UCLA

UCLA Electronic Theses and Dissertations

Title

Crime and Demographics: An Analysis of LAPD Crime Data

Permalink

https://escholarship.org/uc/item/2v76v571

Author

Cung, Bianca

Publication Date

2013

Peer reviewed|Thesis/dissertation

eScholarship.org

Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Crime and Demographics:
An Analysis of LAPD Crime Data

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Bianca Cung

2013

© Copyright by

Bianca Cung

2013

ABSTRACT OF THE THESIS

Crime and Demographics: An Analysis of LAPD Crime Data

by

Bianca Cung

Master of Science in Statistics

University of California, Los Angeles, 2013

Professor Robert Gould, Co-Chair

Professor Frederic Paik-Schoenberg, Co-Chair

The Los Angeles Police Department (LAPD) often faces the task of predicting crime before it happens to better safeguard the community. The LAPD has a historical record of over a million incidences. Statistical analysis of some of these incidences can aid the LAPD in crime prevention by identifying patterns and trends. The LAPD can then allocate resources accordingly. This thesis looks at crimes in Los Angeles from 2005 to 2009 though focuses on crimes in the most recent year. It uses clustering and regression techniques in addition to contingency tables and Chi-squared tests to find clear crime trends. Historical weather data and demographic data from the U.S. 2010 Census are included as part of the analysis in order to determine some external factors that may affect crime. This paper looks particularly at trends over the course of the year as well as victim characteristics.

This thesis of Bianca Cung is approved.

Yingnian Wu

Frederic Paik Schoenberg, Committee Co-Chair

Robert Gould, Committee Co-Chair

University of California, Los Angeles

2013

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

It is the goal of the LAPD to protect the lives and property of the Los Angeles community. One of the major milestones in fulfilling this goal is to reduce crime each year. It comes to question, however, how much crime can be reduced and if there is a point where crime rate can no longer be cut down. One of the difficulties in answering this question pertains to how the number of crimes varies per day. To address this difficulty, temperature, rain, time of day, and day of the week can be considered potential predictors of crime.

The people involved in crimes are also of interest as it is believed that certain ethnicities and genders are more likely to be victims of a specific kind of crime. It is just as important to identify who crimes target in addition to when crimes occur. Finding crime trends and relations can help better direct the LAPD's efforts in safeguarding the Los Angeles community.

This thesis aims to explore relationships between demographic background, weather, and crime. It hopes to identify crime patterns so that incidents can be foreseen and therefore better prepared for. Key variables are identified through contingency tables and Chi-squared tests. Clustering and regression also make up a large part of the analysis. In particular, this thesis analyses crimes in 2009, which consist of over 300,000 incidents.

CHAPTER 2

The Three Datasets

The data used for this thesis all pertain to Los Angeles. It consists of historical weather data from NOAA's National Climatic Data Center [3], demographic information from the US 2010 Census [4], and a record of crimes in 2005 to 2009 from the LAPD. Because demographics on April 1, 2010 (Census Day) is believed to not have changed by much in one year, analysis focuses on incidences in 2009 which consists of 339,685 recorded incidences, though 2005 to 2008 data was used for the analysis of holidays. Many observations have missing data and are omitted from some analyses due to insufficient information. All of the data preprocessing is done in R and Microsoft Excel, while only R is used for the analysis.

In addition to each incident's beginning date, beginning and ending times as well as ending dates are also included in each observation. Only the beginning times and dates are considered in this analysis since the idea is to be able to foresee the incident before it happens. In figures, day 1 refers to the first day of 2009 while day 365 refers to the last day. Crime counts for each day are matched with daily average, high, and low temperature and precipitation amount in inches as measured by the weather station located at the Los Angeles International Airport. However, due to geographic variation, weather characteristics are only considered in the clustering portion of the analysis.

The data set includes fourteen types of crime. The crime types are listed below:

AGG	Aggravated Assault
ARSON	Arson
BTFV	Burglary/Theft from Vehicle
BURG	Burglary
GTA	Grand Theft Auto
GTP	Grand Theft Property
HOM	Homicide
KID	Kidnapping
NON	Non-aggravated Assault
OTH	Other
RCVD	Received Stolen Goods
ROBB	Robbery
THEFT	Theft
VAND	Vandalism

Up to five crime types are assigned to each incident, though 91.5% of incidents have only one type of crime. Multiple crime types are accounted for when considering frequency of each crime by day and by ZIP code for demographic data.

Frequencies of crime and victim characteristics are accounted for by ZIP code. Victim characteristics include ethnic background and gender. Each ZIP code is considered one observation. The U.S. Census data provides population counts per ZIP code. Only 210,557 incidents have valid ZIP codes while the remaining 129,128 are excluded from analyses involving demographic background. Figure 2.1 shows, however, that the proportions of excluded crime types are not the same as that of included ones. Figure 2.2 shows that the victim’s ethnic background in excluded incidences due to a lack of ZIP code is slightly disproportionate. While the 2010 Census data set has a separate variable for Hispanic and non-Hispanic background in addition to ethnic background, the crime data set has only one variable represent the victim’s descent.

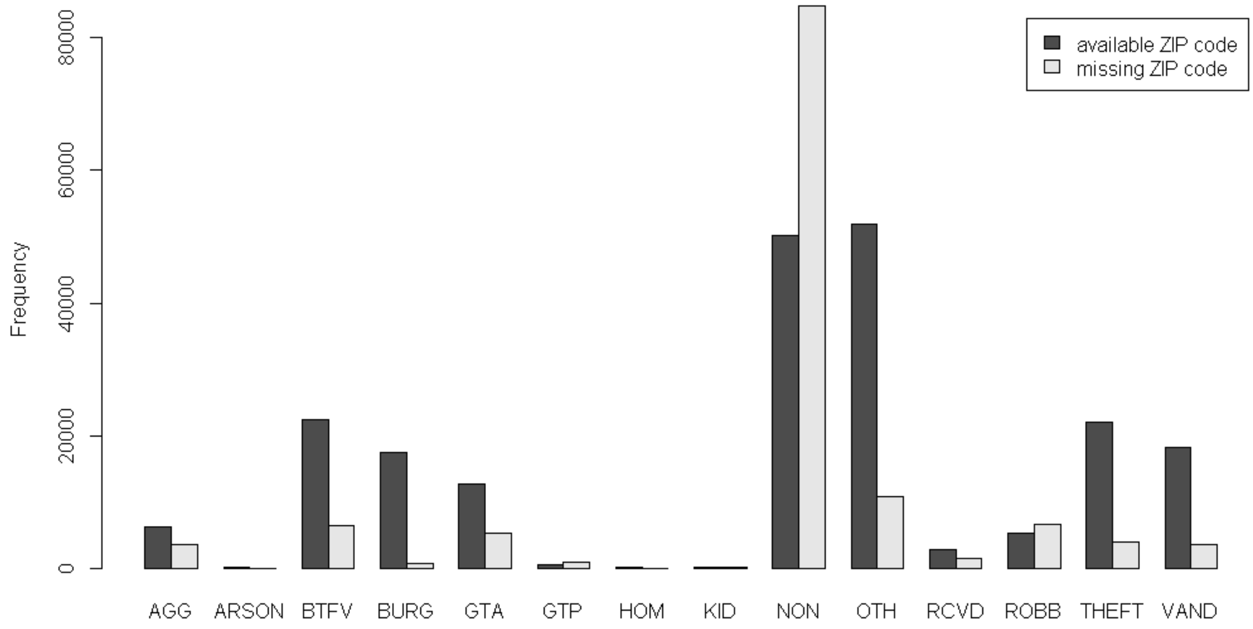


Figure 2.1: Bar Chart of Crime Types by Available or Missing ZIP Code

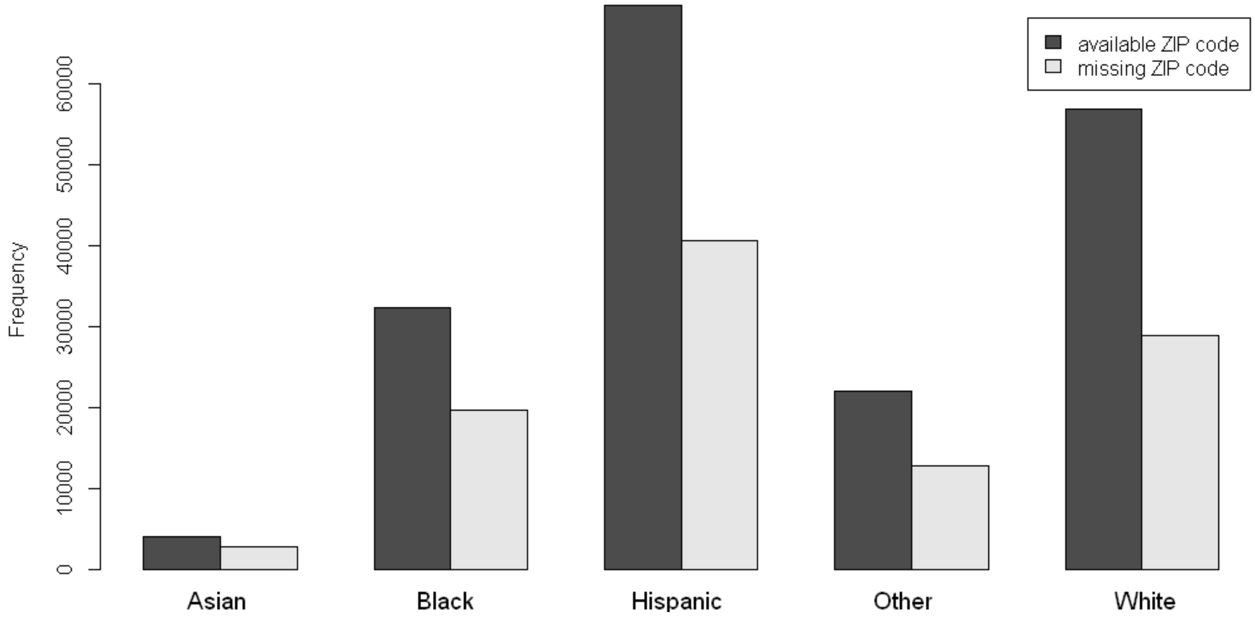


Figure 2.2: Bar Chart of Victim Ethnicity by Available or Missing ZIP Code

CHAPTER 3

Crime Type Overview

3.1 Crime Over the Year

Among the 339,685 recorded incidents in 2009, more than a third of the events are non-aggravated assaults (Table 3.1). The least frequent are homicides, arsons, and kidnappings. We can get sense of trends of each crime type in Los Angeles over the year by fitting linear and Poisson regression models. Although the nature of a linear model suggests a negative number of crimes per day if used for prediction beyond 2009, a linear model works for simplicity and exploratory analysis. The Poisson model is, nevertheless, more realistic.

Table 3.2 shows summaries of the two regression methods estimating the crimes per day over the year for each crime type. January 1, 2009 incidents are excluded from the regression due to its extremely large number of crimes and high leverage. It is suspected that January 1 is the default date for crimes with unknown dates. Because of the possibility that there are crimes with unknown dates, the models may give underestimates for the number of crimes per day. Nevertheless, most crime types see a decrease over the year. Linear and Poisson regression both suggest that burglaries, burglaries/thefts from vehicle, homicides, and kidnappings increase over the year. The estimated values for crimes with increases are insignificant. Although most of the crime types decrease by only a couple crimes per day by the end of the year, that small decrease constitutes at least 6% of the crimes. The lowest decrease, 6%, is GTA beginning with 53 crimes per day and decreasing to about 50 crimes per day.

Table 3.1: Crime Type Overall Frequency

Crime Type	Frequency	Proportion
AGG	9,805	0.0289
ARSON	346	0.0010
BTFV	28,905	0.0851
BURG	18,261	0.0538
GTA	18,729	0.0551
GTP	1,433	0.0042
HOM	311	0.0009
KID	503	0.0015
NON	147,680	0.4348
OTH	63,282	0.1863
RCVD	20,141	0.0593
ROBB	12,163	0.0358
THEFT	26,255	0.0773
VAND	22,334	0.0657
Total	339,685	

After taking into consideration day of the week to estimate the number of each type of crime that will occur in a day, a model is constructed for each crime type (Table 3.3). Sundays generally see fewer crimes in total, but more there are also more aggravated assaults and arsons in particular. Saturdays tend to have more crimes that Sundays, though not as much as weekdays.

In total, Mondays have fewer crimes than Saturdays, though for many crime types, it follows closely along the lines of the other weekdays. Fridays see the most crime counts, with an estimate of over 1,000 crimes per day. Fridays are estimated to have the highest number of burglaries, burglaries/thefts from vehicle, grand theft auto, grand theft property, non-aggravated assaults, theft, and *other*. Day of week seems to be an important factor that affects crime rate over the year.

Table 3.2: Projecting Crime over the Year

	Linear Model		**Poisson Model**	
Crimes	**Intercept**	**Day**	**Intercept**	**Day**
Total	982.3286	-0.2824	6.8909	-0.0003
AGG	29.4102	-0.0139	3.3841	-0.0005
ARSON	1.2237	-0.0015	0.2250	-0.0016
BTFV	78.4020	0.0043	4.3619	0.0001
BURG	49.4601	0.0031	3.9012	0.0001
GTA	53.0935	-0.0097	3.9724	-0.0002
GTP	4.1373	-0.0012	1.4210	-0.0003
HOM	0.8170	0.0002	-0.2016	0.0002
KID	1.3669	0.0001	0.3126	0.0000
NON	430.4610	-0.1413	6.0662	-0.0003
OTH	187.7164	-0.0784	5.2371	-0.0005
RCVD	60.3079	-0.0280	4.1022	-0.0005
ROBB	34.8046	-0.0081	3.5504	-0.0002
THEFT	72.9546	-0.0056	4.2899	-0.0001
VAND	67.5031	-0.0345	4.2155	-0.0006

Table 3.3: Prediction by Day of Week

	Prediction for	Change Relative to Sunday						
Crimes	**Sunday**	**Mon.**	**Tue.**	**Wed.**	**Thu.**	**Fri.**	**Sat.**	**R-squared**
Total	819.08	82.23	104.00	140.62	141.07	223.67	88.81	0.42
AGG	36.08	-11.85	-12.62	-14.27	-13.28	-9.33	-3.08	0.34
ARSON	1.02	-0.04	-0.19	0.00	-0.25	-0.02	0.00	0.01
BTFV	74.94	-1.73	3.48	4.67	5.66	10.33	7.31	0.11
BURG	32.13	20.29	22.19	22.42	22.53	32.21	5.54	0.56
GTA	52.37	-3.50	-2.87	-2.79	-2.48	3.27	1.02	0.08
GTP	3.37	0.44	0.67	0.44	0.86	1.13	0.37	0.03
HOM	1.06	-0.21	-0.35	-0.31	-0.40	-0.19	0.02	0.02
KID	1.23	0.13	0.04	0.19	0.11	0.12	0.44	0.01
NON	332.37	47.12	74.71	101.08	104.26	131.29	46.60	0.49
OTH	158.90	22.50	10.58	16.19	13.59	28.56	9.90	0.08
RCVD	50.54	5.02	4.85	4.90	4.97	8.44	4.31	0.06
ROBB	32.81	1.58	-0.94	-0.42	-1.81	2.02	3.23	0.06
THEFT	61.96	6.52	8.83	14.69	10.59	18.44	10.71	0.21
VAND	64.48	-7.87	-8.40	-10.54	-7.16	4.21	6.79	0.28

3.2 When Crime Happens

The frequencies of beginning times for each crime type make sense in that vandalism, grand theft auto, and burglary and theft from vehicle occur most often in the late evenings while other burglaries are more common during working hours (Figure 3.1). To make better sense of the burglary crimes' relation to working hours, Figure 3.2 shows similar patterns on weekdays and a decline in such crimes on weekends. Strangely, there is a spike in burglaries on weekdays as with non-aggravated assault and other crimes at noon. A closer look at this attributes most of the counts to noon, which indicates that noon might be the default time assigned to unknown start times.

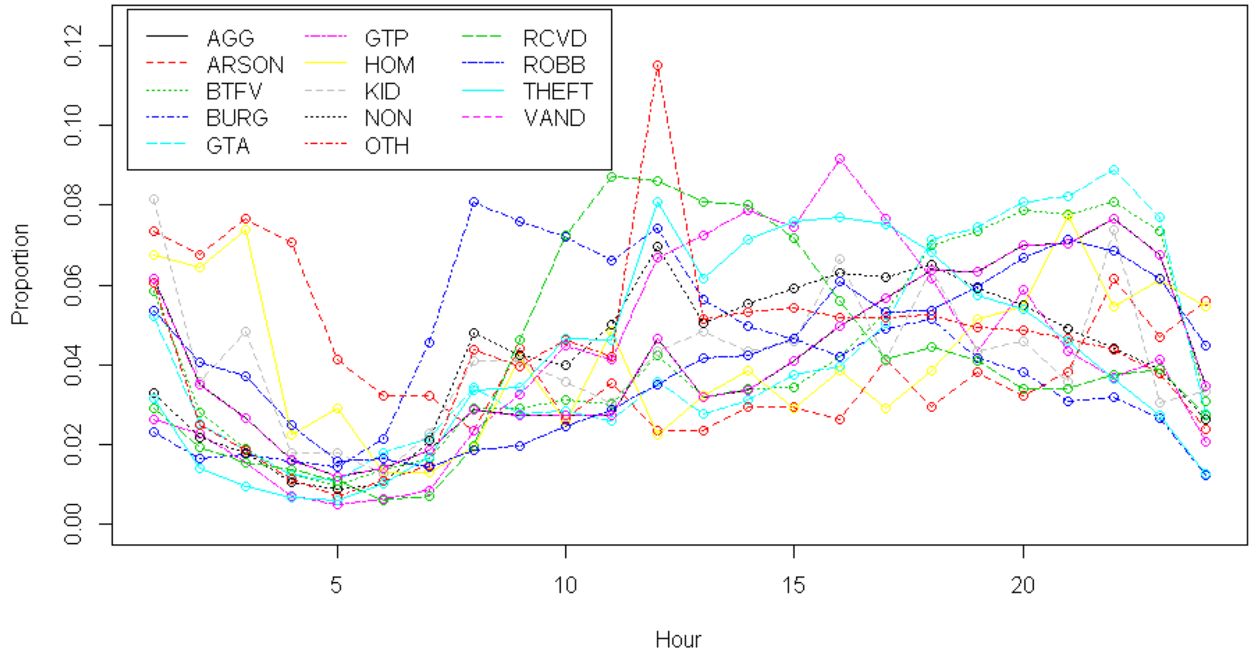


Figure 3.1: Frequency of Crime Start Times by Crime Type in 2009

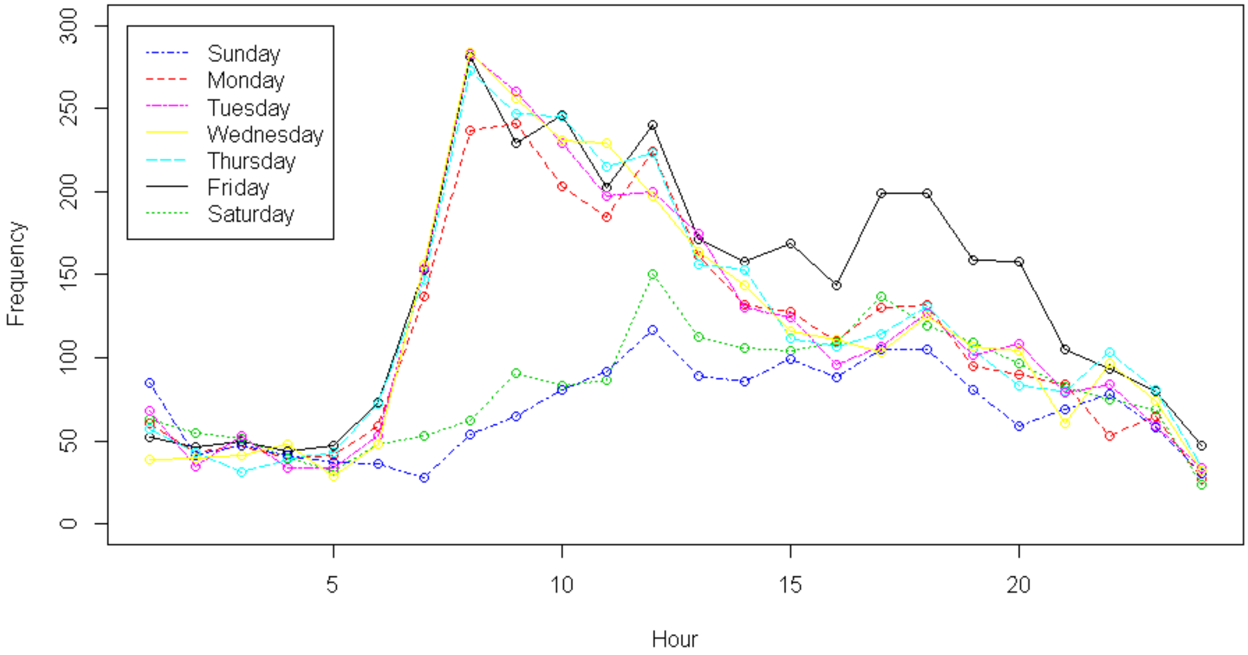


Figure 3.2: Frequency of Crime Start Times by Day of the Week in 2009

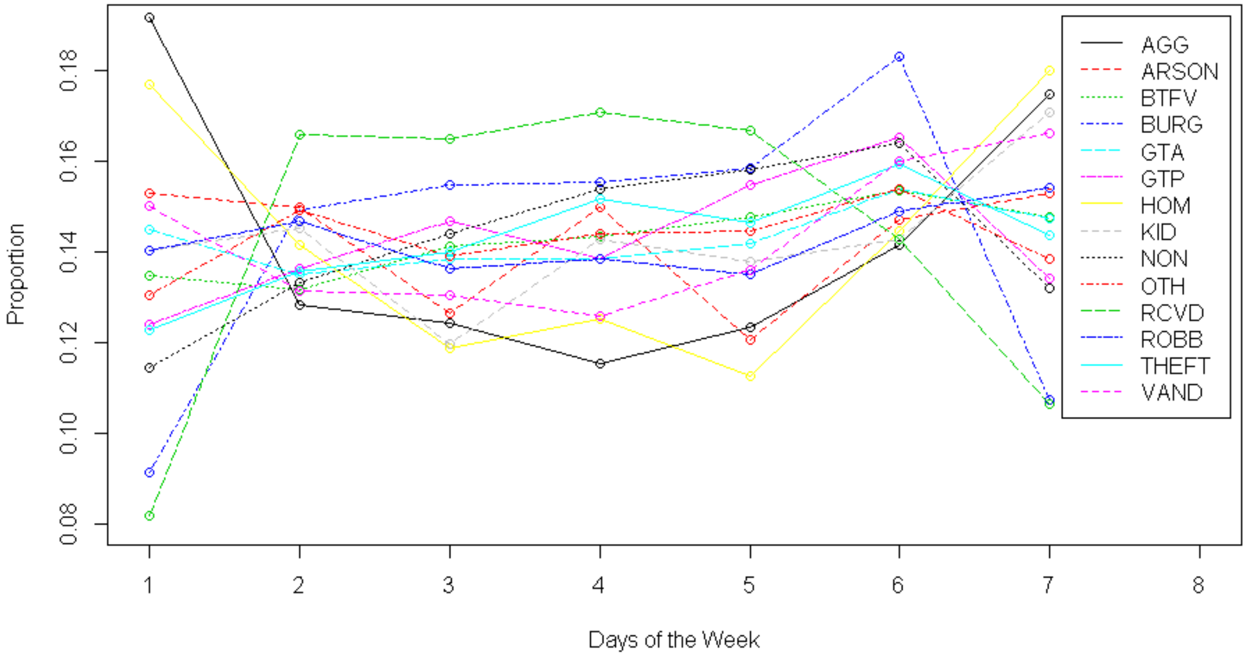


Figure 3.3: Proportion of Crime Types by Day of the Week in 2009

Differences can be found when comparing number of incidences on weekdays to that on weekends for other crime types. Figure 3.3 compares the proportion of each crime that occurs over the course of the week beginning with Sunday for the different crime types. As in the burglaries case, non-aggravated assault, other, receipt of stolen goods, and theft tend to happen more often on the weekdays than on the weekends. Aggravated assault and grand theft auto see an opposite pattern. A chi-squared test (Table 3.4) for the proportion of crimes on weekends equal to 2/7 (0.2857) shows that aggravated assault and vandalism are more likely to occur on weekends ($p < 0.0036$ after adjusting for multiple comparisons). Burglaries, non-aggravated assault, other, receipt of stolen goods, and theft are more likely to occur on weekdays. There is no significant difference in crime rate on weekdays or weekends for the other crime types.

Table 3.4: Chi-Square Test for Differences in Crime Frequency on Weekdays and Weekends

Crime Type	Weekday	Weekend	Prop. on Weekend	Chi-Square Statistic	P-Value
AGG	6,183	3,576	0.3664	311.5503	< 2.2e-16
ARSON	236	104	0.3059	0.6776	0.4104
BTFV	20,723	8,166	0.2827	1.3135	0.2518
BURG	14,625	3,630	0.1988	674.9384	< 2.2e-16
GTA	12,753	5,270	0.2924	3.9524	0.0468
GTP	1,045	364	0.2583	5.1739	0.0229
HOM	200	111	0.3569	7.7251	0.0054
KID	270	122	0.3112	1.2500	0.2636
NON	101,645	33,260	0.2465	1014.2397	< 2.2e-16
OTH	45,867	16,876	0.2690	86.1950	< 2.2e-16
RCVD	3,559	827	0.1886	202.8794	< 2.2e-16
ROBB	8,582	3,580	0.2944	4.4540	0.0348
THEFT	19,213	6,982	0.2665	47.1932	6.43e-12
VAND	14,982	6,934	0.3164	101.0515	< 2.2e-16

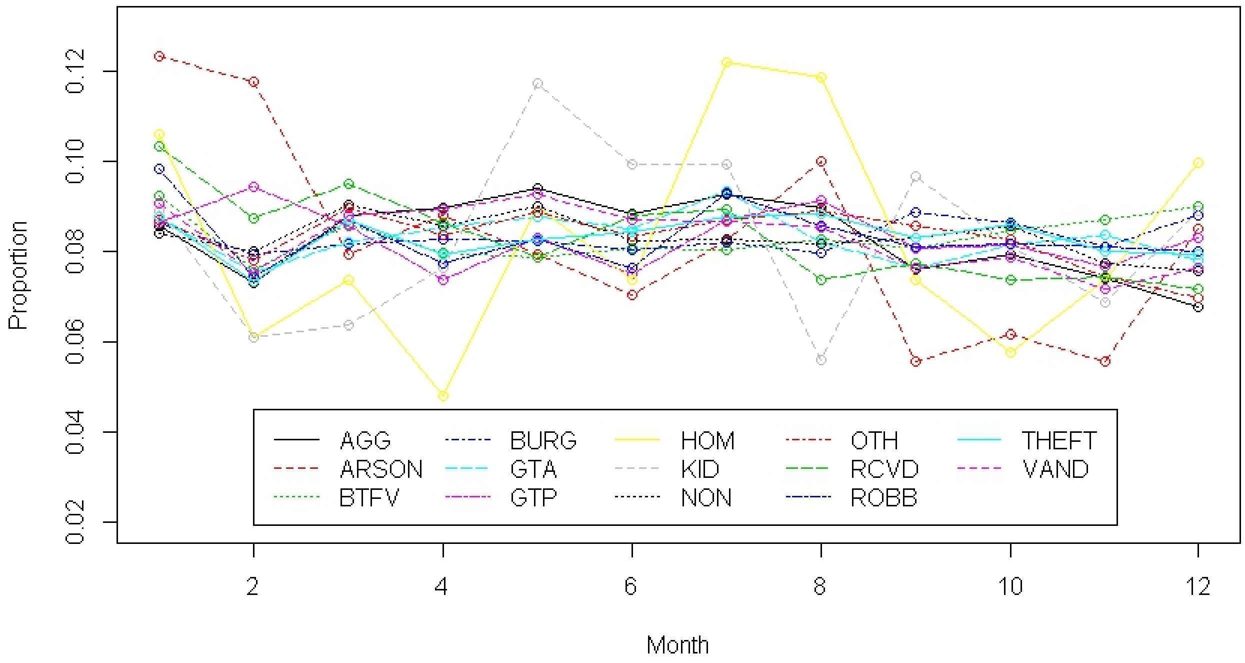


Figure 3.4: Proportion of Crime Types by Month in 2009

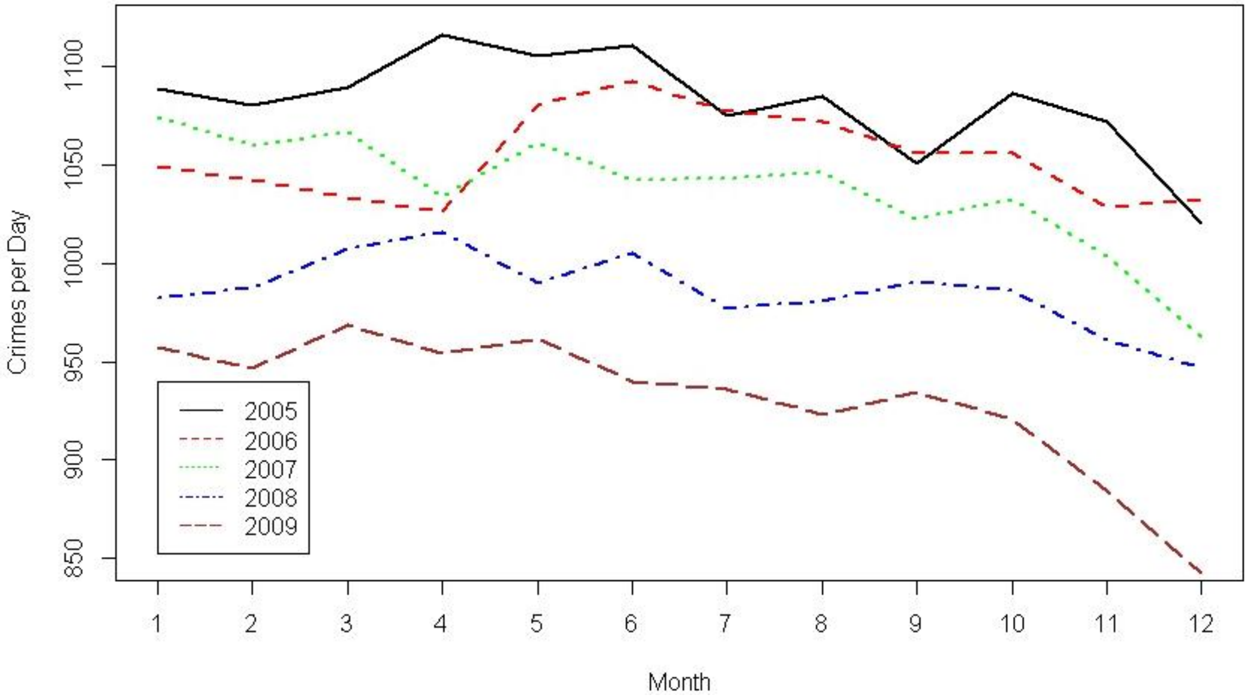


Figure 3.5: Average Crimes per Day Across Months in 2005-2009

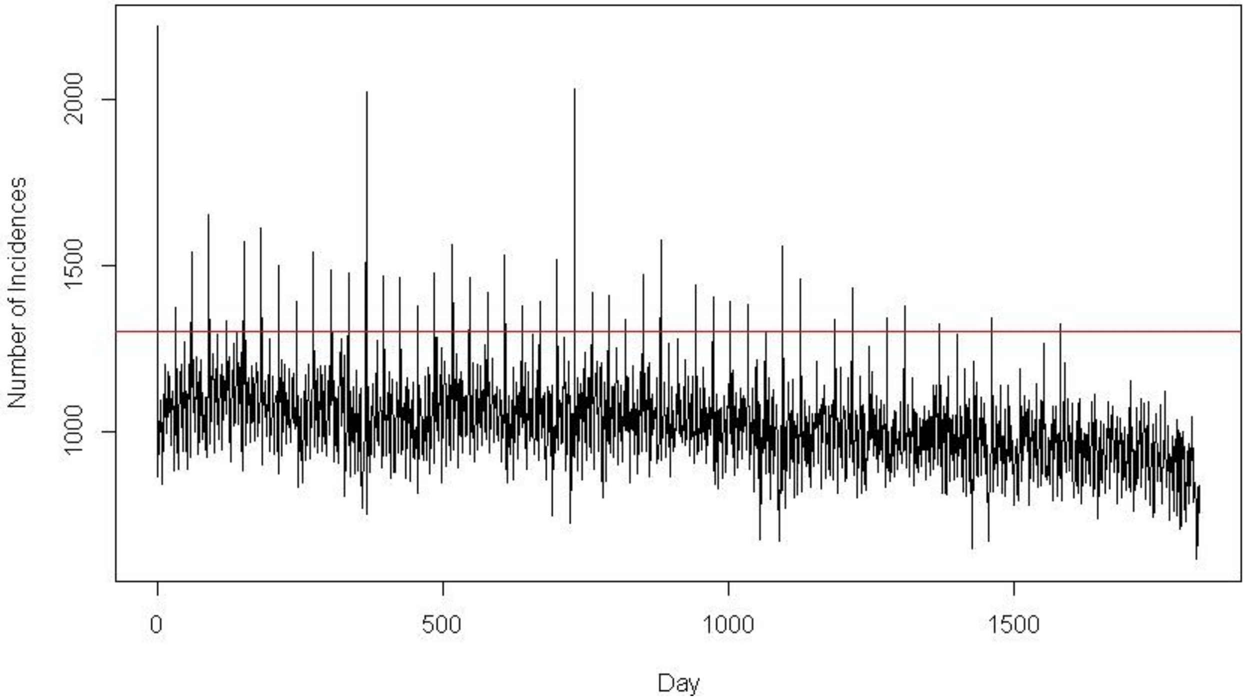


Figure 3.6: Crimes per Day in 2005-2009

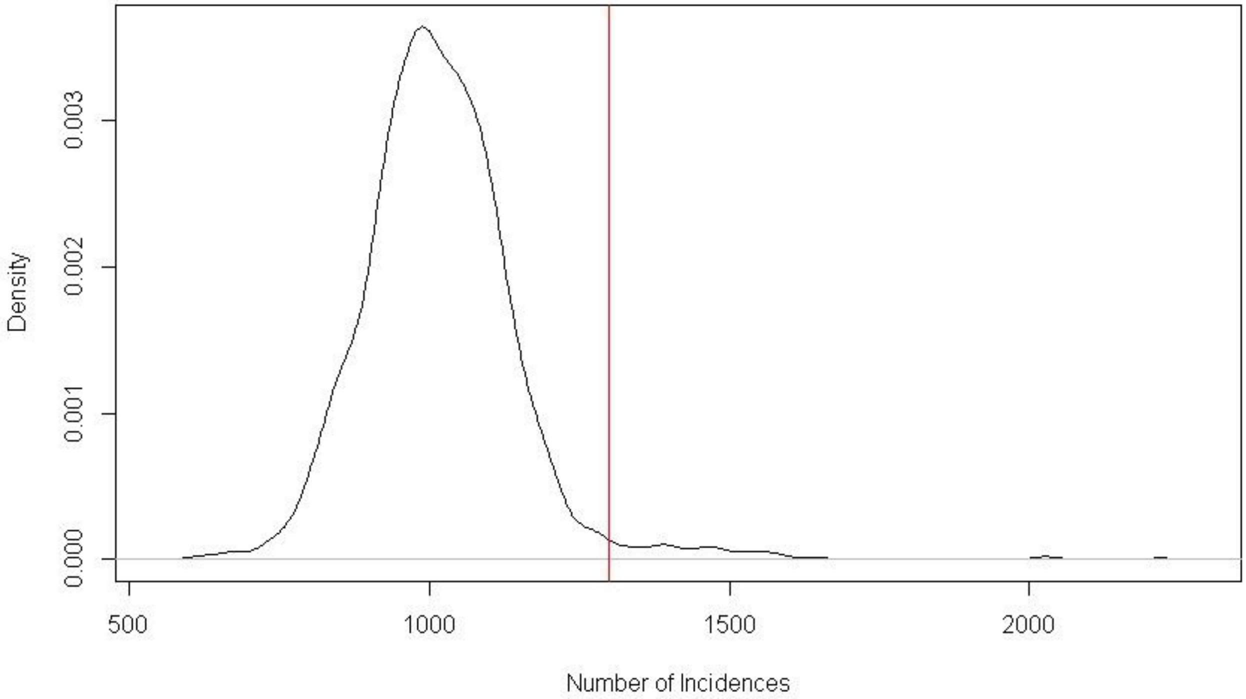


Figure 3.7: Density of Crimes per Day Across Months in 2005-2009

Across months in 2009, there is no distinguishable pattern for all crime types over the year (Figure 3.4). There are notably a higher proportion of homicides in the late summer months than other months, a higher proportion of kidnappings in spring, and a higher proportion of arson in January and February. The drop in crimes from January to February for all crime types can be attributed to fewer days in February (Figure 3.5). The year 2008 even sees an increase in crimes per day from January to February even though leap year has been accounted for in that month. Though it may just be coincidence, November has fewer crimes per day than October throughout 2005 to 2009.

Despite the sharp increase in crimes per day in May 2006, crime has generally been decreasing, reaching an average of about 850 crimes per day in December 2009. Figure 3.6 gives the recorded number of crimes per day over the five years. Though decreasing, there is a recurring spike in crimes. It turns out that the days that have more than 1,300 crimes per day are the first of every month. We suspect that the first of the month is a default date given to crimes with unknown dates. After removing the first of every month, the density of crimes per day is approximately normal (Figure 3.7).

We can use kernel smoothing to obtain an estimate of number of crimes per day relative to other days. Kernel smoothing is a nonparametric regression model created from a weighted moving average of nearby points. In this case, kernel regression calculates an average number of crimes per day within a week, giving an average estimate over two weeks. The first seven days of the year take the same estimate as the eighth day of the year while the last seven days of the year take the same estimate as the eighth to last day. The kernel estimates are taken to be the expected number of crimes on the given date. The estimate can be seen in red on Figure 3.8.

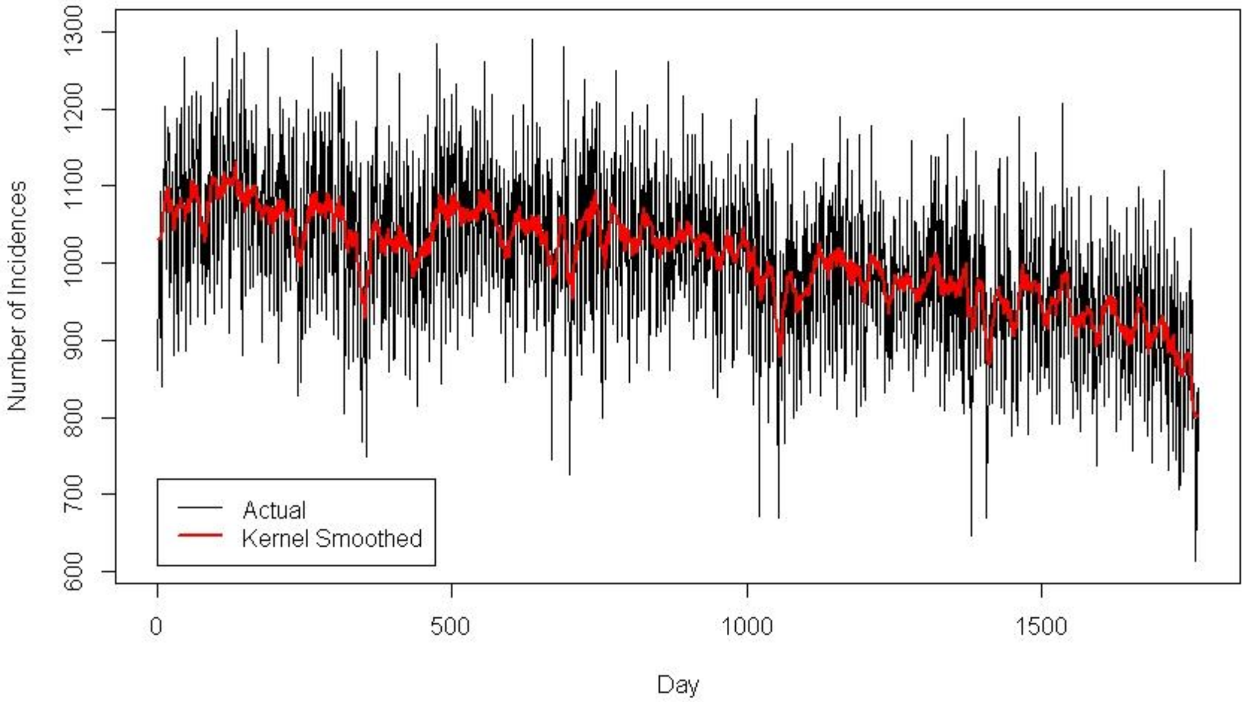


Figure 3.8: Crimes per Day in 2005-2009 with First Day of the Month Removed

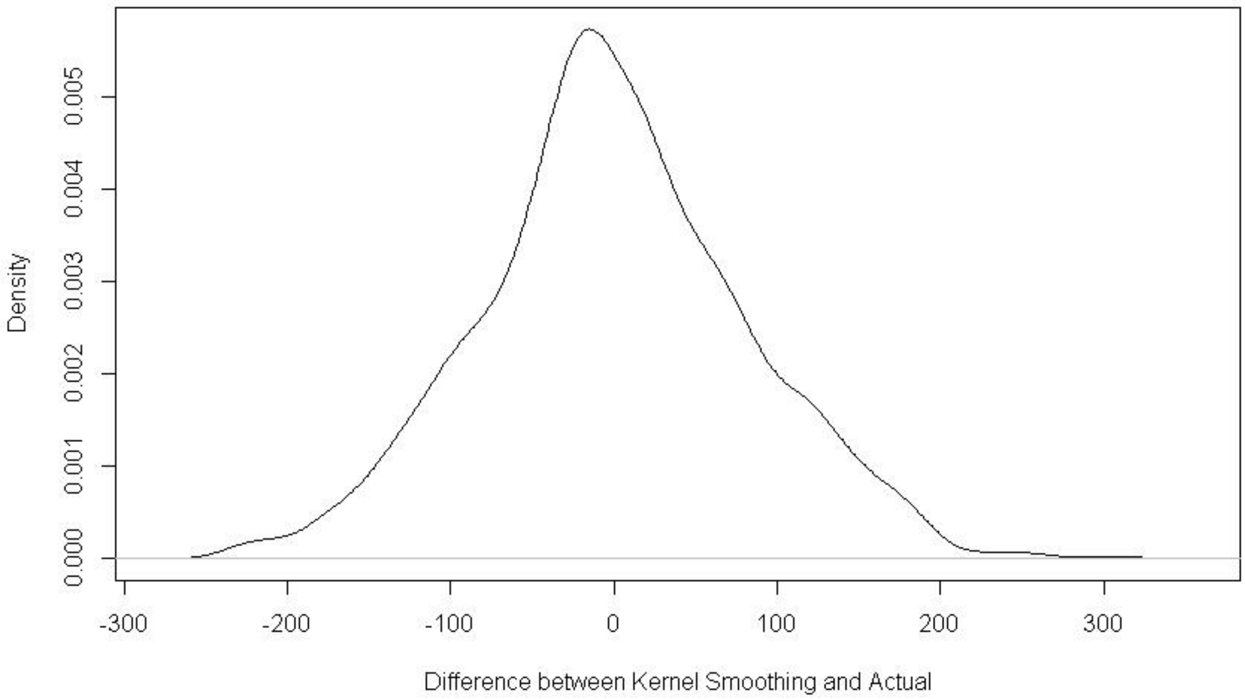


Figure 3.9: Difference in Expected and Actual Crimes per Day in 2005-2009

Figure 3.9 gives a density curve of the difference between the kernel regression estimate and actual number of crimes per day. There are 88 days which make up the top 5% of days with more crimes than expected. Among the 88 days, 78 are Fridays and have at least 134 more crimes than estimated from kernel regression. Nine of the remaining 10 days are also weekdays, while the only Saturday happens to be on Halloween. Since there is only one weekend Halloween observation, we remain inconclusive about Halloween's effect on crime. Similar to how crimes tend to occur more on Fridays, we can also expect fewer crimes on Sundays.

Among the 88 days that make up the top 5% of days with fewer crimes than expected, 63 are Sundays and 13 are Mondays. The remaining 12 observations are spread throughout the week. Most of the non-Sunday observations happen to be around the time of holidays such as Thanksgiving and Christmas. Monday observations also include Martin Luther King, Jr. Day, President's Day, Memorial Day, and Labor Day. This suggests that holidays may influence crime.

3.3 Holidays

Since holidays have different ways of celebration, some more popular than others, it comes to question which holidays stand out in terms of crime behavior.

It was found that Monday holidays such as Martin Luther King, Jr. Day, President's Day, Memorial Day, and Labor Day tend to have fewer crimes on average in comparison to other days within a week prior and after. The crime beginning time proportions on these holidays tend to follow that of Sundays, where there fewer incidences occurring in the morning during working hours. Since for these holidays, people usually have the day off from work or school, it makes sense there would be fewer crimes during working hours (Figure 3.10). The proportion of crimes on Mondays that also happen to be holidays are closer to Sunday proportions if not between that of Sundays and that of all other Mondays (Figure 3.11). While holidays that fall on a Monday do not follow Sundays' trends exactly, we can expect some kind of middle ground on such three day weekends.

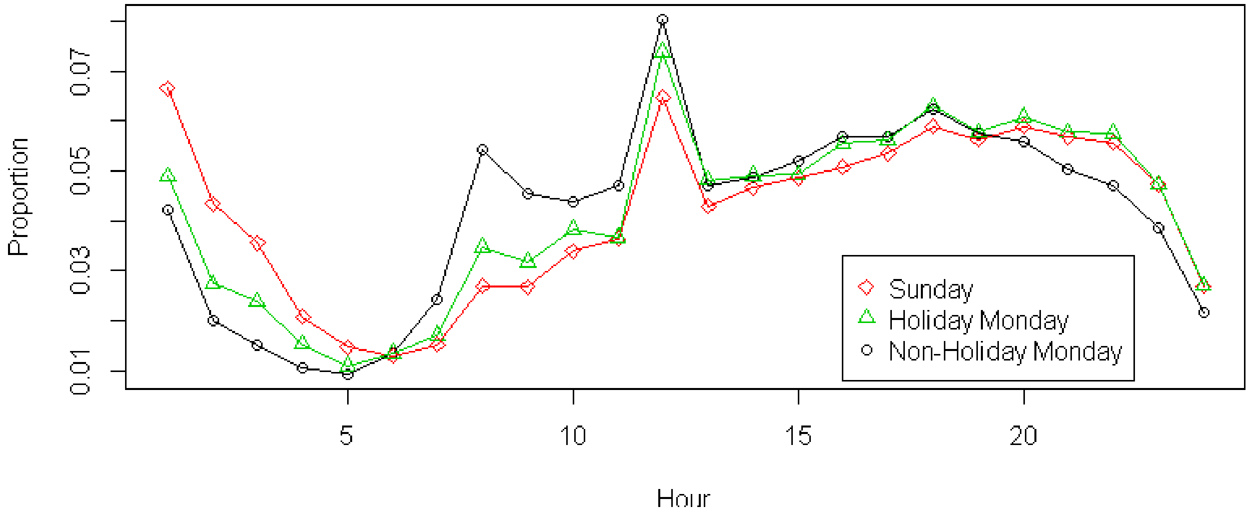


Figure 3.10: Crime Time Proportions of Sundays and Mondays

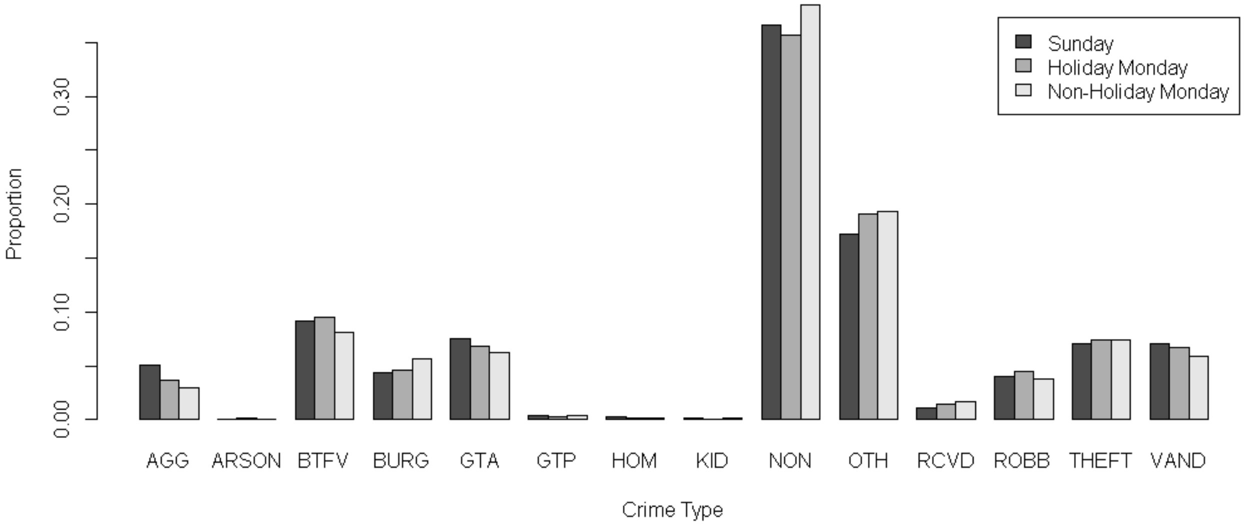


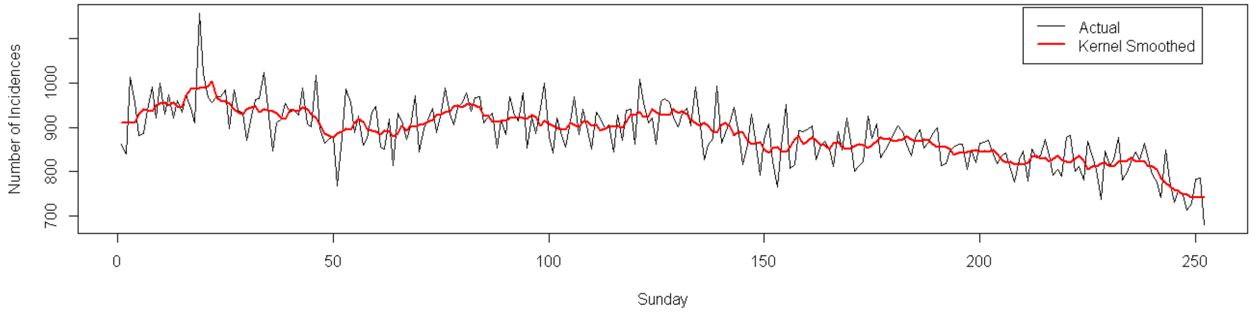
Figure 3.11: Crime Type Proportions of Sundays and Mondays

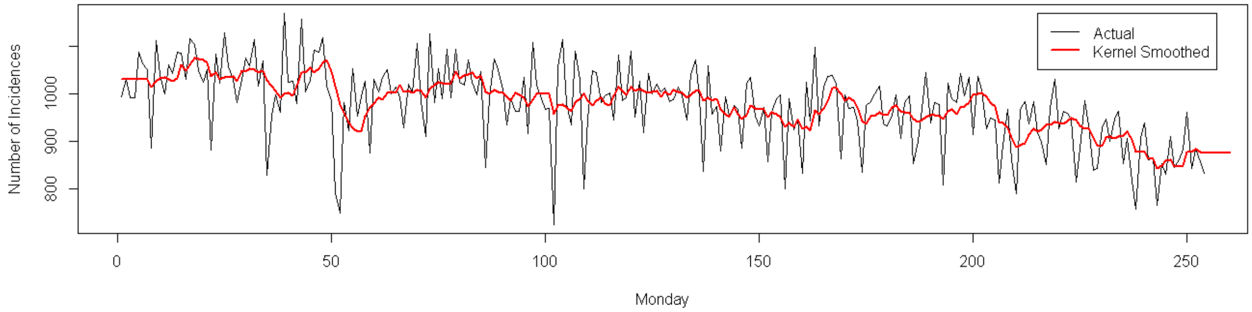
The overall crime trend and day of the week are taken into consideration by applying kernel regression smoothing to each day of the week over the five years. The kernel estimates for Sundays are obtained from using the two available Sundays prior to and the two available Sundays after the date of interest. The same can be applied to other days of the week. The estimates can be found in Figure 3.12. The first day of each month remain excluded due to the unusually large number of incidents reported.

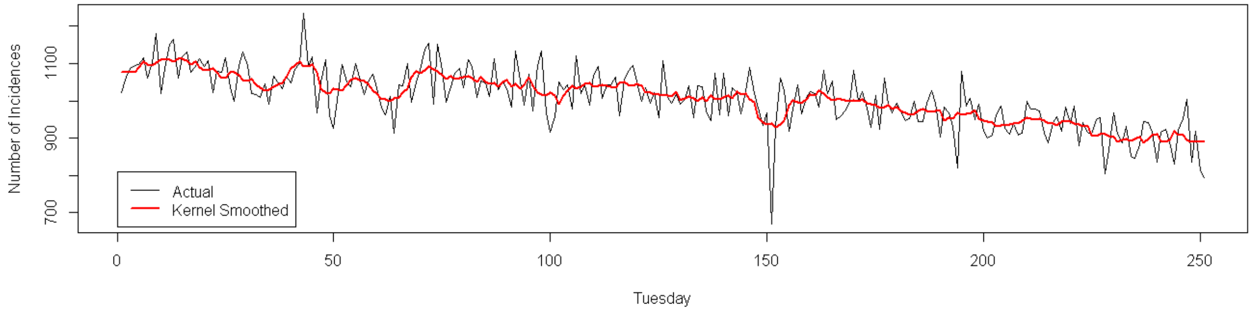
November and December stood out most among days that had more or less than expected crime counts. On Thursdays, there are two recurring drops in crime each year. These drops happen to be Thanksgiving Day and the week of Christmas. Extreme drops in number of incidences from expected on other days of the week also reflect the week of Christmas and the days around Thanksgiving. Thanksgiving and Christmas are thus crucial holidays that have distinctly different crime patterns from the rest of the year. However, since both holidays are celebrated differently and only five years of data is currently available, we omit further analysis of how the crime counts drop.

On the other hand, crime rate is higher than estimated on days one week prior to Thanksgiving and Christmas. Since these days still follow the general trend of crimes per day, we conclude that crime rate is not actually higher but rather the kernel estimate is influenced by the low crime count on Thanksgiving or Christmas. The findings further confirm a drop in crime around Thanksgiving and Christmas.

A clear pattern has yet to be identified for other days with higher or lower than expected crime counts. With the exception of Wednesday, November 23, 2005 the peaks in Figure 3.12 happen to fall on 15th of the months including January, May, and November in 2005, as well as July in 2006 and August in 2007. Among the top 5% of days with fewer than expected crimes, though still unclear, none fall on the 15th. This brings suspicion that the 15th may also be a default date for unknown crimes, as in the case for the 1st of the month.







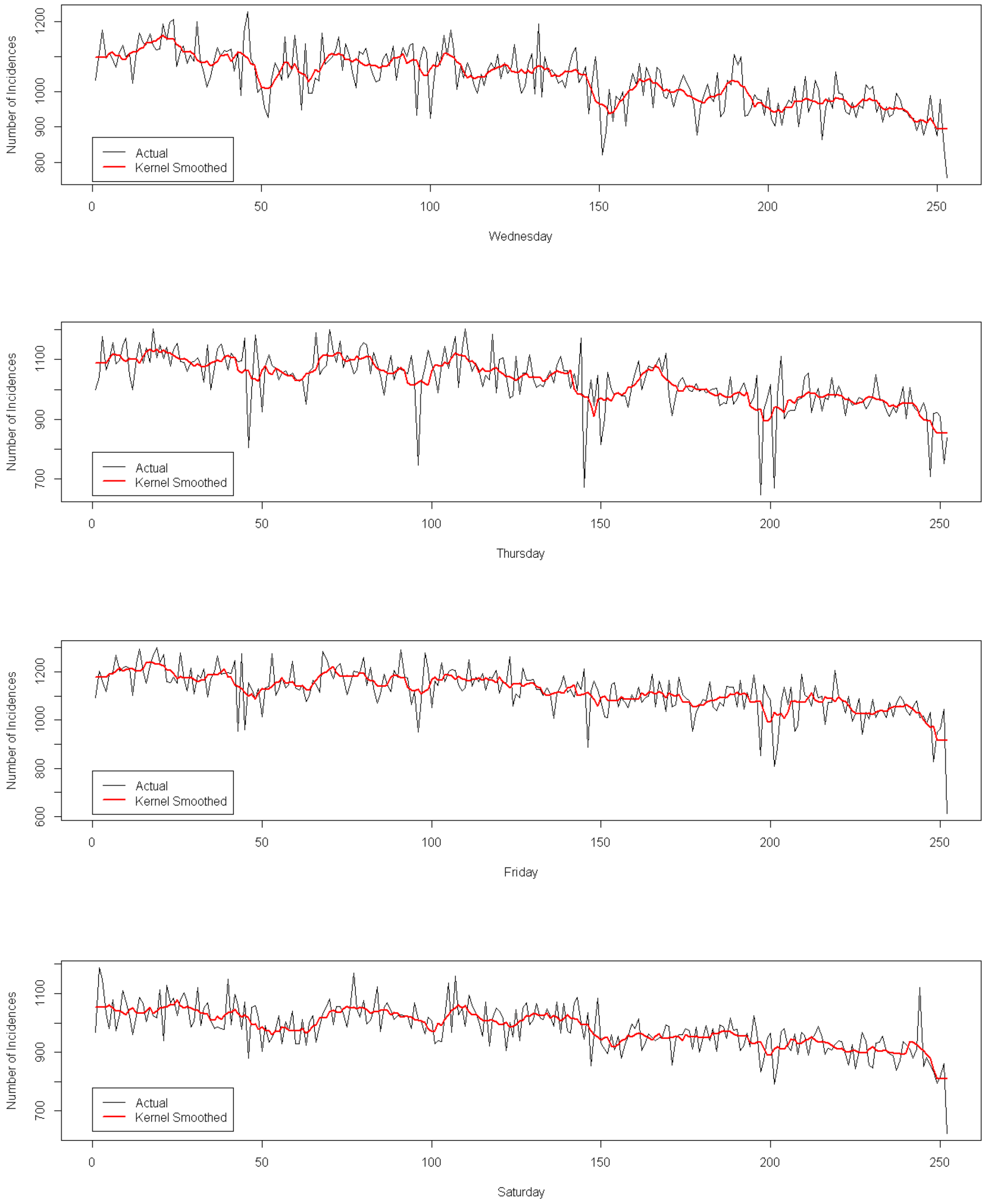


Figure 3.12: Kernel Smoothed Estimates for Number of Recorded Crimes by Day of the Week

3.4 Where Crime Happens

Table 3.5 gives the count for each of the top ten most common premises, which make up about 89% of crimes with recorded premises. 25,637 incidents do not have recorded premises. More than a third of incidences in 2009 occur on streets and parkways, which is also the most common location for many crime types. Single family dwellings are the most common location for burglaries, theft, and "other" crimes. Vehicles are the most common premise for vandalism. Note that the definition of stores is a bit arbitrary in that it only includes businesses such as jewelry or liquor stores, but not nail salons or pawn shops.

Geographically, LAPD's jurisdiction is divided into 21 areas, where Areas 3 and 12 have the highest crime count and Areas 4 and 8 have the lowest crime count (Figure 3.13). The proportions of different crimes look mostly consistent across the 21 locations. However, the division of Los Angeles' 21 areas are quite arbitrary. Further analysis will look at crime by ZIP code as the U.S. Census data gives population count for each ZIP code.

Table 3.5: Top 10 Common Crime Premises in 2009

Premise	**Number of Incidents**
Streets/Parkways	134,851
Single Family Dwelling	47,398
Multi-Unit Dwelling	25,637
Parking Lot	25,378
Store	12,465
Sidewalk	12,082
Vehicle	8,981
Police Facility	4,577
Driveway	4,401
Garage/Carport	3,340

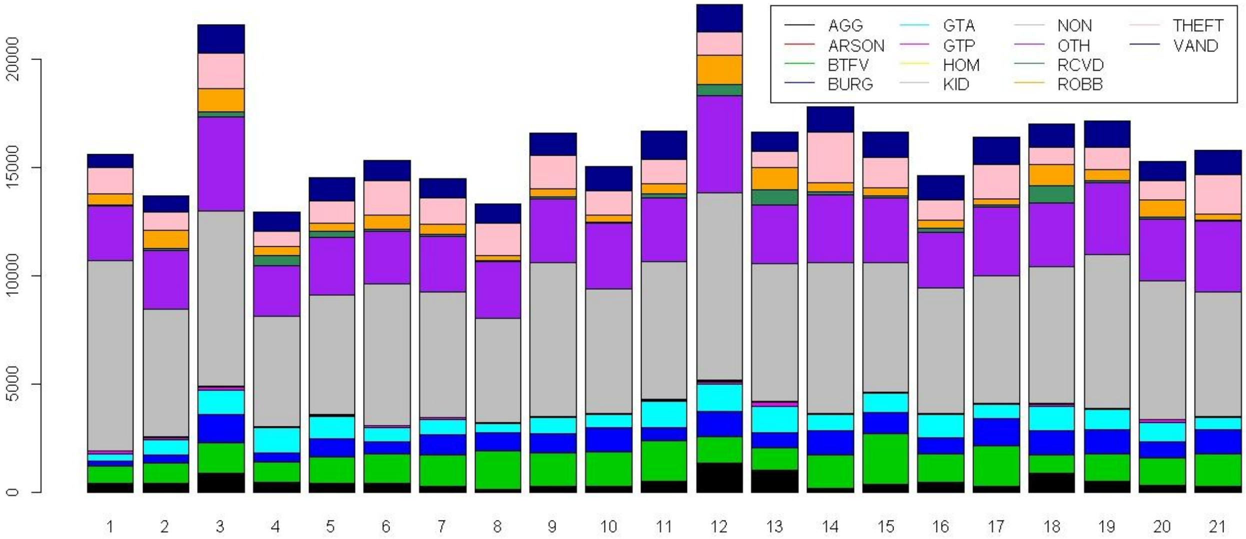


Figure 3.13: Bar Chart of Crime Types in 21 Areas in 2009

CHAPTER 4

U.S. Census

The LAPD crime data for 2009 is merged with 2010 U.S. Census data to examine demographic patterns in crime. A total of 133 valid ZIP codes are found within the LAPD crime data, assigning 210,557 incidents to their respective regions. The total population for the 133 areas is 4,713,903, with 90% of the areas ranging from 3,500 to 67,500 people (Figure 4.1). Three valid ZIP codes stand out due to extremely low population. These ZIP codes are 91608 with a population of 0, 90095 with a population of 3, and 90071 with a population of 15. The ZIP codes correspond to Universal Studios, UCLA, and part of the financial district respectively. Other low populated ZIP codes include parts of CSU Northridge (91330) and University of Southern California (90089), both of which have populations of at least 2,000. Crimes in these low-populated ZIP codes are not removed.

U.S. Census data include gender and ethnicity counts of different neighborhoods. Table 4.1 gives the ethnic background and gender summary of the overall population and of recorded crime victims (including incidences with unmatched ZIP codes). The 2010 U.S. Census reports *Hispanic* or *non-Hispanic* as a separate category from ethnicity. However, the LAPD crime data set reports the two under the same variable. According to the U.S. Census, about a little less than half the population in Los Angeles is of Hispanic descent. About half the population in Los Angeles is White, and gender is almost evenly split.

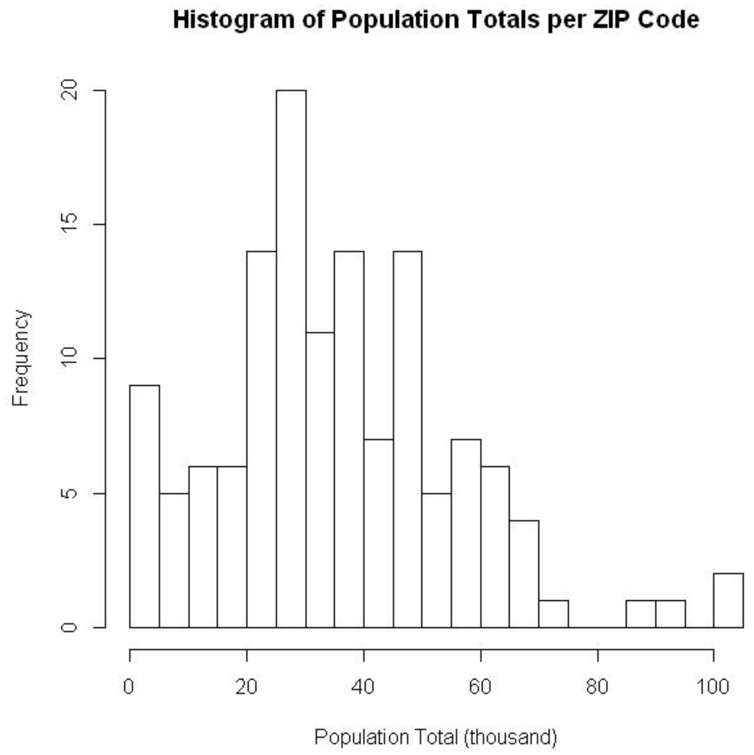


Figure 4.1: Histogram of Population for 133 ZIP code areas

Table 4.1: Demographic and Victim Summary*

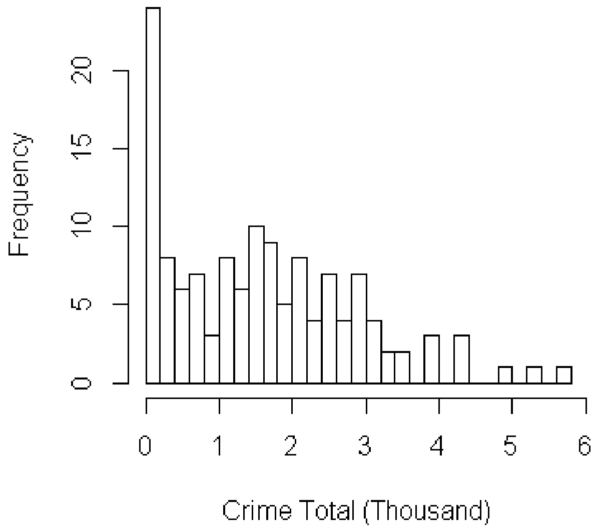
	White	Black	Asian	Other
Population	2,362,680	460,125	540,091	1,351,007
Pop. Proportion	0.5012	0.0976	0.1146	0.2866
Victims	85,951	52,089	9,030	44,566
Victims per 1,000	36.4	113.2	16.7	33.0

	Hispanic	Non-Hispanic	Male	Female
Population	2,253,227	2,460,676	2,339,904	2,373,999
Pop. Proportion	0.4780	0.5220	0.4964	0.5036
Victims	110,374	81,262	108,931	81,671
Victims per 1,000	49.0	33.0	46.6	34.4

*Hispanic is asked as a separate category on the 2010 Census form while Hispanic is categorized as an ethnicity in the LAPD dataset.

4.1 Population and Crime

Figure 4.2 shows a histogram of the total number of crimes per ZIP code. Like the case of population count per ZIP code, it is positively skewed. However, one interesting finding is that many areas have low crime count. A plausible reason for this is that the areas are bordering neighborhoods where jurisdiction is divided between the LAPD and another police department. For example, the 90022 ZIP has only 1 crime for a population of 67,179 people. The area actually lies in Monterey Park and Montebello, both of which have their own police department. The 91732 ZIP code, which maps to El Monte, also has only 1 crime for a population of 61,386. These are the two areas that fall far below the 95% prediction interval in Figure 4.3.



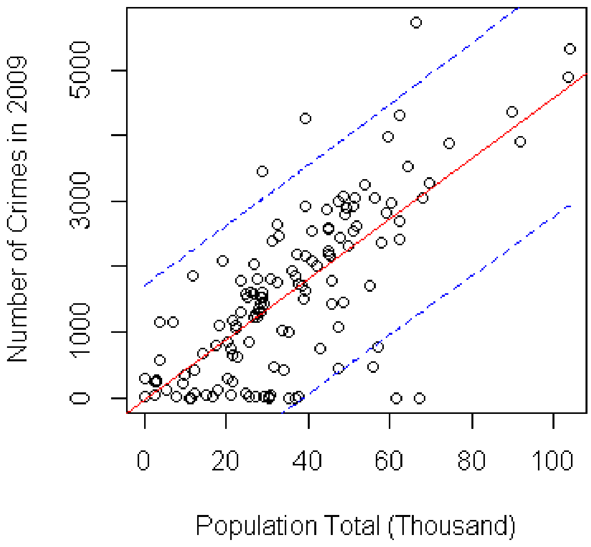


Figure 4.2 (left): Histogram of Total Crimes per ZIP Code

Figure 4.3 (right): Linear Model of Crime from Population Total

Table 4.2: LAPD Crime Count for ZIP Codes Outside of Los Angeles in 2009

ZIP	Total	BTFV	BURG	GTA	NON	OTH	RCVD	THEFT	VAND
90022	1	0	0	0	0	1	0	0	0
90405	2	0	0	0	0	2	0	0	0
91105	1	0	0	0	0	0	0	0	1
91205	10	1	0	1	5	3	1	0	1
91302	2	0	0	0	0	1	0	1	0
91732	1	0	0	0	0	0	0	1	0
91803	1	0	1	0	0	0	0	0	0

Among the 133 ZIP codes, 7 lie outside of Los Angeles boundaries while 45 are only partially in Los Angeles County. Partial ZIP codes are obtained from the Los Angeles Housing Department [5]. Among the 7 included areas that lie outside of Los Angeles boundaries, in addition to 90022 and 91732, other places include 90405 in Santa Monica, 91105 in Pasadena, 91205 in Glendale, 91302 in Calabasas, and 91803 in Alhambra. Table 4.2 gives the breakdown of the crime types in the 7 ZIP codes outside of Los Angeles in 2009. While most of these areas have only one or two incidents in 2009, the ZIP code 91205 is interesting in that it has 10 reported crimes in the LAPD dataset. Figure 4.4 shows that the 91205 portion of Glendale falls in a crevice of Los Angeles boundaries.

Figures 4.5 and 4.6 give the histogram and scatter plot of total crimes in ZIP codes that are wholly in Los Angeles. The overlaid least squares regression line in Figure 4.6 estimates an increase of about 50 incidents for every one thousand people in the population. The model explains about 76% of the variation in the data. The more populous a place is, the more crimes there will be. There are a few ZIP codes that fall outside of the prediction interval of the linear model. The one ZIP code that lies fully in Los Angeles yet with extremely low incident to population ratio is 90024. The 90024 area corresponds to the dormitories and apartments surrounding UCLA. Three ZIP codes have high incident to population ratio: 90003, 90028, and 90045. It turns out that 90045 contains the Los Angeles International Airport (LAX). It comes to question whether other factors, such as unemployment and population density, have anything to do with crime rate in a specific area.

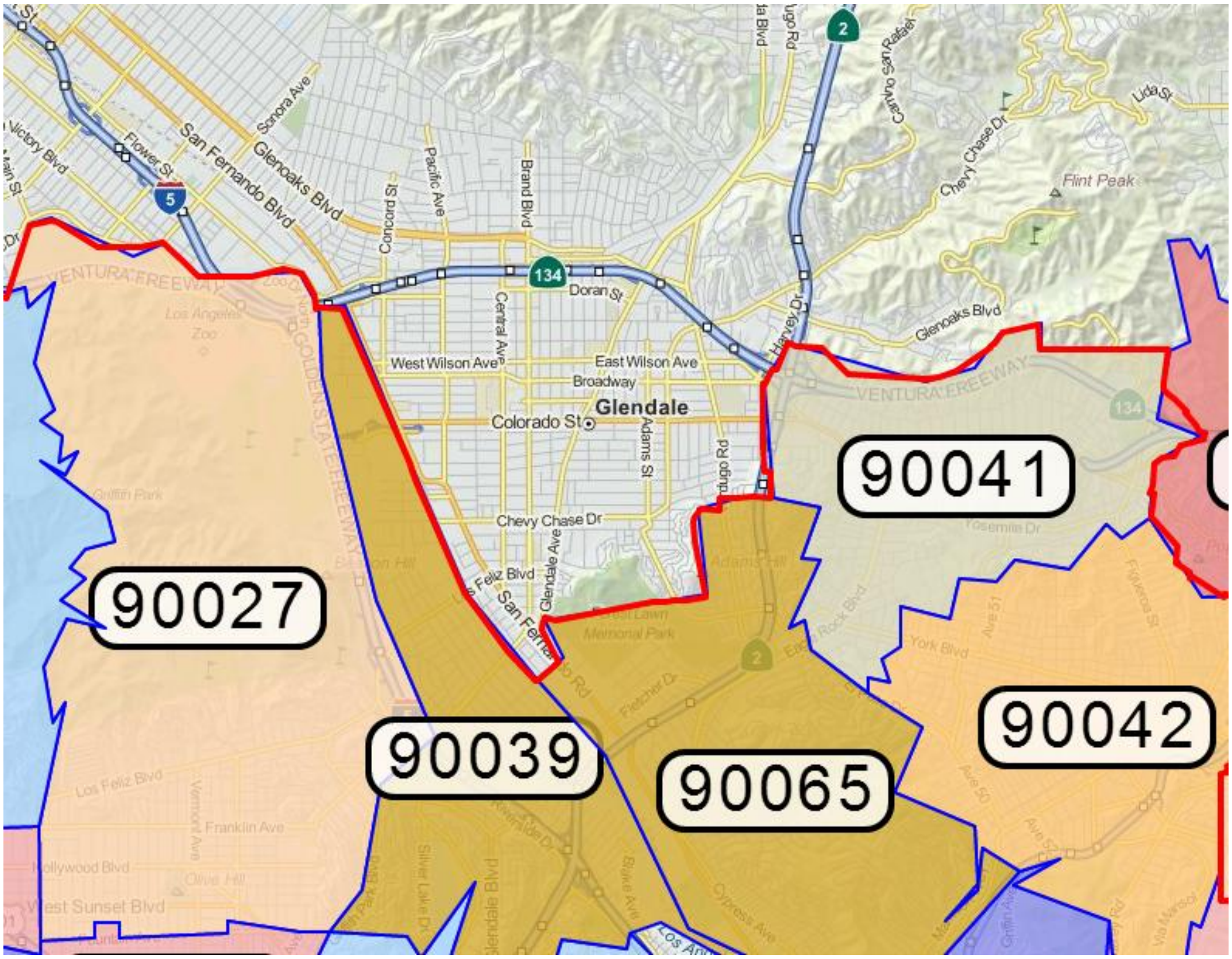
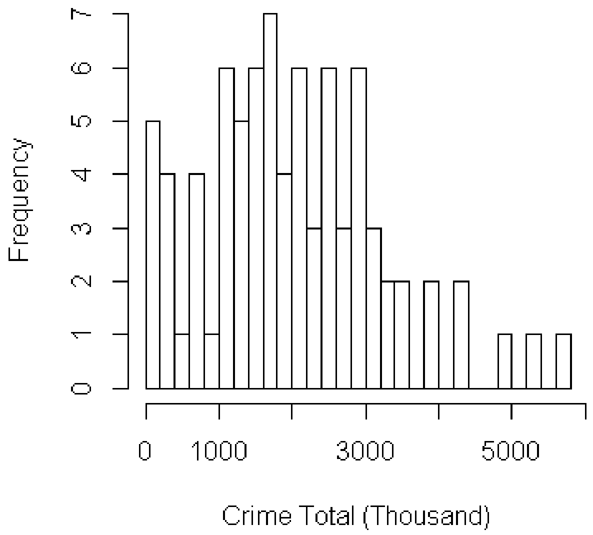


Figure 4.4: The 91205 Zip Code



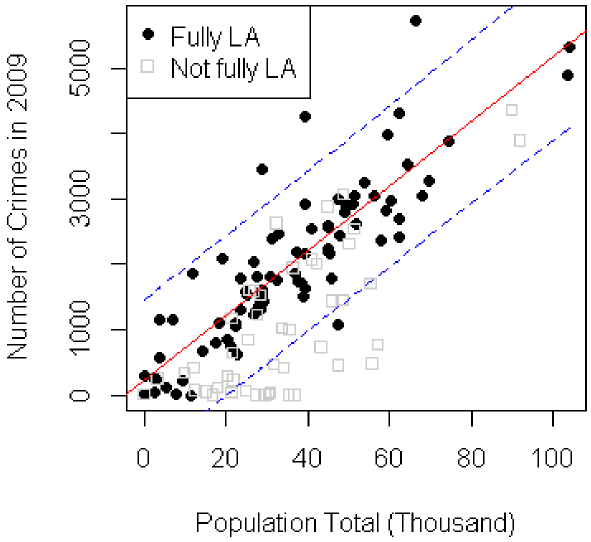


Figure 4.5 (left): Histogram of Total Crimes per ZIP Code (Fully in Los Angeles)

Figure 4.6 (right): Linear Model of Crime from Population Total (Fully in Los Angeles)

4.2 Density, Unemployment, and Poverty

Population density, unemployment rate, and percent of people with incomes below poverty are obtained from USZip.com for ZIP codes that are fully in Los Angeles. According to the site, the ZIP codes corresponding to CSUN and USC both had 100% poverty rate while the ZIP code corresponding to UCLA had 0% unemployment and poverty rate. The five area codes related to the universities and LAX were removed due to the unique nature of the locations. The correlations between density, unemployment rate, and poverty rate with the number of crimes in 2009 are 0.3218, 0.2934, and 0.4874 respectively. However, density, unemployment rate, and poverty rate are also correlated (Figure 4.7), which means that one variable may mask another when considering a regression model. Although unemployment is not significant in the full model, alone it is able to predict the number of incidences in one ZIP code (Table 4.3).

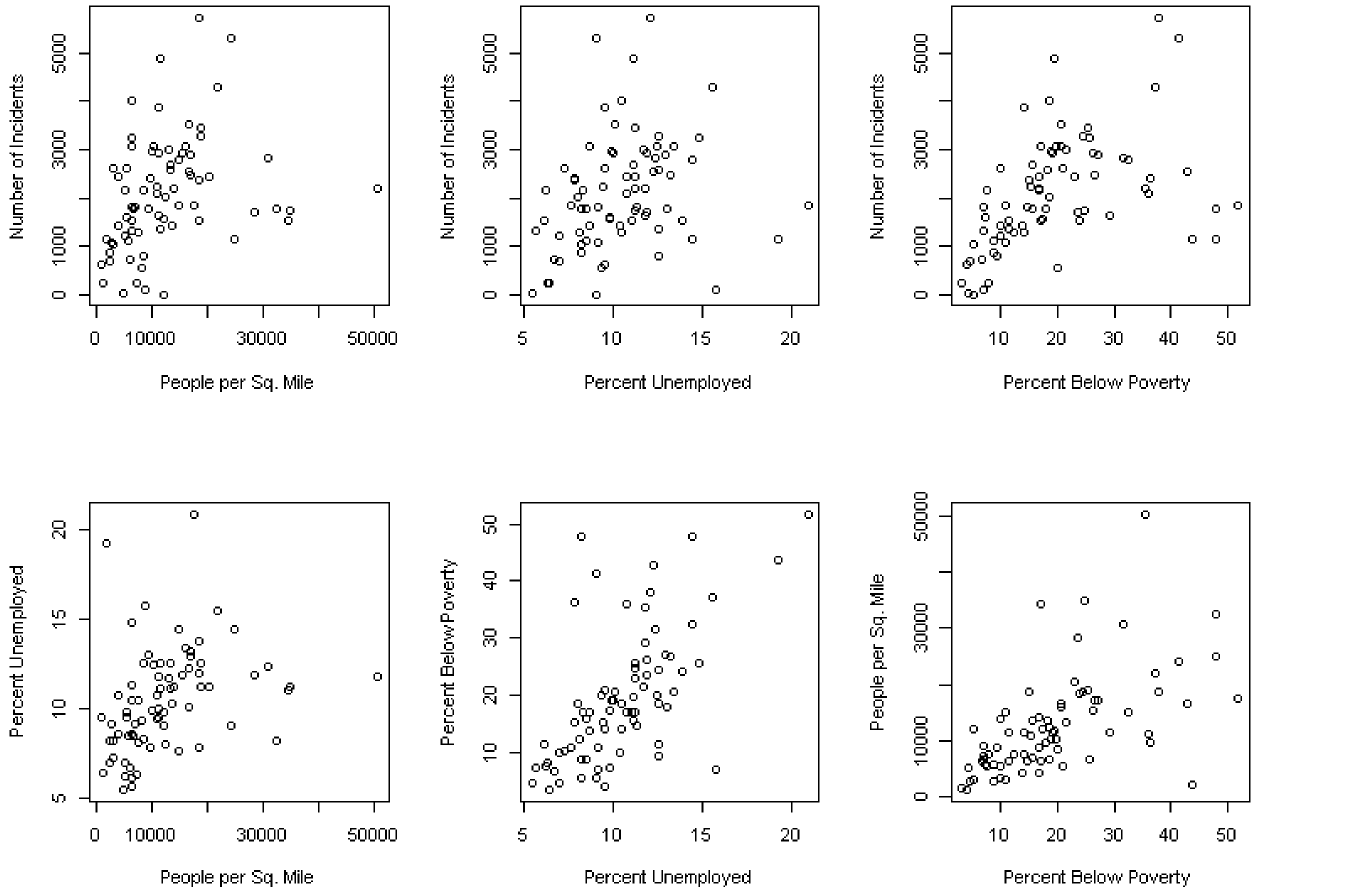


Figure 4.7: Scatterplot for Density, Unemployment, and Poverty

Table 4.3: Regression Models for Number of Crimes in 2009

	Full Model	**Reduced Model**	**Unemployment Only**
Intercept	1.83		1037.19 (*)
Unemployment	-15.76		100.26 (*)
Poverty	38.54 (***)	33.98 (***)	
Density	-0.02 (**)	-0.02 (**)	
Population Total	46.75 (***)	45.45 (***)	
	0.8692	0.9674	0.0618

However, we are also interested in indentifying situations and places of higher crime risk. One approach is to identify places with high incident to square mile ratio. The number of square miles contained in each ZIP code is calculated by dividing the total population by the density. There are five areas (90013, 90014, 90017, 90028, and 90057) with more than 2000 crimes per square mile and four areas (90056, 90077, 90212, 90272) with less than 100 crimes per square mile (Figure 4.8). The greatest contrast between low crime areas and high crime areas is the percent of people with incomes below poverty. Three of the five areas with more than 2000 crimes per square mile have the highest percent of poverty, all of which are almost 50%. About 25.5% of the 90028 population and 35.5% of the 90057 population have incomes below poverty. In contrast, the four areas with less than 100 crimes per square mile make up four of the five lowest poverty areas, the highest being 5.4%. Table 4.4 gives a summary that shows the contrast between the low crime areas and high crime areas, which confirms our findings in Table 4.3.

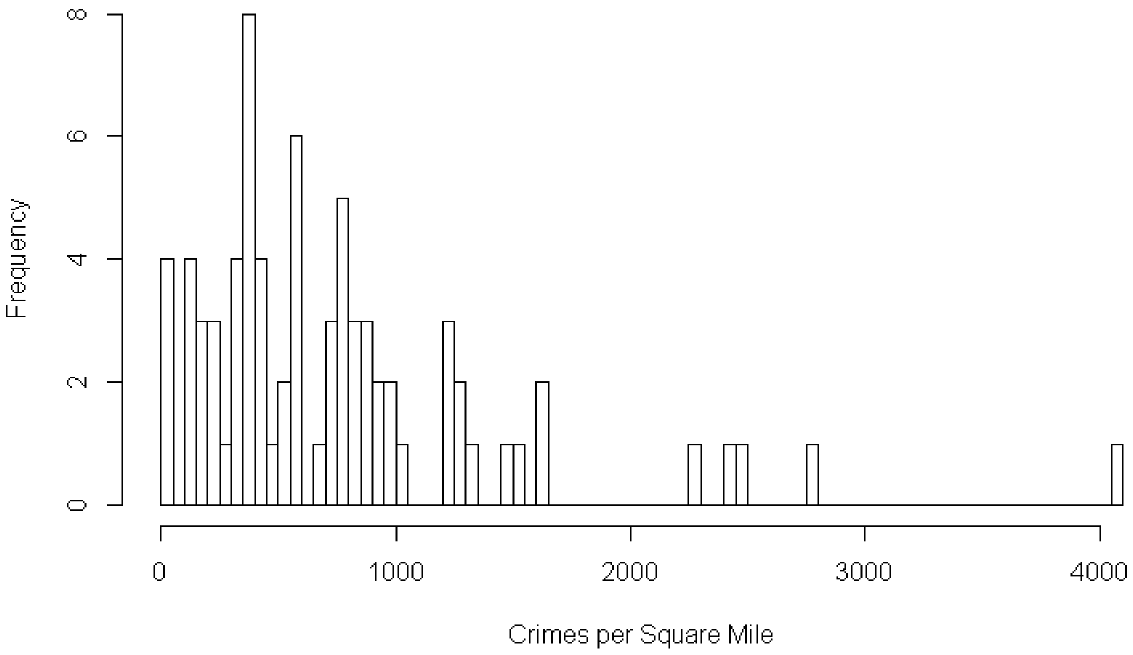


Figure 4.8: Histogram of Crimes per Square Mile

Table 4.4: Comparison of High Crime Areas and Low Crime Areas

	ZIP Code	Percent Poverty	Percent Unemployed	Density
High Crime Area	**90013**	51.8	20.9	17,570.15
	90014	47.9	14.4	25,017.86
	90017	47.9	8.2	32,558.90
	90028	25.5	11.2	18,890.79
	90057	35.5	11.8	50,559.55
Low Crime Area	**90056**	4.5	5.4	4,953.80
	90077	3.3	6.4	1,279.26
	90212	5.4	9.0	12,036.46
	90272	4.1	9.5	1,006.83

Another method for identifying high risk crime regions is to look at the number of crimes per thousand people. A high number of crimes per thousand people would indicate high risk. Figure 4.9 gives the histogram of crimes per thousand people. ZIP codes with over 100 crimes per thousand people include 90010, 90013, 90014, 90517, 90021, and 90028. These areas, like in the case of crimes per square mile, tend to have a high percent of people with incomes below poverty, as well as a high percent of unemployed and high population density. 90021 is the area with almost 300 crimes per thousand people. While the high number of crimes per thousand people may be due to how the population size is only 3,951, USZip.com tells us that more than half the households have an income of less than $15,000 per year. In contrast to high crime count per thousand people, the opposite applies for areas with less than 30 crimes per thousand people (90056, 90077, 90094, 90212, 90272). It is not surprising that these areas are the same places as the crimes per square mile case since as seen in Figure 4.7 ZIP codes with a higher percent of the population unemployed or in poverty tend to be denser areas.

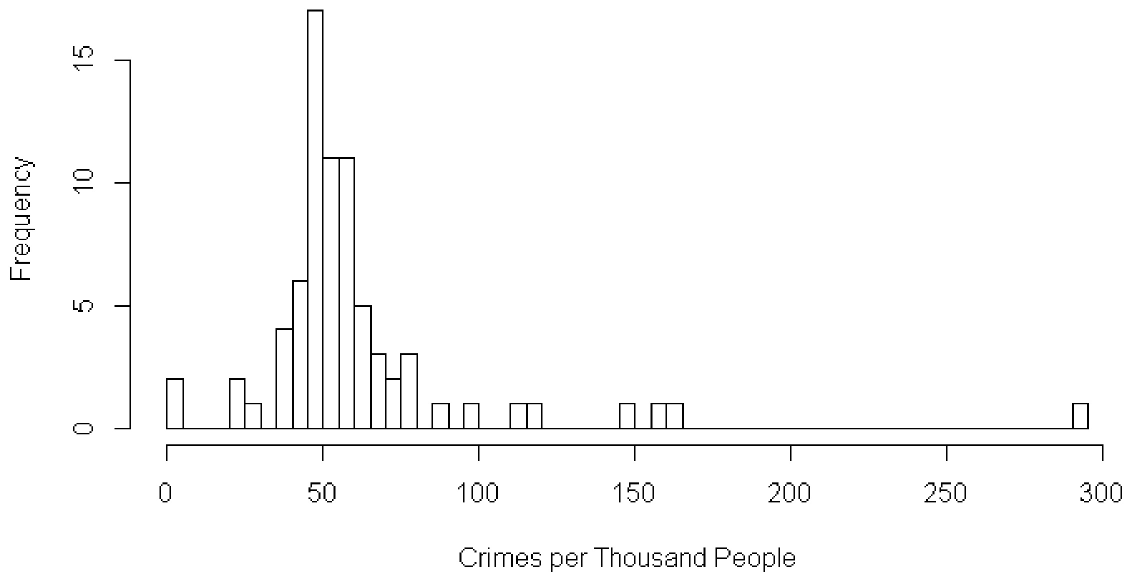


Figure 4.9: Histogram of Crimes per Thousand People

4.3 Gender

Dividing gender-related cases gives similar trends between population and crime count in 2009 to that of Figure 4.6. In the data set, there are a total of 108,931 male victims and 81,671 female victims known. For every 1,000 females in one ZIP code, there is an estimated 37.2 female victims. This estimate is higher for males, reaching about 49 victims per 1,000 males. These values are underestimates due to unknown gendered victims. At least 85% of variation in the data is explained by the two linear regression models. A similar estimate can be obtained by dividing the total number of crimes by the total population, though this method gives an underestimate from counting extra people outside of LAPD's jurisdiction. The estimates are 34.4 female victims per thousand females and 46.6 male victims per thousand males. A Chi-Squared test for the difference in proportion shows that it is not by chance that males are more likely to be victims than females.

Table 4.5: Victims by Crime and Gender

	Male	**Female**	**Total**	**Proportion Male**	**P-Value**
AGG	6985	2813	9,798	0.7129	< 2.2e -16
ARSON	235	107	342	0.6871	1.27 e -12
BTFV	14334	11520	25,854	0.5544	< 2.2e -16
BURG	10220	7973	18,193	0.5618	< 2.2e -16
GTA	534	231	765	0.6980	< 2.2e -16
GTP	392	1041	1,433	0.2736	< 2.2e -16
HOM	271	40	311	0.8714	< 2.2e -16
KID	136	348	484	0.2810	< 2.2e -16
NON	89416	43329	132,745	0.6736	< 2.2e -16
OTH	30160	32441	62,601	0.4818	1.32 e -13
RCVD	377	105	482	0.7822	< 2.2e -16
ROBB	8574	3297	11,871	0.7223	< 2.2e -16
THEFT	15959	10071	26,030	0.6131	< 2.2e -16
VAND	12740	9544	22,284	0.5717	< 2.2e -16

Certain genders are more likely to be victims of different crime types. Table 4.5 separates the number of incidences for each crime type by gender. Incidences are double counted if multiple crime types are listed, though results do not change by much. In total, there are slightly fewer males (0.4964 in proportion), though many of the crimes have a higher proportion of male victims. The highest proportion of male victims is among homicide cases. On the other hand, despite males being more likely victims of crime in general, females are more likely to be victims of grand theft property, kidnapping, and *other*.

4.4 Ethnicity

A similar analysis can be done for ethnic background. Table 4.6 reveals a significant difference in the proportion of victims between White, Black, Asian, and other. There is also a significant difference in proportion of Hispanic and non-Hispanic victim. We assume that the number of Hispanic and non-Hispanic victims is proportionally representative in the crime data set, though a look at the numbers in Table 4.1 is odd, so no further analysis is done regarding the *Hispanic-origin* variable.

Table 4.6: Difference in Proportion Test for Victims According to Background

	χ^2	**P-Value**
Male vs. Female	4484.29	< 2.2e-16
White, Black, Asian, Other	73178.58	< 2.2e-16
Hispanic vs. Non-Hispanic	7682.13	< 2.2e-16

With the exception of *other*, people of White descent make up the highest number of victims though people of Black descent are more likely to be victims (Table 4.7). For every 1,000 ethnic Blacks, there are about 113.2 Black victims in 2009. For every 1,000 ethnic Whites, there are about 36.4 victims. In the case of ethnic Asians, there are 16.7 victims for every 1,000 people. These values are acceptable estimates since overestimation from including incidences without valid ZIP codes approximately balances out with underestimation from including populations may that are most likely out of the LAPD's jurisdiction. For the case of other ethnicities, there are 72.3 victims per 1,000 people though this may be an overestimate since victims with unknown descent information are also categorized under *other*.

Table 4.7: Victims by Crime and Ethnicity

	Victim Ethnicity Count					**Incident Per 1,000 People**			
	White	**Black**	**Asian**	**Other**	**Total**	**White**	**Black**	**Asian**	**Other**
AGG	1215	3038	137	481	4871	0.5142	6.6026	0.2537	0.3560
ARSON	76	57	7	82	222	0.0322	0.1239	0.0130	0.0607
BTFV	10349	3047	960	3277	17633	4.3802	6.6221	1.7775	2.4256
BURG	6698	2644	880	2473	12695	2.8349	5.7463	1.6294	1.8305
GTA	207	90	28	64	389	0.0876	0.1956	0.0518	0.0474
GTP	238	267	64	114	683	0.1007	0.5803	0.1185	0.0844
HOM	16	107	3	11	137	0.0068	0.2325	0.0056	0.0081
KID	53	107	8	24	192	0.0224	0.2325	0.0148	0.0178
NON	35101	22741	3911	25306	87059	14.8564	49.4235	7.2414	18.7312
OTH	16882	12903	1540	6052	37377	7.1453	28.0424	2.8514	4.4796
RCVD	142	64	15	67	288	0.0601	0.1391	0.0278	0.0496
ROBB	1770	1852	411	1163	5196	0.7491	4.0250	0.7610	0.8608
THEFT	9897	3280	806	4460	18443	4.1889	7.1285	1.4923	3.3012
VAND	7074	3654	608	3114	14450	2.9941	7.9413	1.1257	2.3049

Across the different crime types, people of Black descent are more likely to be victims than people of any other descent. Except for burglary/theft from vehicle and theft, ethnic Blacks are at least two times more likely than ethnic Whites to be victims of the different crime types. Asians are the least likely to be victims of crime, though they are more likely to be victims of grand theft auto cases than *other* ethnicities and more likely to be victims of grand theft property cases than ethnic Whites and *other* ethnicities. The low rate of crime targeting Asians is questionable.

Reasons for low crime rate targeting Asians may include how victim descent is recorded or geography. The LAPD data set originally record Black as one category and White as one category. However, Chinese, Cambodian, Filipino, Japanese, Korean, Laotian, Vietnamese, and Asian Indian, and Other Asian are all recorded as separate categories, implying difficulty in recording Asian victims. Thus a larger proportion of Asian victims may be unclear or may be recorded as "Unknown."

Areas shared between LAPD and other police organizations' jurisdiction may justify the low crime rate targeting Asian people. However, 28.6% of the population in the ten lowest crime ZIP codes (those with less than five recorded crimes) are Asian, while only 4.8% are black (Table 4.8). Asians have the lowest increase in crimes per 1,000 people after subtracting the population count of these ten areas from the total population, which implies the opposite of suspicions.

Crime rate varies for different races and genders. In 2009, the most targeted victims were males. Most victims were of White descent, though people of Black descent are more likely to be a victim of crime. Specific types of crimes tend to target certain ethnicities and genders. Asians are among the least targeted victims. Explanations for low crime rate targeting Asians may be found by exploring other geographical or sociological phenomenon.

Table 4.8: Victim Proportion Excluding ZIP Codes with 10 Lowest Crime Count

	White	**Black**	**Asian**	**Other**
10 ZIP	180,603	8,986	51,684	82,000
Population less 10 ZIP	2,182,077	451,139	488,407	1,269,007
Victims Per 1,000	39.4	115.5	18.5	35.1

	Hispanic	**Non-Hispanic**	**Male**	**Female**
10 ZIP	153,036	170,237	158,843	164,430
Population less 10 ZIP	2,100,191	2,290,439	2,181,061	2,209,569
Victims Per 1,000	0.0332	0.0615	0.0499	0.0370

CHAPTER 5

Clustering

5.1 AGNES Clustering

When looking at crime, there are many factors, including temperature and demographics, which may or may not affect crime. Clustering can tie together weather and demographic data with the crime data provided by the LAPD to better help determine some of the outstanding factors that affect crime. AGNES clustering is a hierarchical method that is able to work with categorical variables [1]. The method works from a bottom-up approach in which all observations are initially their own clusters. At each step, the two closest clusters are merged. Closeness, in the analysis of the crime incidents, is defined by the Euclidean distance. The algorithm repeats until all observations are in one cluster.

AGNES clustering is applied to 3,397 randomly selected incidents in 2009 (1% of the data) based on primary crime type, day of the week, location, premise, number week of the year, whether a victim was involved, and the time of day (sectioned into six hour blocks: dark hours, morning, afternoon, and evening). Five clusters are formed when a cut is made around the height 55, where height is an indicator for similarity. The clusters are numbered as 1 to 5 from left to right. Cluster 1 is the most dissimilar cluster, but also the largest cluster with 2020 observations. Clusters 2 and 3 are more similar to each other than other groups. Clusters 4 and 5 are also more similar to each other than other groups. Clusters 4 and 5 make up the smallest clusters with 134 and 62 observations respectively. Table 5.1 and Figure 5.1 show summaries of the observations in the five clusters.

Half of the observations in Cluster 1 are non-aggravated assaults, and half of the victim ethnicities are *other*. These proportions are much smaller in the other four clusters. While all crimes are likely to occur during the daytime, the crimes found in Cluster 1 are more likely to occur in the later evening that crimes in other clusters. There is also a larger proportion of burglary/theft from vehicle crimes in Cluster 1 than the other four clusters. Correspondingly, there is a smaller proportion of *other* and theft (not from vehicle) crimes. In this sense, non-aggravated assaults are more similar to burglary/theft from vehicle crimes than to *other* or theft crimes. Cluster 1 crimes occur mostly in outdoor areas.

Comparing Clusters 2 and 3 to Clusters 4 and 5, but there are fewer female victims and consequently more male victims in proportion in last two clusters. Clusters 4 and 5 also have a larger proportion of theft and vandalism cases, which is consistent with findings in Table 4.7. That is, males are more likely to be victims of theft and vandalism. Cases in Clusters 2 and 3 happen more frequently in private dwellings in the mornings and afternoons. Cases in Clusters 4 and 5 happen more frequently in the evenings in public locations including stores and restaurants. Cluster findings hint at a relationship between time and victim gender. Table 5.2 shows the victim gender count and proportions according to 2 hour time blocks. Although a higher proportion of males are victims than females, the gap in proportion is smaller during the morning and early afternoon.

Table 5.1: AGNES Five-Cluster Summary

	Cluster 1		**Cluster 2**		**Cluster 3**		**Cluster 4**		**Cluster 5**	
AGG	62	*0.03*	21	*0.03*	6	*0.01*	5	*0.04*	0	*0*
ARSON	2	*<0.01*	2	*<0.01*	0	*0*	0	*0*	0	*0*
BTFV	271	*0.13*	21	*0.03*	11	*0.02*	5	*0.04*	1	*0.02*
BURG	4	*<0.01*	74	*0.11*	79	*0.15*	9	*0.07*	9	*0.15*
GTA	196	*0.10*	1	*<0.01*	1	*<0.01*	1	*0.01*	1	*0.02*
GTP	15	*0.01*	3	*<0.01*	0	*0*	1	*0.01*	1	*0.02*
HOM	1	*<0.01*	0	*0*	1	*<0.01*	0	*0*	0	*0*
KID	2	*<0.01*	0	*0*	0	*0*	0	*0*	0	*0*
NON	1001	*0.50*	140	*0.21*	102	*0.20*	26	*0.19*	10	*0.16*
OTH	152	*0.08*	253	*0.38*	182	*0.36*	30	*0.22*	21	*0.34*
RCVD	39	*0.02*	0	*0*	1	*<0.01*	0	*0*	0	*0*
ROBB	100	*0.05*	17	*0.03*	13	*0.03*	6	*0.04*	1	*0.02*
THEFT	54	*0.03*	96	*0.14*	82	*0.16*	34	*0.25*	12	*0.19*
VAND	121	*0.06*	42	*0.06*	33	*0.06*	17	*0.13*	6	*0.10*
White	449	*0.27*	145	*0.22*	194	*0.39*	43	*0.33*	23	*0.40*
Black	315	*0.19*	174	*0.27*	45	*0.09*	14	*0.11*	5	*0.09*
Asian	50	*0.03*	22	*0.03*	6	*0.01*	8	*0.06*	1	*0.02*
Other	842	*0.51*	72	*0.11*	93	*0.19*	37	*0.28*	11	*0.19*
Sun	252	*0.12*	88	*0.13*	60	*0.12*	15	*0.11*	7	*0.11*
Mon	264	*0.13*	96	*0.14*	57	*0.11*	19	*0.14*	6	*0.10*
Tues	312	*0.15*	102	*0.15*	75	*0.15*	10	*0.07*	10	*0.16*
Wed	295	*0.15*	88	*0.13*	74	*0.14*	18	*0.13*	9	*0.15*
Thurs	283	*0.14*	95	*0.14*	84	*0.16*	20	*0.15*	13	*0.21*
Fri	300	*0.15*	111	*0.17*	102	*0.20*	24	*0.18*	12	*0.19*
Sat	314	*0.16*	90	*0.13*	59	*0.12*	28	*0.21*	5	*0.08*
Male	1076	*0.53*	318	*0.47*	260	*0.51*	90	*0.67*	40	*0.65*
Female	574	*0.28*	332	*0.50*	227	*0.44*	39	*0.29*	17	*0.27*
Rain	132	*0.07*	56	*0.08*	32	*0.06*	8	*0.06*	5	*0.08*
Premise	Street/Pkwy		1-Family Home		1-Family Home		Business (Store)		Business (Store)	
	1340	*0.73*	252	*0.39*	220	*0.45*	80	*0.62*	38	*0.64*
	Parking Lot		Multi-Unit		Multi-Unit		Transport Fac.		Restaurant	
	163	*0.09*	167	*0.26*	97	*0.20*	17	*0.13*	7	*0.12*
	Sidewalk		Police Facility		Police Facility		Restaurant		Transport Fac.	
	133	*0.07*	26	*0.04*	22	*0.04*	14	*0.11*	3	*0.05*

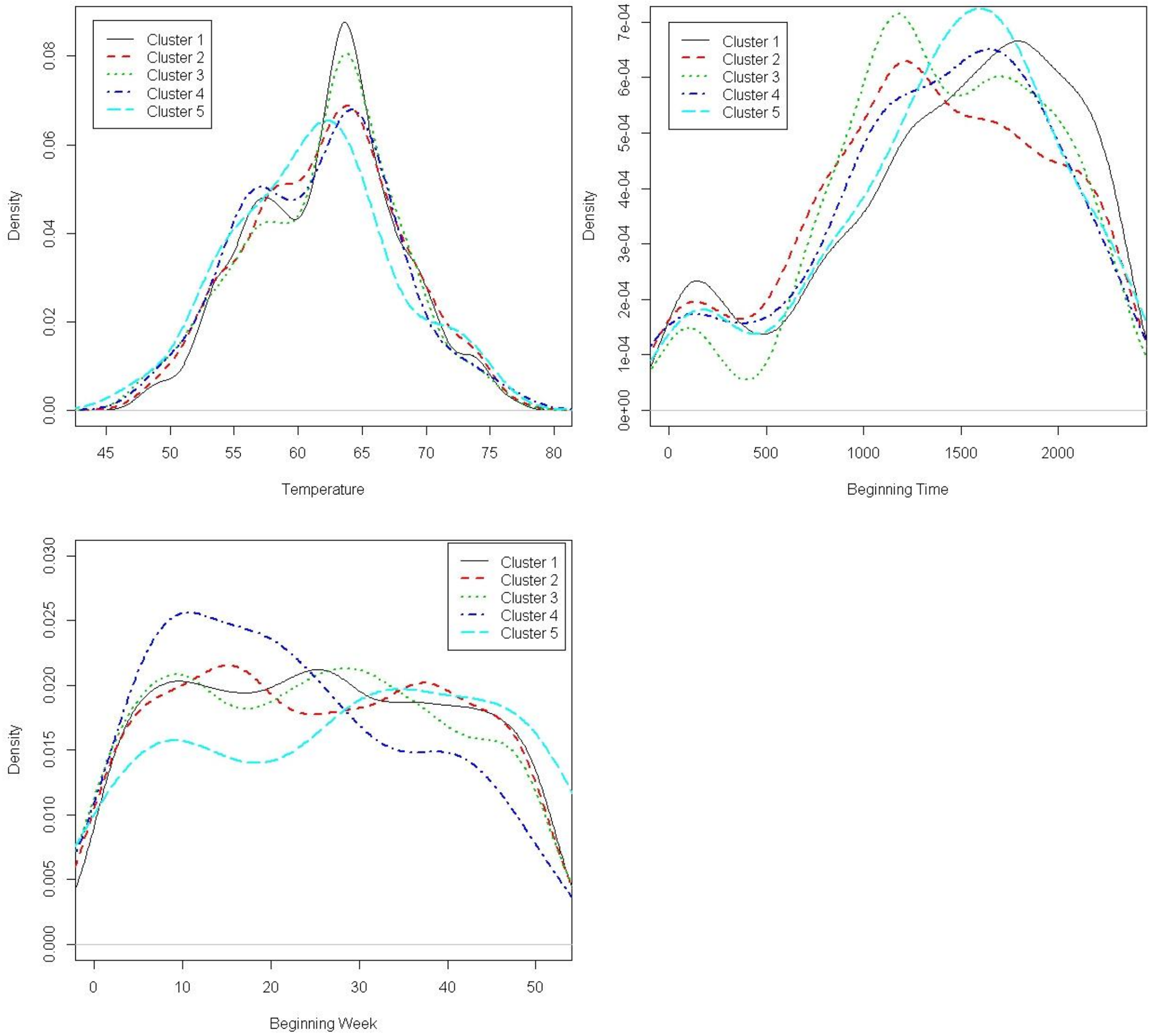


Figure 5.1: AGNES Five-Cluster Density Characteristics

Table 5.2: Victim Gender in Two-Hour Time Blocks

	Females	Males	Prop. Females	Prop. Males
Midnight-2 am	6724	11936	0.3603	0.6397
2-4 am	3906	7510	0.3422	0.6578
4-6 am	1998	3763	0.3468	0.6532
6-8 am	5536	7023	0.4408	0.5592
8-10 am	11142	13793	0.4468	0.5532
10-noon	11454	15196	0.4298	0.5702
Noon-2pm	16238	21605	0.4291	0.5709
2-4 pm	13384	19988	0.4011	0.5989
4-6 pm	13236	21178	0.3846	0.6154
6-8 pm	13446	21780	0.3817	0.6183
8-10 pm	12002	19767	0.3778	0.6222
10-midnight	9953	17034	0.3688	0.6312

Clusters 2 and 3 are very similar. It turns out that the two clusters are separated primarily based on ZIP code. Interestingly, while the two clusters are not built based on victim ethnicity, there is a larger proportion of White victims and a smaller proportion of Black victims in Cluster 3 than Cluster 2. This confirms the tendency for difference ethnicities to live in separate communities as seen in the U.S. Census data set. Clusters 4 and 5 are also separated based on location.

A second run of AGNES clustering on a different sample of 3,397 incidents yields similar results, reflecting consistency of the algorithm on the data set.

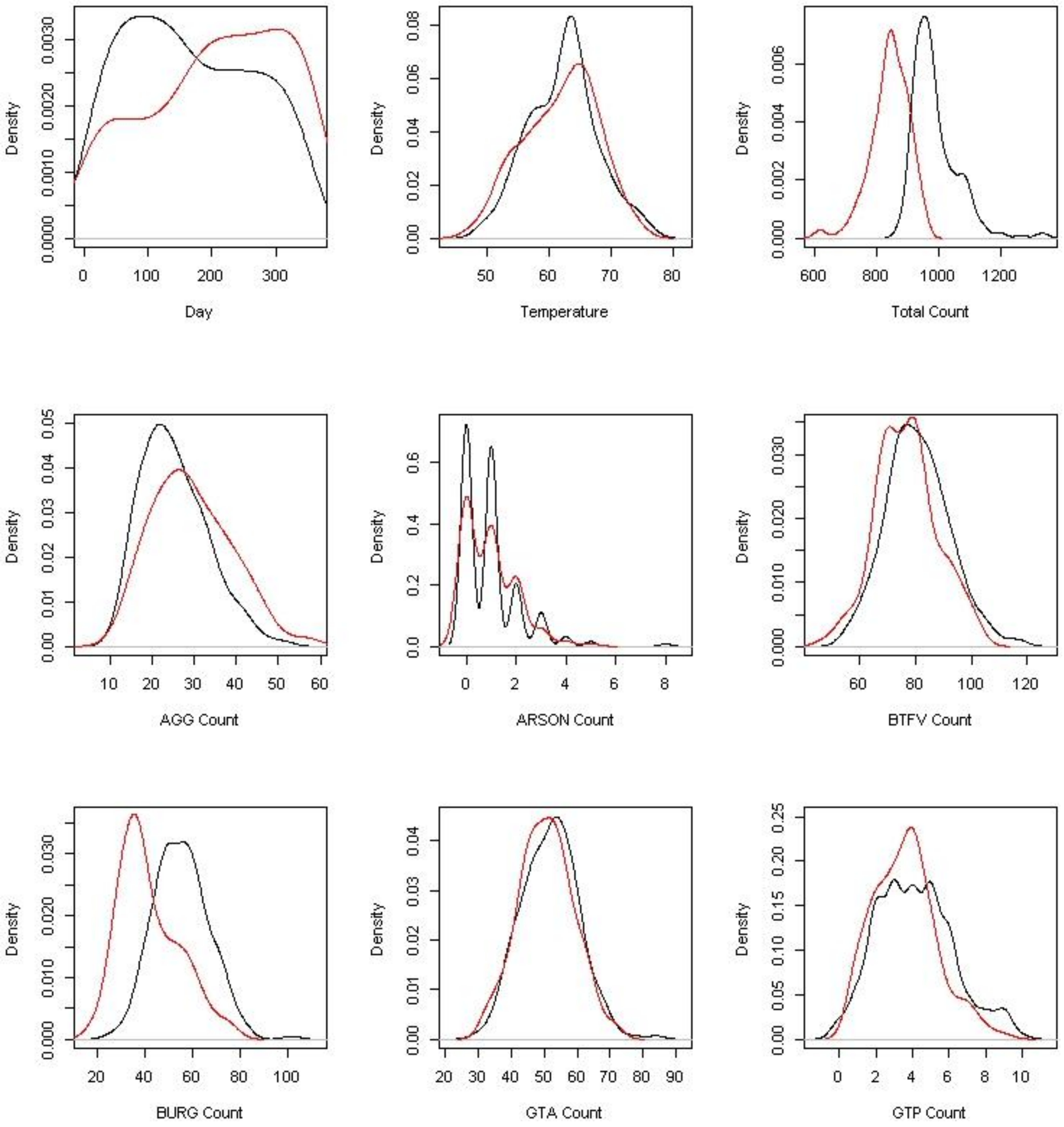
5.2 *k*-Means Clustering

k-means is a popular clustering method that picks *k* centers within the data set and assigns points to the closest center, forming clusters. Again, distance is defined by Euclidean distance. The cluster centers are recalculated and observations are reassigned a number of times or until convergence. Because the method is quick, sampling a fraction of the data is not necessary. The *k*-means method is applied to daily weather data and crime counts in 2009, as well as to demographic background across ZIP codes. Unlike AGNES clustering, a number of clusters must be specified beforehand.

Applying *k*-means clustering with *k* = 2 to daily temperatures and crime count yields clusters of size 222 and 143. The first cluster (Cluster 1) contains mostly days in the first half of the year while the smaller cluster (Cluster 2) contains days in the second half of the year (Figure 5.2) despite how the date is not a factor used for clustering. Since temperature and precipitation result in similar distributions in both clusters, weather is not a factor leading to separation by the first half and second half of the year. The formation of the two clusters is instead due to the drop in number of incidents on specific days of the week and towards the end of the year. Cluster 2 is made up predominantly of weekends and Monday (Table 5.3). The few weekdays found in Cluster 2 are mostly holiday-related, such as the last two weeks of the year. On the other hand, the few weekends found in Cluster 1 are generally at the beginning of the year or the first of the month. As mentioned in Chapter 3, the numbers for the first of the month may be inflated. There are fewer burglaries, non-aggressive assaults, and theft in Cluster 2 than Cluster 1. The average difference in non-aggressive assaults between Clusters 1 and 2 is more than 100. Consequently, Cluster 1 also has a higher victim count. The two clusters tell us that overall, there are fewer crimes towards the end of the year, on weekends, and on holidays.

Table 5.3: Day of Week for 2 Clusters from *k*-Means Clustering

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Cluster 1	49	31	15	7	5	3	33
Cluster 2	3	21	37	45	48	49	19



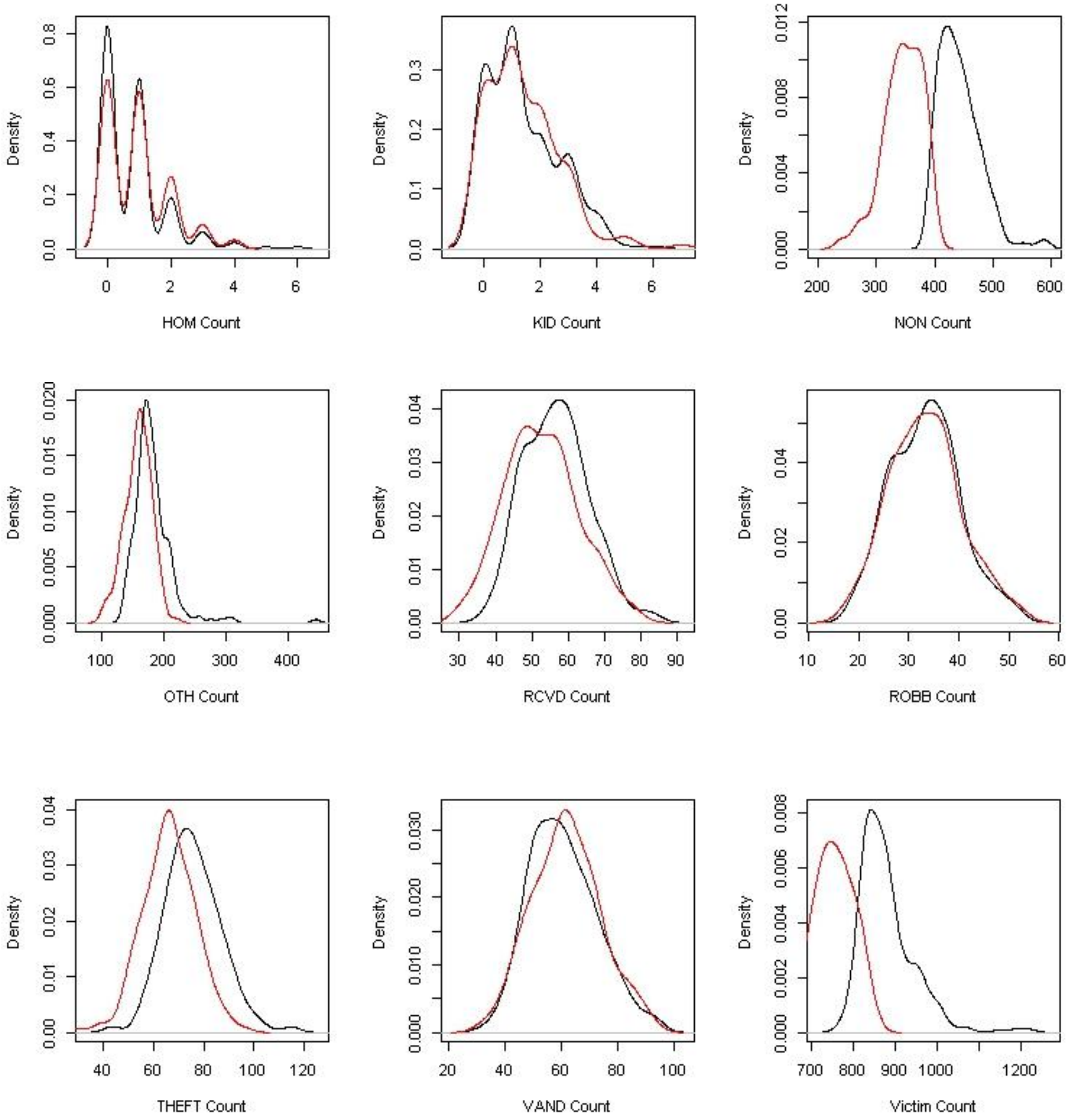


Figure 5.2: Comparison of Clusters in *k*-Means Clustering with *k* = 2

The findings suggest a decrease in crime over the year. A linear regression model estimates about 982 crimes at the beginning of the year and a decrease of about 0.28 crimes per day. At the end of the year, there is an estimated 879 crimes a day, which means about a 10% decrease in crimes over the year. However, the linear regression model is not plausible for long-term projection as the model predicts -49 crimes per day at beginning of 2019. Poisson regression gives a more plausible (though not perfect) model, in which the beginning of 2019 will still have 325 crimes per day. The model is given as:

$$y = e^{6.891+0.0003035x}$$

where y is the number of crimes per day and x is the number of days after December 31, 2008.

We apply *k*-means clustering to the climate data set, again. *k*-means clustering can start with either randomly selected centers or specific centers. In this case, to identify any uniqueness on rainy days, we start with all sunny days in one cluster and all rainy days in another. The algorithm converges in one iteration, where there are 138 days in the first cluster and 227 days in the other. Thirteen rainy days are in the smaller cluster while twelve rainy days are in the other. This suggests that there is no difference (p = 0.1927) in crime count in general on rainy days, or rather there are not enough rainy days draw any meaningful conclusions. A basic summary of the two clusters can be seen in Figure 5.3. The summary is very similar to that of *k*-means with randomly selected centers (Figure 5.2).

Increasing the number of clusters to 5 and reapplying *k*-means clustering with random centers to daily temperature and crime count, gives clusters of size 8, 52, 76, 125, and 104 (Figure 5.4). Call these Clusters 1 through 5. We overlook Cluster 1 since the cluster is made up of the first of eight different months. As in Chapter 3, we assume that the first is the default day of the month recorded if the exact date is unknown.

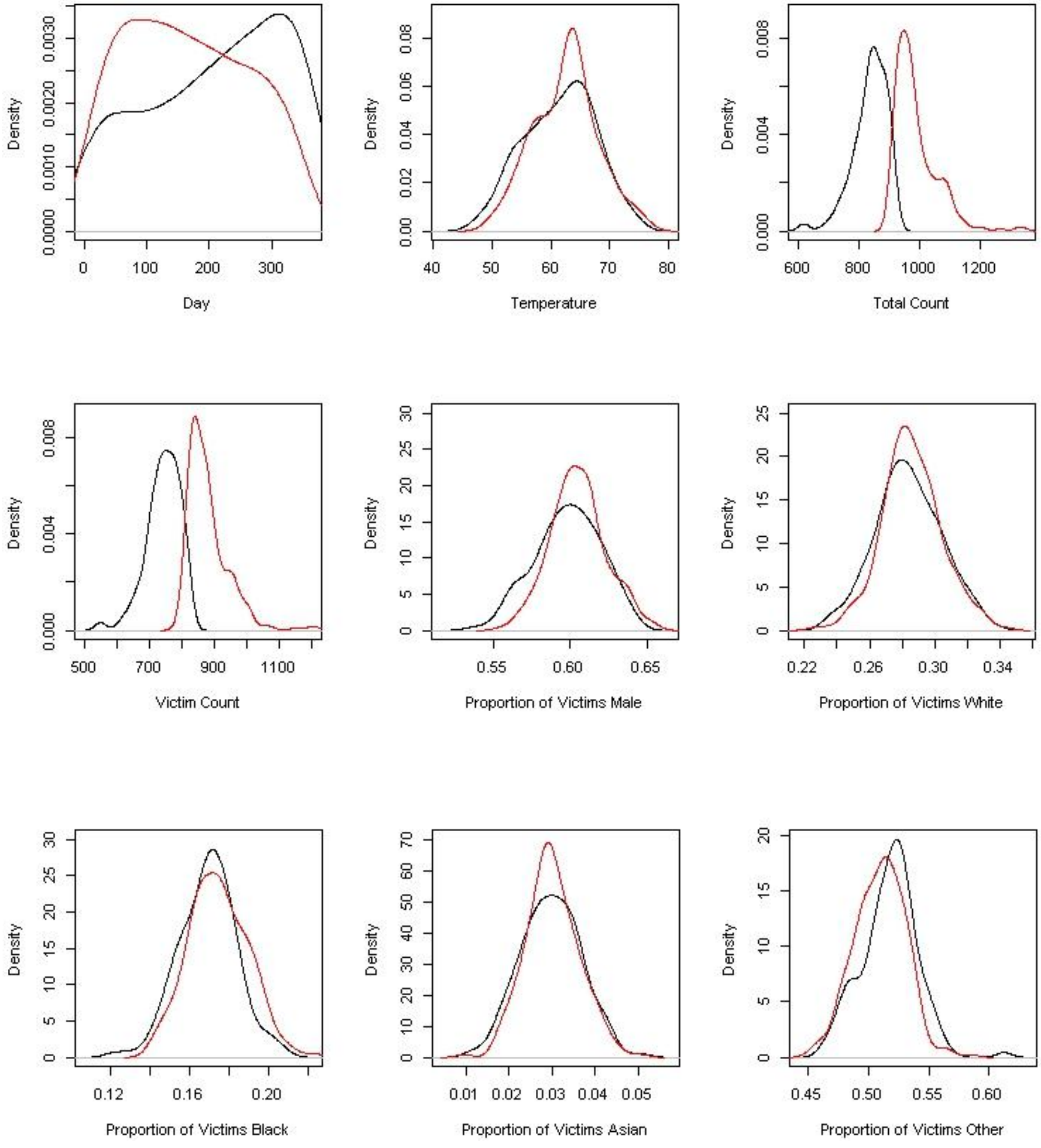
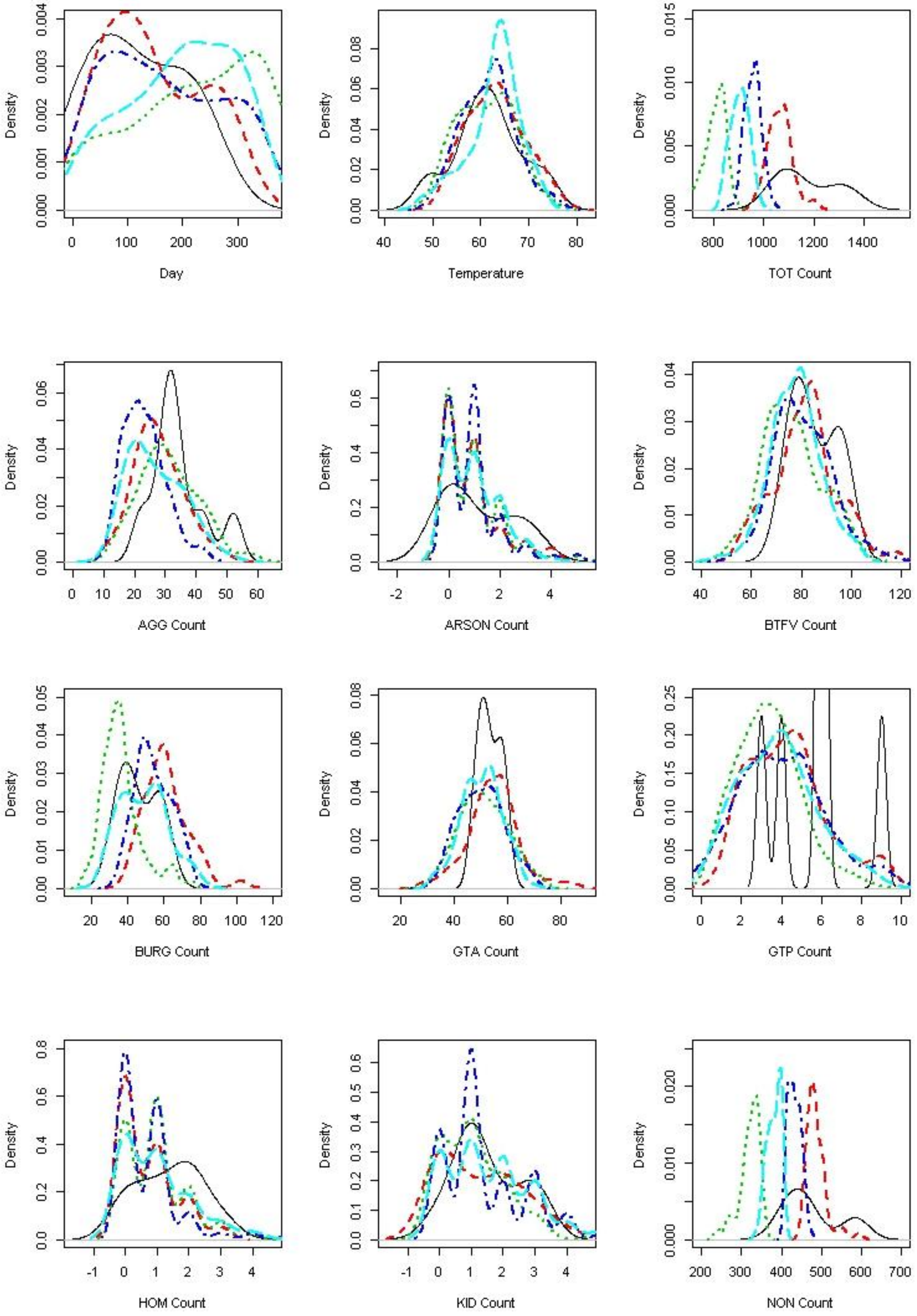


Figure 5.3: Clusters in *k*-Means Clustering Starting with Rain vs. No Rain Center


Density
0.000 0.001 0.002 0.003 0.004
0 100 200 300
Day
Density
0.00 0.02 0.04 0.06 0.08
40 50 60 70 80
Temperature
Density
0.000 0.005 0.010 0.015
800 1000 1200 1400
TOT Count
Density
0.00 0.02 0.04 0.06
0 10 20 30 40 50 60
AGG Count
Density
0.0 0.2 0.4 0.6
-2 0 2 4
ARSON Count
Density
0.00 0.01 0.02 0.03 0.04
40 60 80 100 120
BTFV Count
Density
0.00 0.01 0.02 0.03 0.04 0.05
20 40 60 80 100 120
BURG Count
Density
0.00 0.02 0.04 0.06 0.08
20 40 60 80
GTA Count
Density
0.00 0.05 0.10 0.15 0.20 0.25
0 2 4 6 8 10
GTP Count
Density
0.0 0.2 0.4 0.6 0.8
-1 0 1 2 3 4
HOM Count
Density
0.0 0.1 0.2 0.3 0.4 0.5 0.6
-1 0 1 2 3 4
KID Count
Density
0.000 0.010 0.020
200 300 400 500 600 700
NON Count

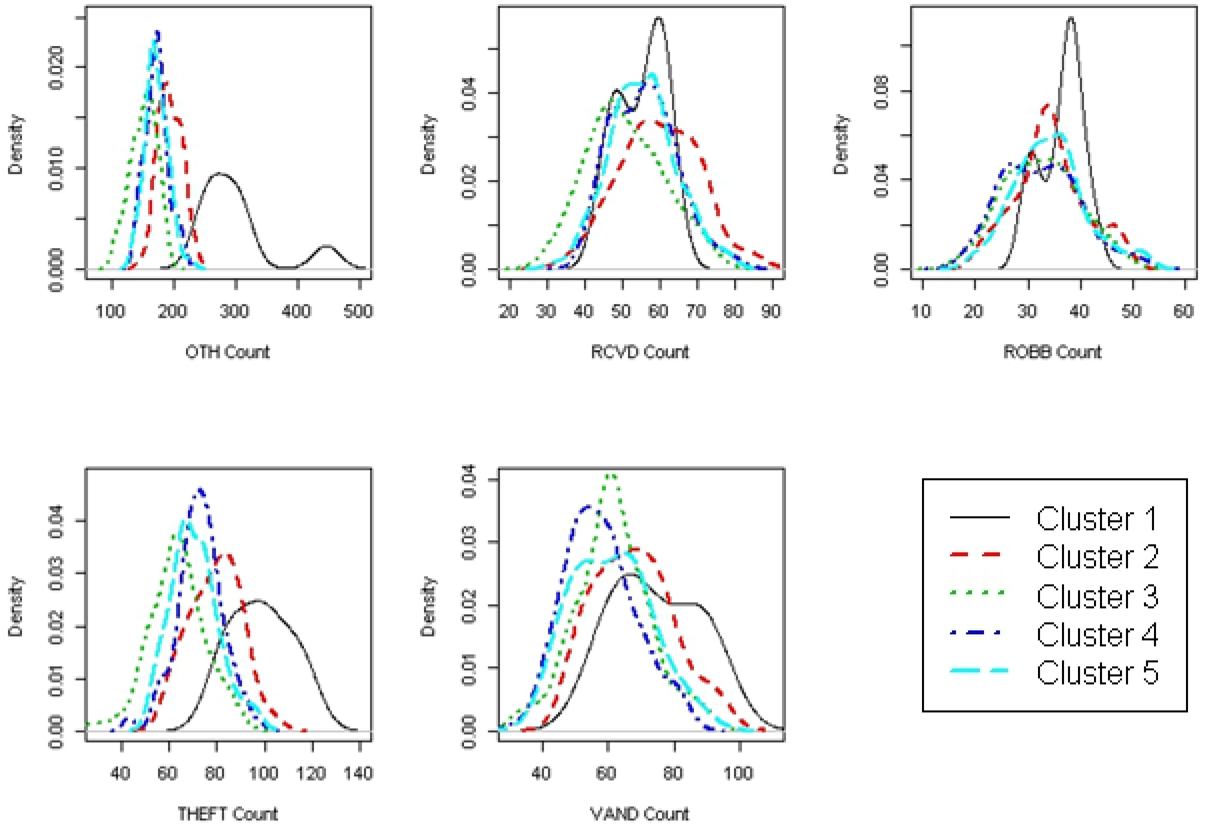


Figure 5.4: Cluster Summaries from *k*-Means Clustering with *k* = 5

Table 5.4: Day of Week for 5 Clusters from *k*-Means Clustering

	Sunday	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
Cluster 1	2	0	1	2	1	1	1
Cluster 2	0	2	1	7	10	31	1
Cluster 3	44	10	5	2	3	3	9
Cluster 4	0	10	24	29	36	17	9
Cluster 5	6	30	21	12	3	0	32

A look at the days of the year and days of the week of each of the five different clusters shows some seasonal and weekly patterns (Table 5.4). Cluster 2 shows that springs and Fridays in 2009 tend to have high crime counts compared to the rest of the year. More specifically, Cluster 2 has a high number of non-aggravated assaults and burglaries. Cluster 4 follows closely behind Cluster 2. Both cover most of spring, though Cluster 4 tends to have fewer crimes in total due to fewer burglaries, non-aggravated assaults, and thefts per day than Cluster 2, though still more than the other clusters. While Fridays make up most of Cluster 2, Cluster 4 consists primarily the other days in the middle of the week.

Cluster 3 is made up of Sundays and the end of the year, namely the days near Christmas. As found, Sundays and days near Christmas and Thanksgiving have the least number of crimes. There are fewer burglaries, non-aggravated assaults, and thefts. Cluster 5 is made up mainly of Mondays and the summer. The crime count for days found in Cluster 5 is fewer on average than those in Cluster 4 but more than those in Cluster 3. Mondays and summer days follow a pattern that falls in between weekends and weekdays. This is generally true for burglaries, non-aggravated assaults, and thefts. As for other crime types, *k*-means clustering in this run is unable to find a distinct trend. Holidays, the day of the week, and the time of the year are all important predictor for the number of each crime type that will occur in a given day.

Figure 5.5 shows cluster results of *k*-means clustering with $k = 2$ on demographic background of various ZIP codes. The two clusters are divided mostly by population size, where the dividing point is at roughly 40 thousand people in population. In highly populated areas, there is a lower White population proportion, but a higher Hispanic population proportion. In low populated areas, it is the reverse. Low populated areas have a higher proportion of burglary/theft from vehicle, burglaries, and thefts, but a slightly lower proportion of arsons, grand theft autos, kidnappings, receipt of stolen goods, and robberies. Places with a larger Hispanic population have a wider range in population size and therefore a wider range in number of crimes. Figure 5.6 shows similar findings when clustering into $k = 5$ groups, where the clearest distinction is in population size.

5.3 Cluster Findings

From cluster findings, an estimate of the number of each type of crime can be obtained by looking at the day of the week, season, and whether the day is a holiday. To estimate victims count by gender and descent, time of day may be a helpful predictor. Temperature distribution is approximately the same between clusters, suggesting that the temperature as recorded by the weather station at LAX is not a good predictor for the number of crimes in a day. However, this does not mean that temperature is not a good predictor for crime rate, but rather geographical variations in weather may not have been captured with the given dataset. Rain, which may be more consistent than temperature across locations may not be consistent over the course of a day. Furthermore rain only accounts for less than 10% of the days in a year, making it difficult to draw a definite conclusion.

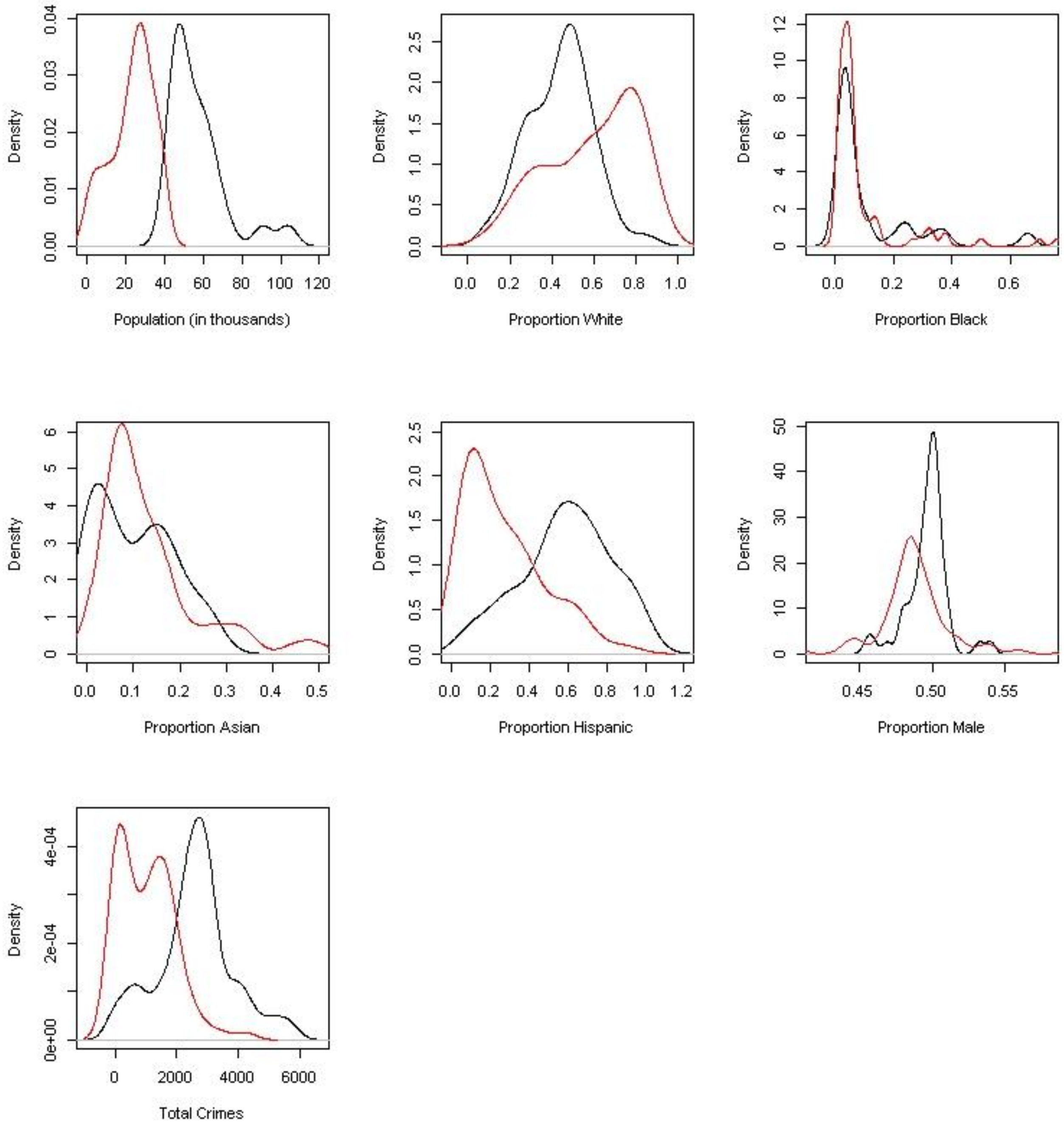


Figure 5.5: *k*-Means Clustering with *k* = 2 on Demographic Background

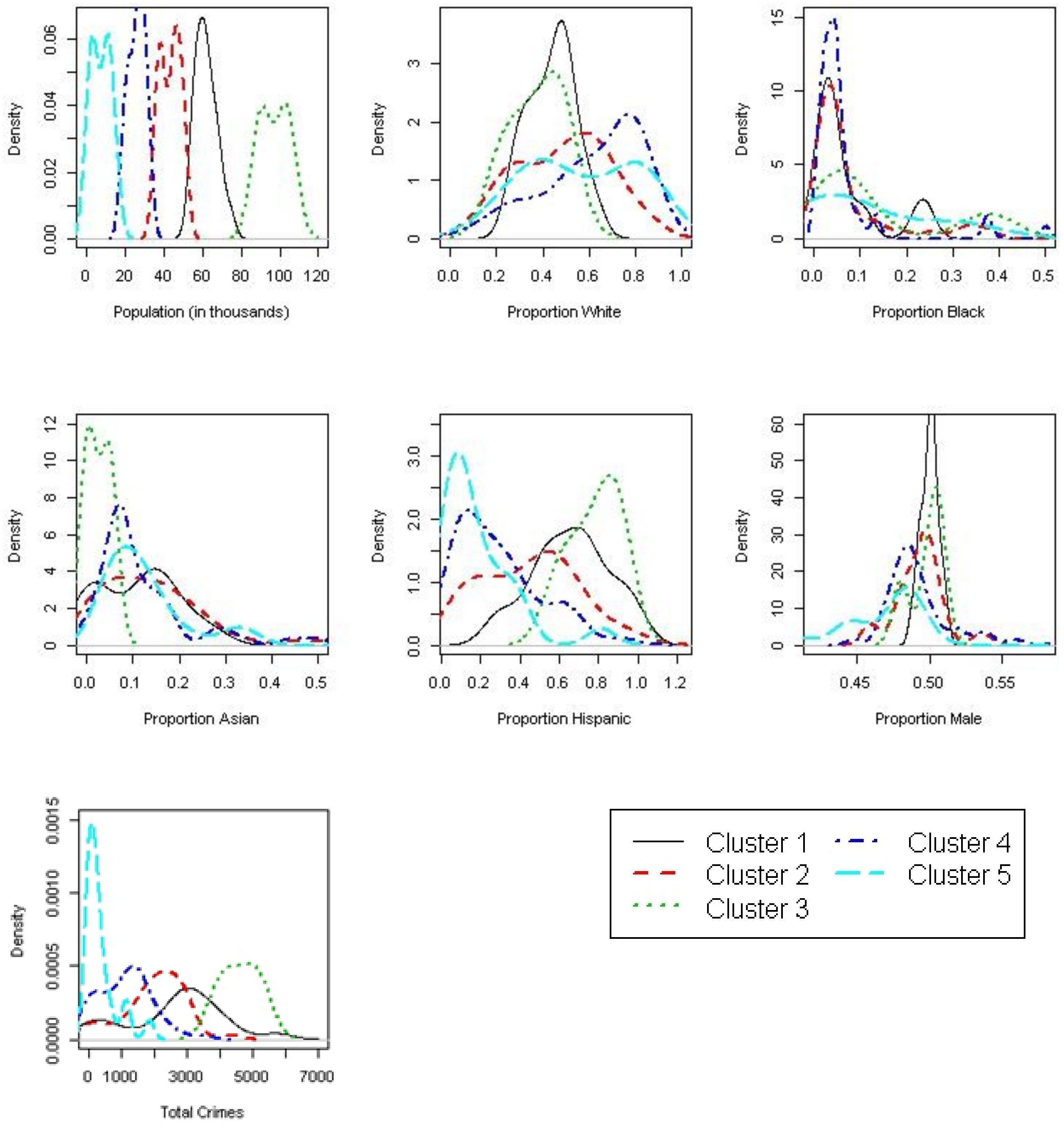


Figure 5.6: *k*-Means Clustering with *k* = 5 on Demographic Background

CHAPTER 6

Concluding Remarks

6.1 Conclusion

Despite how different crime types have varying trends, the day of the week is an excellent predictor of crime count while gender and ethnicity could be used as precursors to determine victims of specific crimes. Aggravated assaults, for example, tend to happen more on weekends. Males and people of black descent are more likely to be victims of aggravated assaults. Many crimes have a larger proportion of male victims. On the other hand, females are more likely to be victims of kidnappings and grand theft property, which occur more on weekdays.

Demographics are also involved in crime patterns, as certain places see higher rates of crime. Across the board, people who are ethnically Black are more likely to be victims of crimes while people who are ethnically Asians are least likely to be victims of crimes. Population size is an important factor for the number of crimes that occur in an area. The more populous the ZIP code, the more crimes there are. However, defining high crime rate in terms of crimes per square mile or crimes per thousand people, we find that poverty rate is an important predictor. Places with high percents of people with income below poverty level have a higher crime rate. Correspondingly, high crime areas have denser populations and higher percent of unemployed.

It is expected that crime rate will be high on Fridays at densely populated places where poverty rate is high. On the other hand, crime rate is expected to be low on days around Christmas and Thanksgiving. Monday holidays resemble Sundays. Sunday is expected to have the lowest crime rate compared to the rest of the week.

6.2 Future Studies

For most of this analysis, only 2009 data is explored as it is the closest year available to relate to the 2010 Census data. However, as annual patterns were found for holidays in 2005 to 2009, future studies may want to look at seasonal patterns. *k*-means clustering with five centers (Figure 5.4) showed had one cluster which consisted of Mondays and summer days. It is suspected that summer days will exhibit a different crime trend from the rest of the year despite how temperature as reported by the LAX weather station is not able to capture crime trends.

Similar to seasons, temperature may still be an important predictor of crime. Further studies may want to consider multiple weather stations to address geographical variations in temperature and rain. Temperature could be recalculated and weighted based on proximity to the weather station. Furthermore, weighted clustering may also be a key method for future analyses of crime and its relation to temperature.

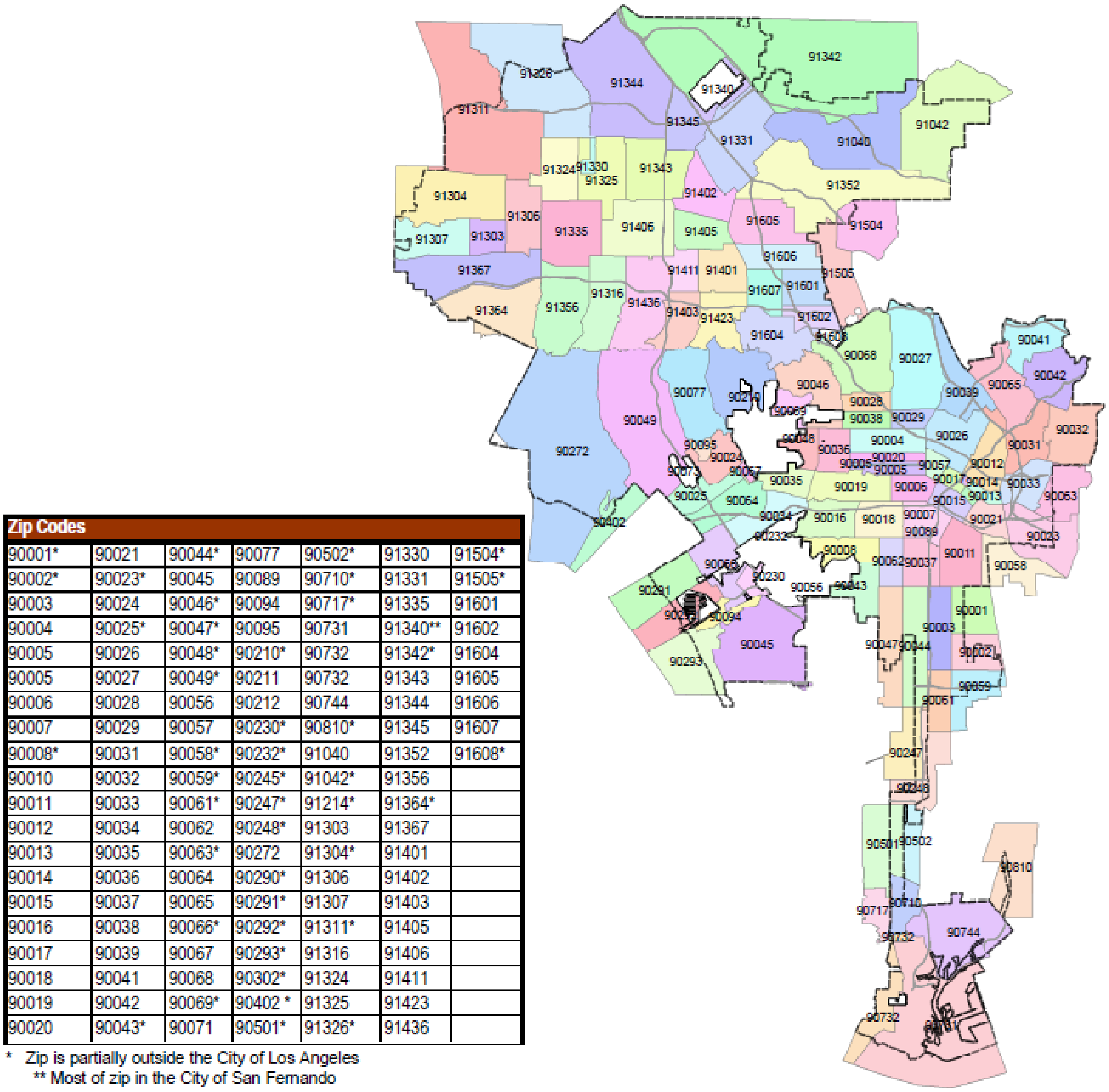
Since there are only a few rainy days per year in Los Angeles, one may want to examine crime on rainy days across multiple years or zoom in on rainy hours throughout the day. Since rain is not necessarily consistent throughout the day, it would be interesting to see how crime is affected while it’s raining and only a rainy day when it’s not pouring. Using multiple years can provide more crime rate information for rainy days so that a more concrete conclusion can be drawn. Multiple year data allows for a comparison of annual trends, which is especially important when considering holidays.

This paper looked mainly at ethnicity and gender for demographic background. However, the U.S. 2010 Census also has background information on the age distribution of different neighborhoods. Future studies may was to look at victims' ages and what age groups tend to be victims of certain crimes. Studies may also want to look at the relationship between age distribution and crime rate in a specific area. Although CSUN, UCLA, and USC locations were excluded from part of the analysis, age distribution may be able to capture college-towns and better explain crime in relation to age.

The scope of crime analyses extends far and wide. This thesis provides only a starting point to crime data analysis. Future work may want to include a deeper analysis of 2005-2008 or take a look at more recent years. ZIP codes worked well in breaking down the 21 LAPD districts, though it faces the problem of overlapping with places outside of the LAPD's jurisdiction. For part of the analyses, these overlapping ZIP codes were treated in the same manner as ZIP codes fully in Los Angeles. Otherwise, they were excluded from other analyses. Future studies ought to include consideration of distance and weighting for ZIP codes on the border of Los Angeles and for crime incidences that occur far from the weather station.

APPENDIX A

Map of Los Angeles



Zip Codes						
90001*	90021	90044*	90077	90502*	91330	91504*
90002*	90023*	90045	90089	90710*	91331	91505*
90003	90024	90046*	90094	90717*	91335	91601
90004	90025*	90047*	90095	90731	91340**	91602
90005	90026	90048*	90210*	90732	91342*	91604
90005	90027	90049*	90211	90732	91343	91605
90006	90028	90056	90212	90744	91344	91606
90007	90029	90057	90230*	90810*	91345	91607
90008*	90031	90058*	90232*	91040	91352	91608*
90010	90032	90059*	90245*	91042*	91356	
90011	90033	90061*	90247*	91214*	91364*	
90012	90034	90062	90248*	91303	91367	
90013	90035	90063*	90272	91304*	91401	
90014	90036	90064	90290*	91306	91402	
90015	90037	90065	90291*	91307	91403	
90016	90038	90066*	90292*	91311*	91405	
90017	90039	90067	90293*	91316	91406	
90018	90041	90068	90302*	91324	91411	
90019	90042	90069*	90402 *	91325	91423	
90020	90043*	90071	90501*	91326*	91436	

* Zip is partially outside the City of Los Angeles
** Most of zip in the City of San Fernando

* Map from Los Angeles Housing Department [5]

REFERENCES

[1] Kaufman, L. and Rousseeuw, P. J. (2009), “Agglomerative Nesting Program AGNES”, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley & Sons, New Jersey. 199-209.

[2] “The Mission Statement of the LAPD,” *Los Angeles Police Department*, available at http://www.lapdonline.org/inside_the_lapd/content_basic_view/844

[3] “Quality Controlled Local Climatological Data,” *National Oceanic and Atmospheric Administration*, available at http://cdo.ncdc.noaa.gov/qclcd/QCLCD

[4] U.S. 2010 Census Data available at http://www.census.gov/2010census/data/

[5] “Zip Codes Within the City of Los Angeles,” *Los Angeles Housing Department*, available at http://lahd.lacity.org/lahdinternet/Portals/0/Policy/LAZipCodes.pdf