Final Report

Group B

September 29, 2017

Contents

1 Introduction

1.1 Context

Los Angeles (L.A) is one of the most vibrant cities in the United States with a high level of economic activity. It is the 2nd most populous city in the US after New York city and is home to a wide range of business and professional interests. Criminal activity in L.A is often heavily scrutinised given its diverse population and importance as a cultural hub.

L.A is considered as one of the most dangerous cities in the US, with a high crime rate of 6.45 incidents per 1000 residents compared to the national median of 3.8 in 2015. According to Los Angeles Police Department (LAPD), crime rate in the city has fallen for 10 consecutive years from 2002 to 2012. However, in 2015, it was reported that for the first time in more than a decade, all categories of crime rose across the city (violent crime rose 20.2% and all crime was up 12.6% compared with 2014). The upward trend continued in 2016, with higher crime across four categories compared to 2015.

Officials have attributed the increase to a variety of factors, including increased gang activity and a rise in the number of homeless people. With a high chance of roughly 1 in 33 to become a victim of either violent or property crime in L.A, it is important to focus efforts on crime prevention and education.

In order to learn more about patterns of crime in L.A and to come up useful suggestions to aid efforts in crime prevention, this paper looks into three main questions of crime patterns in L.A; firstly, who is more likely to be a victim; secondly, where are the accident-prone areas in L.A and thirdly, when are crimes more likely to occur. The dataset used in this paper contains information of crime occurred in L.A since 2010, with weekly update by City of Los Angeles.

1.2 Theory

Studies of criminal patterns and activities have in the past belonged to the domain of psychology and sociology. However, with the rise of criminal activities and increased sophistication of criminal operations, there is a need for more targeted crime prevention efforts. In the past decades there have been increased efforts in data collection of criminal activites which have allowed for quantitative analysis to better predict incidence of crime, likely victims of crime and a myriad of other questions. This paper aims to adopt statistical tools and visualisation efforts to paint a clearer picture of criminal activities in L.A and provide some useful insights and policy recommendations based on the analysis done. Our analysis will also be useful for crime prevention efforts from the victims' point of view, as we aim to highlight demographic and location factors that will make one more susceptible to crime.

2 Data Description

The dataset contains information on crime incidents in Los Angeles. The source description (LAPD) states that the dataset may have some inaccuracies due to information being transferred from original crime statements that are typed on paper. As crime data contains sensitive personal information, fields that contain address information are only provided to the nearest hundred block. The dataset includes crime data from 2010 through September 2017.

Two separate csv files are provided with Crime_Data_2010_2017 containing the main data and MO_Codes containing the description of MO codes that are included in the crime data for classification. The dataset is made up of 1.584.316 observations and 26 variables. To structure the variables, we split them into four groups: demographic, geographic, crime and time type variables. Demographic category consists of the victims age, sex and descent. The type of variables and a detailed description can be found in the table below:

Demographics

Variable	Type	Description
Victim.Age	dbl	Two character numeric
Victim.Sex	fctr	F - Female M - Male X - Unknown
Victim.Descent	fctr	Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian

The geographical category contains variables relating to the geographical occurrence of crime and reporting of crimes. Variables include the area ID, the name of the area, the district reporting the crime, the premise code, a premise description, the address, the cross street of rounded address and the location where the crime occurred. The type of variables and a detailed description can be found in the table below:

Geographic

Variable	Type	Description
Area.ID	int	The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21.
Area.Name	fctr	The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. For example 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles.
Reporting.District	int	A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons.
Premise.Code	dbl	The type of structure, vehicle, or location where the crime took place.
Premise.Description	fctr	Defines the Premise Code provided
Address	fctr	Street address of crime incident rounded to the nearest hundred block to maintain anonymity.
Cross.Street	fctr	Cross Street of rounded Address.
Location	fctr	The location where the crime incident occurred. Actual address is omitted for confidentiality. XY coordinates reflect the nearest 100 block.

The crime category contains all variables that describe the incidents of crime in more detail such as the type of crime, the weapons used or the current status of the crime. The variables we assorted to this category include the number of the crime on record, the weapons used and their description, the status of the crime (code and description), then crime codes and their description as well as the modus operandi associated with the suspect in commission of the crime.

Crime

Variable	Type	Description
DR.Number	int	Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits
Weapon.Used.Code	dbl	The type of weapon used in the crime.
Weapon.Description	fctr	Defines the Weapon Used Code provided.

Variable	Type	Description
Status.Code	fctr	Status of the case. (IC is the default)
Status.Description	fctr	Defines the Status Code provided.
Crime.Code.1	fctr	Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious.
Crime.Code.2	fctr	May contain a code for an additional crime, less serious than Crime Code 1.
Crime.Code.3	fctr	May contain a code for an additional crime, less serious than Crime Code 1.
Crime.Code.4	fctr	May contain a code for an additional crime, less serious than Crime Code 1.
Crime.Code	int	Indicates the crime committed. (Same as Crime Code 1)
Crime.Code.Description	fctr	Defines the Crime Code provided.
MO.Codes	fctr	Modus Operandi: Activities associated with the suspect in commission of the crime.

The final category groups all variables related to time. It contains a variable that contains the reported date, the date and time occurrence.

Time

Variable	Type	Description
Date.Reported	fctr	MM/DD/YYYY
Date.Occurred	fctr	MM/DD/YYYY
Time.Occurred	int	In 24 hour military time.

2.1 Strengths and Weaknesses of Dataset

2.1.1 Strengths:

1. Large number of observations

The dataset is comprehensive and documents criminal activities in L.A. from 2010 to September 2017. The large number of observations means that our analysis will be more robust and enable us to more accurately make predictions on future criminal activties.

2. Many variables regarding each criminal offence were collected and documented

We can conduct analysis on different areas with respect to crime in L.A with the data that we have. The data collected was extensive and detailed, with information on demographics, type of crime, and timings amongst other things. This will again allow us to study and identify interesting patterns with regards to criminal activities in L.A.

2.1.2 Weaknesses:

1. Lack of socio-economic information

While the dataset was comprehensive, it did not contain information on the socio-economic status of victims or perpetrators of crimes. Having such information would allow us to make more interesting inferences on likely victims of crime and also the type of crime committed. It is highly likely that socio-economic information will be as powerful as demographics, if not more powerful, in predicting criminal activities.

2. Lack of proper classification for different types of crime

The classification of types of crime in the dataset was messy and was not conducive for analysis, making it difficult to study the correlation between types of crime being committed and other variables such as area and demographics. This is because the crime description in the original dataset differs even for crimes that are broadly similar in nature. We thus have to go through the dataset and make classifications that are logical and at the same time, conducive for analysis.

3. Lack of continuous variables

The dataset contains mainly variables that are mainly discrete or categorical, making it diffcult for simple linear regression analysis to be conducted. This means that we have to rely mainly on hypothetical testing and visualisation for analysis and logistic regression for predictions.

2.2 Data Cleaning Description

As the data is provided in two different csv files. The first step that we took was to merge the crime_data and code_data into one single dataset. Then we looked at the date columns (date occured and date reported) and reformated each date to three new columns. Thus creating a new colum for the day occured, month occured and year occured as well as day reported, month reported and year reported. For the completeness of full information on every year, we have omitted data from the year 2017 and will focus on the data from 2010 to 2016 included.

In order to make an interesting analysis we wanted to work with the information on victim descent. The abbreviations for victim descent in the original dataset, however, made analysis a little impractical as they weren't intuitively understandable (e.g. "D" standing for "Cambodian"). This is why we labeled the descent variables with the full description. (e.g. "W" being "White" and "D" being "Cambodian").

As there were some very specific subcategories for America (such as Alaskan Native) we also added an additional more global column on victim descent to be able to work with both the more detailed descent information and the more global descent variables. We reduced the global descent column to contain six groups:

Victim Descent

Victim Descent Global Variables
White
Black
Hispanic
Other
Asian
Pacific

In order to simplify the mapping of location information, we also extracted longitude and latitude from the location column and created additional, separate columns for longitude and latitude.

The data also covers and categorizes different types of crime in extensive detail. In order to extract valuable insights, we simplified the crime type variables by grouping them into more general categories. This reduced the 135 variables to sixteen more general variables.

As we put the focus of our analysis on understanding Los Angeles' victim profiles we decided not to use the data that is provided on crime suspects.

3 Demographic

Firstly,we want to find out what kind of people are more likely to be a victim of crime and thus look at some demographic variables such as age,sex and ethnicity.

3.1 Age and sex distribution

Here we begin with a summary of the age among all victims.The summary of the victim age shows the youngest victim is only 10 years old and the oldest being 99 years old. The mean age of victim is around 36 years old.

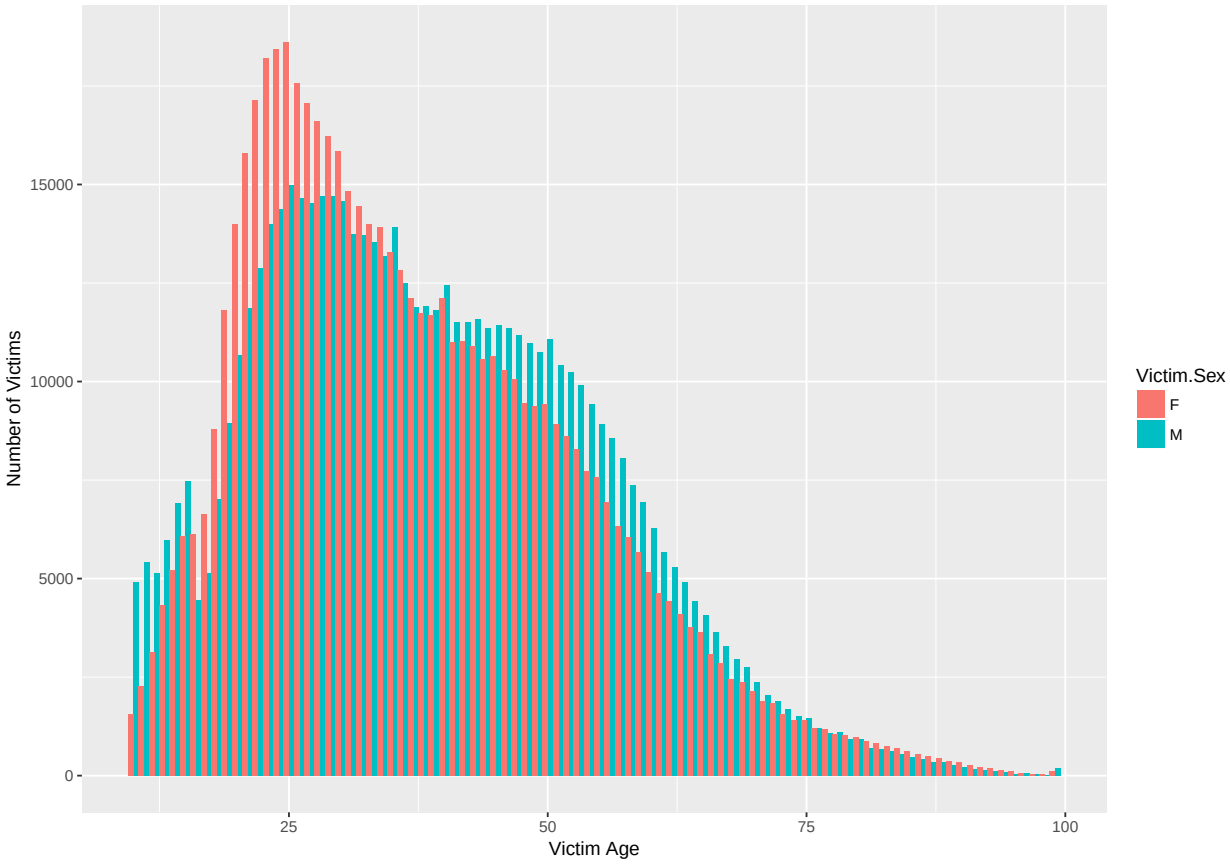
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.00   23.00   34.00   35.91   48.00   99.00  118059
```

Typically the age pattern of victims between male and female are not the same.We then construct a summary for male and female respectively and a histogram plot to see if it is the case.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.00   25.00   35.00   37.52   48.00   99.00   13927
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.00   26.00   37.00   38.72   50.00   99.00   71043
```

```
## Warning: Removed 84970 rows containing non-finite values (stat_bin).
```



The minimum and maximum of victim age are the same for both sex. However,the female victim age is smaller(37.52 years old) on average than for male(38.72 years old). From the overlapping pattern of the

histogram plot above, it is obvious that male are more likely to be a victim of crime at a young age(10-15 years old) or from 40 to 70 years old while female tend to be more in danger between around 15 to 35 years old, especially aged between 18 and 25. For example, at age 25,number of female victims are approximately 7500 higher than male victims.

Hypothesis testing

As we have seen a different pattern in the victim age for male and female above, we want to use student t-test to check if the difference exists statistcally or is due to random variation.

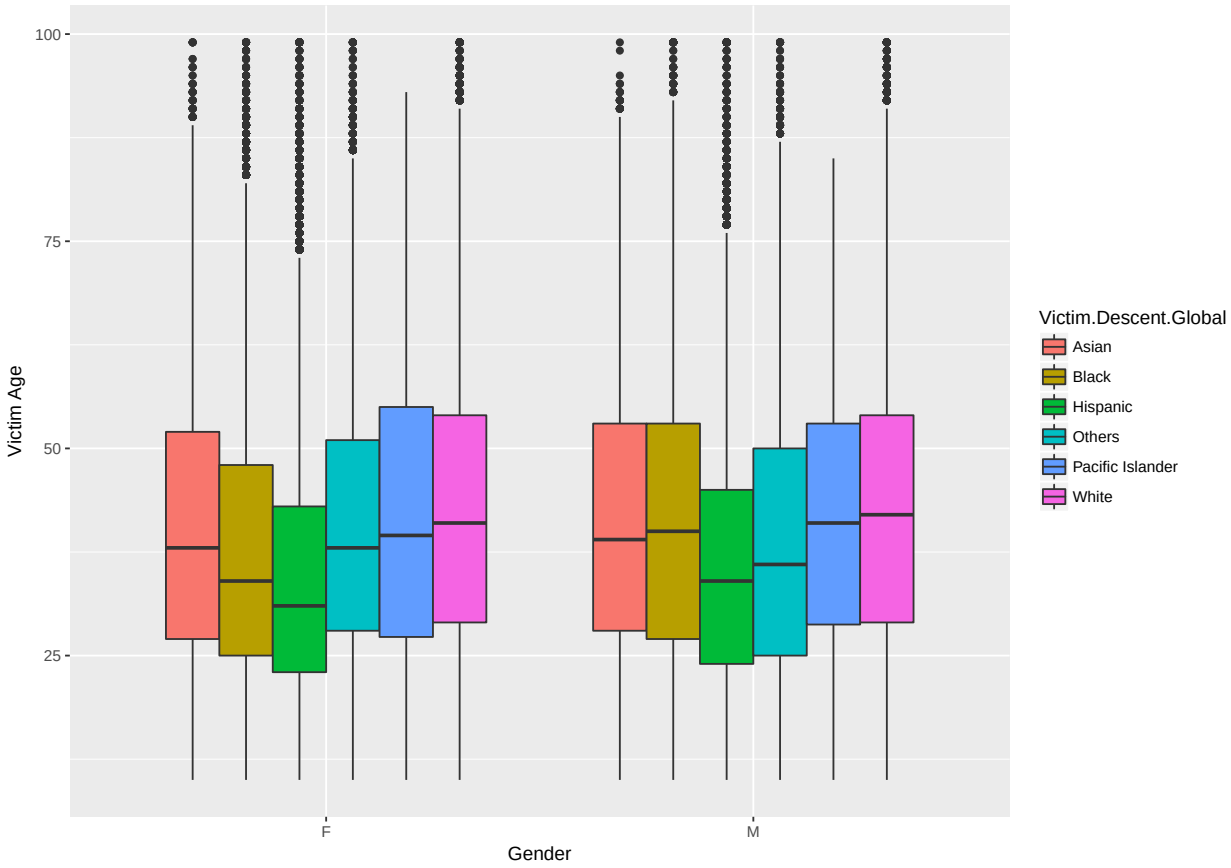
```
##
##  Welch Two Sample t-test
##
## data:  female_age and male_age
## t = -41.384, df = 1199700, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.258436 -1.144626
## sample estimates:
## mean of x mean of y
##  37.51671  38.71824
```

A t-test is conducted build on the null hypothesis that there is no significant difference in the mean victim age between male and female.Based on the t-test,a p-value of less than 2.2e-16 is really low and we can conclude that there is strong evidence in rejecting the null hypothesis that the difference in means is 0.Thus,mean age of victim is significantly smaller for female.

3.2 Ethnicity

Now that we have some clue about the relationship between age and gender,we want to dig more by adding another ethnicity variable.Below is a boxplot of victim age of male and female based on different ethnicity.

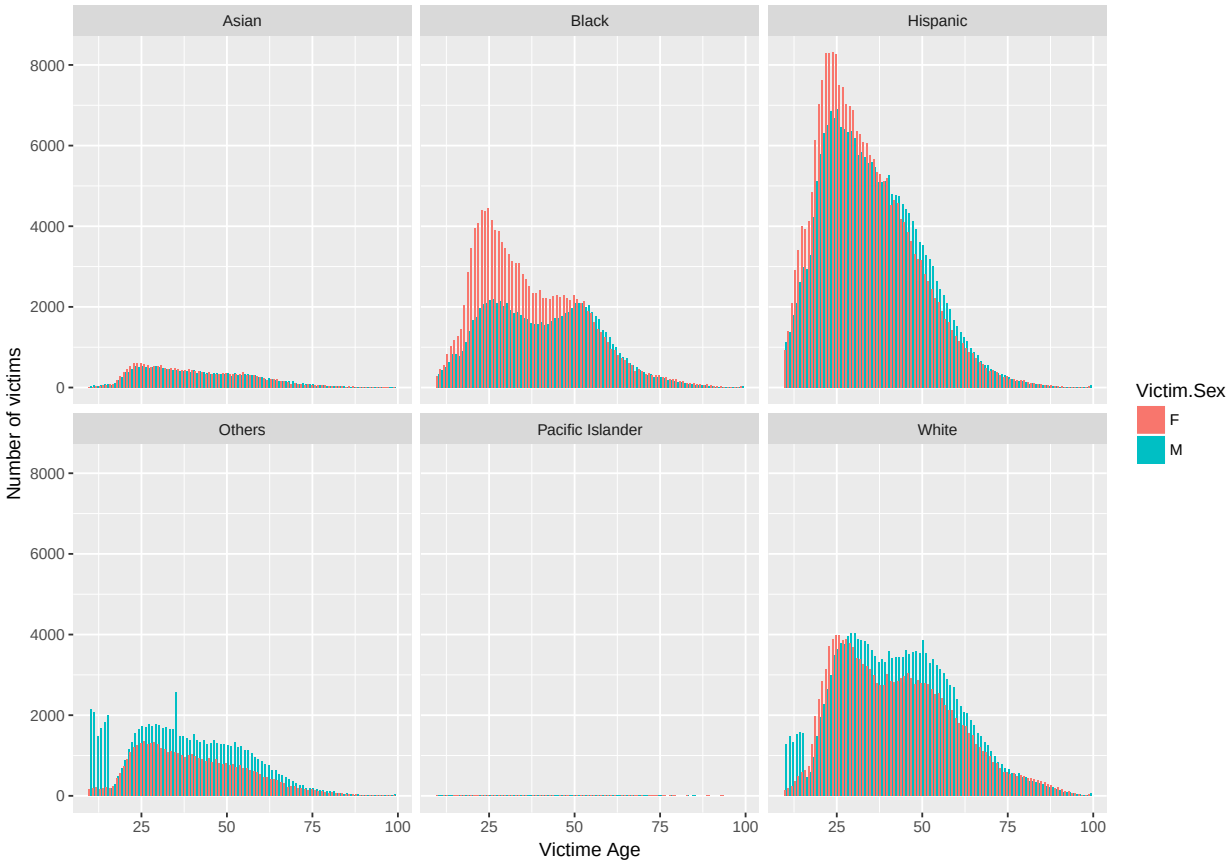
```
## Warning: Removed 84962 rows containing non-finite values (stat_boxplot).
```



From the boxplot based on ethnicity and gender,Hispanic have the lowest median age among all ethnicity groups and female Hispanic is the lowest(around 33 years old).For Blacks and Hispanic,female victim tends to be younger while it is not that obvious for other ethinicty groups.

We also have a histogram plot of victims counts with different gender and ethnicity below.

```
## Warning: Removed 84962 rows containing non-finite values (stat_count).
```



From the above plot,it is obvious that Hispanic account for the largest part of victims,followed by black and white victims. Also,there are observable difference in the age distribution between male and female for the three ethnicity group.Focus on hispanic victims,female aged below around 40 are much more likely to be a victim than male.

Hypothesis testing We want to carry out hypothesis testing to determine whether it is indeed the case according to the plot.

```
##
##  Welch Two Sample t-test
##
## data:  Hispanic_female and Hispanic_male
## t = -39.04, df = 475000, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.675325 -1.515150
## sample estimates:
## mean of x mean of y
##  33.78245  35.37769
```

```
##
##  Welch Two Sample t-test
##
## data:  black_female and black_male
## t = -51.339, df = 194590, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -3.607523 -3.342203
## sample estimates:
## mean of x mean of y
##  37.11240  40.58726
```

```
##
##  Welch Two Sample t-test
##
## data:  white_female and white_male
## t = 2.3716, df = 322380, p-value = 0.01771
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02422969 0.25494992
## sample estimates:
## mean of x mean of y
##  42.57392  42.43433
```

In the t-tests above,all three p-values are smaller than 0.05 which means that the difference in mean age is significant for all three ethnicity groups.Female Hispanic victims are around 6 years younger on average than male Hispanic victims.For black people,female victims are around 3 years younger on average than male victims.However,as for white people,the average age is approximately the same(with a difference of between 0.02 to 0.25).

Also,we want to know if the proportion of crime victims based on different ethnicity are the same as in the real population.Therefore,we calculate the proportion of white,black,hispanic victims below.

```
## [1] 0.2743819
```

```
## [1] 0.1782434
```

```
## [1] 0.3824481
```

According to the population of LA, non-Hispanic white people was 28.7% while black people was 9.6% of the population. While in the crime data,white people was 27.4% and black people was 17.8%.Thus,there are huge difference between the population proportion with the crime proportion for black people.

We can see that the variables we discussed above are correlated with each other from the plots,now we construct a chi-squared test to verify the dependence shown in the plots.

```
## Warning in chisq.test(crime_data$Victim.Descent, crime_data$Victim.Age):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  crime_data$Victim.Descent and crime_data$Victim.Age
## X-squared = 903010, df = 1780, p-value < 2.2e-16
```

```
## Warning in chisq.test(crime_data$Victim.Descent, crime_data$Victim.Sex):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  crime_data$Victim.Descent and crime_data$Victim.Sex
## X-squared = 2216100, df = 100, p-value < 2.2e-16
```

```
## Warning in chisq.test(crime_data$Victim.Age, crime_data$Victim.Sex): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  crime_data$Victim.Age and crime_data$Victim.Sex
## X-squared = 847320, df = 445, p-value < 2.2e-16
```

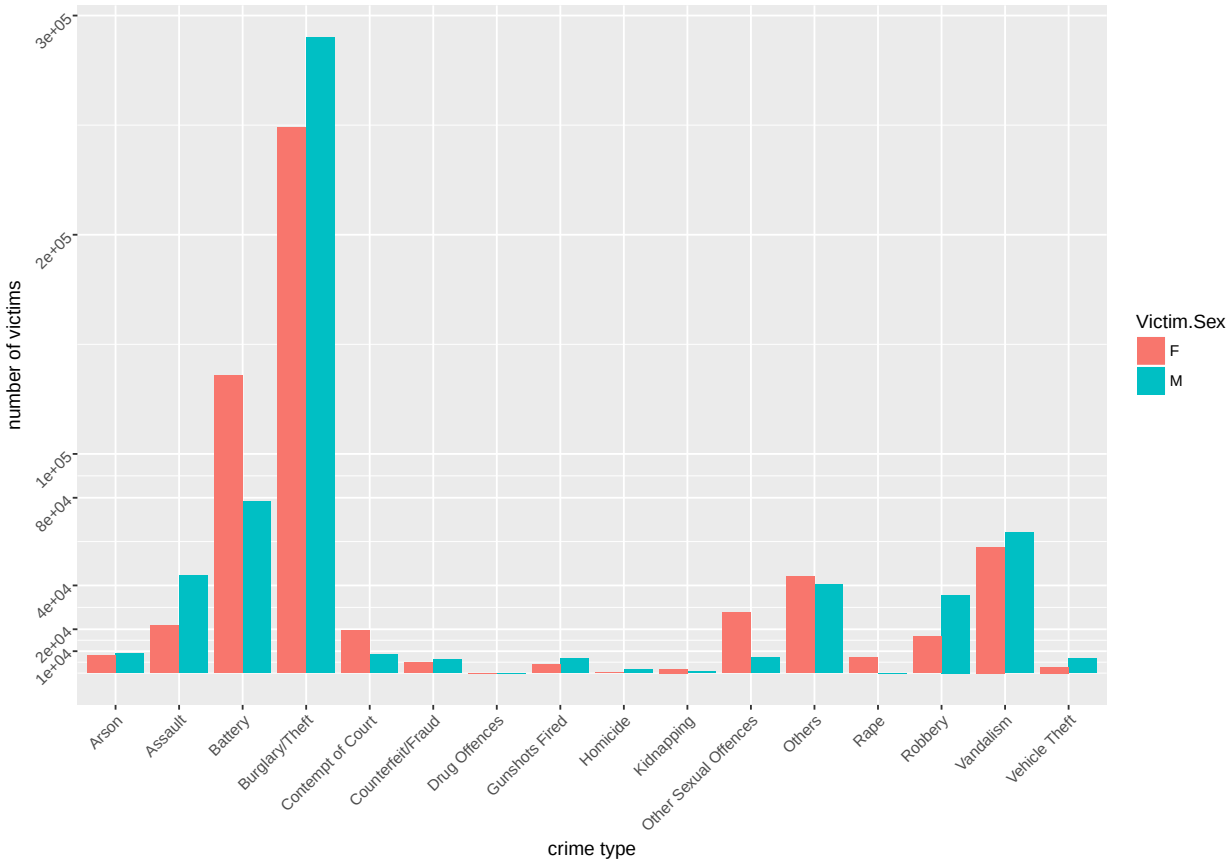
From the chi-squared tests above,it is obvious that there are certain degree of dependence between age and sex,age and ethnicity,sex and ethnicity.

summary

From the demographic analysis carried out so far,we found that people with certain demographic characteristics are more likely to be a victim of crime. For example,Hispanic young female aged aroung 25 years old are more likely to become crime victim.

3.3 relationship with type of crime

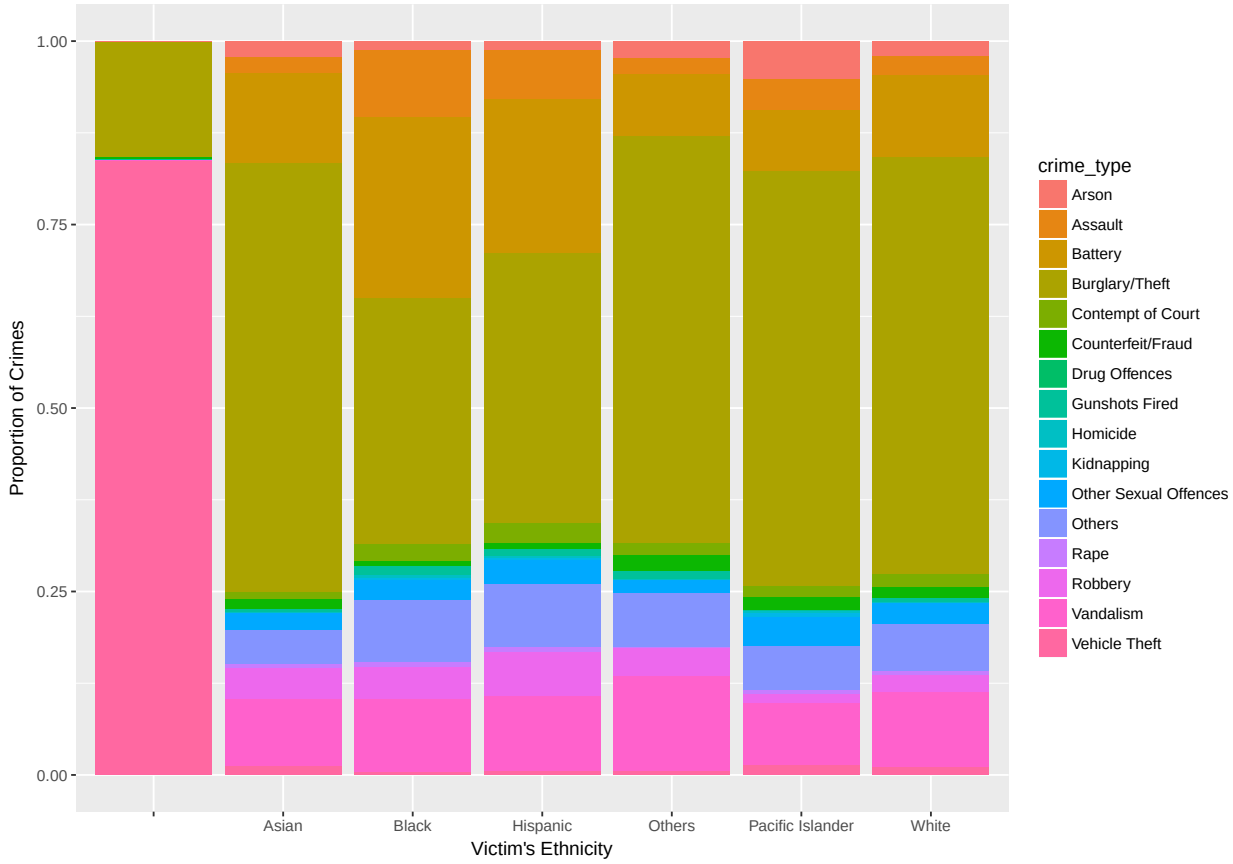
As we now have some understanding on the victim information,we want to find out whether the characteristics of victims differs in different types of crime.



From the barplot on crime type above,burglary,battery,assault and vandalism are the most likely crime types.Burglary and assault tends to be aiming more on male while battery and other sexual offences have more female victims.

We are also intersted in the pattern of crime type with different ethnicity.Overall,burglary/theft occurs most frequently for all ethnicity groups.The burglary/theft proportion is the lowest for black victims.As for the

proportion of assault and battery,asians,white,asian,pacific islanders share similar proportion pattern whereas black and hispanic people have higher proportion in this two crime types.



4 Area distribution

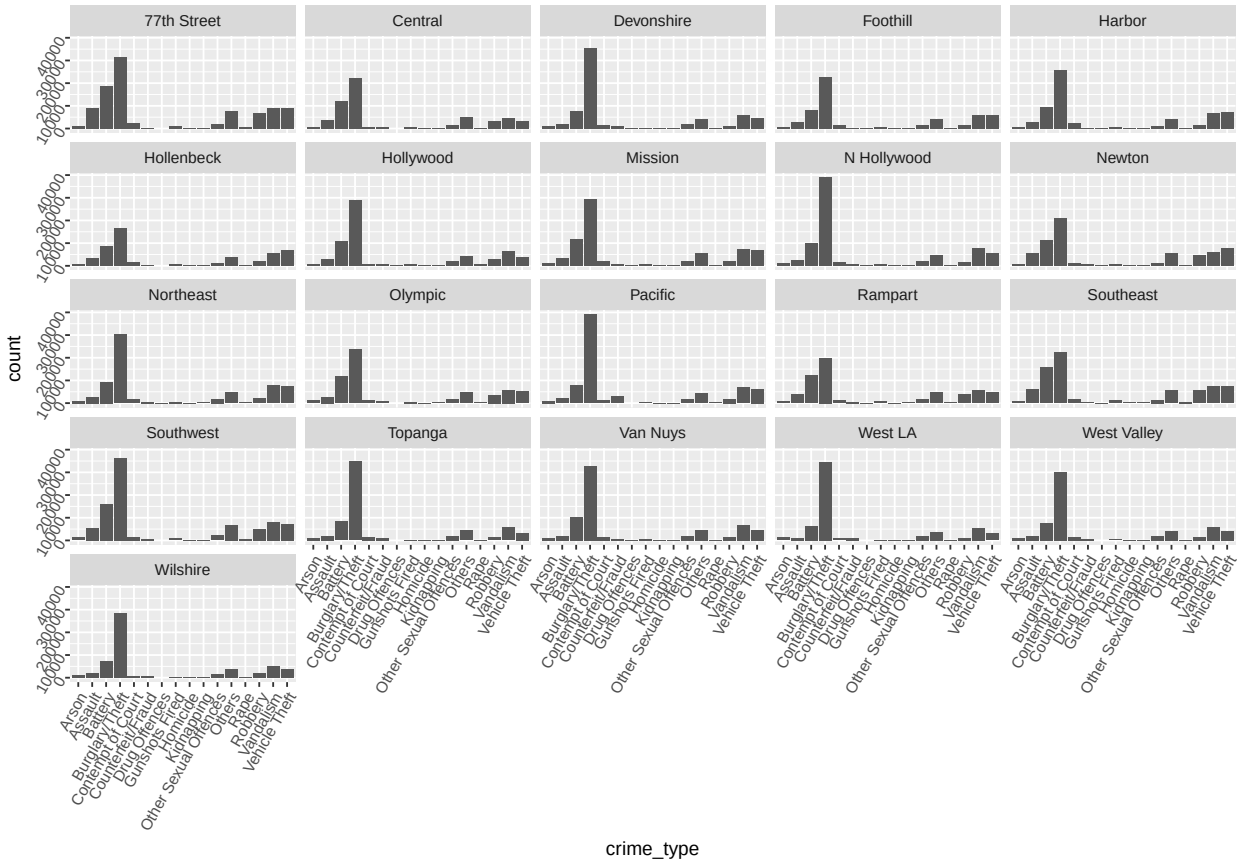
We have seen who are the victims of the crime, now let's focus on the location of these crimes. Of course, we can make the hypothesis that the location depends on the type of crime and on the victim.

4.1 number of crimes in different areas

First, we want to study the distribution of crime in the different areas.

```
## # A tibble: 80 x 3
## # Groups:   crime_type [16]
##    crime_type   Area.Name count
##        <fctr>      <fctr> <int>
##  1      Arson  Devonshire  1169
##  2      Arson N Hollywood  1230
##  3      Arson   Southwest  1295
##  4      Arson    Van Nuys  1230
##  5      Arson     West LA  1285
##  6    Assault 77th Street  9036
##  7    Assault      Newton  5369
```

```
##  8    Assault     Rampart  3835
##  9    Assault   Southeast  5873
## 10    Assault   Southwest  5424
## # ... with 70 more rows
```



From these graphs, we can see that the distribution of the type of crime is slightly similar in the different areas.

Hypothesis testing

Let's conduct a X-squared test to test the hypothesis whether the crime area is independent of the type of a crime at .05 significance level.

```
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 83714, df = 300, p-value < 2.2e-16
```

As the p-value is lower than the .05 significance level, we reject the null hypothesis that the crime area is independent of the type of a crime.

Another way to address the problem is to look at the crime density for different type of crime in Los Angeles.

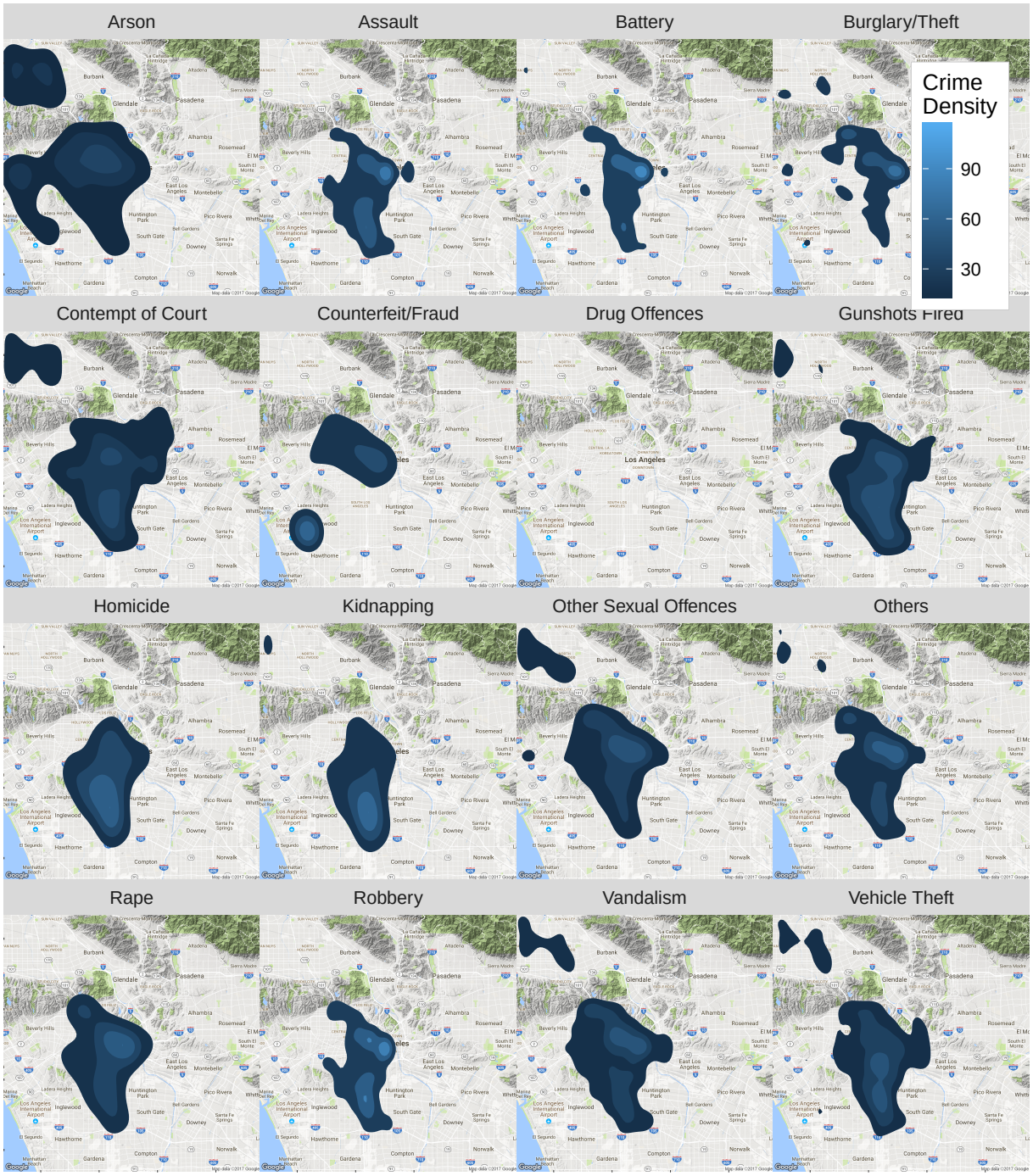
4.2 Area distrubution per type of crime

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=Los+Angeles&zoom=11&size=640x640
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=Los%20Angeles&sensor=
```

```
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```

```
## Warning: Computation failed in `stat_density2d()`:
## bandwidths must be strictly positive
```


Arson
Assault
Battery
Burglary/Theft
Crime Density
90
60
30
Contempt of Court
Counterfeit/Fraud
Drug Offences
Gunshots Fired
Homicide
Kidnapping
Other Sexual Offences
Others
Rape
Robbery
Vandalism
Vehicle Theft

We can see from the plot above the different area densities for each type of crime. As we will expect, for each type of crime, the region with the highest density of crime is close to central L.A.. There are a few main observations that we can make based on the plot.

Firstly, we observe that homicides, gun shot crimes and kidnapping have fairly similar area distributions, with the highest density located around the area south of downtown Los Angeles.

Secondly, we can see that rape crimes, burglary, robbery, vandalism and other sexual offences are most concentrated in downtown Los Angeles.

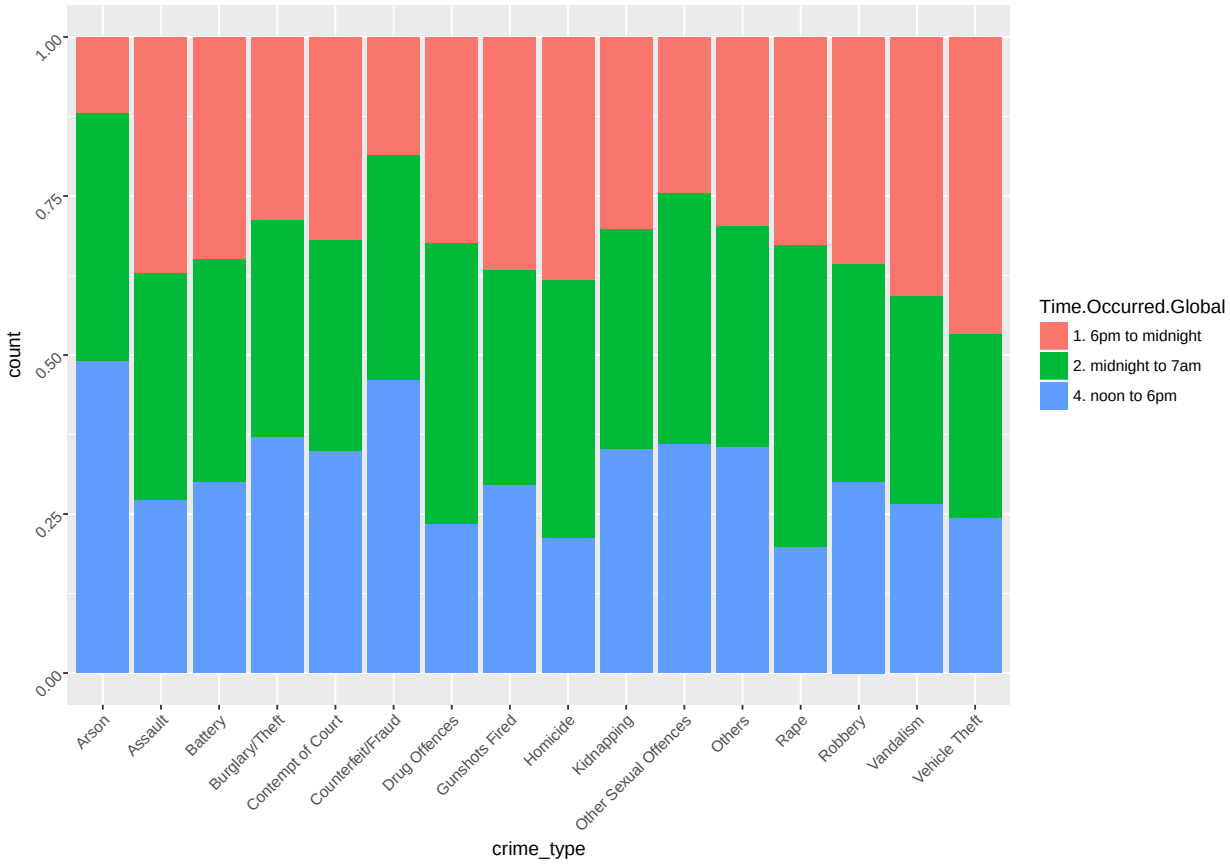
Lastly, we observe that crimes like arson and contempt of court have a large area distribution.

5 Time

In people's intuition,we always believe more crimes happen during night compared to day time and women are more likely to be a victim during night compared to men.In this part,we will try to look into these thoughts and find if this intuition consistent with factual data in terms of different types of crime.Besides we will also focus on the time between case occurred and reported to see if victim's quick reaction has a positive effect on solving the cases.

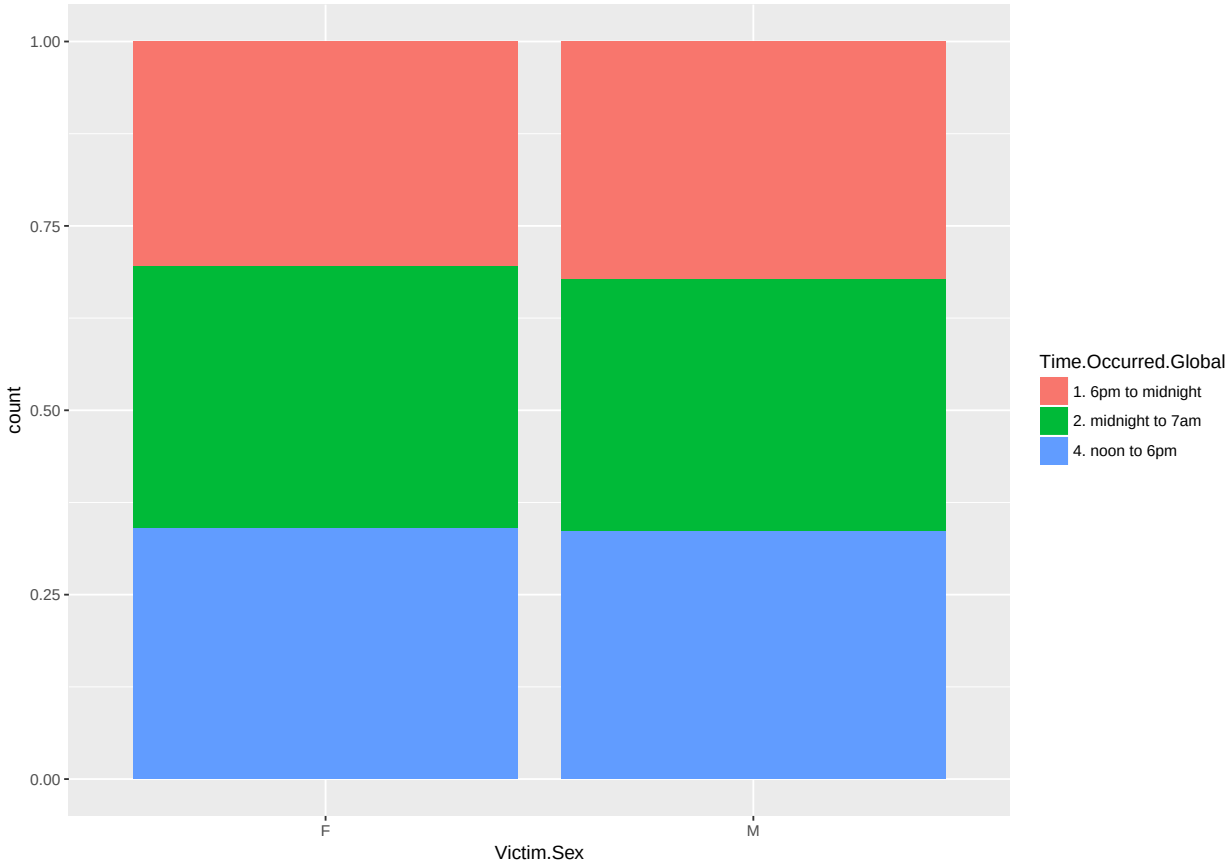
5.1 Occurred Time according to type of crime

(ex : fraud in the day, rape a night, burglary during the day etc. . .)



We can learn from the graph that unsurprisingly over 70% rape cases occured during night,from 6 pm to 7am.Homicide,assault,battery,gunshots,robbery,vandalism and vehicle theft are also more like to happen during night.On the contrary, Arson,fraud,and drug offences are more likely to happen during day time.So what is consistent with our intuition is that physical injury cases are more likely to happen during night.

5.2 Occurred Time according to sex



From the graph, there is no significant difference of probobilities of being a victim between males and females st a specific time during a day. What surprised us is that it seems that actually males are more likely to be a victim in night(6pm to 7am) compared to females.In order to see if the difference is significant,we use prop.test to analyse the data.

Hypothesis testing

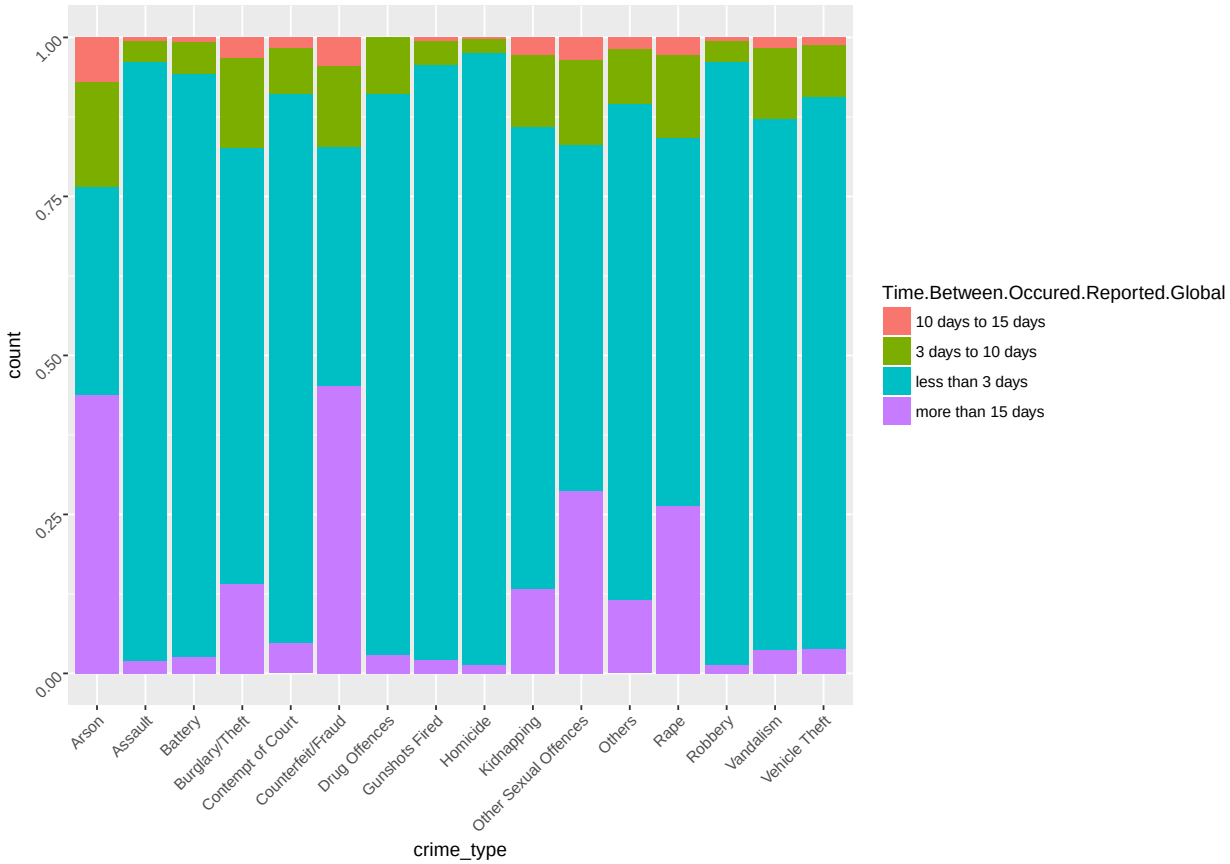
```
##    occurred.time
## sex day time  night
##   F   209361 405120
##   M   225756 444820
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table2
## X-squared = 23.491, df = 1, p-value = 1.255e-06
## alternative hypothesis: two.sided
```

```
## 95 percent confidence interval:
##  0.002412369 0.005691814
## sample estimates:
##    prop 1    prop 2
## 0.3407119 0.3366598
```

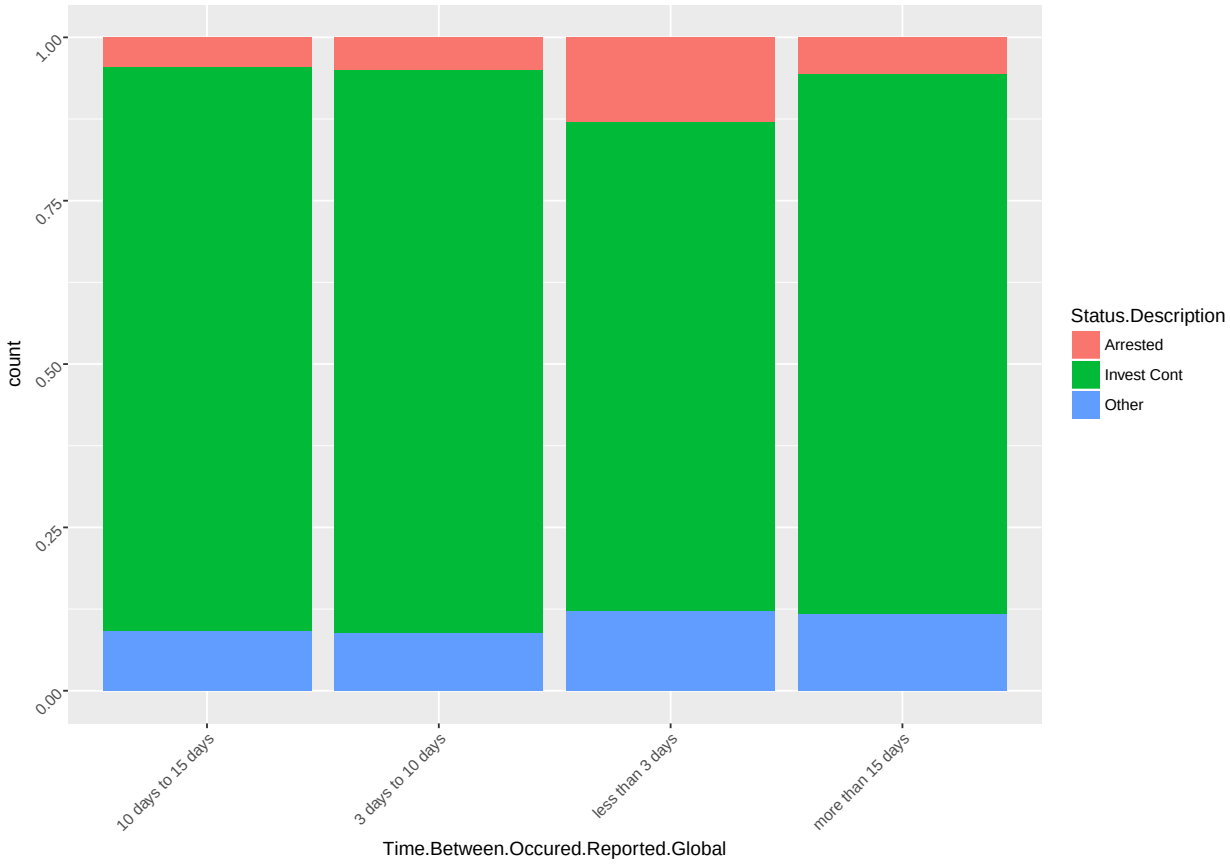
To do the prop.test,we combine the four periods time of a day into two categroies:day time(7am to 6pm) and night(6pm to 7am).The probabilities we get here is the probabilities of crime happening during day time for females and males,respectively.The P value is smaller than 0.05, which implies that there is a siginificant difference of the proportions of cases occurring during day time between men and women.Similar to our findings through graph above, women are more likely to be victims during day time while men are more likely to be attacked during night,which is inconsistent with our expectation. A possible reason behind this is that women are probably more tend to stay at home during night for their safety, which seems indeed a good idea.

5.3 Correlation between time reported - time occured and type of crime



It's shown from the graph that people usually react quickly to physical attack such as homicide,gunshots,assault.It's not surprising that victims report much later on fraud(may need time to realize that they are cheated),rape and other sexual offences(may feel too shameful to report) ,while it's quite hard to explain that why there is a delay in reporting arson cases.

5.4 Correlation between time reported - time occured and case solved or not



From this graph,we find that crimes reported in less than three days have a highest clear-up rate compared to that with longer reporting time.But the difference is not obvious so we need to do a prop.test on this to see if the difference is significant.

Hypothesis testing

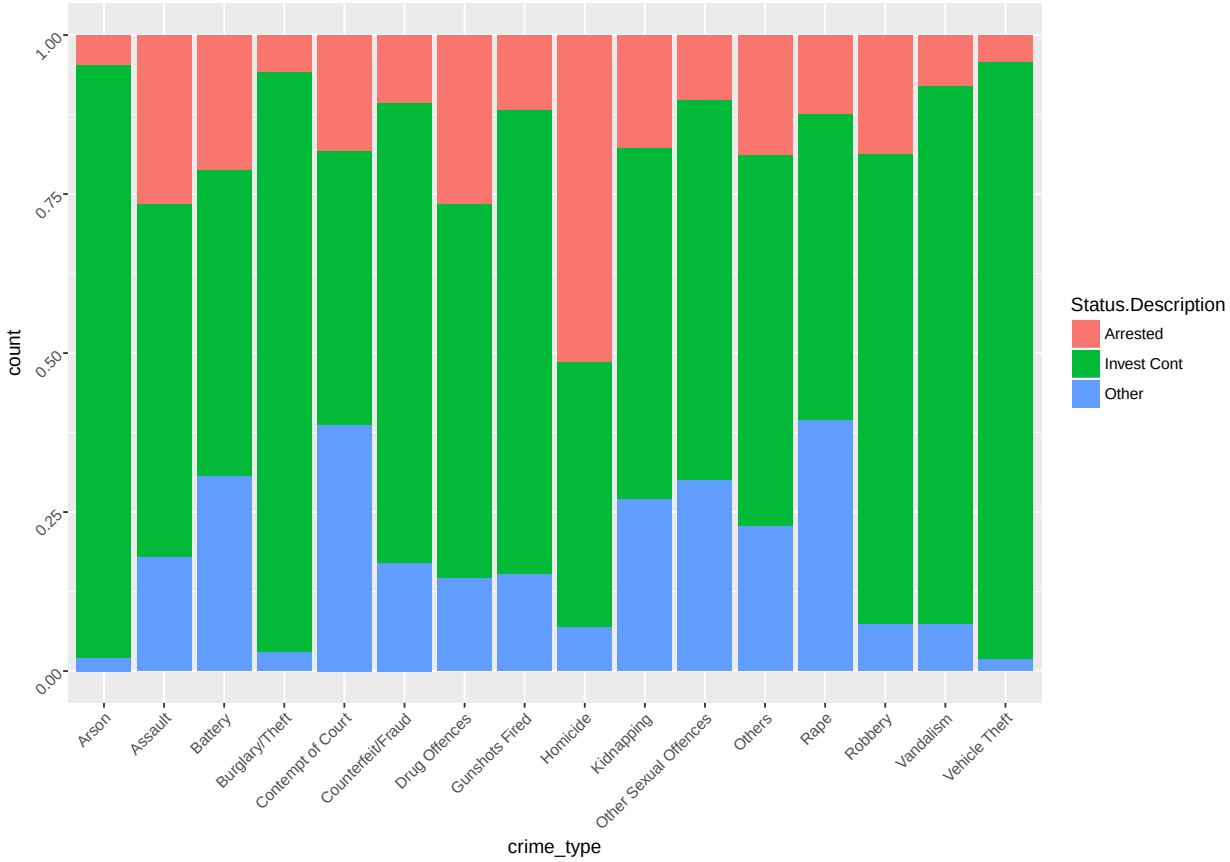
```
##                     status
## time                  FALSE   TRUE
##   10 days to 15 days  30263   1448
##   3 days to 10 days  140536   7474
##   less than 3 days   964896 143739
##   more than 15 days  136332   8042
```

```
##
##  4-sample test for equality of proportions without continuity
##  correction
##
## data:  table1
## X-squared = 15088, df = 3, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4
## 0.9543376 0.9495034 0.8703460 0.9442974
```

In the hypothesis testing, we combine the three types of cases status(Arrested,invest Cont and others) into

two status,where arrested equals to TRUE while invest Cont and others equal to FALSE.The probabilities we get here represents the proportion of unsolved cases of all cases categorized by the time between cases occurred and reported,which is 10 days to 15 days,3 days to 10 days, less than 3 days and more than 15 days, respectively. As p value in the proportion test is smaller than 0.05, it tells us that the unsolved rate of cases reported within less than 3 days(0.8703460) is significantly lower than the that of cases with longer report time, which is consistent with the intuitive findings according to the graph.This shows us that the quicker victims report, the more likely for police to solve the cases.

5.5 Correlation crime type and case solved or not



We can see from the graph that of all different types of crime, homicide has the highest clear-up rate, which is not surprising as police departments around all over the world always attach more emphasis and put more energy in homicide cases,and LAPD,according to what we find,is not an exception. Arson,burglary and vihecle theft have the three lowest detection rate, which may result from the lack of security cameras in streets in L.A. or the little attention paid on these cases.Generally speaking, the cases that invovle physical attack,or so called violent crimes,like assault,battery,have higher probability to be solved while the property crimes are less likely to be solved in L.A..

Based on all the above analysis,we learn that generally people report quicker on physical injury cases and police officers take these cases more seriously,leading to a higher probability of sovling these types cases.

5.6 Public holidays impact

Police always releases crime prevention tips before holidays, which gives us an impression that the number of crime occured on a holiday is generally higher than that on a regular day. We are interested in knowing whether our dataset support this impression to be true or not.

```
## # A tibble: 2 x 2
##        public_holiday Average.Crime.per.Day
##                 <fctr>                <dbl>
## 1     Public Holidays                577.5
## 2 Non-Public Holidays                559.8
```

To begin with, we have extract a list of holiday dates (not including Sundays) in LA from 2010 - 2016 from the Internet to catagorise crime occured dates to holiday and non-holiday. The avereage number of crime per day is on public holidays is 577.5 and that on non-public holidays is 559.8, which is a good start as it looks align with the impression.

```
## # A tibble: 1 x 4
##         Jan1 Average.Crime.per.Day Average.Report.Time SD.Report.Time
##        <chr>                 <dbl>               <dbl>          <dbl>
## 1 Other Days                 560.3                17.7          108.2
```

However, we also found a pattern when exploring the data; number of crime on 1st Janurary in each year is extremly high comparing to othere days, and as 1st Jan is usually a public holiday, this would distort the picture. Besides, there is also a significant difference in the average reporting time (number of day between when the crime ocurred and reported) and the standard deviation of the distribution of reporting time.

Unfortunately we could not find any information on the Internet or from the source of the dataset that could explain the peaks on 1st Janurary. The most logical reason of this could be some of the cases were occured in the year but on an unknown day, and instead of keeping the occured day NA or missing, it was set as the 1st day of the year.

In order to have a meaningful comparison between number of crime cases on holidays and regular days, we have to remove the cases that were not truly occured on 1st Jan. As there is no indicator in the dataset pointing out those cases, the best estimator we have is number of days between occured and reported. Assuming the crimes on 1st Janurary would have the same distribution as other dates in a year, we decided to exclude cases pccured on 1st Jan with time difference between occured and reported more than 14.4 + 2*88.7 = 192 days. In theory, this range should have covered over 95% of the cases on 1st Jan.

```
## # A tibble: 1 x 4
##         Jan1 Average.Crime.per.Day Average.Report.Time SD.Report.Time
##        <chr>                 <dbl>               <dbl>          <dbl>
## 1 Other Days                 560.3                17.7          108.2
```

The way we use to exclude cases on 1st Jan is definiely not prefer, the standard derivation of cases on 1st Jan is lower than that on other days as we have a sharp cut at 192 days. But it is the best way we have to reduce some of the 'noise' and have excluded half of the cases on 1st Jan.

```
## # A tibble: 2 x 2
##        public_holiday Average.Crime.per.Day
##                 <fctr>                <dbl>
## 1     Public Holidays                577.5
## 2 Non-Public Holidays                559.8
```

And the average number of crime per day on public holidats is now less than that of that on non-public holidays, after excluding some of the 1st Jan cases. Before moving on to the hypothesis testing on whether there is more or less crime cases on a public holidays, we still have one question, are the natural of Sundays the same as public holidays, which most of people are not required to work and shop closes early on the day, in terms of number of crime cases per day?

```
## # A tibble: 3 x 4
##   Sundays.PH.Regular Average.Crime.per.Day Average.Report.Time
##                <chr>                 <dbl>               <dbl>
## 1     non-PH Sundays                 536.5                15.7
## 2   not PH or Sunday                 563.8                15.9
## 3    Public Holidays                 577.5                75.1
## # ... with 1 more variables: SD.Report.Time <dbl>
```

```
##
##  Welch Two Sample t-test
##
## data:  Sunday.count$count and PH.count$count
## t = -1.3548, df = 78.619, p-value = 0.1794
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -101.14211   19.22379
## sample estimates:
## mean of x mean of y
##  536.5014  577.4605
```

Not surprisingly, the mean number of crime cases per day on Public Holidays and on Sundays are not much different. And the t-test also indicate that there is no difference between the mean number of crime cases in Sundays and Public Holidays. Therefore, Sundays and Public Holidays could be grouped in testing whether there is a difference between number of crime cases in Holidays(Public Holidays and Sundays) and the other days.

```
## # A tibble: 2 x 4
##   Sundays.PH.Regular2 Average.Crime.per.Day Average.Report.Time
##                 <chr>                 <dbl>               <dbl>
## 1            Holidays                 536.5                15.7
## 2    not PH or Sunday                 564.3                18.0
## # ... with 1 more variables: SD.Report.Time <dbl>
```

```
##
##  Welch Two Sample t-test
##
## data:  Holiday.count$count and Regular.count$count
## t = -5.5728, df = 491.41, p-value = 2.068e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -19.56582
## sample estimates:
## mean of x mean of y
##  536.5014  564.2824
```

The average number of crime cases on Holidays (Public Holidays and Sundays combined) is less than that on regular days, shown by the t-test, in 99% confidence level, which is different from the general perception that there are more crime cases on holidays.

```
## # A tibble: 32 x 3
## # Groups:   crime_type [?]
##           crime_type Sundays.PH.Regular2 average_per_day
##                <fctr>               <chr>           <dbl>
##  1            Arson            Holidays        4.300836
##  2            Arson    not PH or Sunday        8.366545
##  3          Assault            Holidays       33.983562
##  4          Assault    not PH or Sunday       25.937500
```

```
##  5           Battery            Holidays      102.008219
##  6           Battery    not PH or Sunday       83.947080
##  7    Burglary/Theft            Holidays      196.895890
##  8    Burglary/Theft    not PH or Sunday      243.197536
##  9 Contempt of Court            Holidays       11.320548
## 10 Contempt of Court    not PH or Sunday       11.547901
## # ... with 22 more rows
```

```
##                        Sundays.PH.Regular2
## crime_type                Holidays not PH or Sunday
##   Arson                    4.300836         8.366545
##   Assault                 33.983562        25.937500
##   Battery                102.008219        83.947080
##   Burglary/Theft         196.895890       243.197536
##   Contempt of Court       11.320548        11.547901
##   Counterfeit/Fraud        4.149296         6.273312
##   Drug Offences            1.000000         1.523810
##   Gunshots Fired           5.636364         4.924115
##   Homicide                 1.621622         1.481210
##   Kidnapping               1.727660         1.747253
##   Other Sexual Offences   12.797260        14.244069
##   Others                  36.158904        39.813869
##   Rape                     3.762857         3.031453
##   Robbery                 22.528767        22.226734
##   Vandalism               55.772603        52.911040
##   Vehicle Theft           45.452055        46.367245
```

Most of crime type has less ocuuring rate on public holidays, but it likes look assault, battery, gunshots, rape and vandalism has a higher occuring rate in public holidays.

6 Prediction and logistic regression

We have studied the correlations between the different variables of our datasets, and we have tried to find some patterns about the crime type distribution according to the area and the victims.

+Now, from these analysis, we would like to make some prediction about the type of a crime or the crime area according to different parameters. To do so, we need to run a multinomial logistic regression since our dependant variable is categorical (crime type or area). Unfortunately, the results are not satisfactory with few variables, and the algorithm is too slow when we add other variables. This is something that should be done in an deeper analysis.

Last but not least, We want to run a simple logistic regression to predict whether a case is likely to be solved or not.

Data Cleaning

Training set and learning set

We train our algorithm with the data until 2015. Our testing dataset are the crimes which occured in 2016.

Logistic Regression

We try to predict the final status of a crime (Solved or not) according to different parameters such as the age of the potential victim, the sex, the ethnicity, the area, the type of crime and the time between occured and reported crime.

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 11, 2017 - 7:20:35 PM

From this model, we can see that all the variables are significant at a 95% confidence with a p-value lower than 0.05. When comparing the different independant variables, we realize that the ethnicity of a victim has a smaller impact on finding the suspect of a crime than the area or the type of crime. Moreover, the estimate of Time.Between.Occured.Reported confirms being negative, it confirms that the sooner a crime is reported, the more likely it will be solved.

Prediction

```
## [1] "Accuracy 0.886482527153563"
```

We run our algorithm with the data from 2016 and have an accuracy of 0.88, which is very good. We can predict with 88% of precision whether a crime will be solved even before the investigation !

7 Discussion

7.1 What results did you find?

We have studied the patterns of crime in L.A. in three different aspects. Looking at the demographics of victims, we have information about the characteristics of people that are most likely to be a victim. Younger females are more likely to be victims of crime compared to males of similar ages. Also, Hispanics account for the largest proportion of victims compared to other ethnic groups. More specifically, Hispanic young women around the age of 25 are the most susceptible to crime in L.A.. In addition, males are more likely to be victims in crimes like burglary and assault while female victims account for a larger proportion in battery and other sexual offences.

Secondly, while analysing the area distribution of crimes in L.A., we find that 77th Street has the highest number of crimes in 2016. We also find that some crimes have similar distributions. Specifically, more serious crimes like homicide and kidnapping are concentrated at the area south of central L.A. while less serious crimes like arson and vandalism are concentrated in central L.A..

From the time analysis part,firstly we find that physical injury cases like homicide,assault,battery,rape are more likely to happen during night.Also,women are more likely to be victims during day time while men are more likely to be attacked during night,which is inconsistent with our expectation. What else, crimes reported within less than three days have a highest clear-up rate compared to those with longer reporting time, which shows us that the quicker victims report, the more likely for police to solve the cases.Besides the reporting time, the nature of cases also affects the clear up rate.The cases that invovle physical attack,or so called violent crimes,like assault,battery,have higher probability to be solved while the property crimes are less likely to be solved in L.A.

Although the general perception is that there are more crime cases occurred on holiday, we found this is not true in our dataset. The average number of crime cases on a regular day is 5% higher than that on a holiday. However, certain types of crime that could be associated with festivals and alcohol occur more often during holidays. Thereford, it is better to stay away from the crowd when you do not want to be a victim on holidays.

7.2 Why is this interesting?

With the information extracted from the L.A crime dataset,it is possible to provide useful suggestions for LA police in order to reduce the number of either citizens or tourists of L.A from being hurt.In addition,we are interested in the general pattern of crime and curious about what will happen in the future.Thus,a prediction of the crime based on the recent dataset is crucial for the future crime prevention. Help LA police, predict the crime, prevention for the potential victims

Table 6:

	Dependent variable:
	Status.Description.Global
crime_typeAssault	2.069***
	(0.045)
crime_typeBattery	1.756***
	(0.044)
crime_typeBurglary/Theft	0.208***
	(0.044)
crime_typeContempt of Court	1.569***
	(0.047)
crime_typeCounterfeit/Fraud	1.054***
	(0.056)
crime_typeDrug Offences	2.417***
	(0.454)
crime_typeGunshots Fired	0.977***
	(0.056)
crime_typeHomicide	3.155***
	(0.067)
crime_typeKidnapping	1.378***
	(0.072)
crime_typeOther Sexual Offences	0.854***
	(0.048)
crime_typeOthers	1.632***
	(0.044)
crime_typeRape	1.058***
	(0.059)
crime_typeRobbery	1.450***
	(0.045)
crime_typeVandalism	0.437***
	(0.045)
crime_typeVehicle Theft	0.289***
	(0.069)
Time.Between.Occured.Reported	−0.001***
	(0.0001)
Victim.SexM	−0.093***
	(0.007)
Victim.Age	−0.010***
	(0.0002)
Area.NameCentral	0.079***

7.3 What would be the next steps?

In this paper, we focus on the victim of crime in order to help prevent certain crimes. The next step could be carrying out investigation in the suspect of crime to help police find suspects and solve crime more efficiently. crime type prediction area crime prediction suspect