Analysis on Policing Data, Dallas 2016

Mathew, Reethu

April, 2023

Introduction

Data Visualisation is critical for comprehending and analysing complex data sets. It enables us to communicate information derived from the data effectively and efficiently, allowing us to find patterns, relationships, and trends that may not be apparent in raw data. Visual representations of data provide a more intuitive way of interpreting information, making it easier to identify outliers and anomalies, compare different variables, and extract relevant insights. With the alarmingly increasing volume and complexity of data, the importance of data visualisation in understanding data cannot be overlooked.

The Data

For this report, we are analysing a policing dataset from Dallas, Texas in 2016. The source file is acquired from https://www.kaggle.com/center-for-policing-equity/data-science-for-good and modified to the version `37-00049_UOF-P_2016_prepped.csv` which is the input to this analysis. Analysing this data can help us obtain insights into numerous areas of policing, such as crime rates, the demographics of those involved in criminal activities, and the success of various policing strategies. By visualising the results obtained from the analysis, we can identify patterns and trends that can assist in improving policing decisions and the general quality of policing in Dallas.

Now let's load the data. The data is loaded from the working directory using the read.csv() function. The input file 37-00049-UOF-P_2016_prepped.csv contains two title rows. So, the second row, which is unwanted, is removed to get the working dataset named RawData. The method of removing the second header row is referred from stackoverflow.com.

```
BaseData <- readLines("37-00049-UOF-P_2016_prepped.csv")
skip_2 <- BaseData[-2]
RawData <- read.csv(textConnection(skip_2), header = TRUE, stringsAsFactors = FALSE)
```

To get a brief idea about the structure of the data, we'll use the `head()` and `str()` functions. The head() function by Deault returns the first five rows of the data, and the str() function returns a list of the objects and their structure. The 2016 Dallas Policing data contains 2383 observations of 47 variables.

```
head(RawData)
```

```
##   INCIDENT_DATE INCIDENT_TIME    UOF_NUMBER OFFICER_ID OFFICER_GENDER
## 1        9/3/16    4:14:00 AM         37702      10810           Male
## 2       3/22/16   11:00:00 PM         33413       7706           Male
## 3       5/22/16    1:29:00 PM         34567      11014           Male
## 4       1/10/16    8:55:00 PM         31460       6692           Male
## 5       11/8/16    2:30:00 AM  37879, 37898       9844           Male
```

```
## 6       9/11/16    7:20:00 PM         36724       9855           Male
##   OFFICER_RACE OFFICER_HIRE_DATE OFFICER_YEARS_ON_FORCE OFFICER_INJURY
## 1        Black            5/7/14                     2            No
## 2        White            1/8/99                    17           Yes
## 3        Black           5/20/15                     1            No
## 4        Black           7/29/91                    24            No
## 5        White           10/4/09                     7            No
## 6        White           6/10/09                     7            No
##            OFFICER_INJURY_TYPE OFFICER_HOSPITALIZATION SUBJECT_ID SUBJECT_RACE
## 1 No injuries noted or visible                      No      46424        Black
## 2                Sprain/Strain                     Yes      44324     Hispanic
## 3 No injuries noted or visible                      No      45126     Hispanic
## 4 No injuries noted or visible                      No      43150     Hispanic
## 5 No injuries noted or visible                      No      47307        Black
## 6 No injuries noted or visible                      No      46549        White
##   SUBJECT_GENDER SUBJECT_INJURY          SUBJECT_INJURY_TYPE
## 1         Female            Yes      Non-Visible Injury/Pain
## 2           Male             No No injuries noted or visible
## 3           Male             No No injuries noted or visible
## 4           Male            Yes              Laceration/Cut
## 5           Male             No No injuries noted or visible
## 6         Female             No No injuries noted or visible
##   SUBJECT_WAS_ARRESTED SUBJECT_DESCRIPTION          SUBJECT_OFFENSE
## 1                  Yes   Mentally unstable                    APOWW
## 2                  Yes   Mentally unstable                    APOWW
## 3                  Yes             Unknown                    APOWW
## 4                  Yes FD-Unknown if Armed           Evading Arrest
## 5                  Yes             Unknown Other Misdemeanor Arrest
## 6                  Yes             Unknown               Assault/FV
##   REPORTING_AREA BEAT SECTOR      DIVISION LOCATION_DISTRICT STREET_NUMBER
## 1           2062  134    130       CENTRAL               D14           211
## 2           1197  237    230     NORTHEAST                D9          7647
## 3           4153  432    430     SOUTHWEST                D6           716
## 4           4523  641    640 NORTH CENTRAL               D11          5600
## 5           2167  346    340     SOUTHEAST                D7          4600
## 6           1134  235    230     NORTHEAST                D9          1234
##    STREET_NAME STREET_DIRECTION STREET_TYPE
## 1        Ervay               N         St.
## 2    Ferguson            NULL         Rd.
## 3 bimebella dr            NULL         Ln.
## 4         LBJ            NULL       Frwy.
## 5   Malcolm X               S       Blvd.
## 6        Peavy            NULL         Rd.
##   LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION LOCATION_CITY LOCATION_STATE
## 1                             211 N ERVAY ST        Dallas             TX
## 2                            7647 FERGUSON RD        Dallas             TX
## 3                            716 BIMEBELLA LN        Dallas             TX
## 4                              5600 L B J FWY        Dallas             TX
## 5                         4600 S MALCOLM X BLVD        Dallas             TX
## 6                              1234 PEAVY RD        Dallas             TX
##   LOCATION_LATITUDE LOCATION_LONGITUDE INCIDENT_REASON REASON_FOR_FORCE
## 1          32.78220          -96.79746          Arrest           Arrest
## 2          32.79898          -96.71749          Arrest           Arrest
## 3          32.73971          -96.92519          Arrest           Arrest
```

```
## 4                NA                  NA          Arrest           Arrest
## 5                NA                  NA          Arrest           Arrest
## 6          32.83753          -96.69557          Arrest           Arrest
##      TYPE_OF_FORCE_USED1 TYPE_OF_FORCE_USED2 TYPE_OF_FORCE_USED3
## 1  Hand/Arm/Elbow Strike
## 2            Joint Locks
## 3      Take Down - Group
## 4         K-9 Deployment
## 5         Verbal Command     Take Down - Arm
## 6 Hand Controlled Escort
##   TYPE_OF_FORCE_USED4 TYPE_OF_FORCE_USED5 TYPE_OF_FORCE_USED6
## 1
## 2
## 3
## 4
## 5
## 6
##   TYPE_OF_FORCE_USED7 TYPE_OF_FORCE_USED8 TYPE_OF_FORCE_USED9
## 1
## 2
## 3
## 4
## 5
## 6
##   TYPE_OF_FORCE_USED10 NUMBER_EC_CYCLES FORCE_EFFECTIVE
## 1                                 NULL             Yes
## 2                                 NULL             Yes
## 3                                 NULL             Yes
## 4                                 NULL             Yes
## 5                                 NULL         No, Yes
## 6                                 NULL             Yes
```

```
str(RawData)
```

```
## 'data.frame':    2383 obs. of  47 variables:
##  $ INCIDENT_DATE                              : chr  "9/3/16" "3/22/16" "5/22/16" "1/10/16" ...
##  $ INCIDENT_TIME                              : chr  "4:14:00 AM" "11:00:00 PM" "1:29:00 PM" "8:55:0
##  $ UOF_NUMBER                                 : chr  "37702" "33413" "34567" "31460" ...
##  $ OFFICER_ID                                 : int  10810 7706 11014 6692 9844 9855 9881 9058 1038
##  $ OFFICER_GENDER                             : chr  "Male" "Male" "Male" "Male" ...
##  $ OFFICER_RACE                               : chr  "Black" "White" "Black" "Black" ...
##  $ OFFICER_HIRE_DATE                          : chr  "5/7/14" "1/8/99" "5/20/15" "7/29/91" ...
##  $ OFFICER_YEARS_ON_FORCE                     : int  2 17 1 24 7 7 7 9 4 8 ...
##  $ OFFICER_INJURY                             : chr  "No" "Yes" "No" "No" ...
##  $ OFFICER_INJURY_TYPE                        : chr  "No injuries noted or visible" "Sprain/Strain"
##  $ OFFICER_HOSPITALIZATION                    : chr  "No" "Yes" "No" "No" ...
##  $ SUBJECT_ID                                 : int  46424 44324 45126 43150 47307 46549 47555 4417
##  $ SUBJECT_RACE                               : chr  "Black" "Hispanic" "Hispanic" "Hispanic" ...
##  $ SUBJECT_GENDER                             : chr  "Female" "Male" "Male" "Male" ...
##  $ SUBJECT_INJURY                             : chr  "Yes" "No" "No" "Yes" ...
##  $ SUBJECT_INJURY_TYPE                        : chr  "Non-Visible Injury/Pain" "No injuries noted o
##  $ SUBJECT_WAS_ARRESTED                       : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ SUBJECT_DESCRIPTION                        : chr  "Mentally unstable" "Mentally unstable" "Unknow
##  $ SUBJECT_OFFENSE                            : chr  "APOWW" "APOWW" "APOWW" "Evading Arrest" ...
```

```
##  $ REPORTING_AREA                                : int  2062 1197 4153 4523 2167 1134 2049 3122 2072 4
##  $ BEAT                                          : int  134 237 432 641 346 235 132 515 133 614 ...
##  $ SECTOR                                        : int  130 230 430 640 340 230 130 510 130 610 ...
##  $ DIVISION                                      : chr  "CENTRAL" "NORTHEAST" "SOUTHWEST" "NORTH CENTRA
##  $ LOCATION_DISTRICT                             : chr  "D14" "D9" "D6" "D11" ...
##  $ STREET_NUMBER                                 : int  211 7647 716 5600 4600 1234 511 4709 300 18600
##  $ STREET_NAME                                   : chr  "Ervay" "Ferguson" "bimebella dr" "LBJ" ...
##  $ STREET_DIRECTION                              : chr  "N" "NULL" "NULL" "NULL" ...
##  $ STREET_TYPE                                   : chr  "St." "Rd." "Ln." "Frwy." ...
##  $ LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION: chr  "211 N ERVAY ST" "7647 FERGUSON RD" "716 BIMEBE
##  $ LOCATION_CITY                                 : chr  "Dallas" "Dallas" "Dallas" "Dallas" ...
##  $ LOCATION_STATE                                : chr  "TX" "TX" "TX" "TX" ...
##  $ LOCATION_LATITUDE                             : num  32.8 32.8 32.7 NA NA ...
##  $ LOCATION_LONGITUDE                            : num  -96.8 -96.7 -96.9 NA NA ...
##  $ INCIDENT_REASON                               : chr  "Arrest" "Arrest" "Arrest" "Arrest" ...
##  $ REASON_FOR_FORCE                              : chr  "Arrest" "Arrest" "Arrest" "Arrest" ...
##  $ TYPE_OF_FORCE_USED1                           : chr  "Hand/Arm/Elbow Strike" "Joint Locks" "Take Dow
##  $ TYPE_OF_FORCE_USED2                           : chr  "" "" "" "" ...
##  $ TYPE_OF_FORCE_USED3                           : chr  "" "" "" "" ...
##  $ TYPE_OF_FORCE_USED4                           : chr  "" "" "" "" ...
##  $ TYPE_OF_FORCE_USED5                           : chr  "" "" "" "" ...
##  $ TYPE_OF_FORCE_USED6                           : chr  "" "" "" "" ...
##  $ TYPE_OF_FORCE_USED7                           : chr  "" "" "" "" ...
##  $ TYPE_OF_FORCE_USED8                           : chr  "" "" "" "" ...
##  $ TYPE_OF_FORCE_USED9                           : chr  "" "" "" "" ...
##  $ TYPE_OF_FORCE_USED10                          : chr  "" "" "" "" ...
##  $ NUMBER_EC_CYCLES                              : chr  "NULL" "NULL" "NULL" "NULL" ...
##  $ FORCE_EFFECTIVE                               : chr  " Yes" " Yes" " Yes" " Yes" ...
```

The majority of the data set variables are of character type, which we will format as we go along as the character datatype does not fit the format requirements of the visualization functions.

As we now have an idea of the data we have in our hands, we will proceed with the analysis and visualisation. Firstly, we'll take a look at the INCIDENT_DATE variable, which contains the dates on which the events took place. The INCIDENT_DATE column, which is in character type, is converted into date type using the `mdy()` function in the Lubridate package. This date-type data can be subjected to the `weekdays()` function, which extracts the day in which the event takes place. A new column, Days_of_week is created to store these values. This column is converted to a categorical column with the `factor()` function with 7 levels. A bar chart comparing the number of crimes in each day of the week is plotted below :

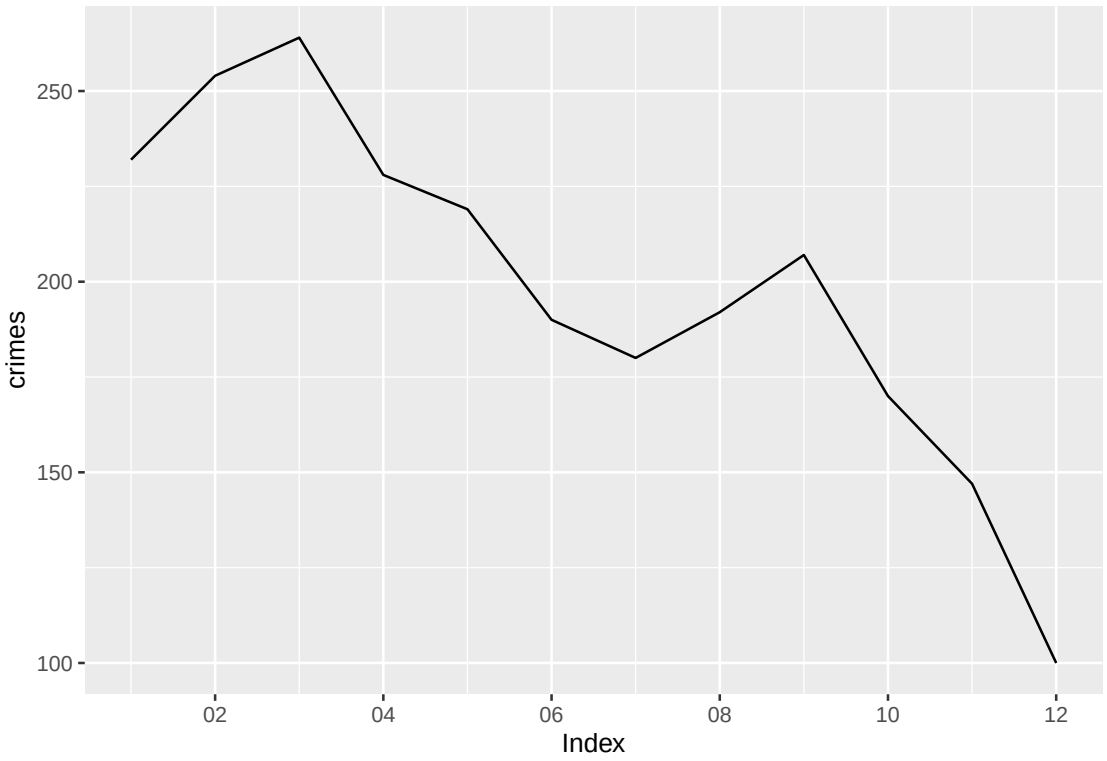
Plots and Analyses

```
RawData <- RawData %>%
  mutate(Days_of_week = weekdays(mdy(INCIDENT_DATE)))
RawData$Days_of_week <- factor(RawData$Days_of_week, levels = c("Monday", "Tuesday", "Wednesday", "Thurs
ggplotly(ggplot(RawData, aes(Days_of_week)) + geom_bar(fill = "#756bb1", width = 0.75, colour = "black"
  ggtitle("Comparison Of Number Of Crimes In Each Day Of The Week") +
  ylab("Number of Crimes") + xlab("Day of the week") +
  theme_bw() + theme_rm)
```

The chart shows a relatively higher number of crimes on Friday, Saturday, and Sunday compared to the other four days of the week. Here, `theme_rm` is a customised theme that formats the plot and the axes

titles. We can take a similar approach to checking the distribution of reported incidents over the year by getting the count for each month. In this case, we're plotting a time series plot.A time series plot displays data points at a regular interval of time, which in our case is a month. Because of their ease of identifying outliers, patterns, and trends over a time period, time series plots are widely used in data visualisation.

```
month <- month(mdy(RawData$INCIDENT_DATE))
a <- data.frame(table(as_datetime(month)))
colnames(a) <- c("Month", "Count")
times <- as_datetime(a$Month)
crimes <- xts(a$Count, times)
autoplot(crimes)
```

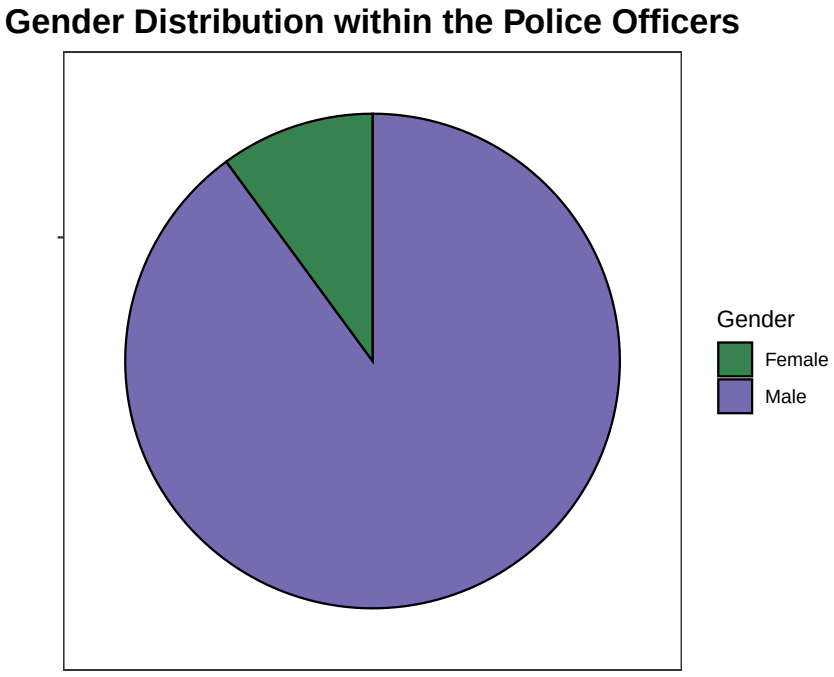


The number of reported incidents starts high at the beginning of the year, with the tallest peak in March. This could be due to relatively warmer weather, which leads to more outdoor events that also become opportunities for criminal activities. After that, the count decreases, maybe due to the strengthening of the police forces due to the increased reports of crimes. Even though there is a slight increase in September, the descent resumes until the bottom is hit in Decemeber.

We will now focus on the various aspects of the police officers and subjects reported in the data. First, we will compare the gender distribution in the police force. The OFFICER_GENDER columns are converted into a factor containing two levels : male and female. The pie chart shows the percentage of male and female police officers.

```
RawData$OFFICER_GENDER <- factor(RawData$OFFICER_GENDER, levels = c("Male", "Female"))
Officer_gender <- table(RawData$OFFICER_GENDER)
ggplot(data.frame(Officer_gender), aes(x = "", y = Freq, fill = factor(Officer_gender))) +
  geom_bar(stat = "identity", width = 3, colour = "black") +
  coord_polar(theta = "y") +
```

```
  scale_fill_manual(values = c("#37834f", "#756bb1"), name = "Gender", labels = c("Female", "Male")) +
  ggtitle("Gender Distribution within the Police Officers") +
  theme_bw() + theme_rm +
  theme(axis.title = element_blank(), axis.text = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```



We can clearly see that there is a majority of male police officers, demonstrating poor gender equality in the police department. We will also take a look at the race distribution of the officers in the given data set. The two-way table below shows the number of officers in six groups of race origin in two gender categories.

```
RawData$OFFICER_RACE <- factor(RawData$OFFICER_RACE, levels = c("American Ind","Asian", "Black", "Hispa
Race_of_officer <- table(RawData$OFFICER_RACE)
two_way_table <- table(RawData$OFFICER_GENDER, RawData$OFFICER_RACE)
kable(two_way_table)
```

	American Ind	Asian	Black	Hispanic	White	Other
Male	6	48	292	440	1336	21
Female	2	7	49	42	134	6

This data is visualised in the horizontal bar chart below.

```
pyr_data <- data.frame(gender = c("Male", "Female"), `American Ind` = c(6,2),
                       `Asian` = c(48,7), `Black` = c(292,49), `Hispanic` = c(440,42),
                       `White` = c(1336,134), `Other` = c(21,6))
long_pyr <- tidyr::pivot_longer(pyr_data, cols = -gender, names_to = "race", values_to = "count") %>%
  arrange(race, desc(count))
```

```
long_pyr <- long_pyr %>% mutate(count = ifelse(gender=="Female", -count, count))

ggplotly(
  ggplot(long_pyr, aes(x=count, y=race, fill=gender)) +
  geom_col(position = "identity") +
  xlim(-500, 1500) + xlab("Number of Officers") + ylab("Race of the officer") +
    ggtitle("Ethnic Distribution with respect to Gender of Officers") +
  scale_fill_manual(values = c("#37834f", "#756bb1"), name = "Gender", labels = c("Female", "Male")) +
  theme_bw() + theme_rm
  )
```

Another important factor to consider is the experience of the police officers involved in the reported incidents. This is important considering the intricacy of the job, where anything could happen at any time. Officers' readiness to face difficult situations is something that can be nurtured with increasing years in the force. The histogram below shows the distribution of officers in the force at the date when the incident was recorded.

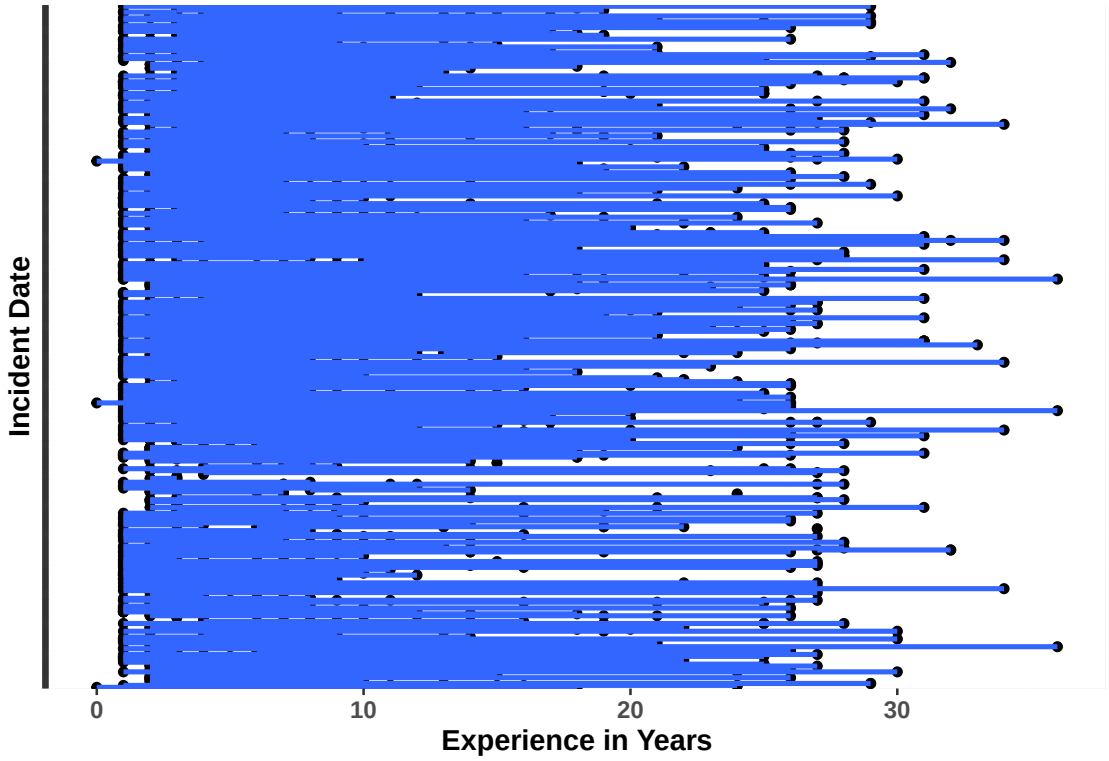
```
ggplotly(
  ggplot(RawData, aes(x = OFFICER_YEARS_ON_FORCE)) +
  geom_histogram(binwidth = 2, color = "black", fill = "#756bb1") +
  ggtitle("Experience of Police Officers") +
  xlab("Experience in years") + ylab("Number of Officers") +
  scale_x_continuous(breaks = seq(0, 36, 2)) +
  theme_bw() + theme_rm +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.major.y = element_blank()))
```

We can clearly see that the police officers in the field generally have less than 10 years of experience. This could be due to the promotions in effect and the increasing seniority in the force.

The following is a smoothed line scatterplot based on the years of experience and the incident date. This plot also confirms the higher density below the 10-year line.

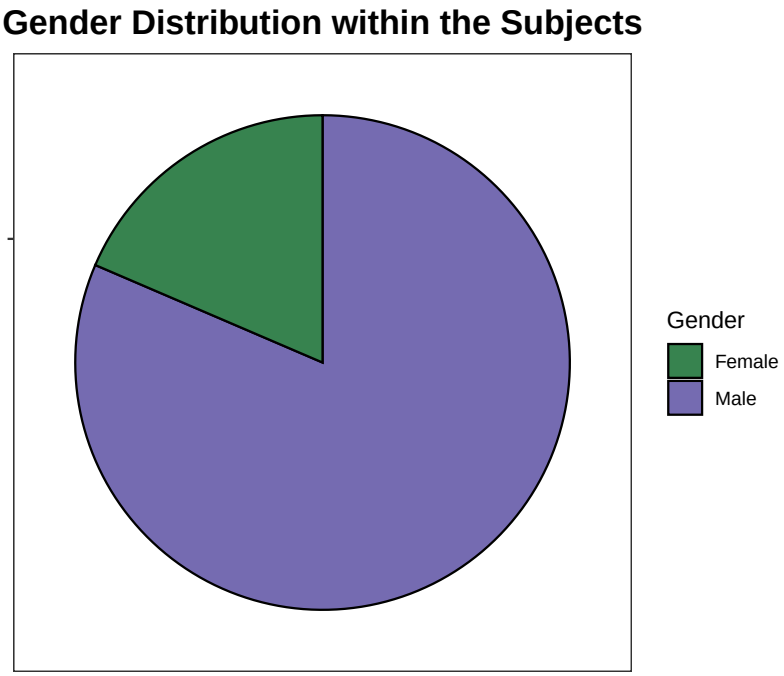
```
ggplot(RawData, aes(x = OFFICER_YEARS_ON_FORCE, y = INCIDENT_DATE)) +
  geom_point() + stat_smooth(method = "lm", se = FALSE) +
  theme(axis.text.y = element_blank()) + theme_rm +
  xlab("Experience in Years") + ylab("Incident Date")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



We will now take a look at the details of the citizens who were recorded in the report. The pie plot below shows the gender distribution of the subjects. The `na_if()` function used to remove the NA values from the data set is referred to in dplyr.tidyverse.org

```
RawData$SUBJECT_GENDER <- na_if(RawData$SUBJECT_GENDER, "Unknown")
RawData$SUBJECT_GENDER <- factor(RawData$SUBJECT_GENDER, levels = c("Male", "Female", NA))
Subject_gender <- na.omit(table(RawData$SUBJECT_GENDER))
ggplot(data.frame(Subject_gender), aes(x = "", y = Freq, fill = factor(Subject_gender))) +
  geom_bar(stat = "identity", width = 3, colour = "black") +
  coord_polar(theta = "y") +
  scale_fill_manual(values = c("#37834f", "#756bb1"), name = "Gender", labels = c("Female", "Male")) +
  ggtitle("Gender Distribution within the Subjects") +
  theme_bw() + theme_rm +
  theme(axis.title = element_blank(), axis.text = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```



In the suspect list as well, there is a majority of men, indicating increased criminal activity among men. We will also check the ethnic distribution among the subjects. Some rows are excluded from the analysis to avoid missing values in the SUBJECT_RACE column.

```
RawData$SUBJECT_RACE <- factor(RawData$SUBJECT_RACE, levels = c("American Ind","Asian", "Black", "Hispa
Race_of_subject <- table(RawData$OFFICER_RACE)
(two_way_t <- table(RawData$SUBJECT_GENDER, RawData$SUBJECT_RACE))
```

```
##
##          American Ind Asian Black Hispanic White Other
##   Male              1     5  1058      455   377    11
##   Female            0     0   274       69    93     0
```

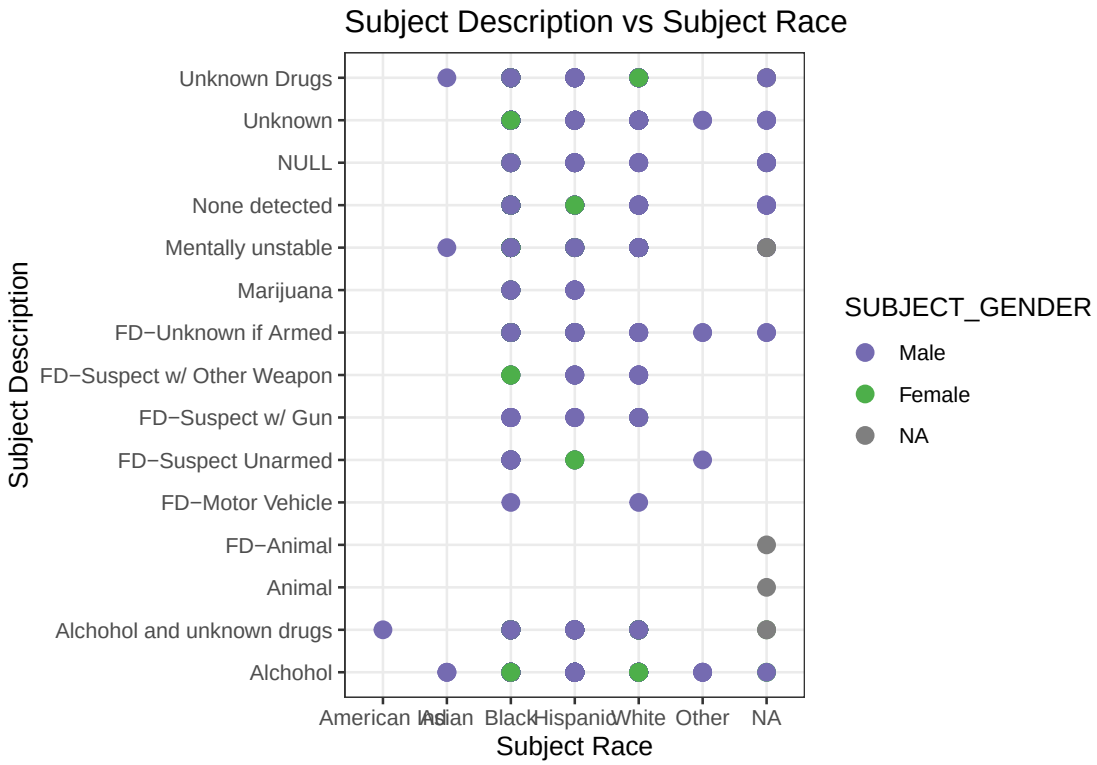
```
# data from the two way table is formatted into another data frame
data_pyr <- data.frame(gender = c("Male", "Female"), `American Ind` = c(1,0),
                       `Asian` = c(5,0), `Black` = c(1058,274), `Hispanic` = c(455,69),
                       `White` = c(377,93), `Other` = c(11,0))
pyr_long <- tidyr::pivot_longer(data_pyr, cols = -gender, names_to = "race", values_to = "count") %>%
  arrange(race, desc(count))
pyr_long <- pyr_long %>% mutate(count = ifelse(gender=="Female", -count, count))

ggplotly(
  ggplot(pyr_long, aes(x=count, y=race, fill=gender)) +
    geom_col(position = "identity") +
    xlim(-500, 1500) + xlab("Number of Subjects") + ylab("Race of the Subject") +
    ggtitle("Ethnic Distribution with respect to Gender of Subjects") +
    scale_fill_manual(values = c("#37834f", "#756bb1"), name = "Gender", labels = c("Female", "Male")) +
    theme_bw() + theme_rm)
```

We can see a significant difference in the race with the highest count of suspects. While the majority of the police force is comprised of white males, a large portion of recorded suspects are black men. Even though

there is some uncertainty caused by the unrecorded ethnicity values, the difference is too big not to consider. It is also notable that suspects from American Indian, Asian, and other races are very few.

We can also look at the following plot, which depicts the subject description with respect to the race of the person. It is again layered with the gender of the subject.

```
RawData$SUBJECT_DESCRIPTION <- factor(RawData$SUBJECT_DESCRIPTION)
ggplot(RawData, aes(x = SUBJECT_RACE, y = SUBJECT_DESCRIPTION, color = SUBJECT_GENDER)) +
  geom_point(size = 3) +
  theme_rm + ylab("Subject Description") + xlab("Subject Race") +
  ggtitle("Subject Description vs Subject Race") + theme_rm +
  theme_bw() + scale_color_manual(values = c("#756bb1", "#4daf4a"))
```



If we ignore the missing values, we can see that the black, hispanic and white races are present in more categories compared to other races. Out of these three, black people are most suspected of being armed or in possession of drugs.

Similarily, in a scatter plot between offense and race, the most cases are recorded against the black ethnicity. The root of this high participation in both cases could be the colonial living tendencies of the black community and the general discrimination they face, which further leads them to criminal behaviour.

```
RawData$RawData$SUBJECT_OFFENSE <- factor(RawData$RawData$SUBJECT_OFFENSE)
ggplotly(ggplot(RawData, aes(x = SUBJECT_RACE, y = SUBJECT_OFFENSE, color = SUBJECT_GENDER)) +
  geom_point(size = 2) + theme_rm + ylab("Subject Description") + xlab("Subject Race") +
  ggtitle("Offenses Distributed Over Subject Race") +
  theme_bw() + theme(axis.text.y = element_blank(),
                     axis.title.y = element_blank()) +
  scale_color_manual(values = c("#756bb1", "#4daf4a")))
```

The relationship between the officers' gender and race and the subjects' gender and race can be visually compared as per the following chart.

```
RawData$OFFICER_GENDER <- as.numeric(RawData$OFFICER_GENDER)
RawData$OFFICER_RACE <- as.numeric(RawData$OFFICER_RACE)
RawData$SUBJECT_RACE <- as.numeric(RawData$SUBJECT_RACE)
RawData$SUBJECT_RACE <- replace(RawData$SUBJECT_RACE, is.na(RawData$SUBJECT_RACE), 0)
RawData$SUBJECT_GENDER <- as.numeric(RawData$SUBJECT_GENDER)
RawData$SUBJECT_GENDER <- replace(RawData$SUBJECT_GENDER, is.na(RawData$SUBJECT_GENDER), 0)
pairs_data <- c("OFFICER_RACE", "OFFICER_GENDER", "SUBJECT_RACE", "SUBJECT_GENDER")
ggplotly(ggpairs(RawData, columns = pairs_data) + theme_bw())
```

The bottom half of the pair plot is displayed in dots as all four variables are categorical values. We see a negative correlation between the officer's race and the subject's gender and a positive correlation between the officer's gender and the subject's gender.

Before we further analyse the correlation values, we will briefly consider the data recorded about injuries that occurred to both the police officer and the subject. The percentages of injury occured to both parties are displayed in the table below.

```
RawData$OFFICER_INJURY <- factor(RawData$OFFICER_INJURY, levels = c("Yes", "No"))
table(RawData$OFFICER_INJURY)
```

```
##
##  Yes   No
##  234 2149
```

```
RawData$SUBJECT_INJURY <- factor(RawData$SUBJECT_INJURY, levels = c("Yes", "No"))
table(RawData$SUBJECT_INJURY)
```

```
##
##  Yes   No
##  629 1754
```

```
# injury dataset is formatted from the tables of the categorical values of officer's injury and subject
injury <- data.frame(Injury = c("Yes", "No"),
                     Officers = c(234, 2149),
                     Subjects = c(629, 1754))
perc_table <- 100*prop.table(injury[,-1])
perc_out <- cbind(injury, perc_table)
colnames(perc_out) <- c("Injury", "Number of Officers", "Number of Subjects", "Percentage of Officers",
perc_out <- perc_out[,c(1,2,4,3,5)]
kable(perc_out)
```

Injury	Number of Officers	Percentage of Officers	Number of Subjects	Percentage of Subjects
Yes	234	4.909778	629	13.19765
No	2149	45.090222	1754	36.80235

The greater percentage of the subject being injured could be due to the fact that force is directed towards them in order to bring them under the control of the police officer.

Now we can look at the relationship between police officer data and subject data through a correlation plot. Correlation plots are extremely useful in identifying patterns, trends, and relationships between variables.

The correlation matrix displays the correlation coefficients between two variables. It is a measure of the strength and direction of the linear relationship between two variables. The value ranges from -1 to 1, with each extreme showing perfect negative and positive correlation, respectively. A correlation coefficient of zero indicates that there is no correlation between the two variables.

```
RawData$OFFICER_INJURY <- as.numeric(RawData$OFFICER_INJURY)
RawData$SUBJECT_INJURY <- as.numeric(RawData$SUBJECT_INJURY)
RawData$SUBJECT_WAS_ARRESTED <- factor(RawData$SUBJECT_WAS_ARRESTED)
RawData$SUBJECT_WAS_ARRESTED <- as.numeric(RawData$SUBJECT_WAS_ARRESTED)
names_col <- c("TYPE_OF_FORCE_USED1",   "TYPE_OF_FORCE_USED2", "TYPE_OF_FORCE_USED3", "TYPE_OF_FORCE_USE
for (col in names(RawData[,names_col])){
  RawData[,col] <- ifelse(is.na(RawData[,col]) | RawData[, col] == "", 0, 1)
}
RawData$LEVEL_OF_FORCE <- rowSums(apply(RawData[,36:45], 2, function(x) as.numeric(as.character(x))), na
cor_df <- RawData[c("OFFICER_GENDER", "OFFICER_RACE", "OFFICER_YEARS_ON_FORCE", "OFFICER_INJURY", "SUBJE
cor_matrix <- cor(cor_df)
cor_matrix
```

```
##                         OFFICER_GENDER OFFICER_RACE OFFICER_YEARS_ON_FORCE
## OFFICER_GENDER             1.00000000 -0.046625287           -0.072806320
## OFFICER_RACE              -0.04662529  1.000000000           -0.036309468
## OFFICER_YEARS_ON_FORCE    -0.07280632 -0.036309468            1.000000000
## OFFICER_INJURY            -0.03482966 -0.008217931           -0.064441067
## SUBJECT_RACE              -0.03331680  0.001892810           -0.010202744
## SUBJECT_GENDER             0.10841991 -0.040622148           -0.012901729
## SUBJECT_INJURY             0.02008415  0.044322749           -0.018146941
## SUBJECT_WAS_ARRESTED       0.03505750  0.003476999            0.006937561
## LEVEL_OF_FORCE             0.02300337  0.063054233           -0.178844436
##                         OFFICER_INJURY SUBJECT_RACE SUBJECT_GENDER
## OFFICER_GENDER          -0.034829664 -0.033316795   0.1084199149
## OFFICER_RACE            -0.008217931  0.001892810  -0.0406221481
## OFFICER_YEARS_ON_FORCE  -0.064441067 -0.010202744  -0.0129017292
## OFFICER_INJURY           1.000000000 -0.012367152   0.0324944688
## SUBJECT_RACE            -0.012367152  1.000000000   0.0293466595
## SUBJECT_GENDER           0.032494469  0.029346660   1.0000000000
## SUBJECT_INJURY           0.160717873 -0.036340729   0.0654727946
## SUBJECT_WAS_ARRESTED    -0.101002367  0.046630730   0.0009402345
## LEVEL_OF_FORCE          -0.131586875 -0.003420194   0.0368055386
##                         SUBJECT_INJURY SUBJECT_WAS_ARRESTED LEVEL_OF_FORCE
## OFFICER_GENDER              0.02008415         0.0350574985    0.023003368
## OFFICER_RACE                0.04432275         0.0034769995    0.063054233
## OFFICER_YEARS_ON_FORCE     -0.01814694         0.0069375606   -0.178844436
## OFFICER_INJURY              0.16071787        -0.1010023669   -0.131586875
## SUBJECT_RACE               -0.03634073         0.0466307302   -0.003420194
## SUBJECT_GENDER              0.06547279         0.0009402345    0.036805539
## SUBJECT_INJURY              1.00000000        -0.1107231855   -0.144484091
## SUBJECT_WAS_ARRESTED       -0.11072319         1.0000000000    0.125156716
## LEVEL_OF_FORCE             -0.14448409         0.1251567161    1.000000000
```

```
ggplotly(ggcorrplot(cor_matrix, hc.order = TRUE, outline.color = "white",
             colors = c("#37834f", "#f7f7f7", "#756bb1")) + theme_bw() + theme_rm +
    scale_x_discrete(expand = c(0,0)) + scale_y_discrete(expand = c(0,0)) +
    theme(axis.title.x = element_blank(),
          axis.title.y = element_blank(),
          panel.grid = element_blank(),
          axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 0.5)))
```

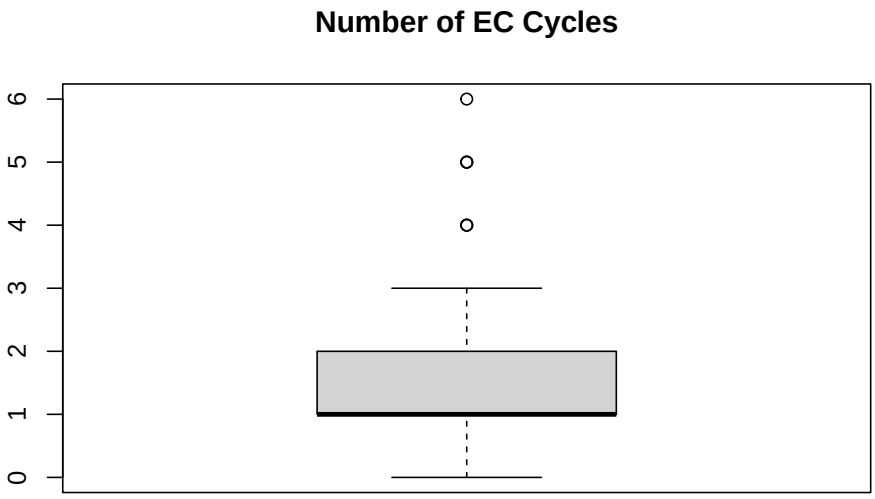
From the correlation matrix, we can make several observations. Following are a few:

- As suggested by the slightly positive correlation between officer race and subject race, male officers are more likely to interact with male individuals, while female officers interact with female subjects.
- Officer race has a weak negative correlation with their experience on the police force, which means that officers of certain races are more likely to have fewer years on the force compared to officers of other races.
- Officer years on force has a moderate negative correlation with the level of force used (correlation coefficient = -0.18), which means that officers with more years on the force are less likely to use force than officers with fewer years on the force.
- Subject gender has a weak positive correlation with subject arrest, which means that male subjects are slightly more likely to be arrested than female subjects.

Similarly, we can easily analyse the relationship between two variables from the correlation matrix.

Electric control devices are often used in situations where force is used in confrontation. Commonly known as a taser, ECs are considered a non-lethal weapon and are used to subdue a resisting subject. The number of EC cycles used could be an indicator of the force used in the event. The following boxplot shows the distribution of the number of EC cycles recorded in the given dataset. As EC cycle data is not present for a large number of observations, the null values are removed from the column.

```
RawData$NUMBER_EC_CYCLES <- as.numeric(RawData$NUMBER_EC_CYCLES)
cycles <- na.omit(RawData$NUMBER_EC_CYCLES)
boxplot(cycles)
title("Number of EC Cycles")
```

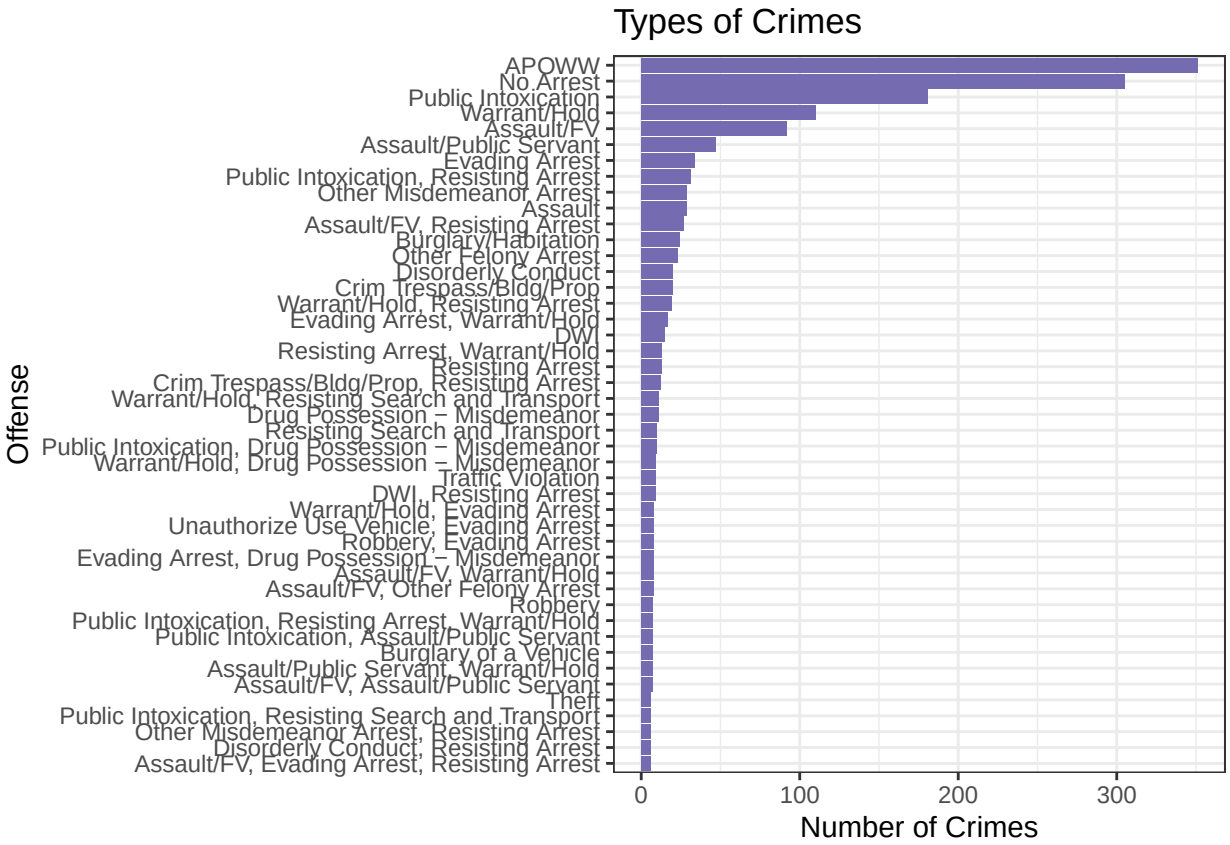


From this, we can see that the EC cycles are used very rarely, and a single use of an EC is effective, which makes a cycle greater than 3 an outlier in the plot.

Finally, we will proceed to the territorial interpretation of the data. For this, we are using the `Leaflet` package. Leaflet is a free open-source library that allows you to make interactive plots. It allows you to add markers, lines, polygons, and other shapes to customise their appearance and behaviour.

A subset of the main dataset is created to focus on crime and location details. From this data, a barplot showing the count of each offense is shown. The plot is filtered to show the categories with a count higher than 5, considering the total number of categories.

```
crime_data <- RawData[c("INCIDENT_DATE",    "INCIDENT_TIME", "SUBJECT_OFFENSE", "REPORTING_AREA",
                        "BEAT", "SECTOR", "DIVISION",   "LOCATION_DISTRICT",    "STREET_NUMBER",
                        "STREET_NAME",  "STREET_DIRECTION", "STREET_TYPE",
                        "LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION", "LOCATION_CITY",
                        "LOCATION_STATE",   "LOCATION_LATITUDE",    "LOCATION_LONGITUDE")]
crime_data$SUBJECT_OFFENSE <- factor(crime_data$SUBJECT_OFFENSE)
crime_data <- crime_data %>% rename(Latitude = LOCATION_LATITUDE,
                                    Longitude = LOCATION_LONGITUDE)
crime_data %>% count(SUBJECT_OFFENSE) %>%
  filter(n > 5) %>%
  ggplot(aes(x = reorder(SUBJECT_OFFENSE, n), y = n)) +
  geom_col(fill = "#756bb1") + coord_flip() + theme_rm + theme_bw() +
  ylab("Number of Crimes") + xlab("Offense") + ggtitle("Types of Crimes")
```



The data displayed in the chart above is plotted on the map below. Here, top 5 categories of crime are marked in different colors. When you click on the rest of the dots (blue), you can see which offense was recorded at that particular place.

The top five categories of offenses recorded in 2016 are:

1. APOWW (Apprehension by Peace Officer Without Warrant) : 351
2. Public Intoxication : 181
3. Warrant/Hold : 110
4. Assault/FV (Family Violence) : 92
5. Assault/Public Servant : 47

```
crime_data %>% na.omit() %>%
  leaflet() %>%
  addTiles() %>%
  addCircleMarkers(popup = ~SUBJECT_OFFENSE) %>%
  addCircleMarkers(data = crime_data[crime_data$SUBJECT_OFFENSE=="APOWW",], group = "APOWW", color = "#
  addCircleMarkers(data = crime_data[crime_data$SUBJECT_OFFENSE=="Public Intoxication",], group = "Publi
  addCircleMarkers(data = crime_data[crime_data$SUBJECT_OFFENSE=="Warrant/Hold",], group = "Warrant/Hol
  addCircleMarkers(data = crime_data[crime_data$SUBJECT_OFFENSE=="Assault/FV",], group = "Assault/FV", 
  addCircleMarkers(data = crime_data[crime_data$SUBJECT_OFFENSE=="Assault/Public Servant",], group = "As
```

The attached dashboard contains separate tabs for the distribution of offenses for each month in 2016.

Conclusion

According to the findings of this report, there is a considerable racial discrepancy in the frequency of occurrences involving officers and individuals. People of black ethnicity, in particular, are presumed to be in possession of weapons or narcotics merely because of their racial background. Furthermore, they suffer most significantly by police officers' use of force and are more likely to be hurt or hospitalised as a result of these instances.

To guarantee that all individuals are treated fairly and justly, law enforcement authorities must recognise and rectify racial inequities. This could include improving police training programmes to encourage impartial policing and holding individuals who engage in discriminatory behaviour accountable for their conduct.

Overall, it is essential to continue monitoring and analyzing police data to identify and address instances of racial discrimination in law enforcement. Only by taking proactive steps to promote fairness and equality can we hope to build a society that is truly just and equitable for all individuals, regardless of their race or ethnicity.