Los Angeles Crime Data Analysis

Zhenjian Wang

2023-10-10

Introduction

Data Source: https://catalog.data.gov/dataset/crime-data-from-2020-to-present (https://catalog.data.gov/dataset/crime-data-from-2020-to-present)

We are analyzing the crime dataset downloaded from DATA.GOV. The dataset includes incidents of crimes in Los Angeles, CA, from 2020 to Present.

As someone who has lived in LA, We are mostly interested in the questions such as what type of people are more likely to be a victim, where the most dangerous place in LA is, and what time most crimes happen. Generally, we want to provide some advice to people who live in LA to help them protect themselves.

```
df = read.csv("Crime_Data_from_2020_to_Present.csv")
str(df)
```

```
## 'data.frame':    811663 obs. of  28 variables:
##  $ DR_NO         : int  10304468 190101086 200110444 191501505 191921269 200100501 200100502
200100504 200100507 201710201 ...
##  $ Date.Rptd     : chr  "01/08/2020 12:00:00 AM" "01/02/2020 12:00:00 AM" "04/14/2020 12:00:0
0 AM" "01/01/2020 12:00:00 AM" ...
##  $ DATE.OCC      : chr  "01/08/2020 12:00:00 AM" "01/01/2020 12:00:00 AM" "02/13/2020 12:00:0
0 AM" "01/01/2020 12:00:00 AM" ...
##  $ TIME.OCC      : int  2230 330 1200 1730 415 30 1315 40 200 1925 ...
##  $ AREA          : int  3 1 1 15 19 1 1 1 1 17 ...
##  $ AREA.NAME     : chr  "Southwest" "Central" "Central" "N Hollywood" ...
##  $ Rpt.Dist.No   : int  377 163 155 1543 1998 163 161 155 101 1708 ...
##  $ Part.1.2      : int  2 2 2 2 2 1 1 2 1 1 ...
##  $ Crm.Cd        : int  624 624 845 745 740 121 442 946 341 341 ...
##  $ Crm.Cd.Desc   : chr  "BATTERY - SIMPLE ASSAULT" "BATTERY - SIMPLE ASSAULT" "SEX OFFENDER R
EGISTRANT OUT OF COMPLIANCE" "VANDALISM - MISDEAMEANOR ($399 OR UNDER)" ...
##  $ Mocodes       : chr  "0444 0913" "0416 1822 1414" "1501" "0329 1402" ...
##  $ Vict.Age      : int  36 25 0 76 31 25 23 0 23 0 ...
##  $ Vict.Sex      : chr  "F" "M" "X" "F" ...
##  $ Vict.Descent  : chr  "B" "H" "X" "W" ...
##  $ Premis.Cd     : int  501 102 726 502 409 735 404 726 502 203 ...
##  $ Premis.Desc   : chr  "SINGLE FAMILY DWELLING" "SIDEWALK" "POLICE FACILITY" "MULTI-UNIT DWE
LLING (APARTMENT, DUPLEX, ETC)" ...
##  $ Weapon.Used.Cd: int  400 500 NA NA NA 500 NA NA NA NA ...
##  $ Weapon.Desc   : chr  "STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)" "UNKNOWN WEAPON/OTHE
R WEAPON" "" "" ...
##  $ Status        : chr  "AO" "IC" "AA" "IC" ...
##  $ Status.Desc   : chr  "Adult Other" "Invest Cont" "Adult Arrest" "Invest Cont" ...
##  $ Crm.Cd.1      : int  624 624 845 745 740 121 442 946 341 341 ...
##  $ Crm.Cd.2      : int  NA NA NA 998 NA 998 998 998 998 NA ...
##  $ Crm.Cd.3      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Crm.Cd.4      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LOCATION      : chr  "1100 W  39TH                         PL" "700 S  HILL
ST" "200 E  6TH                          ST" "5400    CORTEEN                      PL" ...
##  $ Cross.Street  : chr  "" "" "" "" ...
##  $ LAT           : num  34 34 34 34.2 34.2 ...
##  $ LON           : num  -118 -118 -118 -118 -118 ...
```

Our data has 811663 rows and 28 variables. To clarify some column names, DATE.OCC contains dates when the crime occurred, and TIME.OCC represents the time when the crime occurred. AREA is a numerical code representing the AREA.NAME. Crm.Cd is crime code that corresponding to the description column which is Crm.Cd.Desc.

Exploratory Data Analysis

Import two libraries

```
suppressPackageStartupMessages(library(dplyr)) # not show any library conflict error
library(ggplot2)
```

We start by creating a summary of our data.

```
summary(df)
```

```
##      DR_NO            Date.Rptd           DATE.OCC            TIME.OCC
##  Min.   :      817   Length:811663      Length:811663      Min.   :   1
##  1st Qu.:210120732   Class :character   Class :character   1st Qu.: 900
##  Median :220111581   Mode  :character   Mode  :character   Median :1415
##  Mean   :215965221                                         Mean   :1336
##  3rd Qu.:221910224                                         3rd Qu.:1900
##  Max.   :239916487                                         Max.   :2359
##
##       AREA        AREA.NAME          Rpt.Dist.No      Part.1.2
##  Min.   : 1.00   Length:811663      Min.   : 101   Min.   :1.000
##  1st Qu.: 6.00   Class :character   1st Qu.: 622   1st Qu.:1.000
##  Median :11.00   Mode  :character   Median :1142   Median :1.000
##  Mean   :10.71                      Mean   :1118   Mean   :1.414
##  3rd Qu.:16.00                      3rd Qu.:1617   3rd Qu.:2.000
##  Max.   :21.00                      Max.   :2199   Max.   :2.000
##
##      Crm.Cd      Crm.Cd.Desc          Mocodes             Vict.Age
##  Min.   :110.0   Length:811663      Length:811663      Min.   : -3.00
##  1st Qu.:331.0   Class :character   Class :character   1st Qu.:  8.00
##  Median :442.0   Mode  :character   Mode  :character   Median : 31.00
##  Mean   :500.7                                         Mean   : 29.83
##  3rd Qu.:626.0                                         3rd Qu.: 45.00
##  Max.   :956.0                                         Max.   :120.00
##
##    Vict.Sex         Vict.Descent         Premis.Cd     Premis.Desc
##  Length:811663      Length:811663      Min.   :101.0   Length:811663
##  Class :character   Class :character   1st Qu.:101.0   Class :character
##  Mode  :character   Mode  :character   Median :203.0   Mode  :character
##                                        Mean   :305.8
##                                        3rd Qu.:501.0
##                                        Max.   :976.0
##                                        NA's   :9
##  Weapon.Used.Cd   Weapon.Desc           Status           Status.Desc
##  Min.   :101.0    Length:811663      Length:811663      Length:811663
##  1st Qu.:310.0    Class :character   Class :character   Class :character
##  Median :400.0    Mode  :character   Mode  :character   Mode  :character
##  Mean   :362.9
##  3rd Qu.:400.0
##  Max.   :516.0
##  NA's   :528880
##     Crm.Cd.1        Crm.Cd.2         Crm.Cd.3         Crm.Cd.4
##  Min.   :110.0   Min.   :210.0    Min.   :310.0    Min.   :821.0
##  1st Qu.:331.0   1st Qu.:998.0    1st Qu.:998.0    1st Qu.:998.0
##  Median :442.0   Median :998.0    Median :998.0    Median :998.0
##  Mean   :500.5   Mean   :957.5    Mean   :983.6    Mean   :990.8
##  3rd Qu.:626.0   3rd Qu.:998.0    3rd Qu.:998.0    3rd Qu.:998.0
##  Max.   :956.0   Max.   :999.0    Max.   :999.0    Max.   :999.0
##  NA's   :10      NA's   :751848   NA's   :809663   NA's   :811603
##    LOCATION         Cross.Street            LAT             LON
##  Length:811663      Length:811663      Min.   : 0.00   Min.   :-118.7
##  Class :character   Class :character   1st Qu.:34.01   1st Qu.:-118.4
##  Mode  :character   Mode  :character   Median :34.06   Median :-118.3
##                                         Mean   :33.98   Mean   :-118.0  
##                                         3rd Qu.:34.16   3rd Qu.:-118.3  
##                                         Max.   :34.33   Max.   :   0.0  
## 
```

The Age doesn’t look correct because it includes a negative number, so we want to filter any invalid number out. Considering victims may be infants, we set age >= 0.

```
df <- df %>% filter(Vict.Age >= 0)
min(df$Vict.Age)
```

```
## [1] 0
```

Since we have so many columns, we want to focus on the columns that interest us the most.

```
mean_age <- mean(df$Vict.Age, na.rm=TRUE)
sd_age <- sd(df$Vict.Age, na.rm=TRUE)
var_age <- var(df$Vict.Age, na.rm=TRUE)

cat("Mean Age:", mean_age, "\n")
```

```
## Mean Age: 29.83075
```

```
cat("Standard Deviation of Age:", sd_age, "\n")
```

```
## Standard Deviation of Age: 21.76864
```

```
cat("Variance of Age:", var_age, "\n")
```

```
## Variance of Age: 473.8735
```

```
# Frequency table for victim's gender
df <- df %>%
  filter(Vict.Sex %in% c("F", "M"))

gender_table <- table(df$Vict.Sex)
gender_table_percent <- prop.table(gender_table) # we want the percentage to make it more clear
gender_table_summary <- cbind(Number = gender_table, Percentage = gender_table_percent)
print(gender_table_summary)
```

```
##   Number Percentage
## F 299079  0.4713676
## M 335413  0.5286324
```

```
# Create a frequency table for crime descriptions
crime_table <- table(df$Crm.Cd.Desc)

# Sort the table in descending order by frequency
sorted_crime_table <- sort(crime_table, decreasing = TRUE)

# Get the top 10 crime categories
print(sorted_crime_table[1:10])
```

```
## 
##                                 BATTERY - SIMPLE ASSAULT 
##                                                    64175 
##                                        THEFT OF IDENTITY 
##                                                    50849 
##                                    BURGLARY FROM VEHICLE 
##                                                    48674 
##           ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT 
##                                                    45820 
##                        INTIMATE PARTNER - SIMPLE ASSAULT 
##                                                    40833 
## VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 
##                                                    40038 
##                                                 BURGLARY 
##                                                    38060 
##                       THEFT PLAIN - PETTY ($950 & UNDER) 
##                                                    37025 
##      THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER) 
##                                                    28441 
##                                                  ROBBERY 
##                                                    23238
```

```
# Contingency table for Victim Gender vs. Age
gender_Age_table <- table(df$Vict.Sex, df$Vict.Age)
print(gender_Age_table)
```

```
##    
##         0     2     3     4     5     6     7     8     9    10    11    12
##   F  3351   160   213   228   258   260   267   273   331   387   618   929
##   M 26930   182   204   197   227   226   226   231   254   276   400   563
##    
##        13    14    15    16    17    18    19    20    21    22    23    24
##   F  1265  1416  1813  1995  2037  2683  3941  4903  5637  6607  7177  7962
##   M   780  1040  1107  1323  1573  2363  3311  4230  4713  5315  6044  6934
##    
##        25    26    27    28    29    30    31    32    33    34    35    36
##   F  8566  8641  8742  9207  9059  9364  9090  8781  8471  7898  7590  7347
##   M  7265  7409  7928  8164  8570  9106  8594  8434  8238  8024 10489  7685
##    
##        37    38    39    40    41    42    43    44    45    46    47    48
##   F  7124  6764  6436  6254  5773  5488  5442  5200  5037  4583  4586  4324
##   M  7244  7281  6994  6973  6564  6041  5936  5578  5531  5212  5248  4883
##    
##        49    50    51    52    53    54    55    56    57    58    59    60
##   F  4354  4377  4282  4064  3830  3606  3542  3297  3217  3163  3024  2832
##   M  4883  6320  4915  4847  4830  4667  4527  4273  4306  4033  3949  3929
##    
##        61    62    63    64    65    66    67    68    69    70    71    72
##   F  2595  2567  2312  2155  2014  1911  1707  1594  1385  1245  1222  1164
##   M  3623  3380  3290  3102  2768  2521  2252  1916  1751  1666  1423  1330
##    
##        73    74    75    76    77    78    79    80    81    82    83    84
##   F  1056   923   830   748   678   647   522   490   411   441   356   276
##   M  1226  1066   849   804   696   593   531   472   393   334   304   223
##    
##        85    86    87    88    89    90    91    92    93    94    95    96
##   F   285   226   174   153   154   123   127    88    69    58    48    51
##   M   227   199   143   142   103   105    72    55    38    33    28    32
##    
##        97    98    99
##   F    36    31   141
##   M    23    29   152
```

We want to further break down the age by using gender variable.

```
# Mean age by gender
mean_age_by_gender <- aggregate(Vict.Age ~ Vict.Sex, data=df, FUN=mean, na.rm=TRUE)
print(mean_age_by_gender)
```

```
##   Vict.Sex Vict.Age
## 1        F 38.28739
## 2        M 37.49951
```

```
# Standard deviation of age by gender
sd_age_by_gender <- aggregate(Vict.Age ~ Vict.Sex, data=df, FUN=sd, na.rm=TRUE)
print(sd_age_by_gender)
```

```
##   Vict.Sex Vict.Age
## 1        F 16.05367
## 2        M 18.46730
```

```
# Combine mean and standard deviation tables
combined_stats <- cbind(mean_age_by_gender, sd=sd_age_by_gender$Vict.Age)
print(combined_stats)
```

```
##   Vict.Sex Vict.Age       sd
## 1        F 38.28739 16.05367
## 2        M 37.49951 18.46730
```

We could see from the table that, most of the male victims’ ages fall between 19-55(1 sd), and most of the female victims’ ages fall between 22-54. The mean age of male victims is 37.5, and the mean age of female victims is 38.3. From the contingency table, we could see that, among men, 35 years old is the mode. Among women, 30 years old is the mode.

```
# Get top 5 crimes for men
top5_men <- df %>%
  filter(Vict.Sex == "M") %>%
  count(Crm.Cd.Desc) %>%
  mutate(percentage = n / sum(n) * 100) %>%
  arrange(-n) %>%
  head(5)

# Get top 5 crimes for women
top5_women <- df %>%
  filter(Vict.Sex == "F") %>%
  count(Crm.Cd.Desc) %>%
  mutate(percentage = n / sum(n) * 100) %>%
  arrange(-n) %>%
  head(5)

# Display the results
print("Top 5 crimes among men:")
```

```
## [1] "Top 5 crimes among men:"
```

```
print(top5_men)
```

```
##                                               Crm.Cd.Desc     n percentage
## 1                                BATTERY - SIMPLE ASSAULT 33851  10.092334
## 2          ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT 33352   9.943562
## 3                                   BURGLARY FROM VEHICLE 27949   8.332712
## 4                                                BURGLARY 24498   7.303831
## 5 VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 22923   6.834261
```

```
print("Top 5 crimes among women:")
```

```
## [1] "Top 5 crimes among women:"
```

```
print(top5_women)
```

```
##                           Crm.Cd.Desc     n percentage
## 1  INTIMATE PARTNER - SIMPLE ASSAULT 30998  10.364486
## 2           BATTERY - SIMPLE ASSAULT 30324  10.139127
## 3                  THEFT OF IDENTITY 29984  10.025445
## 4              BURGLARY FROM VEHICLE 20725   6.929607
## 5 THEFT PLAIN - PETTY ($950 & UNDER) 17742   5.932212
```

We could see that men and women suffer from different crimes. The biggest number of crimes women suffer from is intimate partner - simple assault, which means domestic violence. For men, we could see that the top 5 crime categories include ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT, which is not included in the top 5 list among women.

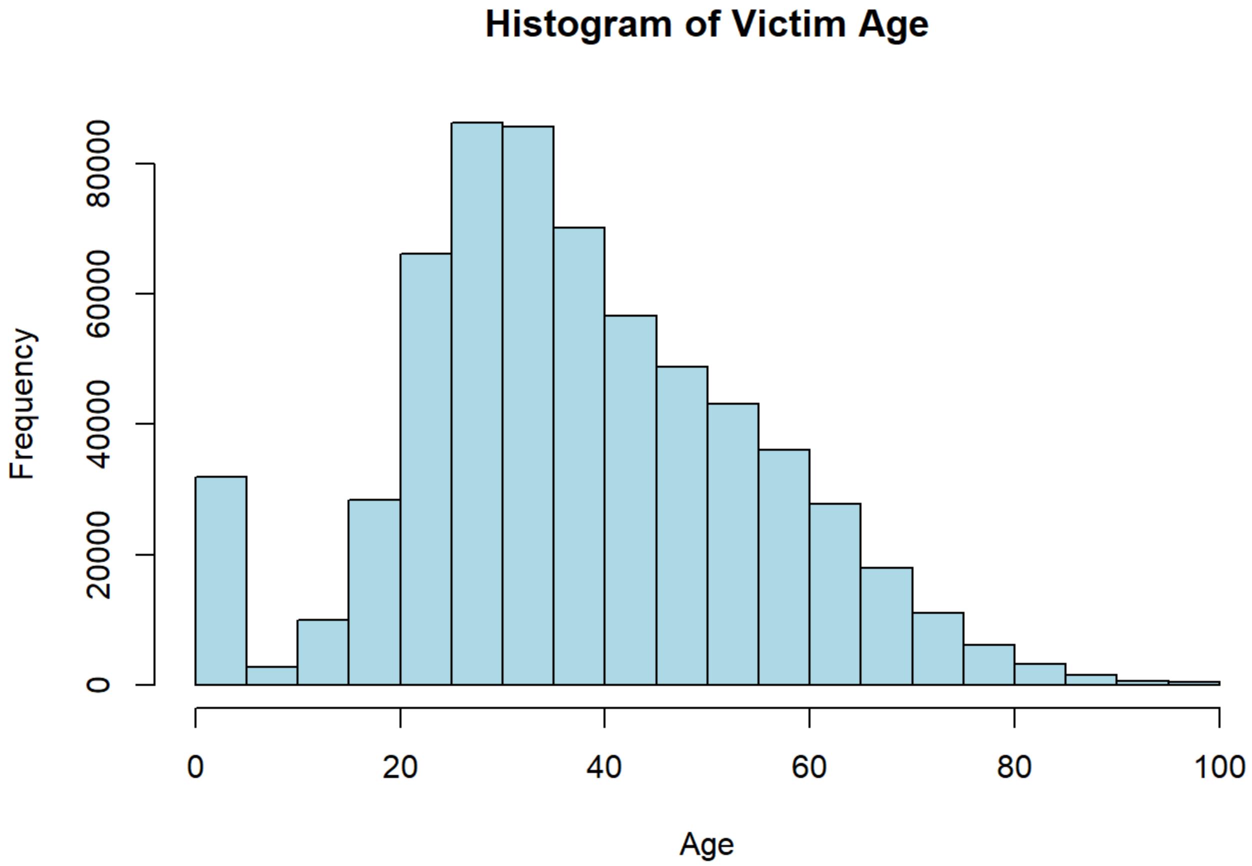
```
# Percentage of crimes where victim's age is below 18
percentage_minors <- df %>%
  summarize(percentage = mean(Vict.Age < 18) * 100) %>%
  pull(percentage)

print(paste("Percentage of victims under 18 among all victims:", round(percentage_minors, 2),
"%"))
```

```
## [1] "Percentage of victims under 18 among all victims: 8.12 %"
```

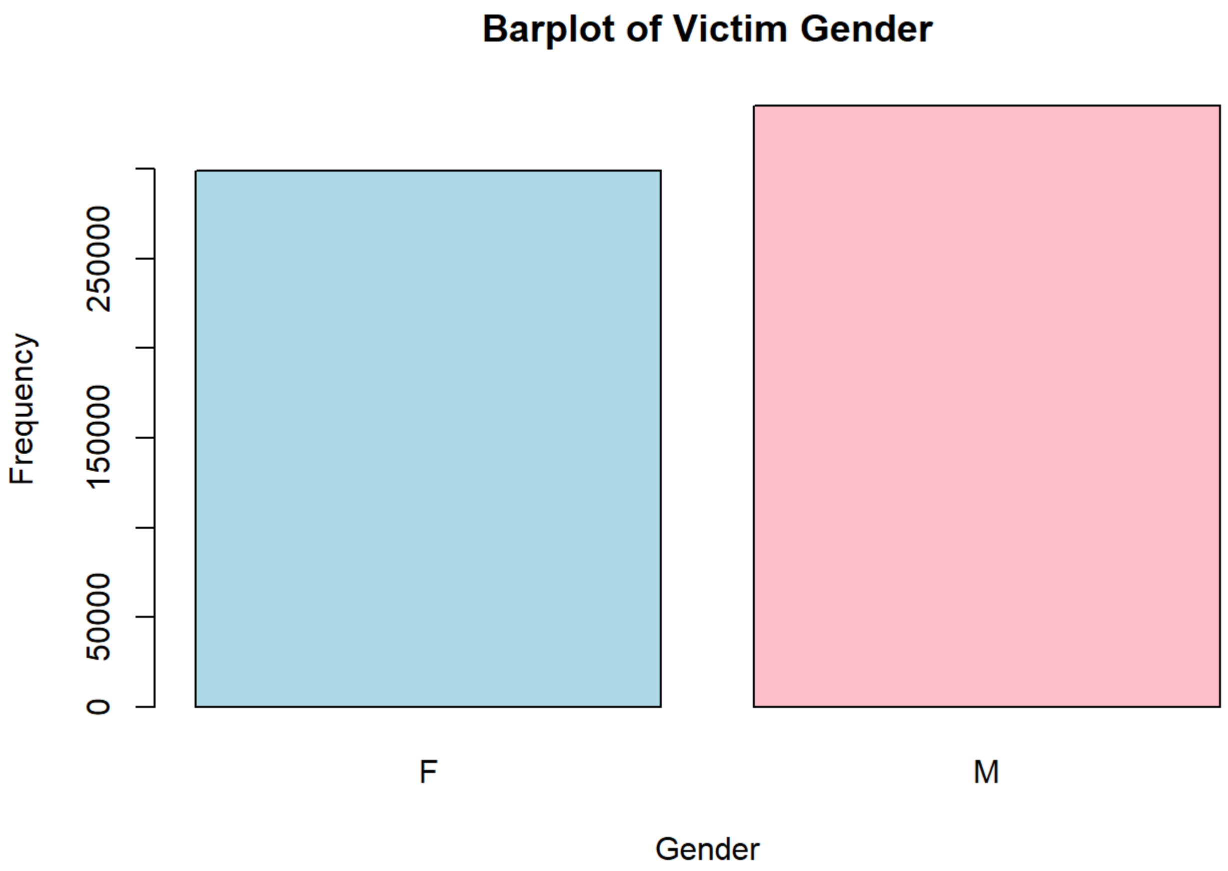
Graphical EDA

A more visual presentation of the distributions of different variables.

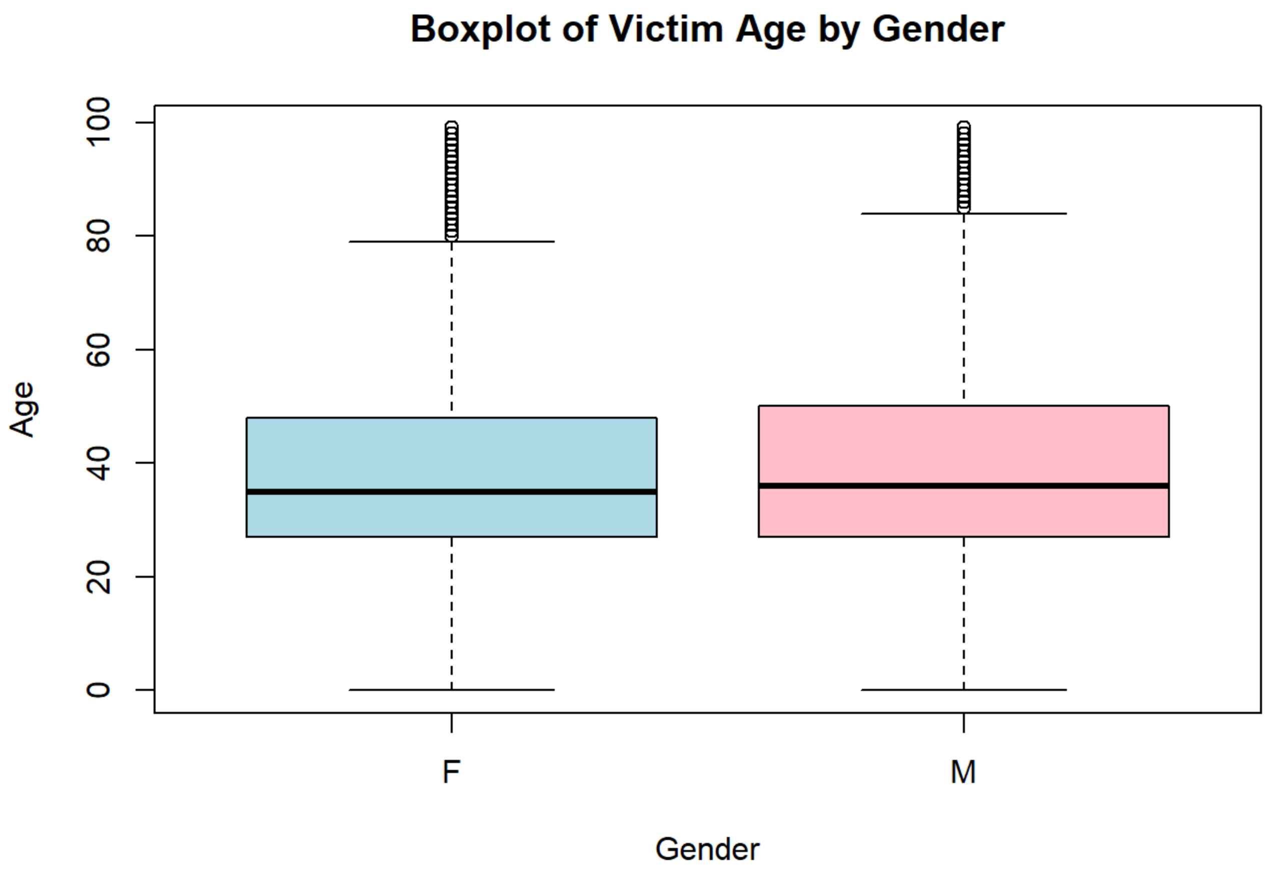
```
hist(df$Vict.Age, main="Histogram of Victim Age", xlab="Age", ylab="Frequency", col="lightblue",
border="black")
```



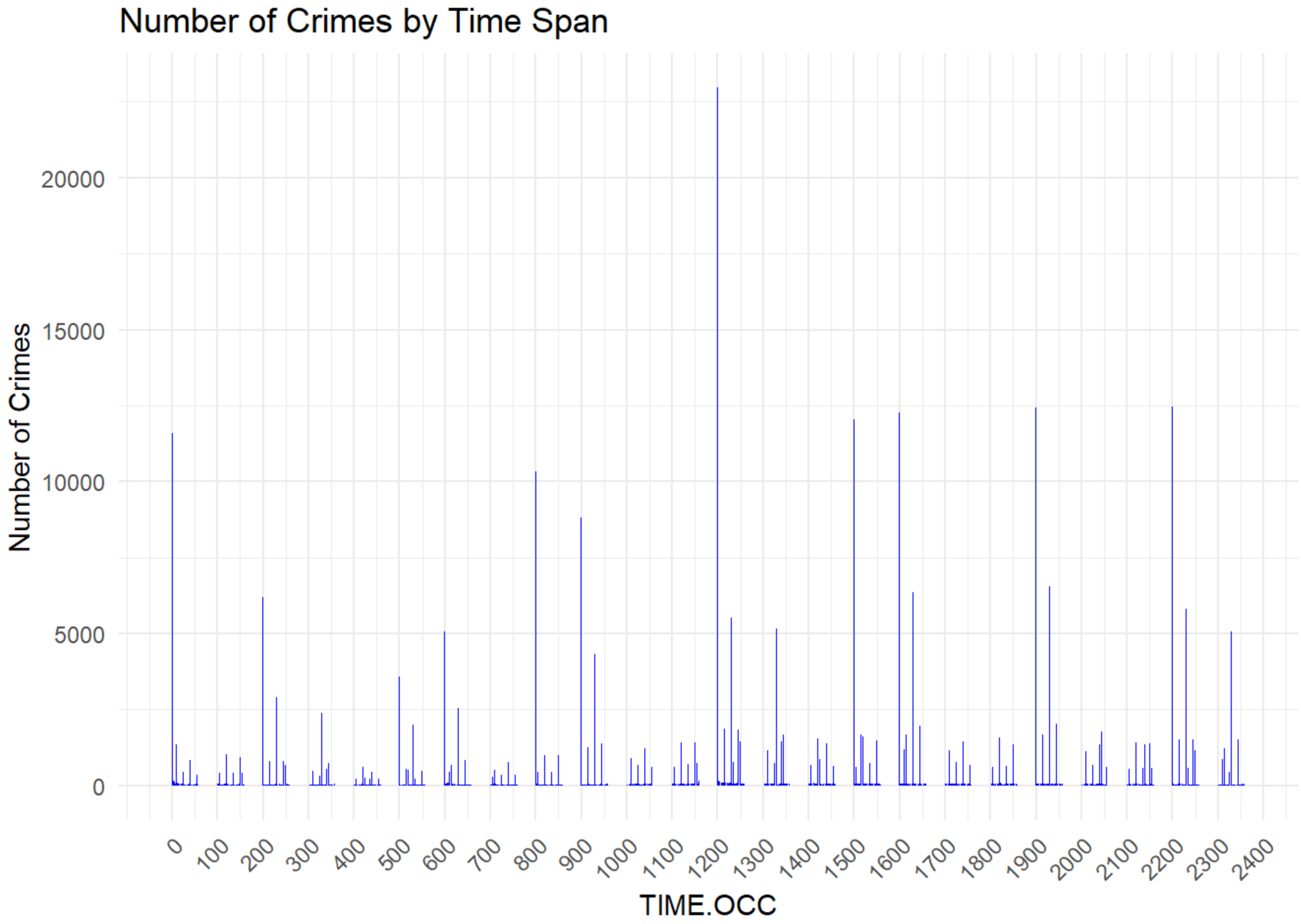
```
gender_freq <- table(df$Vict.Sex)
barplot(gender_freq, main="Barplot of Victim Gender", xlab="Gender", ylab="Frequency", col=c("li
ghtblue", "pink"), border="black")
```



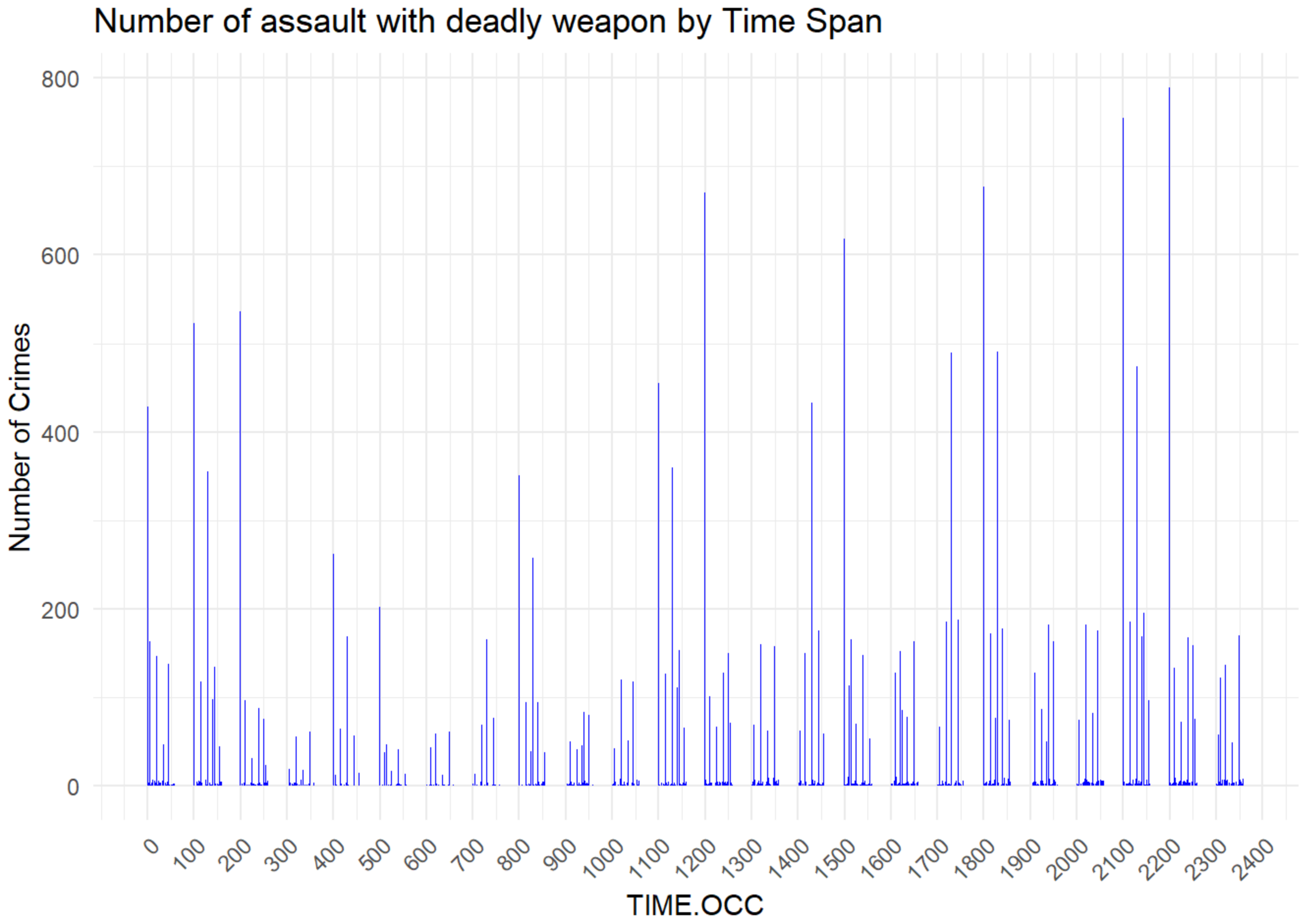
```
boxplot(Vict.Age ~ Vict.Sex, data=df, main="Boxplot of Victim Age by Gender", xlab="Gender", ylab="Age", col=c("lightblue", "pink"))
```



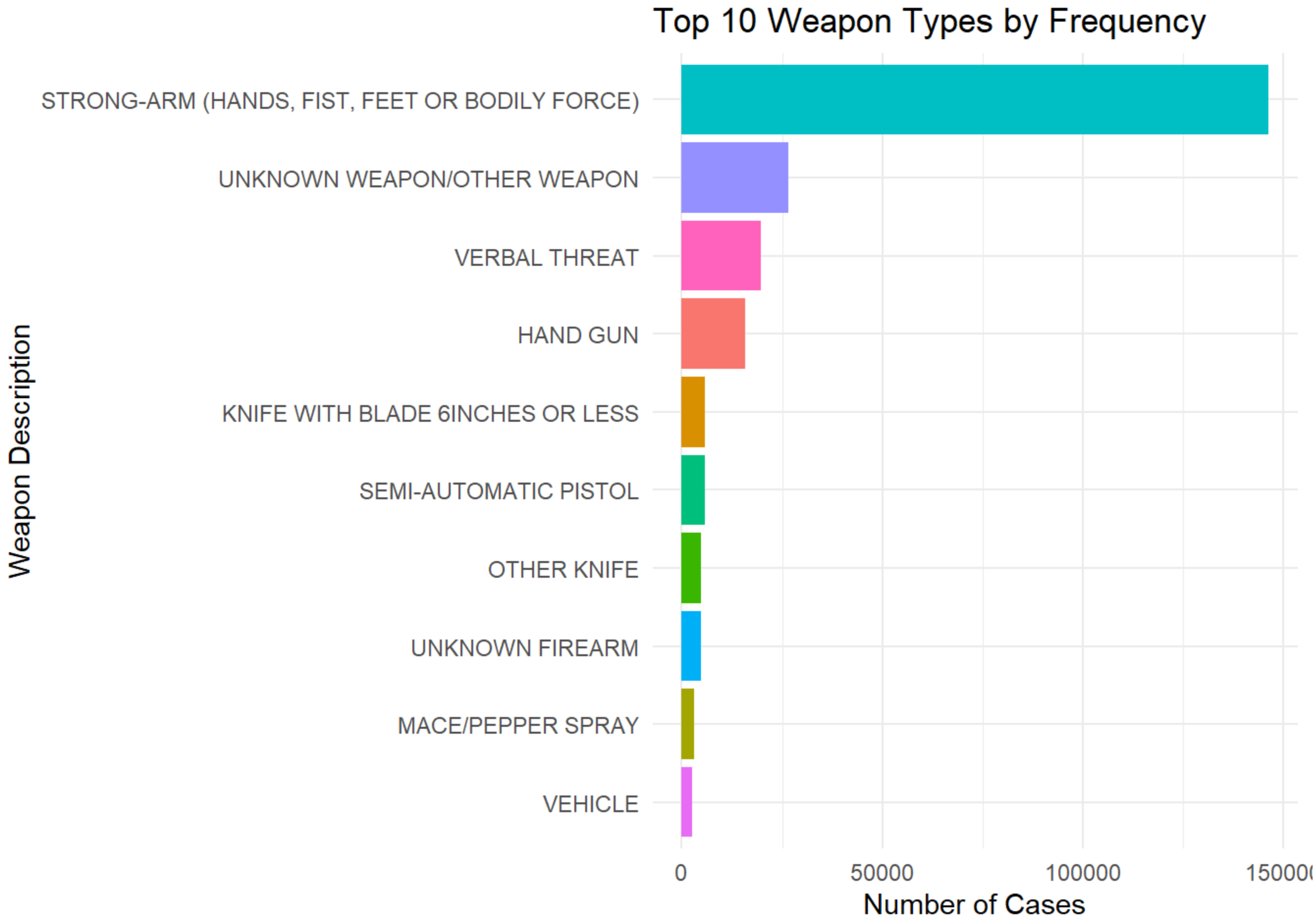
```
# Aggregate the number of crimes by TIME.OCC
crime_counts <- df %>%
  group_by(TIME.OCC) %>%
  summarise(count = n())

# Plot the data
ggplot(crime_counts, aes(x = TIME.OCC, y = count)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Number of Crimes by Time Span", x = "TIME.OCC", y = "Number of Crimes") +
  scale_x_continuous(breaks = seq(0, 2400, by = 100)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# filter only the deadly crime
filtered_data <- df %>%
  filter(Crm.Cd.Desc == "ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT")

# Aggregate the number of crimes by TIME.OCC
crime_counts <- filtered_data %>%
  group_by(TIME.OCC) %>%
  summarise(count = n())

# Plot the data
ggplot(crime_counts, aes(x = TIME.OCC, y = count)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Number of assault with deadly weapon by Time Span", x = "TIME.OCC", y = "Number
of Crimes") +
  scale_x_continuous(breaks = seq(0, 2400, by = 100)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# we exclude the null here because there are many null or empty values in Weapon.Desc
top_weapons <- df %>%
  filter(!is.na(Weapon.Desc) & Weapon.Desc != "") %>%
  group_by(Weapon.Desc) %>%
  summarize(count = n()) %>%
  arrange(-count) %>%
  head(10)

ggplot(top_weapons, aes(y = reorder(Weapon.Desc, count), x = count)) +
  geom_bar(stat = "identity", aes(fill = Weapon.Desc)) +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "Top 10 Weapon Types by Frequency", x = "Number of Cases", y = "Weapon Descriptio
n")
```



Now we have a deeper insight on the time occurrence of crimes. We could see that time around 10:00, 14:00, 17:00, 21:00 have the most number of crimes. This indicates that crimes could not only happen at night, but indeed happen a lot during day time. However, we also find out that the number of assault with deadly weapon cases increase suddenly after 16:00, meaning that people in LA are more likely to encounter severe assault in evening or at night.

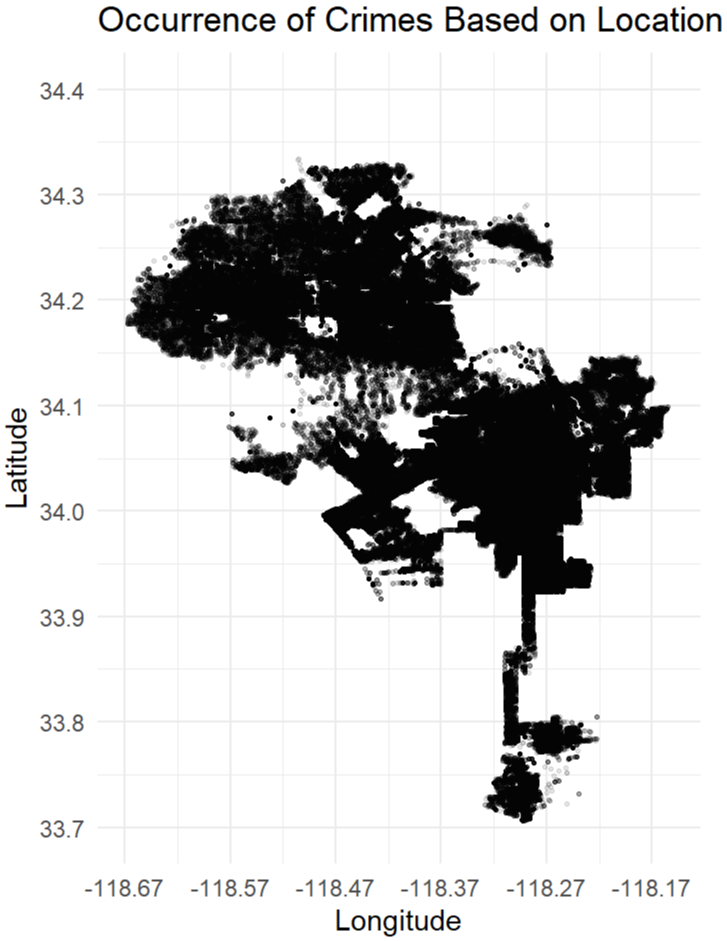
```
# filter out outliers based on https://www.distancesto.com/coordinates/us/los-angeles-latitude-l
ongitude/history/171.html
df_filtered <- df %>%
  filter(LAT > 33 & LAT < 35 & LON > -119 & LON < -118)

lat_min <- min(df_filtered$LAT, na.rm = TRUE)
lat_max <- max(df_filtered$LAT, na.rm = TRUE)

lon_min <- min(df_filtered$LON, na.rm = TRUE)
lon_max <- max(df_filtered$LON, na.rm = TRUE)

cat("Latitude Range: ", lat_min, " to ", lat_max, "\n")
```

```
## Latitude Range:  33.7061  to  34.3343
```

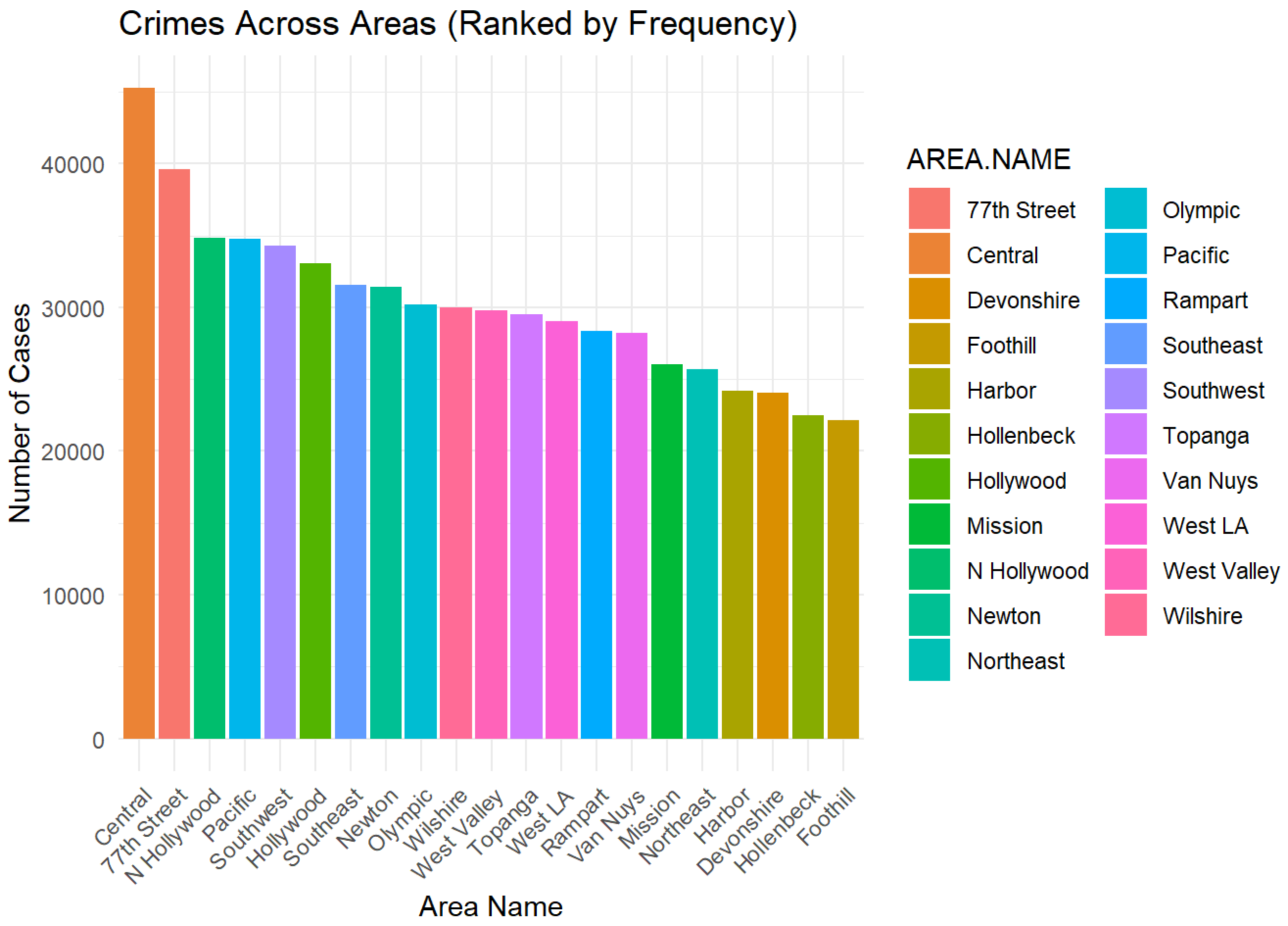
```
cat("Longitude Range: ", lon_min, " to ", lon_max, "\n")
```

```
## Longitude Range:  -118.6676  to  -118.1554
```

```
# Plotting the crime data based on the LAT and LON

ggplot(df_filtered, aes(x = LON, y = LAT)) +
  geom_point(alpha = 0.1, size = 0.5) +
  labs(title = "Occurrence of Crimes Based on Location",
       x = "Longitude",
       y = "Latitude") +
  theme_minimal() +
  coord_fixed(ratio = 1) +
  scale_x_continuous(limits = c(-118.67, -118.15), breaks = seq(-118.67, -118.15, 0.1)) +
  scale_y_continuous(limits = c(33.7, 34.4), breaks = seq(33.7, 34.4, 0.1))
```



```
areas_ranked <- df %>%
  group_by(AREA.NAME) %>%
  summarize(count = n()) %>%
  arrange(-count)

ggplot(areas_ranked, aes(x = reorder(AREA.NAME, -count), y = count)) +
  geom_bar(stat = "identity", aes(fill = AREA.NAME)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Crimes Across Areas (Ranked by Frequency)", y = "Number of Cases", x = "Area Nam
e")
```



we could see that crimes are mostly in the northwest side and the middle east side. These two areas of the city are most dangerous. The Crimes Across Areas further validates our point, where Central, 77th Street, and N Hollywood have the most crimes in LA.

Conclusion

Our data analysis of the crime dataset for Los Angeles from 2020 to the present has yielded several findings:

Victim Age and Gender Distribution:

The most commonly reported age for victims is between 19-55 years. Specifically, 35 years is the mode for males, while 30 years is the mode for females. On average, male victims are approximately 37.5 years old, and female victims are 38.3 years old.

Crimes affecting males and females differ. The dominant crime affecting females is intimate partner - simple assault, indicating serious domestic violence issues in LA. In contrast, males prominently face crimes categorized as assault with deadly weapon, a type which doesn’t rank in the top 5 for females.

Timing of Crimes:

A general assumption might be that crimes mostly occur during the night. However, this analysis has shown that this might not be the case in LA. Crimes are frequently reported around 10:00, 14:00, 17:00, and 21:00, suggesting significant daytime criminal activity.

Yet, it is important to note that the number of severe crimes like assaults with deadly weapons rises after 16:00, indicating that evenings and nights remain very dangerous.

Spatial Distribution:

After eliminating outliers based on latitude and longitude, it is clear that crime is not uniformly distributed across Los Angeles. Concentrations of criminal activities are densest in the northwest and middle eastern regions of the city. People in LA should be particularly cautious in these area, especially Central, 77th Street, and N Hollywood.

This spatial insight can also be used by the police department or community welfare groups. They can utilize this data for patrolling, installing security infrastructure, and launching community awareness programs.

In summary, while Los Angeles is a beautiful city that attracts many tourists and workers, this analysis shows the importance of being aware and prepared. Insights derived can help people to protect themselves in the city.