Final Report

Group B

September 29, 2017

Contents

1 Introduction

Crime is always a hot topic in Los Angeles. With a considerably higher violent crime rate of 6.45 per 1000 residents in 2015, compared to US national median of 3.8 per 1000 residents (based on FBI crime data), L.A is amongst the most dangerous cities in US. According to Los Angeles Police Department(LAPD), from 2002 to 2012 the crime rate in city had declined in 10 consecutive years. However, in 2015, LAPD 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 pointed a range of factors that resulted in the jump, including more gang violence and a growth of homeless people. With a high chance of 1 in 33 to become a victim of either violent or property crime in L.A, it is never too much to emphasize the importance of crime prevention and public safety. In order to find patterns of crime in L.A and come up with some useful suggestions accordingly in crime prevention, this paper looks into three main questions of various crime in L.A: who is more likely to be a victim; where is the accident-prone place in L.A and when the crime is most 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. As this data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data.

1.1 Context

Text here

1.2 Theory

Text here

2 Data Description

Text here

Titre 1	Titre 2	Titre 3
Colonne Alignée à Gauche	Colonne Alignée au Centre	Colonne Alignée à Droite

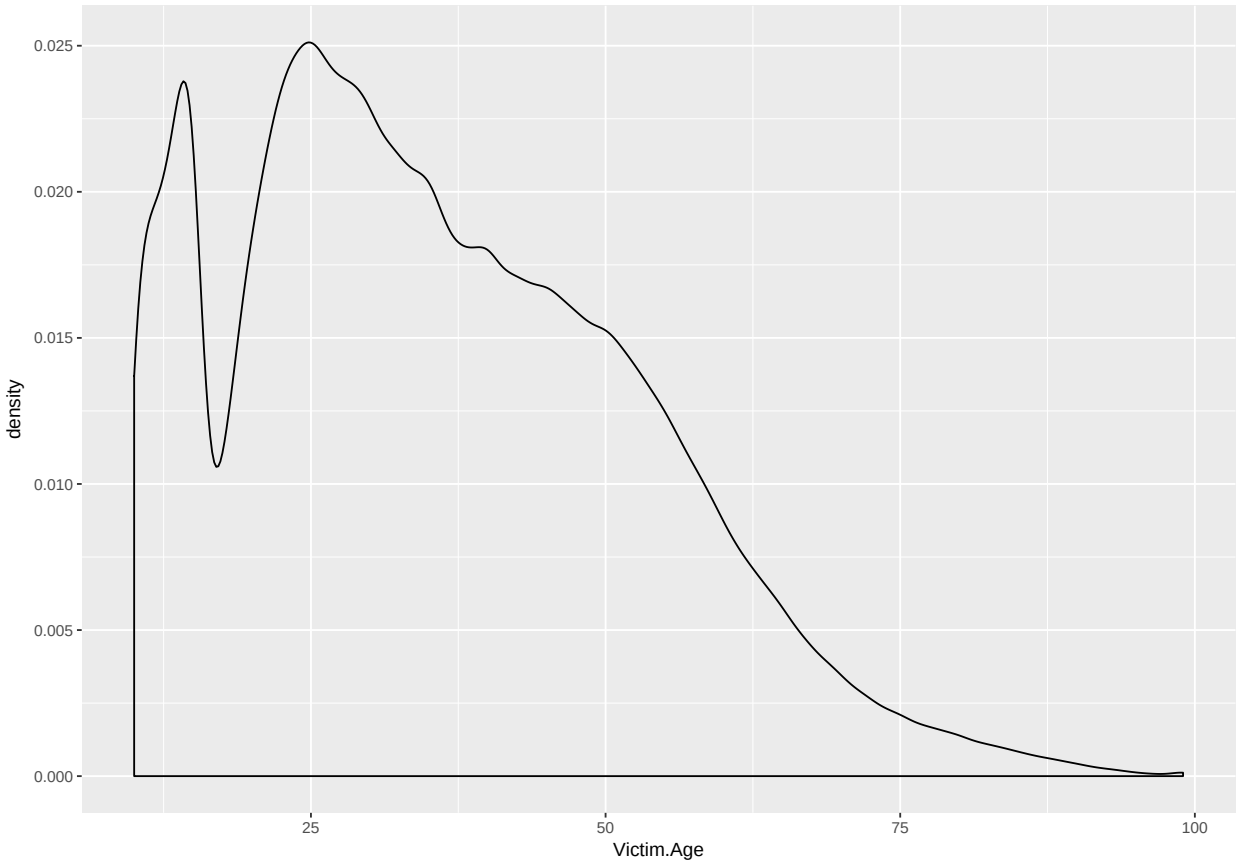
3 Demographic

Text here

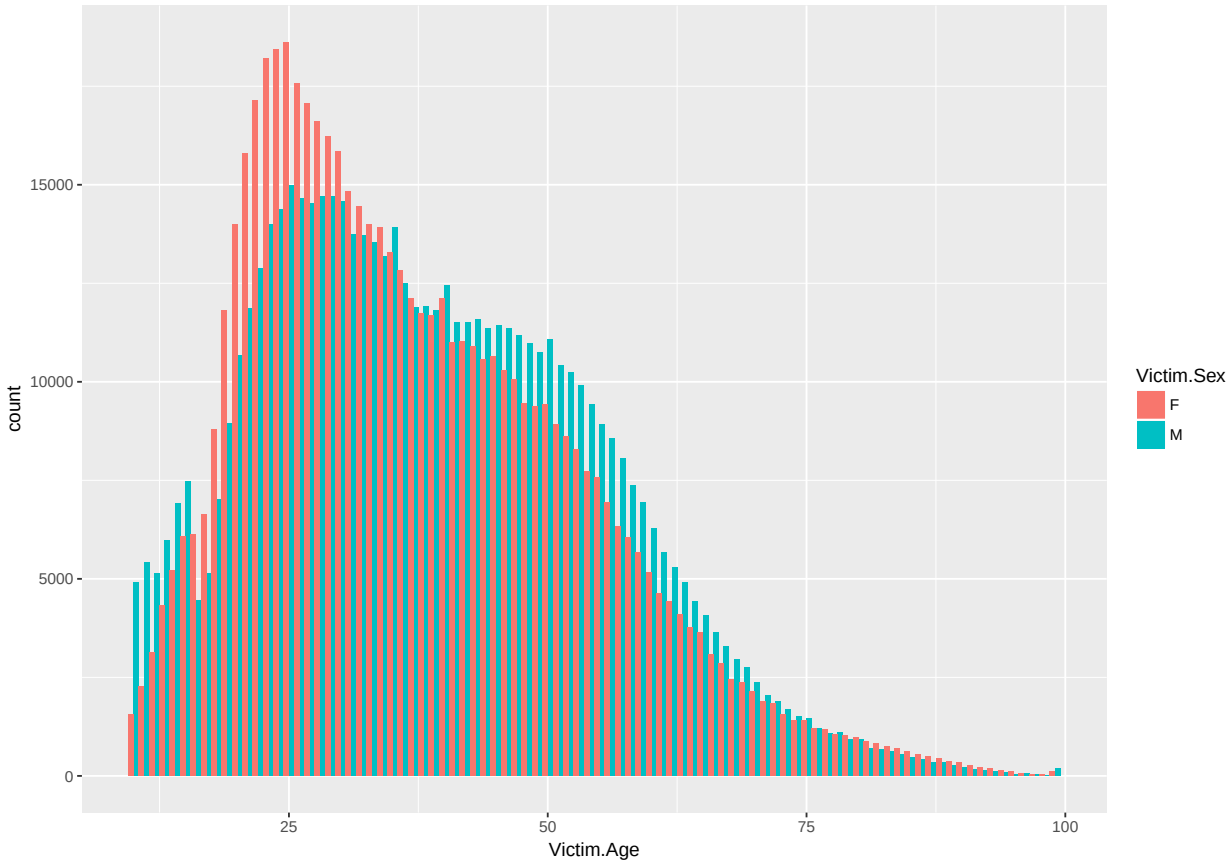
3.1 Age and sex distribution

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

```
## Warning: Removed 118059 rows containing non-finite values (stat_density).
```




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



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.From the density plot of victim age,we can see that there are two peak points at 14 and 25 respectively.In addition,a histogram plot of the victim age of male and female is shown above.From the overlapping pattern,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.

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

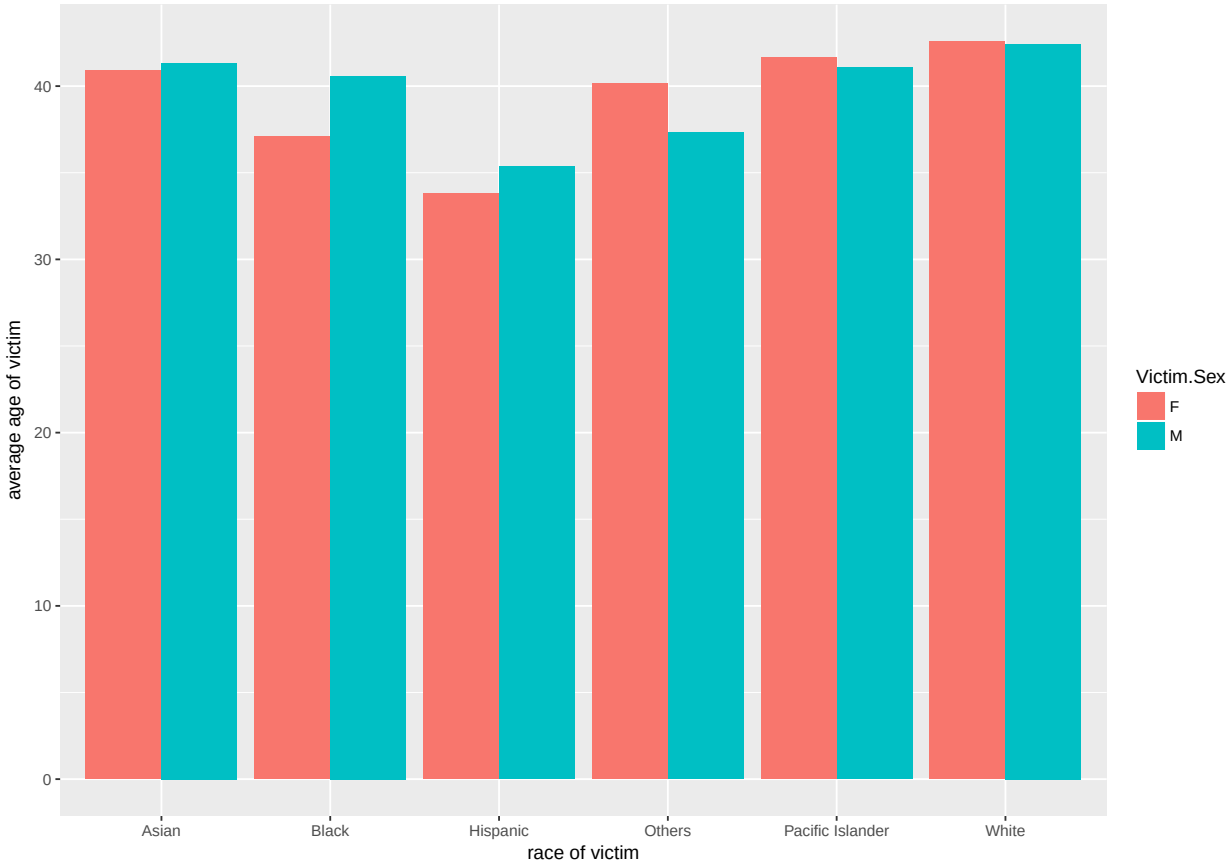
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.00   26.00   37.00   38.72   50.00   99.00   71043
```

Hypothesis testing

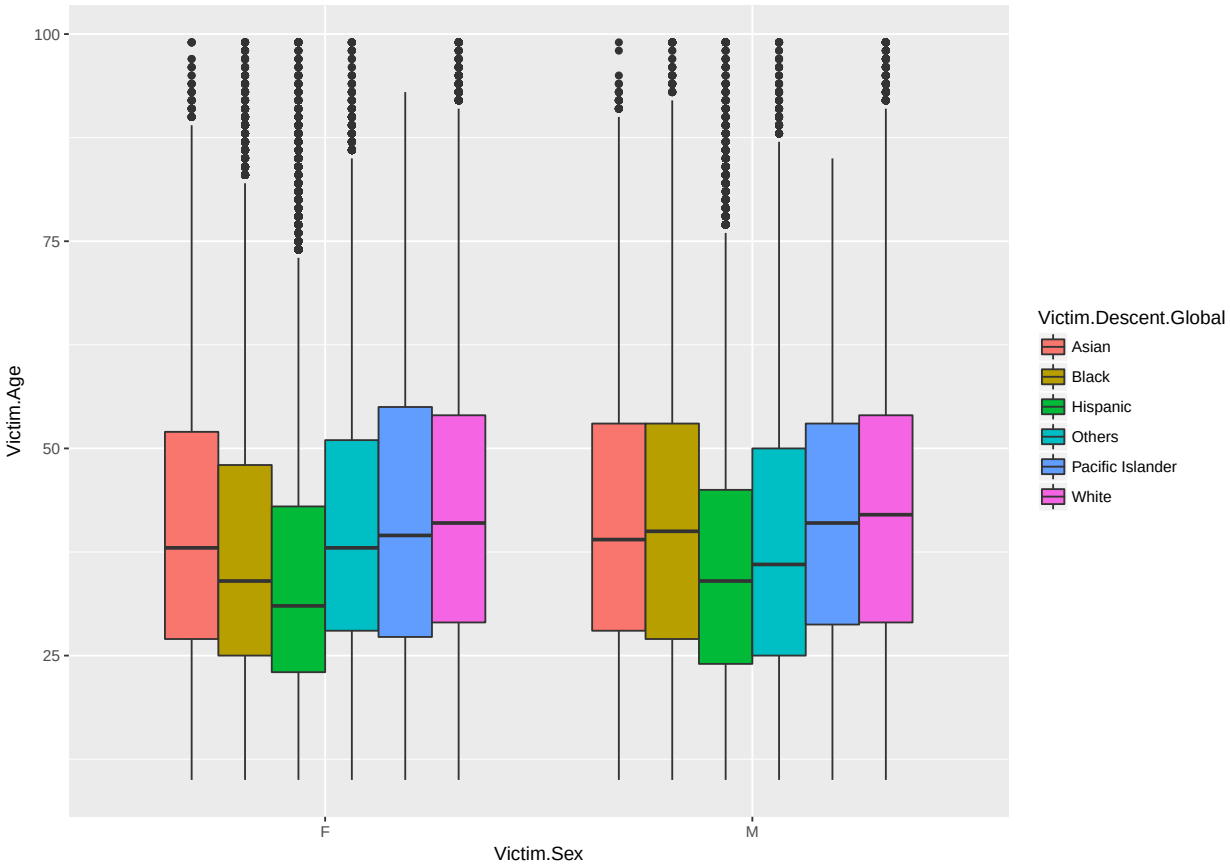
```
##
##  Welch Two Sample t-test
##
## data:  female_age and male_age
## t = -41.384, df = 1199700, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.258436 -1.144626
## sample estimates:
## mean of x mean of y
##  37.51671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

Text here From the summary of both male and female,the mean age of victim is lower for female(37.52 years old).We want to see whether this difference in mean age is due to random variation.A t-test is conducted to find out the results.Based on the t-test,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

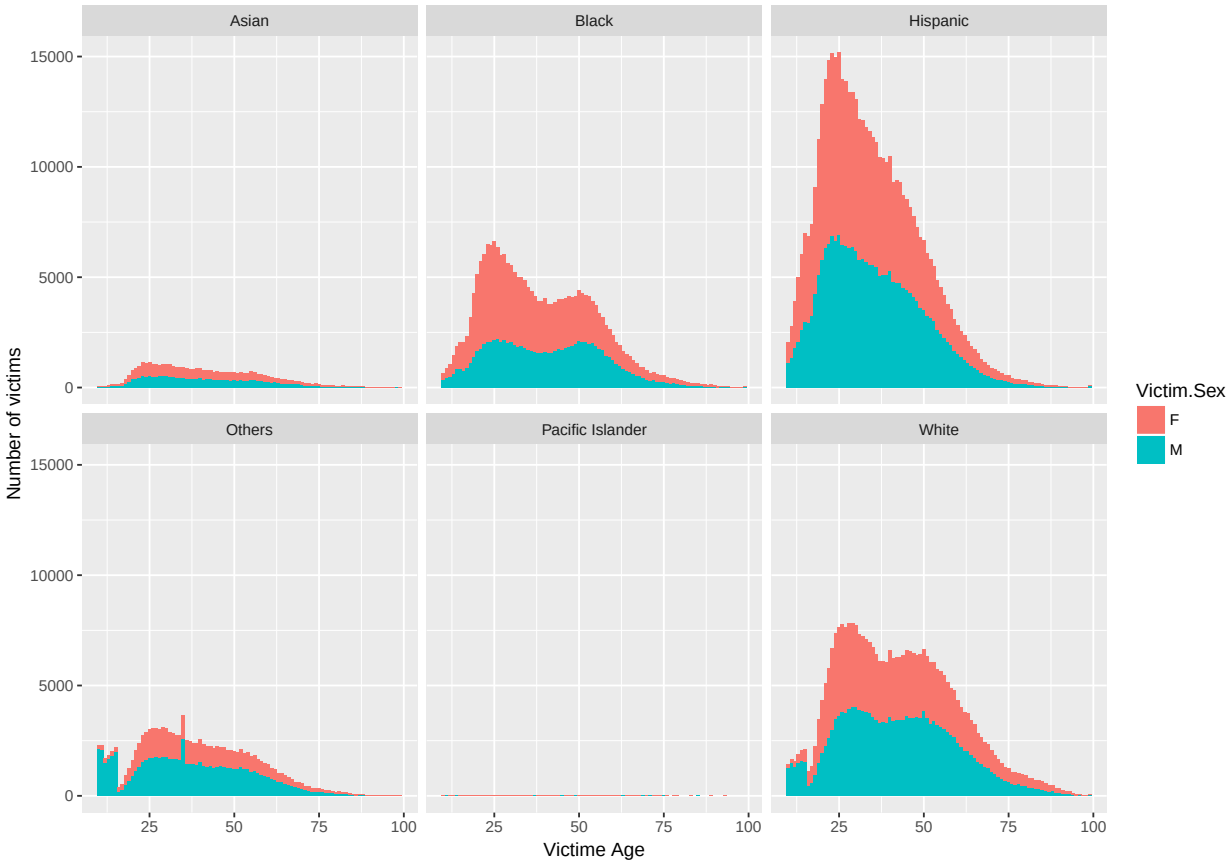



```
## Warning: Removed 84962 rows containing non-finite values (stat_boxplot).
```

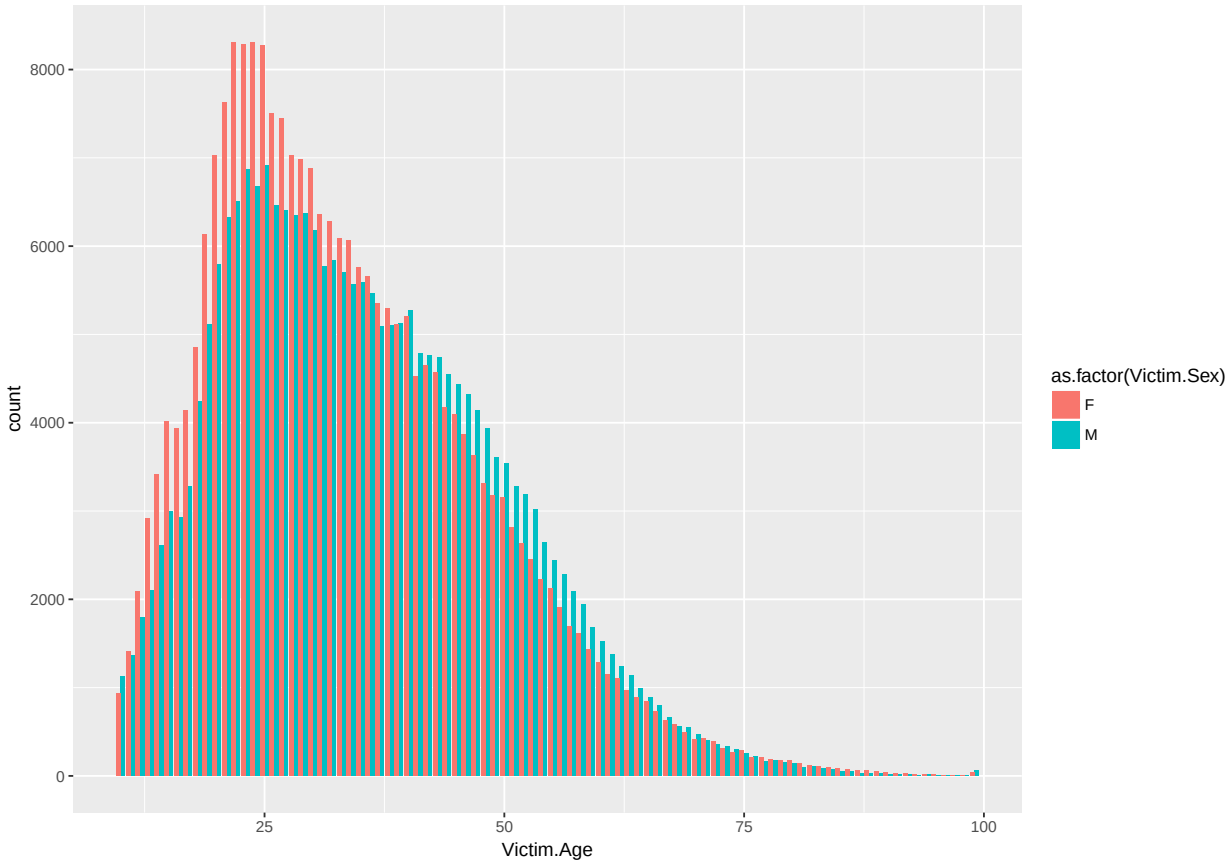


From the bar chart of average victim age based on ethnicity and gender,Hispanic have the lowest average age among all ethnicity groups and female Hispanic is the lowest(around 33 years old).For Blacks and Hispanic,female victim tends to be younger on average while it is not that obvious for other ethinicty groups.We want to carry out hypothesis testing to determine whether it is indeed the case according to the plot.

```
## 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 and we would like to focus on Hispanic victims.



Text here We now have some clue in the most likely victim of a crime:Hispanic young female aged aroung 25 years old. *Hypothesis testing*

```
##
##  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
```

From the t-tests above,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).

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

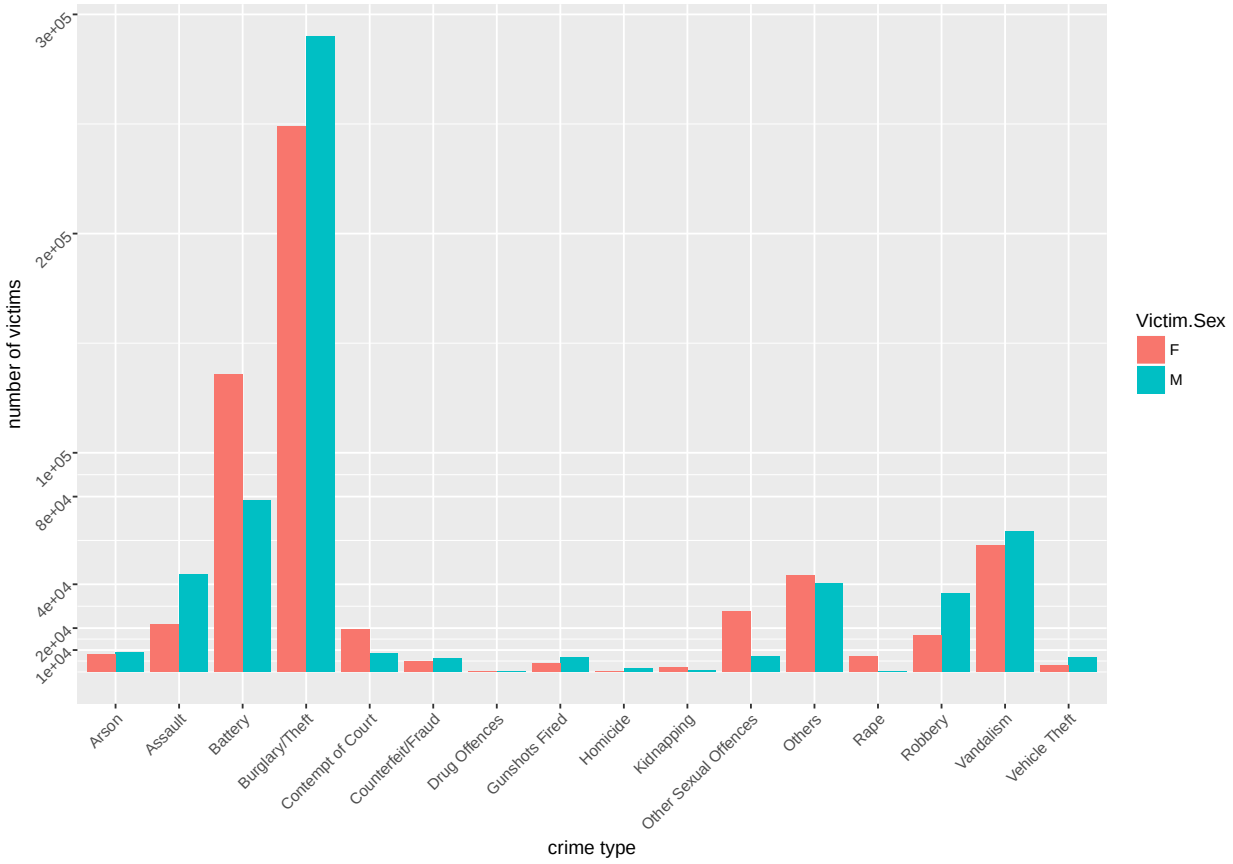
```
## [1] 0.1782434
```

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

We did a proportion comparison with real population data.According to the population of LA,white people was 49.8% 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.

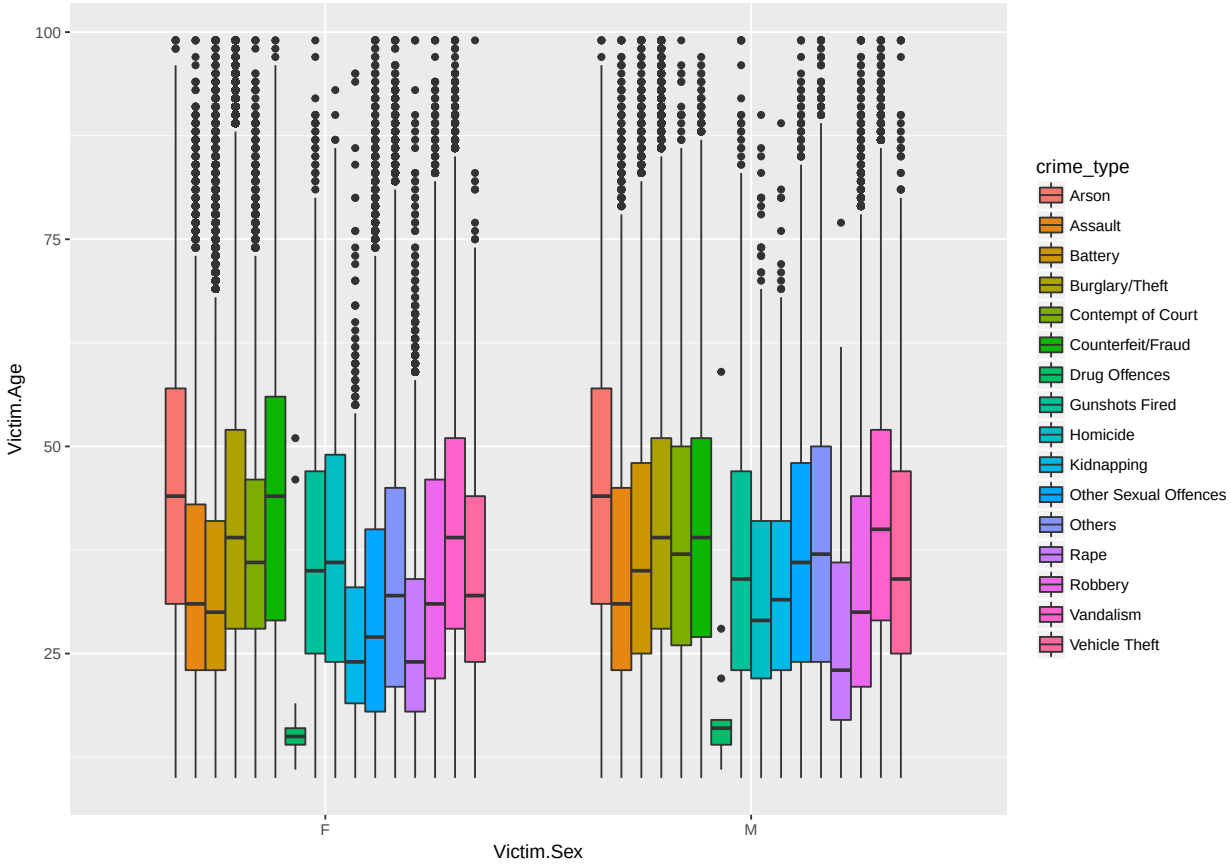
3.3 Correlation between type of crime and age/sex/ethnicity (PLOT + hypothesis Testing !)

3.4 relationship with type 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.

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



4 Area distribution

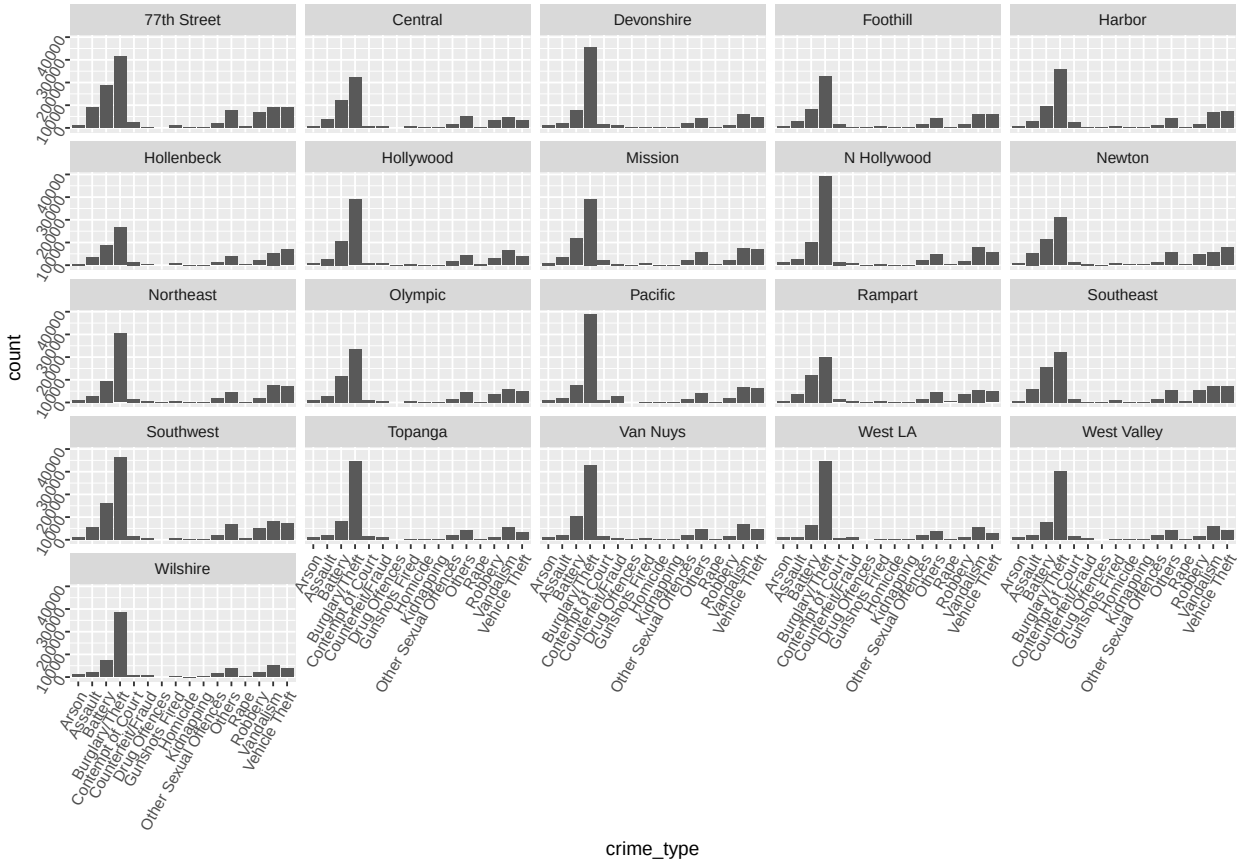
We have seen who are the victims of the crime, now let's focus on the location of these crimes. Of course, we can make the hypothesis that the location depends on the type of crime and on the victim.

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.

Hence, let's delve deeper into the analysis to see the correlation between these two variables.

Let's study more specificaly the distribution of crime between 77th Street and SouthWest.

```
##
##              Arson Assault Battery Burglary/Theft Contempt of Court
##   Hollenbeck   586    3451    8726          16655              1412
##   Topanga     1093    2010    8331          34789              1665
```

```
##
##              Counterfeit/Fraud Gunshots Fired Homicide Kidnapping
##   Hollenbeck               293            715      118        120
##   Topanga                  893            351       32         98
##
##              Other Sexual Offences Others  Rape Robbery Vandalism
##   Hollenbeck                  1228   3742   257    1950      5413
##   Topanga                     1972   4397   298    1295      5716
##
##              Vehicle Theft
##   Hollenbeck          6887
##   Topanga             3427
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(crime_data_77_SW$Area.Name, crime_data_77_SW$crime_type)
## X-squared = 7217.6, df = 14, p-value < 2.2e-16
```

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

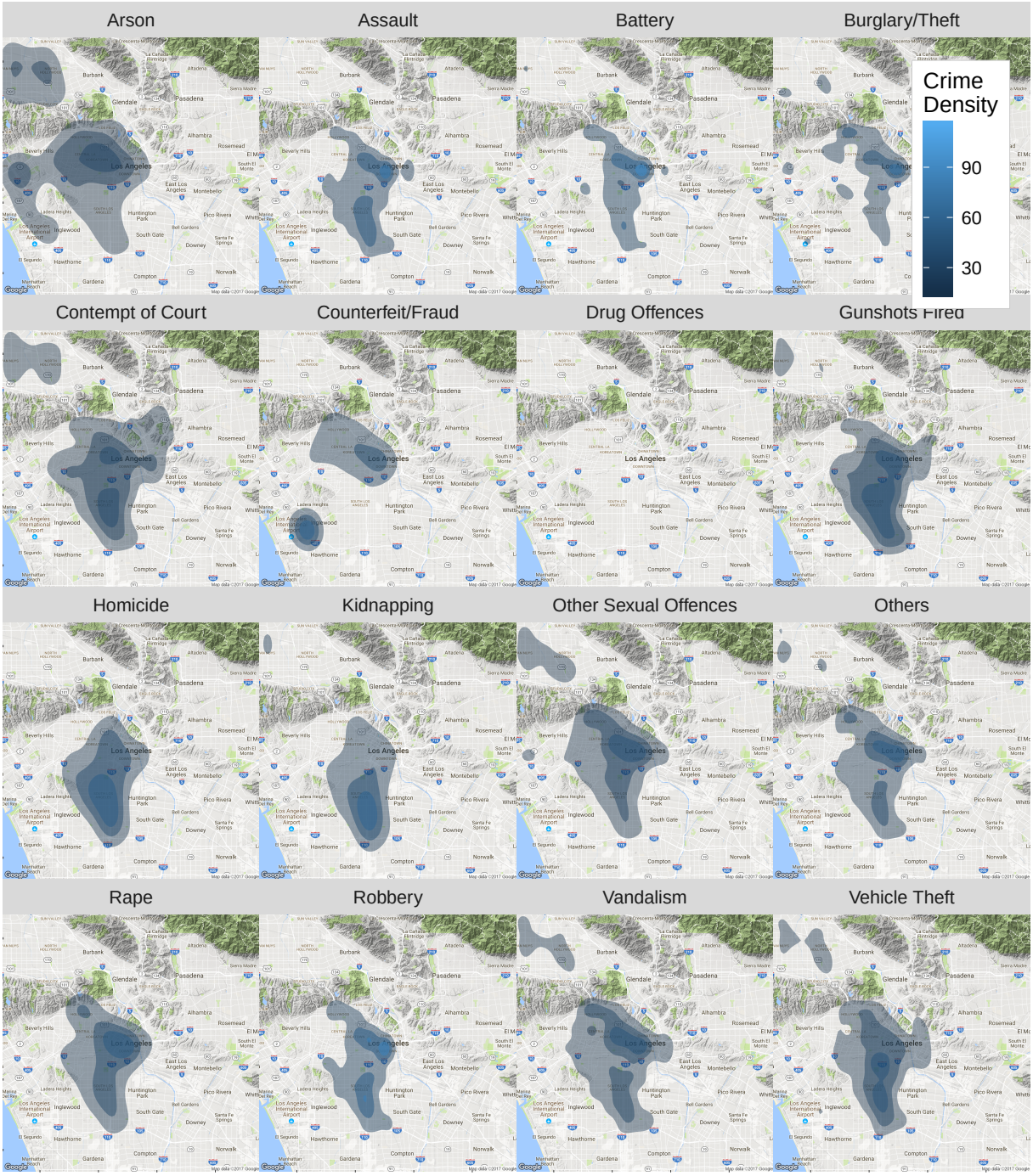
4.2 Area distrubution per type of crime

```
## 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

4.3 Premise of a crime scene per sex

for the main areas with highest crime rate - what are the main types of crime

5 Time

Text here

5.1 Day Time according to type of crime

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

Text here

Hypothesis testing

5.2 Day Time according to sex

Text here

Hypothesis testing

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

Text here

Hypothesis testing

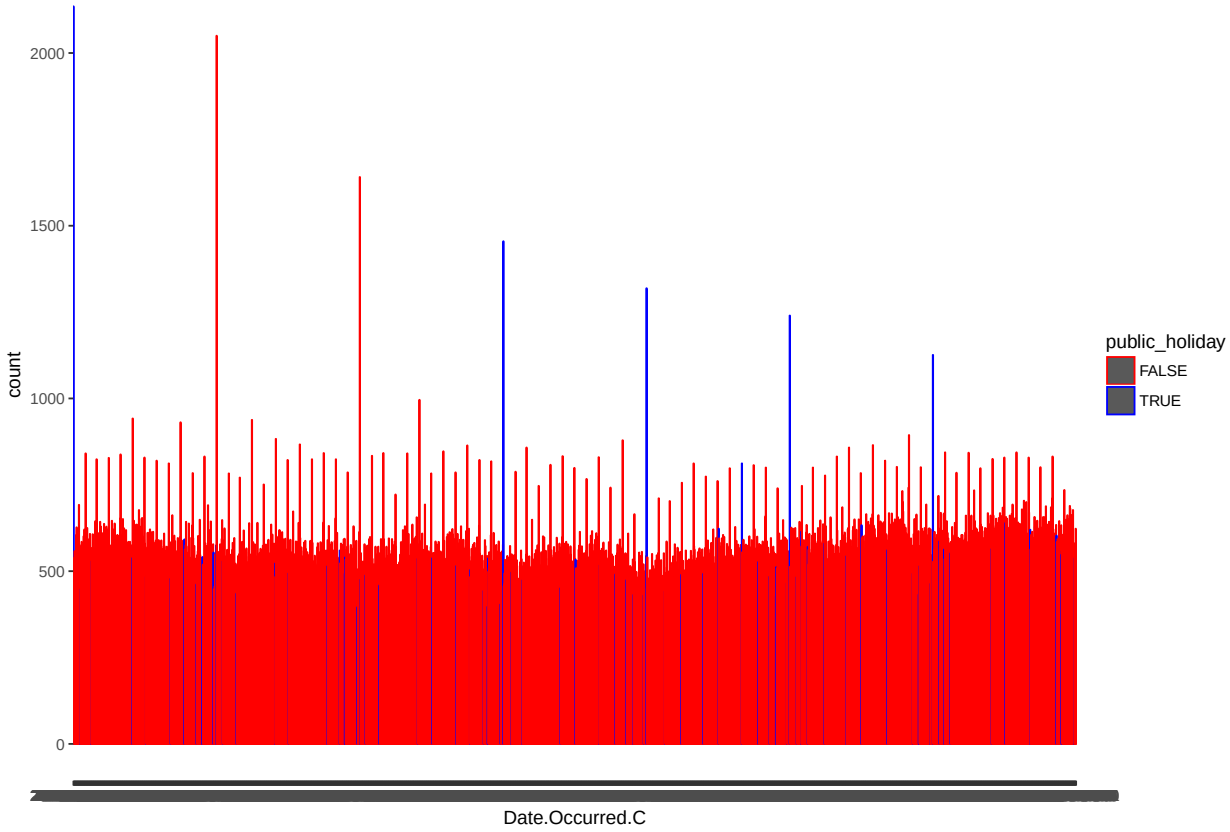
5.4 Correlation between time reported - time occured and case solved or not for a specific month

Text here

Hypothesis testing

5.5 Public holidays impact

Effect of Public Holidays on crime Occurrence and Type.




```
## # A tibble: 2 x 2
##   public_holiday average.crime.per.day
##            <lgl>                 <dbl>
## 1           TRUE              577.4605
## 2          FALSE              559.7916
```

```
## # A tibble: 0 x 2
## # ... with 2 variables: Date.Occurred.C <fctr>, count <int>
```

```
## # A tibble: 1 x 2
##    Jan1 Average.Report.Time
##   <lgl>               <dbl>
## 1 FALSE             17.7277
```

```
## # A tibble: 1 x 1
##   Occured.Reported.Gap.SD
##                     <dbl>
## 1                108.1726
```

Findings from the exploratory plot and tables: i.) The mean of # of crime in LA is higher on public holidays, but ii.) The number of crime is significantly higher on 1st of Jan in every year, and iii.) The length between occured time and reporting time tend to be longer on 1st Jan in every year;

I cannot find information regarding this on the dataLA website, but it looks like they have coded crime incidents occured on other days as 1st Jan for some conditions. So should we exclude cases with reporting time lag for say i.)365 days/ ii. 15 days (which is the average)?

As nature of Sundays is similar to holidays, I will include all Sundays as 'public holidays and Sundays' as the next step.

```
## # A tibble: 2 x 2
##   public_holiday average.crime.per.day
##            <lgl>                 <dbl>
## 1           TRUE              577.4605
## 2          FALSE              559.7916
```

```
## # A tibble: 31 x 3
## # Groups:   crime_type [16]
##           crime_type public_holiday average.crime.per.day
##               <fctr>          <lgl>                 <dbl>
##  1             Arson           TRUE              5.513514
##  2             Arson          FALSE              7.861706
##  3           Assault           TRUE             29.868421
##  4           Assault          FALSE             27.000806
##  5           Battery           TRUE             93.131579
##  6           Battery          FALSE             86.322854
##  7    Burglary/Theft           TRUE            242.210526
##  8    Burglary/Theft          FALSE            236.415961
##  9 Contempt of Court           TRUE             10.605263
## 10 Contempt of Court          FALSE             11.543329
## # ... with 21 more rows
```

Included Sundays and exclude crime occured on 1st Jan and have a time gap between occured and report >15 days

```
## # A tibble: 2 x 2
##   Sundays_public_holiday average.crime.per.day
##                    <lgl>                 <dbl>
## 1                   TRUE              543.5601
## 2                  FALSE              563.8091
```

```
## # A tibble: 32 x 3
## # Groups:   crime_type [16]
##           crime_type Sundays_public_holiday average.crime.per.day
##               <fctr>                  <lgl>                 <dbl>
##  1             Arson                   TRUE              4.508083
##  2             Arson                  FALSE              8.466414
##  3           Assault                   TRUE             33.274376
##  4           Assault                  FALSE             25.796314
##  5           Battery                   TRUE            100.478458
##  6           Battery                  FALSE             83.617202
##  7    Burglary/Theft                   TRUE            204.705215
##  8    Burglary/Theft                  FALSE            243.232987
##  9 Contempt of Court                   TRUE             11.197279
## 10 Contempt of Court                  FALSE             11.581758
## # ... with 22 more rows
```

```
## [1] 563.8091
```

```
## [1] 543.5601
```

```
##
##  Welch Two Sample t-test
##
## data:  Regular_day_crime$count and Holiday_crime$count
## t = 3.0524, df = 497.23, p-value = 0.001196
## alternative hypothesis: true difference in means is greater than 0
```

```
## 95 percent confidence interval:
##  9.31699     Inf
## sample estimates:
## mean of x mean of y
##  563.8091  543.5601
```

Result shows that the average number of crime on holidays and Sundays is lower than that in regular days in 95% significance level.

```
## # A tibble: 32 x 3
## # Groups:   crime_type [?]
##           crime_type Sundays_public_holiday average_per_day
##               <fctr>                 <lgl>           <dbl>
##  1             Arson                 FALSE        8.466414
##  2             Arson                  TRUE        4.508083
##  3           Assault                 FALSE       25.796314
##  4           Assault                  TRUE       33.274376
##  5           Battery                 FALSE       83.617202
##  6           Battery                  TRUE      100.478458
##  7    Burglary/Theft                 FALSE      243.232987
##  8    Burglary/Theft                  TRUE      204.705215
##  9 Contempt of Court                 FALSE       11.581758
## 10 Contempt of Court                  TRUE       11.197279
## # ... with 22 more rows
```

```
##                       Sundays_public_holiday
## crime_type                 FALSE       TRUE
##   Arson                 8.466414   4.508083
##   Assault              25.796314  33.274376
##   Battery              83.617202 100.478458
##   Burglary/Theft      243.232987 204.705215
##   Contempt of Court    11.581758  11.197279
##   Counterfeit/Fraud     6.292718   4.429234
##   Drug Offences         1.523810   1.000000
##   Gunshots Fired        4.901639   5.622426
##   Homicide              1.480076   1.606178
##   Kidnapping            1.748401   1.725352
##   Other Sexual Offences 14.101134  13.732426
##   Others               39.736295  37.160998
##   Rape                  2.980321   3.872038
##   Robbery              22.278828  22.226757
##   Vandalism            52.898393  55.340136
##   Vehicle Theft        46.420605  45.353741
```

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

Data Cleaning

Training set and learning set

Predict the type of crime

First, we try to predict the type of crime according to different parameters such as the age of the potential victim, the sex, the ethnicity and the area

7 Discussion

7.1 What results did you find?

7.2 Why is this interesting?

Help LA police, predict the crime, prevention for the potential victims

7.3 What would be the next steps?