Crime EDA

AUTHOR

Adithya Venghatesan

Dataset: https://www.kaggle.com/datasets/asaniczka/crimes-in-los-angeles-2020-2023

Load required Libraries

```r
library(dplyr)
library(ggplot2)
library(lubridate)
library(caret)
library(leaflet)
library(plotly)
```

Load the dataset

```r
crimeDatacsv <- read.csv("LA Crime Data.csv")
```

```r
summary(crimeDatacsv)
```

```
     DR_NO            Date.Rptd           DATE.OCC            TIME.OCC
 Min.   :      817   Length:944235      Length:944235      Min.   :   1
 1st Qu.:210508145   Class :character   Class :character   1st Qu.: 900
 Median :220620262   Mode  :character   Mode  :character   Median :1419
 Mean   :218877718                                         Mean   :1338
 3rd Qu.:230706134                                         3rd Qu.:1900
 Max.   :249913791                                         Max.   :2359

      AREA       AREA.NAME          Rpt.Dist.No       Part.1.2
 Min.   : 1.00   Length:944235      Min.   : 101   Min.   :1.00
 1st Qu.: 6.00   Class :character   1st Qu.: 622   1st Qu.:1.00
 Median :11.00   Mode  :character   Median :1142   Median :1.00
 Mean   :10.72                      Mean   :1119   Mean   :1.41
 3rd Qu.:16.00                      3rd Qu.:1619   3rd Qu.:2.00
 Max.   :21.00                      Max.   :2199   Max.   :2.00

     Crm.Cd      Crm.Cd.Desc          Mocodes             Vict.Age
 Min.   :110.0   Length:944235      Length:944235      Min.   : -4.0
 1st Qu.:331.0   Class :character   Class :character   1st Qu.:  0.0
 Median :442.0   Mode  :character   Mode  :character   Median : 30.0
 Mean   :500.8                                         Mean   : 29.5
 3rd Qu.:626.0                                         3rd Qu.: 45.0
 Max.   :956.0                                         Max.   :120.0

   Vict.Sex         Vict.Descent          Premis.Cd     Premis.Desc
 Length:944235      Length:944235      Min.   :101.0   Length:944235
 Class :character   Class :character   1st Qu.:101.0   Class :character
 Mode  :character   Mode  :character   Median :203.0   Mode  :character
                                       Mean   :306.6
                                       3rd Qu.:501.0
                                       Max.   :976.0
                                       NA's   :10
 Weapon.Used.Cd   Weapon.Desc          Status            Status.Desc
 Min.   :101.0    Length:944235      Length:944235      Length:944235
 1st Qu.:311.0    Class :character   Class :character   Class :character
 Median :400.0    Mode  :character   Mode  :character   Mode  :character
 Mean   :363.7
 3rd Qu.:400.0
 Max.   :516.0
 NA's   :619758
    Crm.Cd.1        Crm.Cd.2         Crm.Cd.3         Crm.Cd.4
```

```
Min.   :110.0   Min.   :210.0    Min.   :310      Min.   :821.0
1st Qu.:331.0   1st Qu.:998.0    1st Qu.:998      1st Qu.:998.0
Median :442.0   Median :998.0    Median :998      Median :998.0
Mean   :500.6   Mean   :958.1    Mean   :984      Mean   :991.2
3rd Qu.:626.0   3rd Qu.:998.0    3rd Qu.:998      3rd Qu.:998.0
Max.   :956.0   Max.   :999.0    Max.   :999      Max.   :999.0
NA's   :11      NA's   :875977   NA's   :941954   NA's   :944171
  LOCATION         Cross.Street            LAT              LON
Length:944235      Length:944235      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   :33.99   Mean   :-118.1
                                      3rd Qu.:34.16   3rd Qu.:-118.3
                                      Max.   :34.33   Max.   :   0.0
```

```
unique(crimeDatacsv$Crm.Cd.Desc)
```

```
 [1] "VEHICLE - STOLEN"
 [2] "BURGLARY FROM VEHICLE"
 [3] "BIKE - STOLEN"
 [4] "SHOPLIFTING-GRAND THEFT ($950.01 & OVER)"
 [5] "THEFT OF IDENTITY"
 [6] "BATTERY - SIMPLE ASSAULT"
 [7] "SODOMY/SEXUAL CONTACT B/W PENIS OF ONE PERS TO ANUS OTH"
 [8] "CRM AGNST CHLD (13 OR UNDER) (14-15 & SUSP 10 YRS OLDER)"
 [9] "SEX,UNLAWFUL(INC MUTUAL CONSENT, PENETRATION W/ FRGN OBJ"
[10] "ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT"
[11] "LETTERS, LEWD  -  TELEPHONE CALLS, LEWD"
[12] "THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD"
[13] "CRIMINAL THREATS - NO WEAPON DISPLAYED"
[14] "EMBEZZLEMENT, GRAND THEFT ($950.01 & OVER)"
[15] "THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)"
[16] "CHILD ANNOYING (17YRS & UNDER)"
[17] "BURGLARY"
[18] "CONTEMPT OF COURT"
[19] "THEFT PLAIN - PETTY ($950 & UNDER)"
[20] "INTIMATE PARTNER - SIMPLE ASSAULT"
[21] "LEWD CONDUCT"
[22] "THEFT PLAIN - ATTEMPT"
[23] "THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)"
[24] "ROBBERY"
[25] "BUNCO, GRAND THEFT"
[26] "BATTERY WITH SEXUAL CONTACT"
[27] "INTIMATE PARTNER - AGGRAVATED ASSAULT"
[28] "ORAL COPULATION"
[29] "UNAUTHORIZED COMPUTER ACCESS"
[30] "VIOLATION OF RESTRAINING ORDER"
[31] "SHOPLIFTING - PETTY THEFT ($950 & UNDER)"
[32] "VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)"
[33] "OTHER MISCELLANEOUS CRIME"
[34] "BRANDISH WEAPON"
[35] "DOCUMENT FORGERY / STOLEN FELONY"
[36] "SEX OFFENDER REGISTRANT OUT OF COMPLIANCE"
[37] "RAPE, FORCIBLE"
[38] "VANDALISM - MISDEAMEANOR ($399 OR UNDER)"
[39] "CHILD ABUSE (PHYSICAL) - SIMPLE ASSAULT"
[40] "CREDIT CARDS, FRAUD USE ($950.01 & OVER)"
[41] "THREATENING PHONE CALLS/LETTERS"
[42] "SEXUAL PENETRATION W/FOREIGN OBJECT"
[43] "EXTORTION"
[44] "OTHER ASSAULT"
[45] "PICKPOCKET"
[46] "ARSON"
[47] "DISTURBING THE PEACE"
[48] "BUNCO, ATTEMPT"
[49] "HUMAN TRAFFICKING - INVOLUNTARY SERVITUDE"
```

```
 [50] "PEEPING TOM"
 [51] "VIOLATION OF COURT ORDER"
 [52] "FALSE POLICE REPORT"
 [53] "CONTRIBUTING"
 [54] "FALSE IMPRISONMENT"
 [55] "CHILD ABUSE (PHYSICAL) - AGGRAVATED ASSAULT"
 [56] "ATTEMPTED ROBBERY"
 [57] "CREDIT CARDS, FRAUD USE ($950 & UNDER"
 [58] "CHILD STEALING"
 [59] "LEWD/LASCIVIOUS ACTS WITH CHILD"
 [60] "EMBEZZLEMENT, PETTY THEFT ($950 & UNDER)"
 [61] "INDECENT EXPOSURE"
 [62] "CHILD NEGLECT (SEE 300 W.I.C.)"
 [63] "STALKING"
 [64] "DISHONEST EMPLOYEE - GRAND THEFT"
 [65] "TRESPASSING"
 [66] "BURGLARY, ATTEMPTED"
 [67] "RAPE, ATTEMPTED"
 [68] "DISCHARGE FIREARMS/SHOTS FIRED"
 [69] "PIMPING"
 [70] "HUMAN TRAFFICKING - COMMERCIAL SEX ACTS"
 [71] "VEHICLE - ATTEMPT STOLEN"
 [72] "PANDERING"
 [73] "FIREARMS RESTRAINING ORDER (FIREARMS RO)"
 [74] "RESISTING ARREST"
 [75] "BURGLARY FROM VEHICLE, ATTEMPTED"
 [76] "THEFT, PERSON"
 [77] "BATTERY POLICE (SIMPLE)"
 [78] "VEHICLE, STOLEN - OTHER (MOTORIZED SCOOTERS, BIKES, ETC)"
 [79] "THEFT FROM PERSON - ATTEMPT"
 [80] "FAILURE TO YIELD"
 [81] "BOMB SCARE"
 [82] "ASSAULT WITH DEADLY WEAPON ON POLICE OFFICER"
 [83] "BUNCO, PETTY THEFT"
 [84] "SHOTS FIRED AT INHABITED DWELLING"
 [85] "DEFRAUDING INNKEEPER/THEFT OF SERVICES, $950 & UNDER"
 [86] "KIDNAPPING - GRAND ATTEMPT"
 [87] "SHOTS FIRED AT MOVING VEHICLE, TRAIN OR AIRCRAFT"
 [88] "TILL TAP - GRAND THEFT ($950.01 & OVER)"
 [89] "VIOLATION OF TEMPORARY RESTRAINING ORDER"
 [90] "THROWING OBJECT AT MOVING VEHICLE"
 [91] "DOCUMENT WORTHLESS ($200.01 & OVER)"
 [92] "KIDNAPPING"
 [93] "CRIMINAL HOMICIDE"
 [94] "PURSE SNATCHING"
 [95] "THEFT FROM MOTOR VEHICLE - ATTEMPT"
 [96] "DISHONEST EMPLOYEE - PETTY THEFT"
 [97] "CHILD PORNOGRAPHY"
 [98] "WEAPONS POSSESSION/BOMBING"
 [99] "DRIVING WITHOUT OWNER CONSENT (DWOC)"
[100] "REPLICA FIREARMS(SALE,DISPLAY,MANUFACTURE OR DISTRIBUTE)"
[101] "LYNCHING"
[102] "RECKLESS DRIVING"
[103] "SHOPLIFTING - ATTEMPT"
[104] "COUNTERFEIT"
[105] "DEFRAUDING INNKEEPER/THEFT OF SERVICES, OVER $950.01"
[106] "BATTERY ON A FIREFIGHTER"
[107] "CRUELTY TO ANIMALS"
[108] "BOAT - STOLEN"
[109] "ILLEGAL DUMPING"
[110] "PROWLER"
[111] "DRUGS, TO A MINOR"
[112] "THEFT, COIN MACHINE - PETTY ($950 & UNDER)"
[113] "DOCUMENT WORTHLESS ($200 & UNDER)"
[114] "MANSLAUGHTER, NEGLIGENT"
[115] "PETTY THEFT - AUTO REPAIR"
[116] "THEFT, COIN MACHINE - ATTEMPT"
[117] "TILL TAP - PETTY ($950 & UNDER)"
```

```
[118] "PURSE SNATCHING - ATTEMPT"
[119] "LYNCHING - ATTEMPTED"
[120] "BIKE - ATTEMPTED STOLEN"
[121] "GRAND THEFT / AUTO REPAIR"
[122] "CONSPIRACY"
[123] "BRIBERY"
[124] "GRAND THEFT / INSURANCE FRAUD"
[125] "DRUNK ROLL"
[126] "CHILD ABANDONMENT"
[127] "THEFT, COIN MACHINE - GRAND ($950.01 & OVER)"
[128] "DISRUPT SCHOOL"
[129] "PICKPOCKET, ATTEMPT"
[130] "TELEPHONE PROPERTY - DAMAGE"
[131] "BEASTIALITY, CRIME AGAINST NATURE SEXUAL ASSLT WITH ANIM"
[132] "BIGAMY"
[133] "FAILURE TO DISPERSE"
[134] "FIREARMS EMERGENCY PROTECTIVE ORDER (FIREARMS EPO)"
[135] "INCEST (SEXUAL ACTS BETWEEN BLOOD RELATIVES)"
[136] "BLOCKING DOOR INDUCTION CENTER"
[137] "INCITING A RIOT"
[138] "DISHONEST EMPLOYEE ATTEMPTED THEFT"
[139] "TRAIN WRECKING"
```

There have been 139 unique types of crime in Los Angeles.

Lets look at the distribution of the top 10 types of crimes.

```
top_crimes <- crimeDatacsv %>%
  count(Crm.Cd.Desc) %>%
  arrange(desc(n)) %>%
  head(10)
```

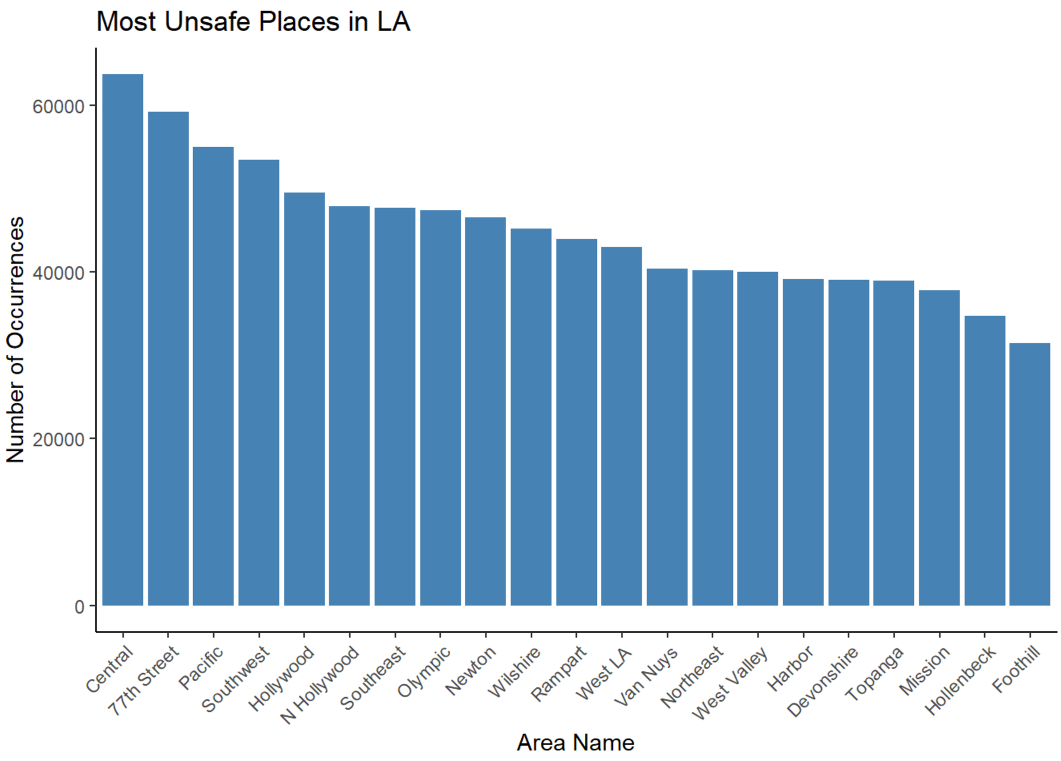
```
top_crimes
```

```
                                               Crm.Cd.Desc      n
1                                          VEHICLE - STOLEN 102036
2                                  BATTERY - SIMPLE ASSAULT  74509
3                                     BURGLARY FROM VEHICLE  58311
4                                         THEFT OF IDENTITY  58240
5                                                  BURGLARY  57497
6  VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)  57194
7           ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT  53192
8                        THEFT PLAIN - PETTY ($950 & UNDER)  48215
9                          INTIMATE PARTNER - SIMPLE ASSAULT  46632
10         THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)  36615
```

Vehicle Theft is the most common type of crime.

```
crime_by_area <- crimeDatacsv %>%
  group_by(`AREA.NAME`) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))

# Plotting the histogram using ggplot2
ggplot(crime_by_area, aes(x =reorder(AREA.NAME, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  # coord_flip() +  # Flip coordinates to make the plot horizontal
  labs(title = "Most Unsafe Places in LA", x = "Area Name", y = "Number of Occurrences") +
  theme_classic()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The areas Central, 77th Street and Pacific are the most crime prone area. Most concerning is the face that even the "safest" place in LA city hass experiences almost 400,000 crimes over these 4 years.

Dropping columns that are not required for EDA is the next order of business. The columns that will be dropped are: DR_NO, Date.Rptd, Rpt.Dist.No, Part.1.2, Mocodes, Status, Status.Desc, Crm.Cd.1, Crm.Cd.2, Crm.Cd.3, Crm.Cd.4, LOCATION, Cross.Street, Weapon.Desc, Weapon.Used.Cd.

We will make the necessary changes and store them in a new dataframe called crimeDataCleaned.

```r
# Store the columns to be removed in a variable
colsToRemove <- c("DR_NO", "Date.Rptd", "Rpt.Dist.No",
                  "Part.1.2", "Mocodes", "Status",
                  "Status.Desc", "Crm.Cd.1", "Crm.Cd.2",
                  "Crm.Cd.3", "Crm.Cd.4", "LOCATION","Premis.Cd","Premis.Desc",
                  "Cross.Street","Weapon.Desc","Weapon.Used.Cd")

# Drop the specified columns and store the result in a new dataframe
crimeDataCleaned <- crimeDatacsv[, !names(crimeDatacsv) %in% colsToRemove]

# Display the structure of the cleaned data to verify
str(crimeDataCleaned)
```

```
'data.frame':	944235 obs. of  11 variables:
 $ DATE.OCC    : chr  "03/01/2020 12:00:00 AM" "02/08/2020 12:00:00 AM" "11/04/2020 12:00:00 AM" "03/10/2020 12:00:00
AM" ...
 $ TIME.OCC    : int  2130 1800 1700 2037 1200 2300 900 1110 1400 1220 ...
 $ AREA        : int  7 1 3 9 6 18 1 3 13 19 ...
 $ AREA.NAME   : chr  "Wilshire" "Central" "Southwest" "Van Nuys" ...
 $ Crm.Cd      : int  510 330 480 343 354 354 354 354 354 624 ...
 $ Crm.Cd.Desc : chr  "VEHICLE - STOLEN" "BURGLARY FROM VEHICLE" "BIKE - STOLEN" "SHOPLIFTING-GRAND THEFT ($950.01 &
OVER)" ...
 $ Vict.Age    : int  0 47 19 19 28 41 25 27 24 26 ...
 $ Vict.Sex    : chr  "M" "M" "X" "M" ...
 $ Vict.Descent: chr  "O" "O" "X" "O" ...
 $ LAT         : num  34 34 34 34.2 34.1 ...
 $ LON         : num  -118 -118 -118 -118 -118 ...
```

DATE.OCC is in dd/mm/yyyy hh:mm:ss AM/PM format. All the times are 12:00:00 since there is a separate TIME.OCC Column. We can get rid of the Time in this column and use lubricate to extract the month and year and store them in separate columns.

```r
# Convert DATE.OCC to a proper datetime format using lubridate
crimeDataCleaned$DATE.OCC <- mdy_hms(crimeDataCleaned$DATE.OCC)

# Extract the month and year from DATE.OCC and store them in new columns
crimeDataCleaned$Month <- month(crimeDataCleaned$DATE.OCC)
crimeDataCleaned$Year <- year(crimeDataCleaned$DATE.OCC)

# Display the structure of the cleaned data to verify
str(crimeDataCleaned)
```

```
'data.frame':   944235 obs. of  13 variables:
 $ DATE.OCC    : POSIXct, format: "2020-03-01" "2020-02-08" ...
 $ TIME.OCC    : int  2130 1800 1700 2037 1200 2300 900 1110 1400 1220 ...
 $ AREA        : int  7 1 3 9 6 18 1 3 13 19 ...
 $ AREA.NAME   : chr  "Wilshire" "Central" "Southwest" "Van Nuys" ...
 $ Crm.Cd      : int  510 330 480 343 354 354 354 354 354 624 ...
 $ Crm.Cd.Desc : chr  "VEHICLE - STOLEN" "BURGLARY FROM VEHICLE" "BIKE - STOLEN" "SHOPLIFTING-GRAND THEFT ($950.01 &
OVER)" ...
 $ Vict.Age    : int  0 47 19 19 28 41 25 27 24 26 ...
 $ Vict.Sex    : chr  "M" "M" "X" "M" ...
 $ Vict.Descent: chr  "O" "O" "X" "O" ...
 $ LAT         : num  34 34 34 34.2 34.1 ...
 $ LON         : num  -118 -118 -118 -118 -118 ...
 $ Month       : num  3 2 11 3 8 12 7 5 12 12 ...
 $ Year        : num  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
```

There are some rows in TIME.OCC that have values such as 1,2,3 etc. We will assume that these mean 0100, 0200, 0300. Also some values are 100, 200, 300. We will assume these values are 0100, 0200, 0300 in military time. We will hence add a new column hour, indicating the hour at which the crime occured.

```r
# Correct TIME.OCC values:
# If TIME.OCC is a single digit (1-9), prepend "0" and append "00" to make it "0100", "0200", etc.
# If TIME.OCC is 2 digits, append "00" to convert it to "HH00"
# If TIME.OCC is already 4 digits, assume it is in the format "HHMM"
crimeDataCleaned$TIME.OCC <- ifelse(nchar(crimeDataCleaned$TIME.OCC) == 1,
                                    sprintf("%02d00", as.numeric(crimeDataCleaned$TIME.OCC)),
                                    ifelse(nchar(crimeDataCleaned$TIME.OCC) == 2,
                                           sprintf("%02d00", as.numeric(crimeDataCleaned$TIME.OCC)),
                                           sprintf("%04d", as.numeric(crimeDataCleaned$TIME.OCC))))

# Extract the hour from TIME.OCC
crimeDataCleaned$hour <- as.numeric(substr(crimeDataCleaned$TIME.OCC, 1, 2))

# Display the structure of the cleaned data to verify
str(crimeDataCleaned)
```

```
'data.frame':   944235 obs. of  14 variables:
 $ DATE.OCC    : POSIXct, format: "2020-03-01" "2020-02-08" ...
 $ TIME.OCC    : chr  "2130" "1800" "1700" "2037" ...
 $ AREA        : int  7 1 3 9 6 18 1 3 13 19 ...
 $ AREA.NAME   : chr  "Wilshire" "Central" "Southwest" "Van Nuys" ...
 $ Crm.Cd      : int  510 330 480 343 354 354 354 354 354 624 ...
 $ Crm.Cd.Desc : chr  "VEHICLE - STOLEN" "BURGLARY FROM VEHICLE" "BIKE - STOLEN" "SHOPLIFTING-GRAND THEFT ($950.01 &
OVER)" ...
 $ Vict.Age    : int  0 47 19 19 28 41 25 27 24 26 ...
 $ Vict.Sex    : chr  "M" "M" "X" "M" ...
 $ Vict.Descent: chr  "O" "O" "X" "O" ...
 $ LAT         : num  34 34 34 34.2 34.1 ...
 $ LON         : num  -118 -118 -118 -118 -118 ...
 $ Month       : num  3 2 11 3 8 12 7 5 12 12 ...
 $ Year        : num  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
 $ hour        : num  21 18 17 20 12 23 9 11 14 12 ...
```

Genders in the reports are only mentioned as M for male, F from Female and X if unknown. For simplicity, let us assume all unknown genders as either male or female distributed equally.

```
# Define a function to randomly replace non 'M' or 'F' values with 'M' or 'F'
replace_invalid_gender <- function(x) {
  # If the value is not 'M' or 'F', replace it with a random 'M' or 'F'
  if (!x %in% c("M", "F")) {
    return(sample(c("M", "F"), 1))
  } else {
    return(x)
  }
}

# Apply the function to the Vict.Sex column
crimeDataCleaned$Vict.Sex <- sapply(crimeDataCleaned$Vict.Sex, replace_invalid_gender)

# Check the unique values in the Vict.Sex column to confirm changes
unique(crimeDataCleaned$Vict.Sex)
```

```
[1] "M" "F"
```

Victim.Age is a bit of a problem. A lot of them are 0. First lets see how many of them are 0.

```
# Count the number of entries where Vict.Age is 0
num_age_zero <- sum(crimeDataCleaned$Vict.Age == 0, na.rm = TRUE)

# Display the result
num_age_zero
```

```
[1] 240110
```

That is 1/4th of data that we have. Upon further exploration of the data, I have found out that in most cases where age is 0, it describes crimes that have not occurred against humans. We will see more about this later.

Since the data for the year 2024 is complete, lets get rid of all records from the year 2024.

```
# Filter out rows where Year is 2024
crimeDataCleaned <- crimeDataCleaned %>%
  filter(Year != 2024)

# Verify the removal
table(crimeDataCleaned$Year)
```

```

  2020   2021   2022   2023 
199700 209703 234975 231642 
```

Now we can begin the EDA.

```
age_zero_data <- crimeDataCleaned %>%
  filter(Vict.Age == 0)

head(age_zero_data)
```

```
    DATE.OCC TIME.OCC AREA   AREA.NAME Crm.Cd
1 2020-03-01     2130    7    Wilshire    510
2 2020-11-01     0130   10 West Valley    510
3 2020-09-09     0630    4  Hollenbeck    510
4 2020-08-14     1300   21     Topanga    668
5 2020-01-18     1600   14     Pacific    420
6 2020-05-26     1200    2     Rampart    420
                                Crm.Cd.Desc Vict.Age Vict.Sex
1                          VEHICLE - STOLEN        0        M
2                          VEHICLE - STOLEN        0        F
3                          VEHICLE - STOLEN        0        F
4      EMBEZZLEMENT, GRAND THEFT ($950.01 & OVER)        0        M
5 THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)        0        F
6 THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)        0        M
```

```
  Vict.Descent     LAT       LON Month Year hour
1            O 34.0375 -118.3506     3 2020   21
2              34.1939 -118.4859    11 2020    1
3              34.0820 -118.2130     9 2020    6
4              34.2105 -118.6157     8 2020   13
5              34.0022 -118.4255     1 2020   16
6              34.0697 -118.2779     5 2020   12
```

```
# Count occurrences of each crime type and arrange them in descending order
top_crimes <- age_zero_data %>%
  count(Crm.Cd.Desc) %>%
  arrange(desc(n)) %>%
  head(10)

# Display the top 10 most common crime types
print(top_crimes)
```

```
                                             Crm.Cd.Desc     n
1                                       VEHICLE - STOLEN 93346
2          THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER) 19043
3                                               BURGLARY 16690
4                 SHOPLIFTING - PETTY THEFT ($950 & UNDER) 14715
5   VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 13111
6                                                ROBBERY  5959
7                       THEFT PLAIN - PETTY ($950 & UNDER)  5828
8  THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD  5762
9                                            TRESPASSING  5290
10                VANDALISM - MISDEAMEANOR ($399 OR UNDER)  4752
```

```
total_crimes <- nrow(age_zero_data)

# Calculate the total number of crimes in the top 10
top_10_total <- sum(top_crimes$n)

# Calculate the percentage
percentage_top_10 <- (top_10_total / total_crimes) * 100

# Display the result
percentage_top_10
```

```
[1] 84.51953
```

As we can see, the top 10 types of crime that have an age of 0 on reports account for about 85% of all crimes that have the age recorded as 0.

```
# Filter the data where Vict.Age is not 0
age_non_zero_data <- crimeDataCleaned %>%
  filter(Vict.Age != 0)

# Count occurrences of each crime type and arrange them in descending order
top_crimes <- age_non_zero_data %>%
  group_by(`Crm.Cd.Desc`) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count)) %>%
  head(10)

# Display the top 10 most common crime types
print(top_crimes)
```

```
# A tibble: 10 × 2
   Crm.Cd.Desc                                          Count
   <chr>                                                <int>
 1 BATTERY - SIMPLE ASSAULT                             68554
 2 THEFT OF IDENTITY                                    54474
 3 BURGLARY FROM VEHICLE                                52351
 4 ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT       47971
 5 INTIMATE PARTNER - SIMPLE ASSAULT                           43265
 6 VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 40018
 7 THEFT PLAIN - PETTY ($950 & UNDER)                          39000
 8 BURGLARY                                                    36714
 9 THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND OVER)         29690
10 ROBBERY                                                     23853
```

```r
total_crimes <- nrow(age_non_zero_data)

# Calculate the total number of crimes in the top 10
top_10_total <- sum(top_crimes$Count)

# Calculate the percentage
percentage_top_10 <- (top_10_total / total_crimes) * 100

# Display the result
percentage_top_10
```

```
[1] 66.27167
```

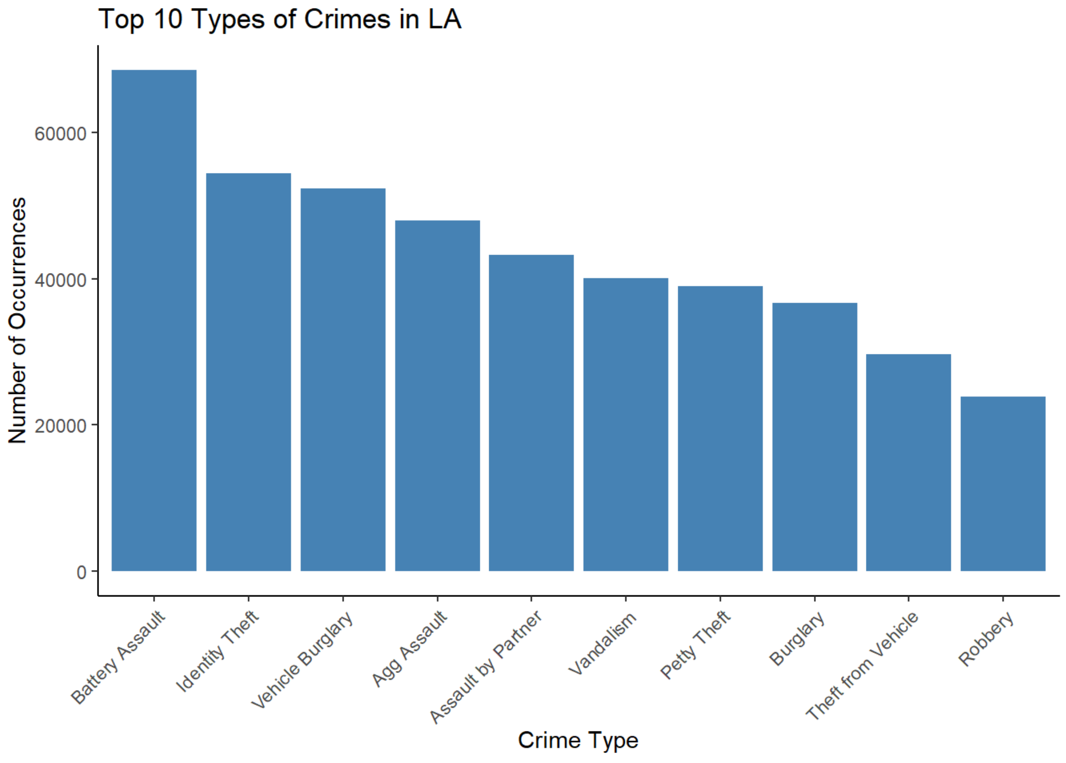
Here the top 10 types of crime account for more than 66% of all crimes that occur.

Lets rename the columns to get a better picture

```r
crimes_personalised <- c("Battery Assault", "Identity Theft", "Vehicle Burglary", "Agg Assault", "Assault by Pa
top_crimes$Crm.Cd.Desc <- crimes_personalised
```

Lets visualise these crimes.

```r
# Plotting the histogram using ggplot2
ggplot(top_crimes, aes(x =reorder(Crm.Cd.Desc, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  # coord_flip() +  # Flip coordinates to make the plot horizontal
  labs(title = "Top 10 Types of Crimes in LA", x = "Crime Type", y = "Number of Occurrences") +
  theme_classic()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

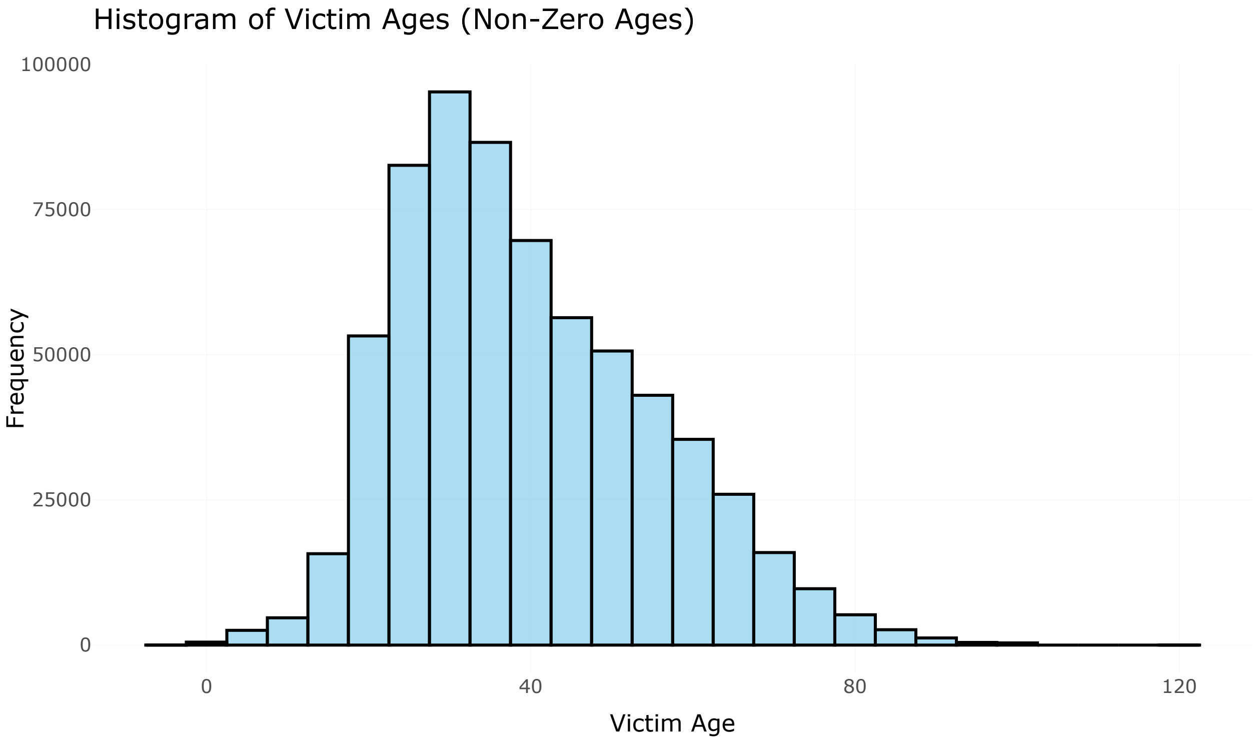


Lets look at the distribution of age of the victims of all crimes.

```r
# Create a ggplot histogram
histogram_plot <- ggplot(age_non_zero_data, aes(x = Vict.Age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Victim Ages (Non-Zero Ages)", x = "Victim Age", y = "Frequency") +
  theme_minimal()

# Convert the ggplot object to an interactive plotly plot
interactive_histogram <- ggplotly(histogram_plot)

# Display the interactive plot
interactive_histogram
```



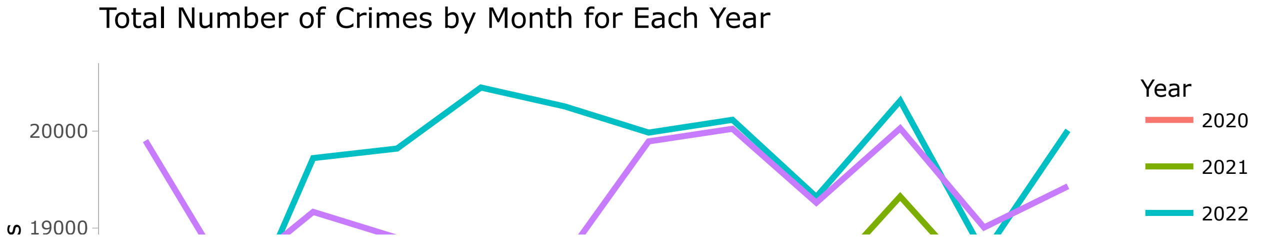
Most victims are of the age group from 25-45.

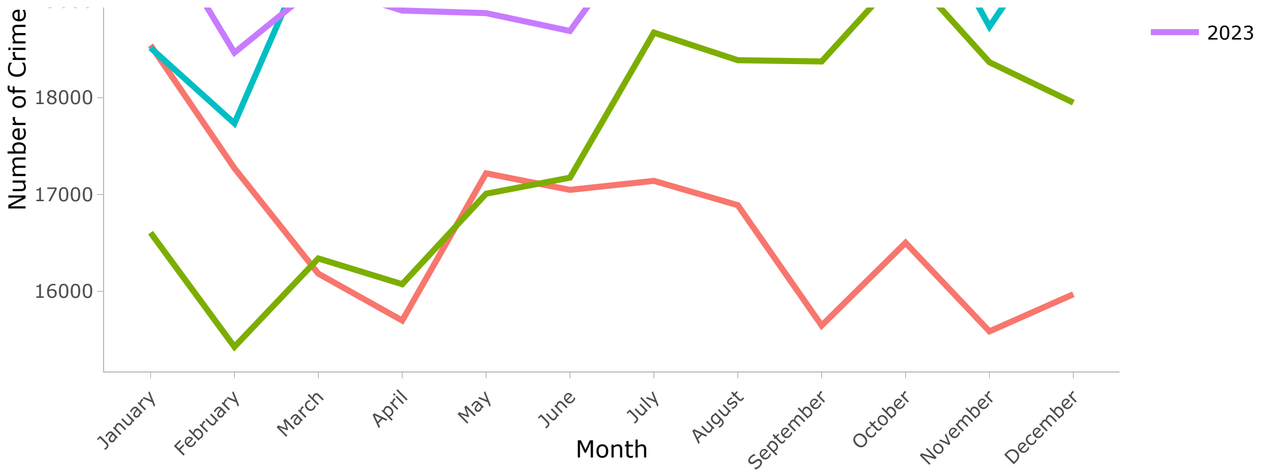
```r
# Aggregate the data by Year and Month
crimes_by_month_year <- crimeDataCleaned %>%
  group_by(Year, Month) %>%
  summarise(Count = n()) %>%
  arrange(Year, Month)

# Create the line plot with different lines for each year
line_plot <- ggplot(crimes_by_month_year, aes(x = Month, y = Count, color = factor(Year), group = Year)) +
  geom_line(size = 1) +
  scale_x_continuous(breaks = 1:12, labels = month.name) +  # Label months by name
  labs(title = "Total Number of Crimes by Month for Each Year",
       x = "Month", y = "Number of Crimes", color = "Year") +
  theme_classic()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),panel.grid = element_blank())

# Convert to an interactive plotly plot
interactive_line_plot <- ggplotly(line_plot)

# Display the interactive plot
interactive_line_plot
```




Number of Crime
18000
17000
16000
January
February
March
April
May
June
July
August
September
October
November
December
Month
2023