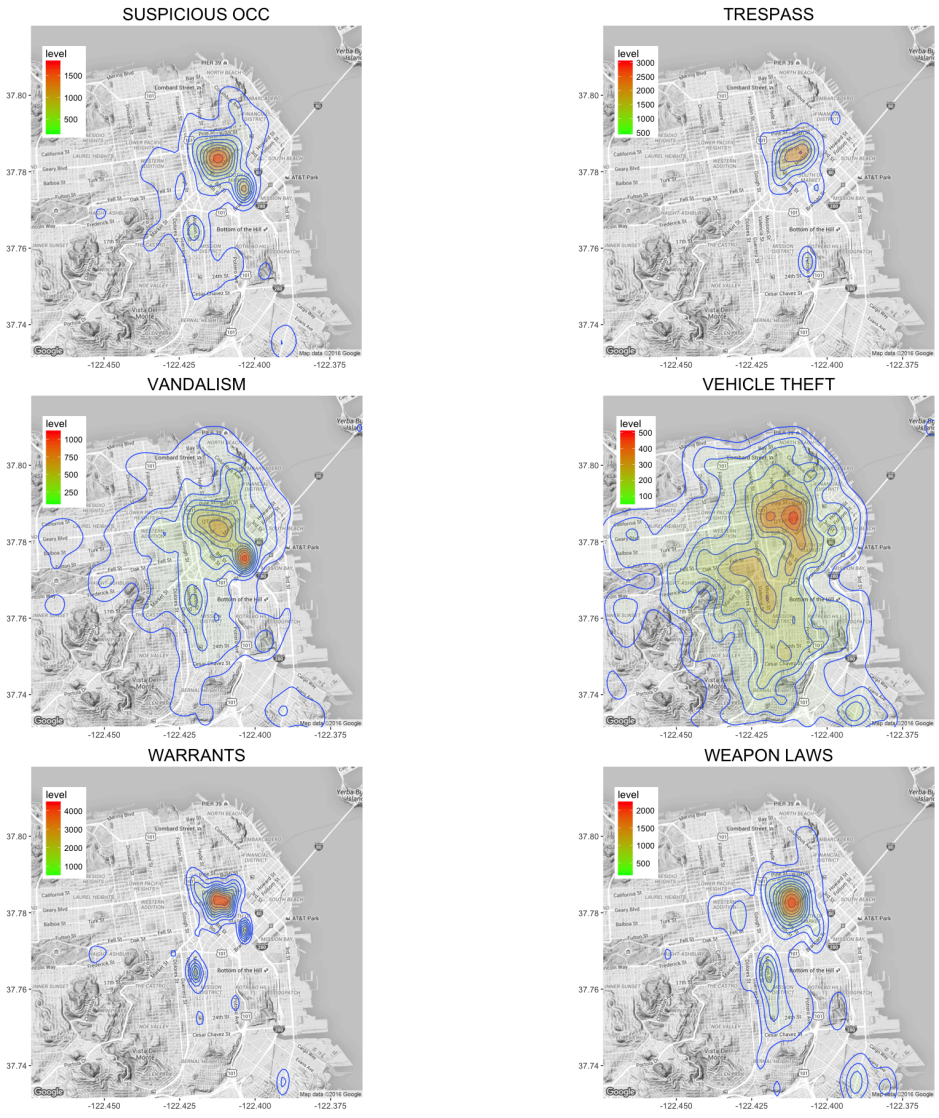


Figure C.6: Crime Distribution Density Over a Map (PART VI)

References

[CS15] City and County of San Francisco. "SF Open Data." https://data.sfgov.org, May 2015.

[Gre10] Jessica Greene. "SFPD Dropping 30 Cases Daily Due to Drug Lab Scandal." http://www.nbcbayarea.com/news/local/SFPD-Dropping-30-Cases-Daily-Amid-Due-to-Lab-Scandal-88551702.html, March 2010.

[Gro08] Public Safety Strategies Group. "San Francisco Police Department District Station Boundary Analysis." http://sanfranciscopolice.org/sites/default/files/FileCenter/Documents/14683-SFPD_DSBAfinal_trnsmtl.pdf, May 2008.

[Gro14] Public Safety Strategies Group. "San Francisco Police Department District Station Boundary Analysis." http://www.sf-police.org/Modules/ShowDocument.aspx?documentID=27425, Dec 2014.

[Jen15] Isaac Jenkins. "Vehicle Thefts or Jerry Rice Jubilation?" https://www.kaggle.com/eyecjay/sf-crime/vehicle-thefts-or-jerry-rice-jubilation, June 2015.

[KW13] David Kahle and Hadley Wickham. "ggmap: Spatial Visualization with ggplot2." *The R Journal*, **5**(1):144–161, 2013.