Data Visualization Project

Dallas Crime Data in 2016

```
### Required Packages
library(tidyverse)
library(ggplot2)
library(dplyr)
library(lubridate)
library(stringr)
library(ggcorrplot)
library(xts)
library(forecast)
library(leaflet)
library(GGally)
library(scales)
library(knitr)
library(gridExtra)
library(captioner)
library(kableExtra)



#setting the working directory path where the data set is present
setwd("C:/Users/sushm/OneDrive/Desktop/reftext")
crime_data<-read.csv("37-00049_UOF-P_2016_prepped (1).csv",stringsAsFactors = FALSE)   # loading the data set
```

Introduction

Racial disparities is potential problem in united states.It not only imbalances the democracy but also it undermines the social fabric. In US according to the research analysis it is said that one among four black people are being targeted by the police. researchers indicate that the race plays a prominent role in shaping policing outcomes. This disparities leads to increased injures,deaths, lost wages, psychological trauma. Police represent the government, if people doesn't trust police then their trust on law enforcement and democracy decreases.This will lead, children who are being harmed or mentally discriminated would most likely to turn up as a criminals. Thus it is important for the government to take the action against these discrepancies in order to have a better democracy.Hence, there are some volunteering associations which will aid in collecting the data useful for stabilizing the law and order in the state.

The CPE (center for policing equity) is a research consortium that promotes the stability between the law enforcement agencies and community they serve. Their main aim is to analyse the crime data and aid the government for enhancing the security in the areas where crime is more. This not only helps government in maintaining law and order but also allows citizens to have a peaceful life and enhances the trust on government.

Collecting the data alone doesn't help to undermine the crime rate, one need to analyse the data for a better result.A popular saying by John W.Tukey "The greatest value of a picture is when it forces us to notice what we never expected to see."."A picture is worth a thousand words. A well-designed chart is worth a million". Pictorial representation of data is like telling a story of the data through pictures and tables. It is true that "When you Visualize then you materialize".A good visual representation tend to enhance the underlying message[1]. Thus the importance of data visualizing pictorially comes into picture. The data visualization is a part of art and part of science.It helps to get the art right without getting the science wrong and vise versa. It is a study of how to transform the data values systematically and logically into a presentable illustrations.Aim is " To Tell a story".Following are the advantages of data visualization in crime analysis :

1.identify patterns 2. Communicates information 3. Aids in decision making 4. Identify the correlations between variables. 5. Easy analysis

The advancements in computer aided technology helps law enforcement department for effective an efficient strategies in dealing with crimes. Now-a-days each and every government is relaying on the data science for better productivity. This helps them to identify the criminal hot-spots.This report aims to analyze the data collected by the CPE in order to analyse the racial disparities in policing outcomes.Through comprehensive understanding of the issue, one could provide the actionable insights for law enforcement agencies, policymakers and community organizations to work together towards the racial disparities in policing and rebuilding the trust of people on government. This report uses a significant R package libraries to enhance the graphical representation.

Data Exploration

The dataset has 2384 records with 47 observational variables. It consists of crime data in Dallas incidents ranging from January 1, 2016 to December 31,2016.Officer id is unique for each officer and the count says the number of cases he/she worked on.The subject ID is uniquie for each subject. The Structure of the data given is in character datatype. Using the as.numeric and as. factor we convert the character data type to a numeric and factor datatypes. Whereas "lubridate" package is used for parsing the the given date and time. As we are taking about the racial discrimination the data gives a deeper insights on race of both police and subjects.The data completely focus on the race, injures, gender of both police and subject. It also contains the area where the incident had happen and the reason for the incident. Thus , the structure and summary of the data set gives us a deeper insight of what the data is talking about.

Structure of the Data

```
dallas_crime_data<-crime_data[-1,]   # removing the first column of the data as it is column names which is already present
as header of the file
str(dallas_crime_data)   # Viewing the structure of the data
```

```
'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" "8:55:00 PM" ...
 $ UOF_NUMBER                                   : chr  "37702" "33413" "34567" "31460" ...
 $ OFFICER_ID                                   : chr  "10810" "7706" "11014" "6692" ...
 $ 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                       : chr  "2" "17" "1" "24" ...
 $ OFFICER_INJURY                               : chr  "No" "Yes" "No" "No" ...
 $ OFFICER_INJURY_TYPE                          : chr  "No injuries noted or visible" "Sprain/Strain" "No injuries noted or v
isible" "No injuries noted or visible" ...
 $ OFFICER_HOSPITALIZATION                      : chr  "No" "Yes" "No" "No" ...
 $ SUBJECT_ID                                   : chr  "46424" "44324" "45126" "43150" ...
 $ 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 or visible" "No injuries
noted or visible" "Laceration/Cut" ...
 $ SUBJECT_WAS_ARRESTED                         : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ SUBJECT_DESCRIPTION                          : chr  "Mentally unstable" "Mentally unstable" "Unknown" "FD-Unknown if Arme
d" ...
 $ SUBJECT_OFFENSE                              : chr  "APOWW" "APOWW" "APOWW" "Evading Arrest" ...
 $ REPORTING_AREA                               : chr  "2062" "1197" "4153" "4523" ...
 $ BEAT                                         : chr  "134" "237" "432" "641" ...
 $ SECTOR                                       : chr  "130" "230" "430" "640" ...
 $ DIVISION                                     : chr  "CENTRAL" "NORTHEAST" "SOUTHWEST" "NORTH CENTRAL" ...
 $ LOCATION_DISTRICT                            : chr  "D14" "D9" "D6" "D11" ...
 $ STREET_NUMBER                                : chr  "211" "7647" "716" "5600" ...
 $ 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 BIMEBELLA LN" "5600 L B J FW
Y" ...
 $ LOCATION_CITY                                : chr  "Dallas" "Dallas" "Dallas" "Dallas" ...
 $ LOCATION_STATE                               : chr  "TX" "TX" "TX" "TX" ...
 $ LOCATION_LATITUDE                            : chr  "32.782205" "32.798978" "32.73971" "" ...
 $ LOCATION_LONGITUDE                           : chr  "-96.797461" "-96.717493" "-96.92519" "" ...
 $ 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 Down - Group" "K-9 Deploym
ent" ...
 $ 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" ...
```

```
dallas_crime_data$INCIDENT_DATE <- parse_date_time(dallas_crime_data$INCIDENT_DATE,orders = c('mdy','dmy'))  # parsing incid
ent date using lubridate package
dallas_crime_data<- dallas_crime_data%>% mutate_at(c('OFFICER_ID','OFFICER_YEARS_ON_FORCE','SUBJECT_ID','REPORTING_AREA','BE
AT','SECTOR','STREET_NUMBER','LOCATION_LATITUDE','LOCATION_LONGITUDE'),as.numeric)  # converting the character datatype vari
ables to numeric data type
dallas_crime_data<- dallas_crime_data %>% mutate_at(c('SUBJECT_RACE','SUBJECT_GENDER','SUBJECT_INJURY','SUBJECT_INJURY_TYP
E','SUBJECT_WAS_ARRESTED','SUBJECT_DESCRIPTION','SUBJECT_OFFENSE','DIVISION','INCIDENT_REASON','REASON_FOR_FORCE','OFFICER_H
OSPITALIZATION','STREET_DIRECTION','STREET_TYPE','LOCATION_DISTRICT','TYPE_OF_FORCE_USED1','TYPE_OF_FORCE_USED2','TYPE_OF_FO
RCE_USED3','TYPE_OF_FORCE_USED4','TYPE_OF_FORCE_USED5','TYPE_OF_FORCE_USED6','TYPE_OF_FORCE_USED7','TYPE_OF_FORCE_USED8','TY
PE_OF_FORCE_USED9','TYPE_OF_FORCE_USED10','LOCATION_CITY','LOCATION_STATE','OFFICER_INJURY','OFFICER_RACE','LOCATION_FULL_ST
REET_ADDRESS_OR_INTERSECTION','NUMBER_EC_CYCLES','FORCE_EFFECTIVE','OFFICER_GENDER','STREET_NAME'),as.factor)   # Converting
the character data to factor
```

The above table shows the structure of the raw data containing the 47 variables with 2383 observations.All the variables are in the character datatype even though the values in it are numeric. So we need to clean the data prior analyzing the data and visualizing the observations.

Summary of the Dallas Dataset

```
summary_dallas<-summary(dallas_crime_data)   # Summarizing the data
summary_dallas
```

```
INCIDENT_DATE                    INCIDENT_TIME       UOF_NUMBER
Min.   :2016-01-01 00:00:00.00   Length:2383        Length:2383
1st Qu.:2016-03-11 12:00:00.00   Class :character   Class :character
Median :2016-05-30 00:00:00.00   Mode  :character   Mode  :character
Mean   :2016-06-10 06:35:48.22
3rd Qu.:2016-09-05 00:00:00.00
Max.   :2016-12-31 00:00:00.00

  OFFICER_ID    OFFICER_GENDER       OFFICER_RACE  OFFICER_HIRE_DATE
Min.   :    0   Female: 240    American Ind:   8   Length:2383
1st Qu.: 8902   Male  :2143    Asian       :  55   Class :character
Median :10115                  Black       : 341   Mode  :character
Mean   : 9572                  Hispanic    : 482
3rd Qu.:10710                  Other       :  27
Max.   :11170                  White       :1470

OFFICER_YEARS_ON_FORCE OFFICER_INJURY OFFICER_INJURY_TYPE
Min.   : 0.000         No :2149       Length:2383
1st Qu.: 3.000         Yes: 234       Class :character
Median : 6.000                        Mode  :character
Mean   : 8.049
3rd Qu.:10.000
Max.   :36.000

OFFICER_HOSPITALIZATION   SUBJECT_ID          SUBJECT_RACE  SUBJECT_GENDER
No :2335                Min.   :    0   American Ind:   1   Female : 440
Yes:  48                1st Qu.:43307   Asian       :   5   Male   :1932
                        Median :44573   Black       :1333   NULL   :  10
                        Mean   :40255   Hispanic    : 524   Unknown:   1
                        3rd Qu.:46088   NULL        :  39
                        Max.   :47972   Other       :  11
                                        White       : 470
SUBJECT_INJURY                   SUBJECT_INJURY_TYPE SUBJECT_WAS_ARRESTED
No :1754       No injuries noted or visible:1622     No : 335
Yes: 629       Abrasion/Scrape             : 209     Yes:2048
               Laceration/Cut              :  56
               Puncture                    :  35
               Non-Visible Injury/Pain     :  34
               Taser Burn Marks            :  30
```

```
               (Other)                    : 397
                SUBJECT_DESCRIPTION              SUBJECT_OFFENSE
Mentally unstable         :412      APOWW                : 351
Alchohol                  :382      No Arrest            : 305
Unknown                   :364      Public Intoxication  : 181
Unknown Drugs             :318      Warrant/Hold         : 110
None detected             :297      Assault/FV           :  92
Alchohol and unknown drugs:280      Assault/Public Servant:  47
(Other)                   :330      (Other)              :1297
REPORTING_AREA      BEAT           SECTOR             DIVISION
Min.   :1001   Min.   :111.0   Min.   :110   CENTRAL      :563
1st Qu.:2014   1st Qu.:216.0   1st Qu.:210   NORTH CENTRAL:319
Median :2231   Median :351.0   Median :350   NORTHEAST    :341
Mean   :3191   Mean   :392.8   Mean   :389   NORTHWEST    :191
3rd Qu.:4327   3rd Qu.:612.0   3rd Qu.:610   SOUTH CENTRAL:310
Max.   :9611   Max.   :757.0   Max.   :750   SOUTHEAST    :362
                                             SOUTHWEST    :297
LOCATION_DISTRICT STREET_NUMBER      STREET_NAME   STREET_DIRECTION
D14     :313       Min.   :    0   Commerce :  48   E    : 120
D2      :310       1st Qu.: 1700   Forest   :  35   N    : 225
D7      :231       Median : 3415   Buckner  :  29   NULL:1728
D4      :222       Mean   : 4904   Lamar    :  25   S    : 187
D6      :213       3rd Qu.: 7532   Elm      :  24   W    : 123
D8      :174       Max.   :54023   Northwest:  20
(Other):920                        (Other)  :2202
 STREET_TYPE  LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION LOCATION_CITY
St.    :557   111 COMMERCE ST      :  25                   Dallas:2383
Rd.    :400   111 W COMMERCE ST    :  22
Ave.   :353   7808 CLODUS FIELDS DR:  21
Dr.    :321   10000 WALNUT ST      :  16
Ln.    :219   1600 CHESTNUT ST     :  12
Blvd.  :160   5200 HARRY HINES BLVD:  12
(Other):373   (Other)              :2275
LOCATION_STATE LOCATION_LATITUDE LOCATION_LONGITUDE
TX:2383        Min.   :32.63     Min.   :-96.96
               1st Qu.:32.74     1st Qu.:-96.82
               Median :32.78     Median :-96.79
               Mean   :32.80     Mean   :-96.78
               3rd Qu.:32.86     3rd Qu.:-96.75
```

```
               Max.   :33.02     Max.   :-96.57
               NA's   :55        NA's   :55
             INCIDENT_REASON                 REASON_FOR_FORCE
Arrest               :1157   Arrest                  :1050
Service Call         : 673   Danger to self or others: 347
Call for Cover       : 131   Active Aggression       : 346
Traffic Stop         :  93   Detention/Frisk         : 206
Crime in Progress    :  82   Weapon Display          : 195
Other ( In Narrative):  70   Other                   : 148
(Other)              : 177   (Other)                 :  91
              TYPE_OF_FORCE_USED1               TYPE_OF_FORCE_USED2
Verbal Command          :818                              :747
Weapon display at Person:329      Verbal Command          :282
Held Suspect Down       :176      Held Suspect Down       :243
BD - Grabbed            :154      BD - Grabbed            :162
Take Down - Arm         :144      Joint Locks             :130
Joint Locks             :140      Weapon display at Person:125
(Other)                 :622      (Other)                 :694
            TYPE_OF_FORCE_USED3             TYPE_OF_FORCE_USED4
                        :1510                           :1996
Held Suspect Down     : 218     Held Suspect Down     : 100
Verbal Command        : 130     Verbal Command        :  48
Joint Locks           :  58     Joint Locks           :  31
BD - Grabbed          :  55     BD - Grabbed          :  30
Hand Controlled Escort:  50     Hand Controlled Escort:  19
(Other)               : 362     (Other)               : 159
            TYPE_OF_FORCE_USED5             TYPE_OF_FORCE_USED6
                        :2226                           :2322
Held Suspect Down     :  38     Held Suspect Down    :  17
Hand Controlled Escort:  12     Verbal Command       :   7
Pressure Points       :  12     Feet/Leg/Knee Strike :   5
Taser                 :  12     Hand/Arm/Elbow Strike:   5
Joint Locks           :  11     Joint Locks          :   4
(Other)               :  72     (Other)              :  23
          TYPE_OF_FORCE_USED7             TYPE_OF_FORCE_USED8
                      :2361                           :2378
Held Suspect Down   :   5     BD - Grabbed         :   1
Feet/Leg/Knee Strike:   2     Handcuffing Take Down:   1
Joint Locks         :   2     Held Suspect Down    :   1
```

```
 OC Spray            :   2     Joint Locks          :   1
 Pressure Points     :   2     Verbal Command       :   1
 (Other)             :   9
     TYPE_OF_FORCE_USED9   TYPE_OF_FORCE_USED10 NUMBER_EC_CYCLES
                :2382                 :2382      NULL   :2226
 Verbal Command:   1     BD - Grabbed:   1      1      :  96
                                                2      :  34
                                                3      :  13
                                                0      :   3
                                                4      :   3
                                                (Other):   8
     FORCE_EFFECTIVE
  Yes          :673
  Yes, Yes     :352
  No, Yes      :276
  No, Yes, Yes:161
  No, No, Yes  :124
  Yes, No      : 85
 (Other)       :712
```

The above table shows the summary of the given Dallas crime data set.This is the formatted data (changed the datatype of each variable by observing its structure).It can be interpreted that the larger portion of the police department jobs are occupied by the males and whites. The number of subjects injured far outrages the number of police injured during this year. Whereas, most of the subjects are black and male.It is analyzed that the most common reason for the crime is subject being mentally unstable,alcoholic.The mean of officer years on force is 8.049 years with a range from 0 to 36 years.

```
# Tabulating the number of cases solved or worked by each officer
officer_count<-dallas_crime_data%>%
  count(OFFICER_ID,sort=T)  # counting by the officer ID

names(officer_count)<-c('OFFICER_ID','Number_of_cases')  # Renaming the columns
```

```
# extracting top 10 officers arranged by the descending order of their cases count
Top_10_officers<-officer_count %>%
  dplyr::filter(Number_of_cases >=10)   # filtering by the number of cases solved greater than or equal to 10

kable(Top_10_officers,caption = "Table 1: Number of cases handled by different officers, ranked in descending order of the t
otal number of cases over an year") %>%
  kable_styling(position="left",font_size = 12)
```

Table 1: Number of cases handled by different officers, ranked in descending order of the total number of cases over an year

OFFICER_ID	Number_of_cases
10724	25
10697	21
10710	18
10818	16
10498	12
11015	12
9925	11
10695	11
10760	11
8525	10
9881	10
10351	10

The table shows the top officers in descending order of the number of cases they solved.There are around 1,042 officers and the maximum number of cases were dealt with the officer with ID 10724 followed by officers with ID 10697,10710,10818. There are 12 officers who have worked on more than 10 cases through out this year. Whereas, around 931 officers worked on less than 5 cases across this 12 months period. Thus , by

tabulating one can interpret the effectiveness of the officers and their capability in dealing with the cases. This helps to identify the effective officers and gives an opportunity for government to appreciate these officers and focus on the efficiency of the other officers who have worked on less than 5 cases.

```
# Tabulating the number of officers categorized by gender and race
officer_gender <- dallas_crime_data %>%
  group_by(OFFICER_GENDER)%>%                 # grouping by gender
  count(OFFICER_GENDER,OFFICER_RACE) %>%      # count no.of officers by race grouped by gender
  arrange(desc(n))                            # arrange in descending order of count
gender_table<-officer_gender %>%
  pivot_wider(values_from = n,names_from = OFFICER_GENDER,values_fill = 0)      # widening the table
gender_table$Total_no_of_officers<- gender_table$Male + gender_table$Female     # computing total number of officers of each
race
gen_tab<-kable(head(gender_table),caption = "Table2: Officers categorized by gender")%>%
  kable_styling(position="left",font_size = 14)

# Plotting the officer gender table for graphical representation

gen_graph<-ggplot(officer_gender,aes(x=reorder(OFFICER_RACE,n),y=n))+        # defining aesthetics of ggplot
  geom_col(position="dodge",aes(col=OFFICER_GENDER,fill=OFFICER_GENDER))+    # using the geom_col for column plot
  ylab("Number of Officers") +                                               # Labeling Y axis
  xlab("Officer race")+                                                      # labeling x axis
  ggtitle("Number of officer classifed by race")+                            # Labeling the title
  coord_flip()+                                                              # flipping the axis
  labs(caption = str_wrap("Figure1: Column plot showing the number of officers classified by race \nand gender, with the X-a
xis representing different officer races and the \nY-axis representing the number of officers. The plot highlights the \ndis
tribution of male and female officers within each racial category",width=90))+
  theme(plot.caption = element_text(hjust=0,size = 10),
        plot.title = element_text(hjust = 0.5,face = "bold"))
gen_tab
```

Table2: Officers categorized by gender

OFFICER_RACE	Male	Female	Total_no_of_officers
White	1336	134	1470
Hispanic	440	42	482

OFFICER_RACE	Male	Female	Total_no_of_officers
Black	292	49	341
Asian	48	7	55
Other	21	6	27
American Ind	6	2	8

```
gen_graph
```

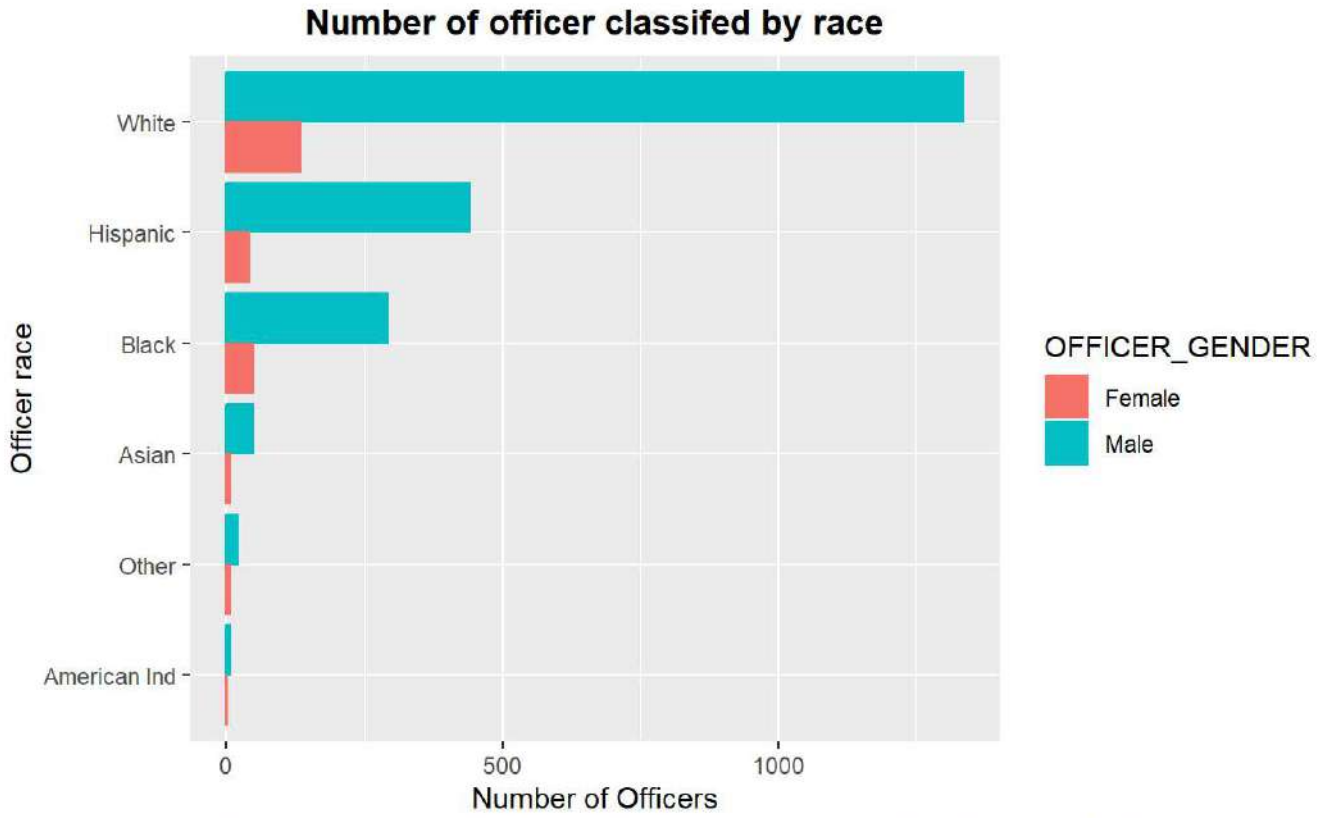


Figure1: Column plot showing the number of officers classified by race and gender, with the X-axis representing different officer races and the Y-axis representing the number of officers. The plot highlights the distribution of male and female officers within each racial category

The above table shows the number of officers classified by race and gender. There are 1470 white race officers among them 1336 are male and 134 are female. whereas , there are only 292 male and 49 female black race officers, which significantly shows the racial discrepancy in job market.The barplot gives the graphical representation of the tabulated data.The above barplot shows the classification of officers by races and differentiated by gender. It can be interpreted that the police department is mostly engaged by the male white race officers. Hispanic race officers are the second majority polices during the year 2016 in Dallas, whereas the number of Black race officers appear to be less than 1/3rd of the total police force. Thus , we can interpret that race and gender plays a prominent role in police job market. This also shows that the women population in police force is far less than male, this could be due to the culture of sexism and harassment. To address these issues, law enforcement agencies could implement some polices and initiatives aimed at increasing diversity and equity within work force.Such as , introducing schemes for women development and support at work place, allocating certain percent share of jobs to minority categories.

```
#calculating the crimes committed by each individual
dallas_subject_count<-dallas_crime_data%>%
  count(SUBJECT_ID,sort=T) %>%          # counting the crimes by subject id
  dplyr::filter(n>5)                    #filtering by n >5

names(dallas_subject_count)<- c("SUBJECT_ID","No_of_crimes_committed")          # renaming the columns
kable(dallas_subject_count,caption="Table3: Number of crimes committed by individuals with unique subject IDs ,who have comm
ited more than 5 crimes")%>%
  kable_styling(position="center",font_size = 14)      # printing the subject crime count data
```

Table3: Number of crimes committed by individuals with unique subject IDs ,who have commited more than 5 crimes

SUBJECT_ID	No_of_crimes_committed
0	147
43676	9
44942	8
45855	7
47548	7
43966	6
44127	6

SUBJECT_ID	No_of_crimes_committed
44918	6
46671	6
46815	6

From the above table it can be interpreted that most of the criminals data is not stored by the police depart database. They lack the id of the criminal. There are almost 147 cases which doesn't have the subject id . There are around 9 subjects who committed more than 5 crimes. This aids in identifying the dangerous individuals who may require an additional government or law enforcement attention. It could help in tracking the subjects who are more likely to commit crime and take required prior actions.

```
# Plotting injured subjects categorized by race and gender

subject_data<-dallas_crime_data %>%
  count(SUBJECT_RACE,SUBJECT_GENDER,SUBJECT_INJURY,sort = T) %>%  # counting the number of subjects injured classified by ra
ce,gender
  filter(n>1)     # Filtering by count greater than 1
injury_status<-c("Not injured","Injured")                          # defining a injury status vector
names(injury_status)<-c("No","Yes")                                # defining the names of the injury status
ggplot(subject_data,aes(x=SUBJECT_RACE,y=n))+                      # defining aesthetics of the ggplot of subject data
  geom_col(position="dodge",aes(fill=SUBJECT_GENDER))+             # using geom_col for the graphical representation
  facet_grid(SUBJECT_INJURY ~ .,labeller = labeller(SUBJECT_INJURY=injury_status)) +  # differentiating plot by injury statu
s
  xlab("Subject race")+                                            # labeling x axis
  ylab("Number of subjects")+                                      # labeling y axis
  ggtitle("Injured subjects categorised by race and gender")+      # labeling the graph
  labs(caption=str_wrap("Figure 2: Column plot showing the number of subjects categorized by race and gender, with the X-axi
s representing different subject races and the Y-axis representing the number of subjects. The plot is differentiated by inj
ury status, with separate facets for injured and non-injured subjects. The plot highlights the distribution of male and fema
le subjects within each racial category.",width=100))+
  theme(plot.caption=element_text(hjust=0,size = 10),
        plot.title = element_text(hjust = 0.5,face = "bold"))
```

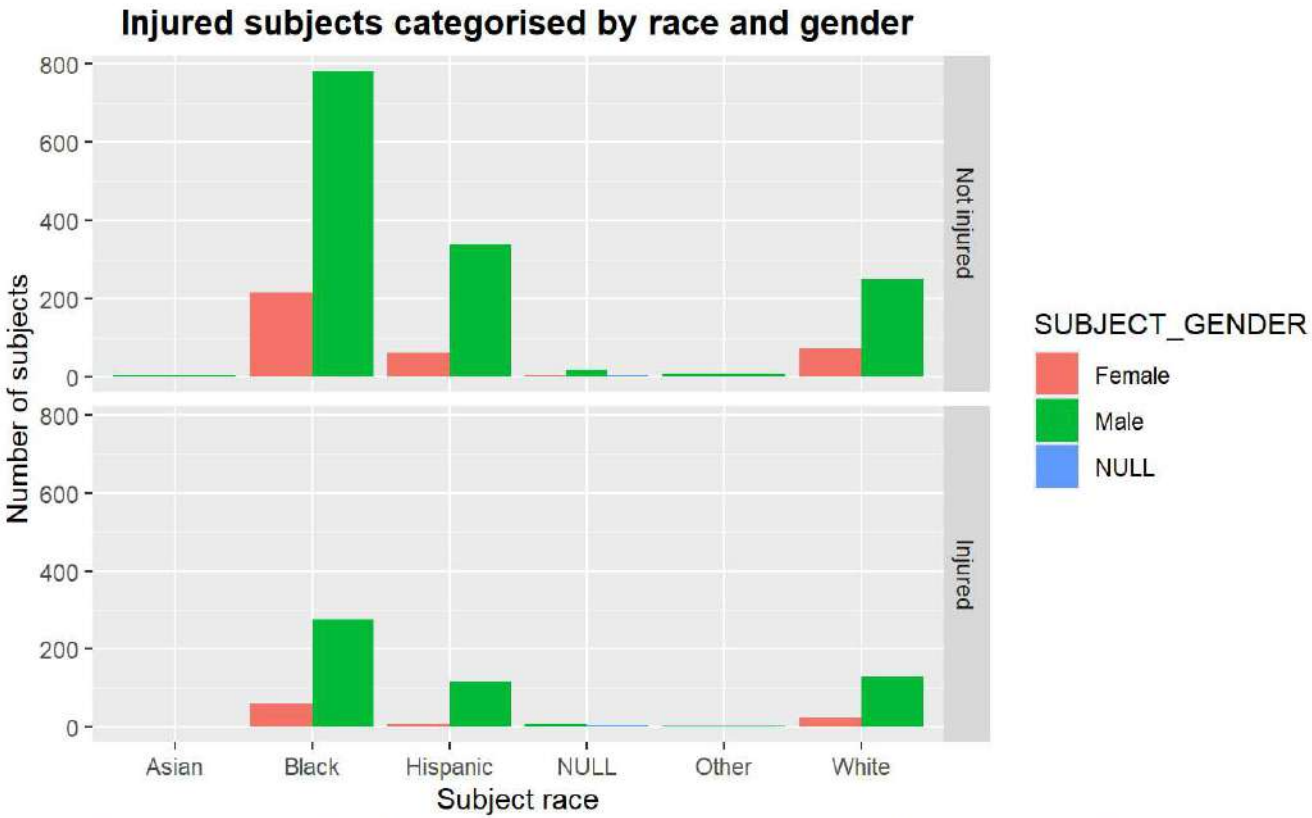


Figure 2: Column plot showing the number of subjects categorized by race and gender, with the X-axis representing different subject races and the Y-axis representing the number of subjects. The plot is differentiated by injury status, with separate facets for injured and non-injured subjects. The plot highlights the distribution of male and female subjects within each racial category.

By visually comparing the graphical distribution of the subjects by race,gender, and injury status, It can be said that most of the subjects were not injured.The black race males were injured mostly when compared with the total injured people followed by the hispanic and white race subjects. Overall, injured male count is higher than the injured women count. This information could be useful in identifying the race who are being suffered by injuries and by the deeper analysis of government agencies one could know whether the injuries are needed to control the law and order.

```
subject_arrest<-dallas_crime_data %>%
  count(SUBJECT_WAS_ARRESTED,SUBJECT_RACE,sort=T) %>%        #counting the number of crimes committed categorized by arrest
and race
  pivot_wider(names_from = SUBJECT_WAS_ARRESTED,values_from = n,values_fill = 0)  # Widening the data
names(subject_arrest)<-c("SUBJECT_RACE","Arrested","Not_Arrested")  # renaming the columns
crime_subjects<-subject_arrest %>%
  mutate(Total_crimes_race=subject_arrest$Arrested + subject_arrest$Not_Arrested)  # muatating the Total crime by each race
kable(crime_subjects,caption = " Table 4:Arrest and non-arrest counts for different racial groups involved in total crimes,
including black, Hispanic, white, other, Asian, and American Indian") %>%
    kable_styling(position="center",font_size = 12)
```

Table 4:Arrest and non-arrest counts for different racial groups involved in total crimes, including black, Hispanic, white, other, Asian, and American Indian

SUBJECT_RACE	Arrested	Not_Arrested	Total_crimes_race
Black	1144	189	1333
Hispanic	451	73	524
White	413	57	470
NULL	26	13	39
Other	8	3	11
Asian	5	0	5
American Ind	1	0	1

Among the 2,383 subjects 2048 subjects were arrested. It can be interpreted that more than 85% of the subjects are arrested by the police. It can be interpreted that their is no disparities in arrests, i.e each race police have arrested more than 85% of the subjects.This illustrates the criminal arrest rate across each race.This helps to discuss the reason for not arresting the criminals and could helpful in developing a strategies for capturing them by the likeliness of crime committing by each subject.

```r
## Graphical representation of the frequency of each type of injury
dallas_crime_data$SUBJECT_INJURY_TYPE<-sapply(str_split(dallas_crime_data$SUBJECT_INJURY_TYPE, ",\\s*"), function(x) x[1])
# cleaning the data
type_injury<-dallas_crime_data %>%
  count(SUBJECT_INJURY_TYPE,sort=T) %>%                                    # frequency count of each injury
  filter(n>10)                                                             # filtering the frequency count by >10

ggplot(type_injury,aes(x=reorder(SUBJECT_INJURY_TYPE,n),y=n))+             # defining the ggplot aesthetics of type inj
ury data
  geom_bar(stat='identity',fill='#9ACD32') +                               # using geom_bar for the graphical represent
ation
  ylab("Number injuries")+                                                 # labeling y axis
  xlab("Injury type")+                                                     # labeling xaxis
  ggtitle("Count of each injury")+                                         # labelling the graph
  coord_flip()+                                                             # flipping the graph
  labs(caption=str_wrap("Figure 3:Bar plot showing the count of each injury type, with the X-axis representing the different
injury types and the Y-axis representing the number of injuries. The plot highlights injury types that have a frequency coun
t greater than 10.",width=80))+
  theme(plot.caption = element_text(hjust=0,size = 10),
        plot.title = element_text(hjust = 0.5,face="bold"))
```

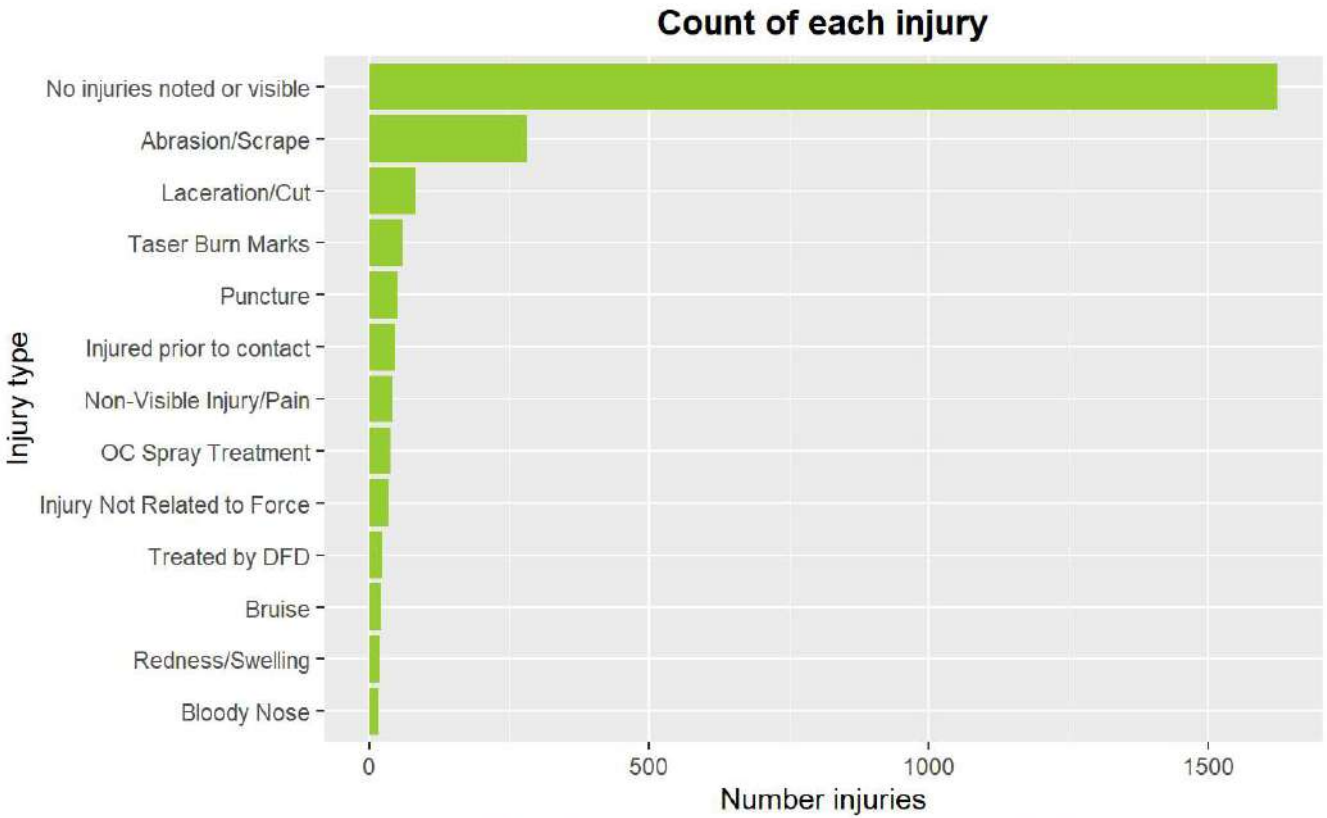


Figure 3:Bar plot showing the count of each injury type, with the X-axis representing the different injury types and the Y-axis representing the number of injuries. The plot highlights injury types that have a frequency count greater than 10.

There are 44 different types of injuries and 10 injures are very common (more than 20).In most cases no injuries are visible or noted among the subjects. Abrasion/scrape is the most commonly observed injury by subjects, followed by laceration/cut,taser burn marks. There were also a significant number of injuries related to nose blood,bruise and swelling.This could give us what type of injuries were faced by the subjects and severity of injuries. This could help in discussing the reason behind the use of force.

```
# representing the boxplot of the officers year categorized by gender

ggplot(dallas_crime_data, aes(x = OFFICER_GENDER, y = OFFICER_YEARS_ON_FORCE)) +    # defining the aesthetics of the ggplot
of Dallas crime data
  geom_boxplot(fill = "lightblue", color = "blue") +                                 # using geom_boxplot for analyzing the
data
  labs(title = "Officer Age Distribution by Gender", x = "Gender", y = "Officer Age",
       caption=str_wrap("Figure4:Box plot showing the distribution of officer age by gender, with the X-axis representing th
e gender of officers and the Y-axis representing the age of officers. The plot highlights the central tendency and distribut
ion of officer age for each gender.",width=100))+                                                  # labeling the x, y and graph
  theme(plot.caption = element_text(hjust=0,size = 10),
        plot.title = element_text(hjust = 0.5,face="bold"))
```

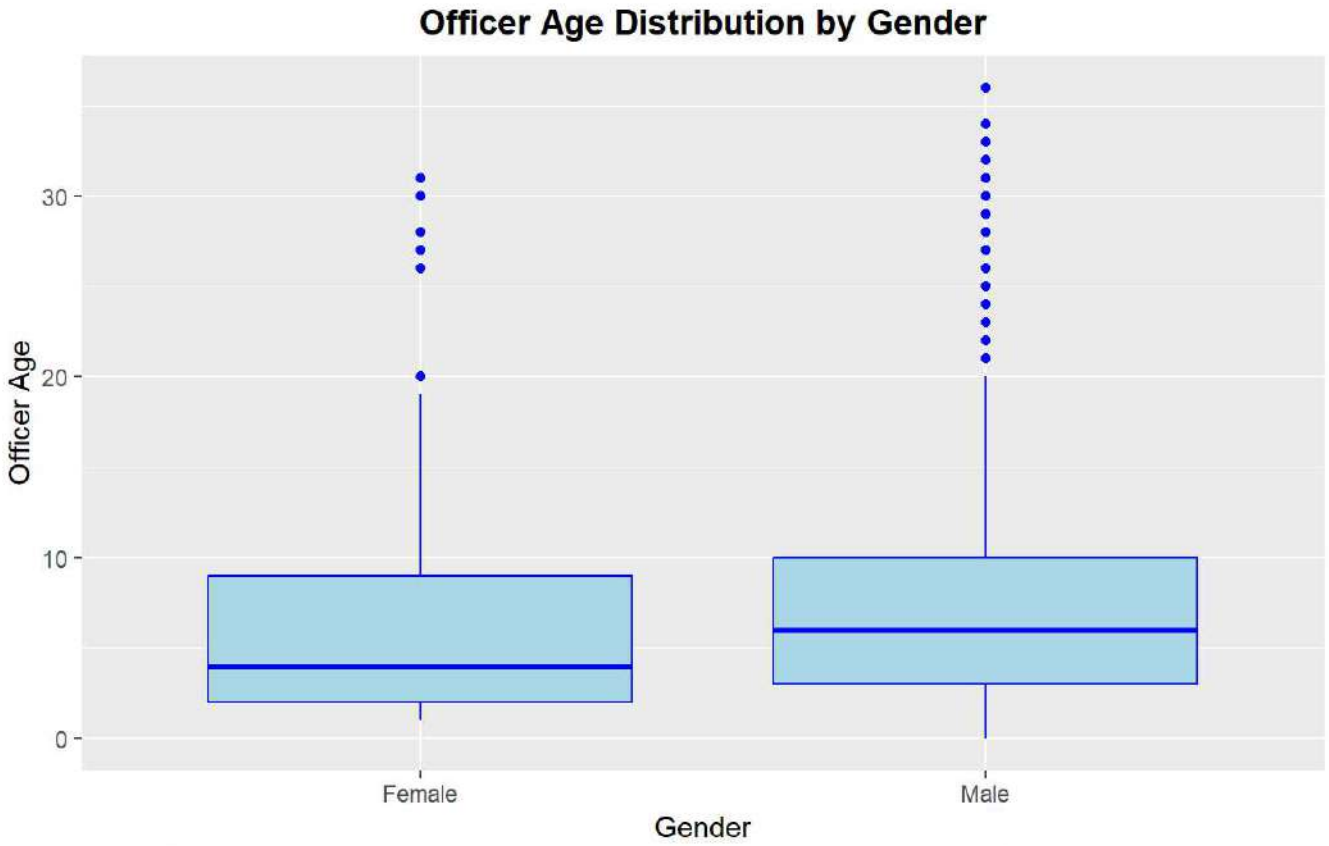


Figure4:Box plot showing the distribution of officer age by gender, with the X-axis representing the gender of officers and the Y-axis representing the age of officers. The plot highlights the central tendency and distribution of officer age for each gender.

The above boxplot shows that there are very thin density of officers in both the sex groups who are having experience more than 20. It shows that the median of female officers is less than the male officers. Whereas, the third quartile range of Male officers is slightly more than the female officers.The inter quartile range of both female and male officers is almost same. There were very few officers who had experience more than 20. it is visualized that male officers have higher years of experience than the female officers. This tells that officers after there 20 years mostly take leave the job this may be due to health issues, lack of strength, injuries etc.

```
# Using the violin plot along with the boxplot for visualizing the density of the officer race and the years of experience
ggplot(dallas_crime_data, aes(x = OFFICER_RACE, y = OFFICER_YEARS_ON_FORCE)) +  # defining the aesthetics of the ggplot of D
allas crime data
  geom_violin(fill = "lightgreen", color = "darkgreen") +                       # using geom_violin for graphical representa
tion
  labs(title = "Officer Years on Force Distribution by Race", x = "Race", y = "Years on Force",
       caption=str_wrap(" Figure 5:Violin plot showing the distribution of officer years on force by race, with the X-axis r
epresenting the race of officers and the Y-axis representing the years of experience. The plot visualizes the density of the
data for each race. The plot also includes a box plot for each race.",width=100))+ # labeling the x axis , y axis and graph
  theme(plot.caption = element_text(hjust = 0,size=10),
        plot.title = element_text(hjust = 0.5,face="bold"))+
    geom_boxplot(width = 0.08)                                          # defining the width of the box plot inside the violin plo
t
```

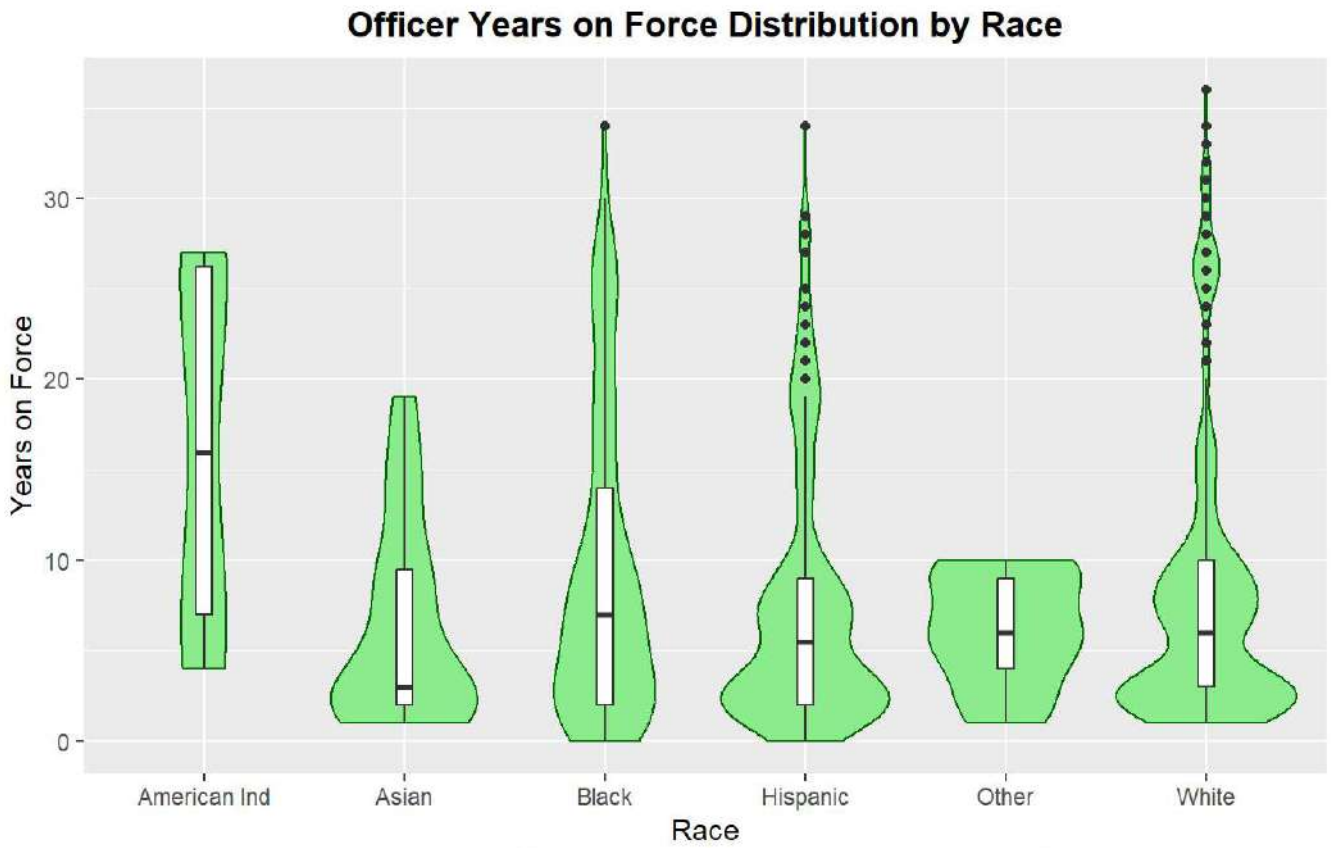


Figure 5:Violin plot showing the distribution of officer years on force by race, with the X-axis representing the race of officers and the Y-axis representing the years of experience. The plot visualizes the density of the data for each race. The plot also includes a box plot for each race.

The above plot shows the viloin plot combined with the boxplot. It shows the officers years of experience across different race. The boxplot helps us to interpret the median and quartile range and outliers of each race.This graph represents the years of experience of the each race.A large number of outliers can be seen by the white race officers have a larger number of outliers, whereas the american ind officers have the wider interquartile range in experience. Asian officers experience has the lowest median. The density of officers having less number of experience is more. It is observed there were no officers from Asian race who are having experience more than 20 and other category race does not have officers having experience more than 10 years. this might be due to the racial discrimination. By using this information government agencies might get an overview of the distribution of the officers across each race and the reason behind the officers with very less years of experience of a specific race.

```
# pie chart for the graphical interpretation of the divisions which were criminal hotspots
division <- dallas_crime_data %>%
  group_by(DIVISION) %>%                                        # grouping by the division
  count() %>%                                                   # frequency of crimes occurred in that division
  ungroup() %>%                                                 # ungrouping the grouped data
  mutate(perc = `n` / sum(`n`)) %>%                             # calculating the percentage distribution of crimes across e
ach reason
  arrange(perc) %>%                                             # arranging in the ascending order of the crimes
  mutate(labels = scales::percent(perc))


ggplot(division, aes(x = "", y = perc, fill = DIVISION)) +     # defining the ggplot aesthetics of the division data
  geom_bar(stat = "identity", width = 5) +                      # using geom_bar for graphical representation
  geom_text(aes(label = labels),                                # printing the percentage value on the pie chart
            position = position_stack(vjust = 0.5),size=4) +
  coord_polar(theta = "y")+                                     # cob=nverting into the pie chart
  ggtitle("Crime percentage in each division")+
  theme_void()+
  scale_fill_manual(values = c("blue", "red", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7"))+
  labs(caption=str_wrap("Figure 6 :Pie chart showing the percentage distribution of different divisions, with percentage val
ue   displayed.",width=100))+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust = 0.5, face="bold"))
```

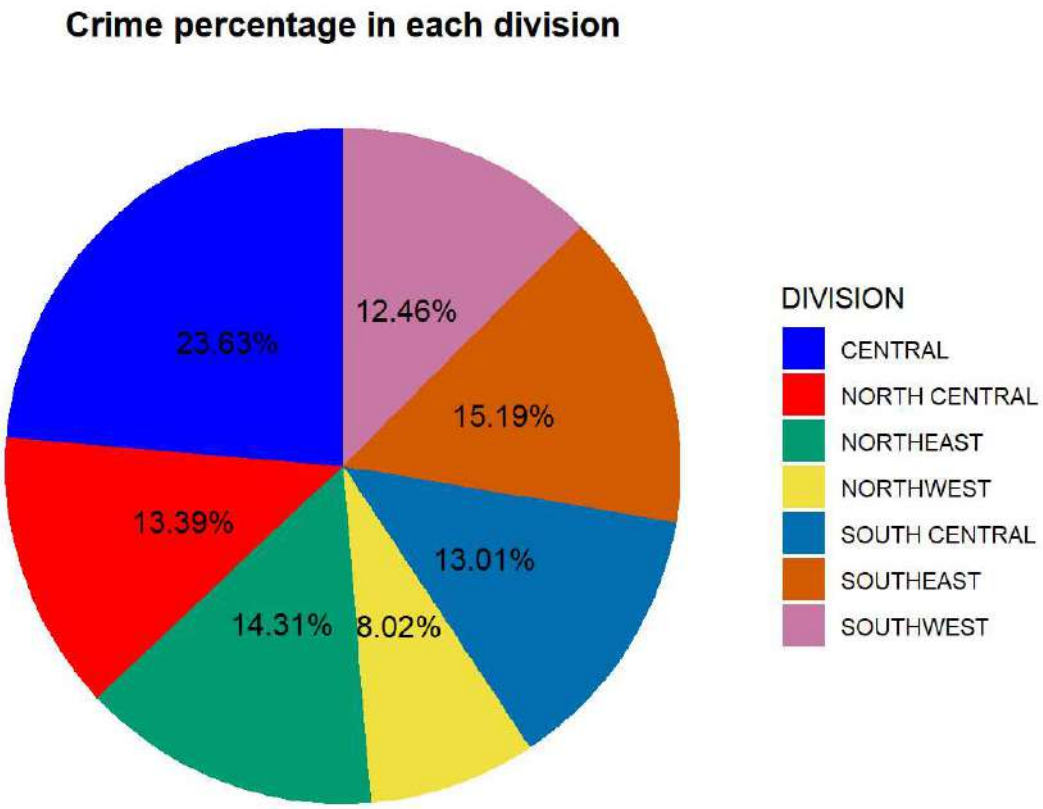


Figure 6 :Pie chart showing the percentage distribution of different divisions, with percentage value displayed.

The Dallas city is divided into 7 divisions(Northwest, Southwest, South Central, North Central, Northeast, Southeast, Central).The pie chart shows that the maximum crimes were recorded in central division whereas the least crimes were recorded in northwest.The Southeast division had the second-highest number of incidents,followed by the northeast. This helps to identify the area having the highest crime rate, which aids the police department to analyse the root cause of the crime rate in that area and helps to develop required strategies to balance the law in the central region.

```
#
ggplot(dallas_crime_data, aes(x = factor(month(INCIDENT_DATE)))) +    # defining the aesthetics of ggplot Dallas crime data
(incident date factoring by month)
  geom_histogram(stat='count',fill="#BC8F8F") +                       # using histogram for graphical representation
  labs(x = "Month", y = "Total Incidents",
       caption=str_wrap("Figure 7: Histogram showing the total number of incidents across the year, with the X-axis represen
ting different months and the Y-axis representing the total count of incidents.",width=100))+ # labeling the x axis and y ax
is
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust=0.5,face="bold"))+
  ggtitle("Number of incidents across the year")                      # labeling the title of the graph
```

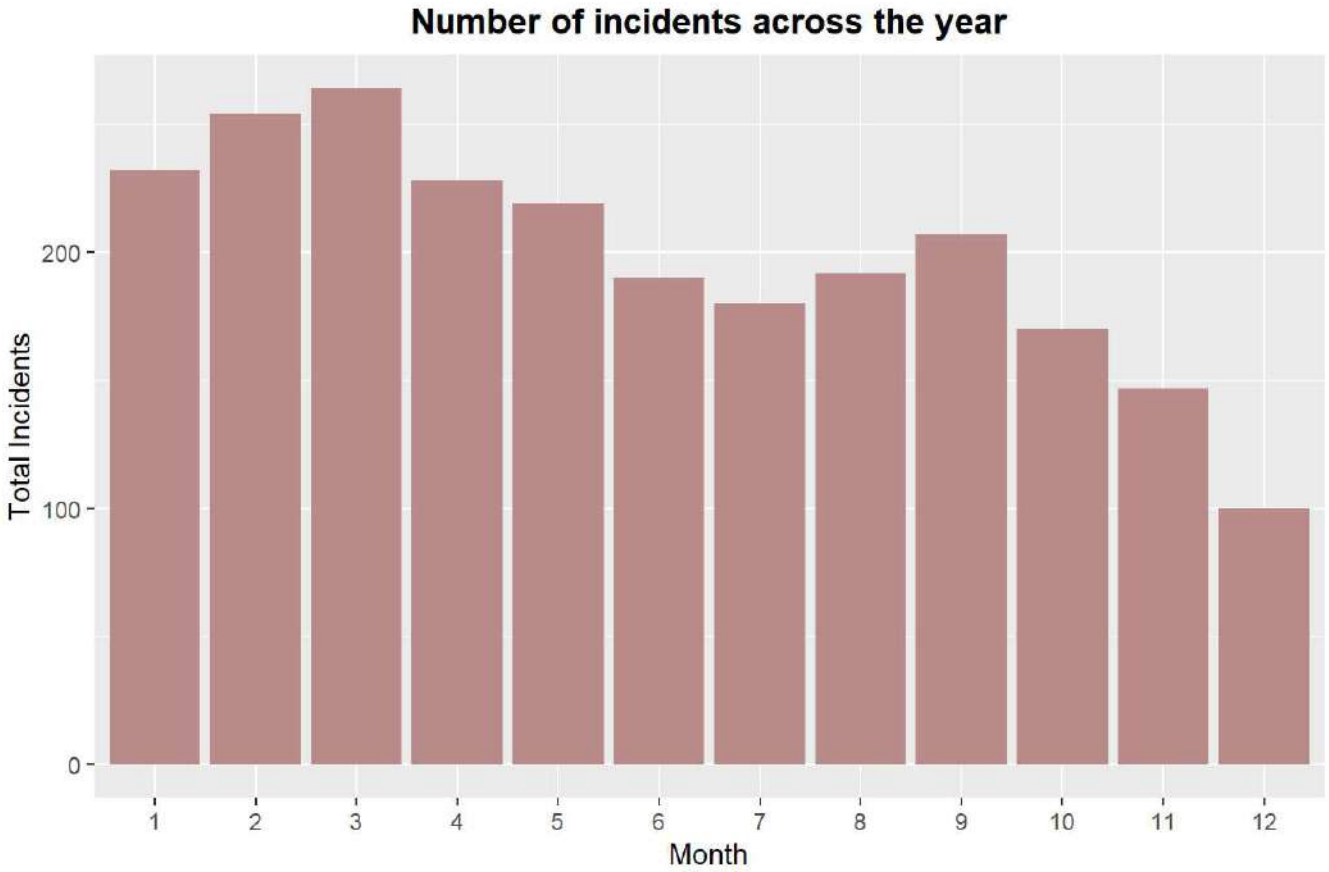


Figure 7: Histogram showing the total number of incidents across the year, with the X-axis representing different months and the Y-axis representing the total count of incidents.

The above histogram represents the total number of incidents per each month.During the first quarter of the year the crime is high where as during the last quarter of the month the crime rate is plummeted. March bags the highest number of incidents whereas the December shows the least number of crimes.June, July and August months had shown the average number of incidents. While remaining months have recorded partially above average number of incidents. [3]This provides the month records of the crime rate , which helps officials to dive deeper into the root cause of the crimes which may include, political riots during the elections, festivals of different races, etc.

```
#constructing the table for the number of crimes per each month

dallas_crime_data$SUBJECT_OFFENSE<-sapply(str_split(dallas_crime_data$SUBJECT_OFFENSE, ",\\s*"), function(x) x[1])
dallas_sub_offence<-dallas_crime_data %>%
  group_by(SUBJECT_OFFENSE) %>%                                          # grouping by the subject offense
  count() %>%                                                            # counting the frequency of subject offense
  arrange(desc(n))                                                       # arranging in the descending order of the f
reqency count

subject_offence<-head(dallas_sub_offence)

# plotting a geom column plot for the incident count of different types of offenses
ggplot(subject_offence,aes(x=reorder(SUBJECT_OFFENSE,n),y=n))+           # defining the aesthetics of the ggplot
  geom_col(fill="#800080")+                                              # using geom_col to visualize the data
  ylab("Frequency")+                                                     # labeling the x axis
  xlab("Subject offence")+                                               # labeling y axis
  ggtitle("Incident count of different types of offenses")+               # labeling the title
  coord_flip()+
  labs(caption=str_wrap("Figure 8:Column plot showing the incident count of different types of offenses, with the X-axis rep
resenting different types of subject offenses and the Y-axis representing the frequency count of each offense.The plot highl
ights the most frequent types of subject offenses based on the available data.",width=90))+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust = 0.5,face="bold"))
```

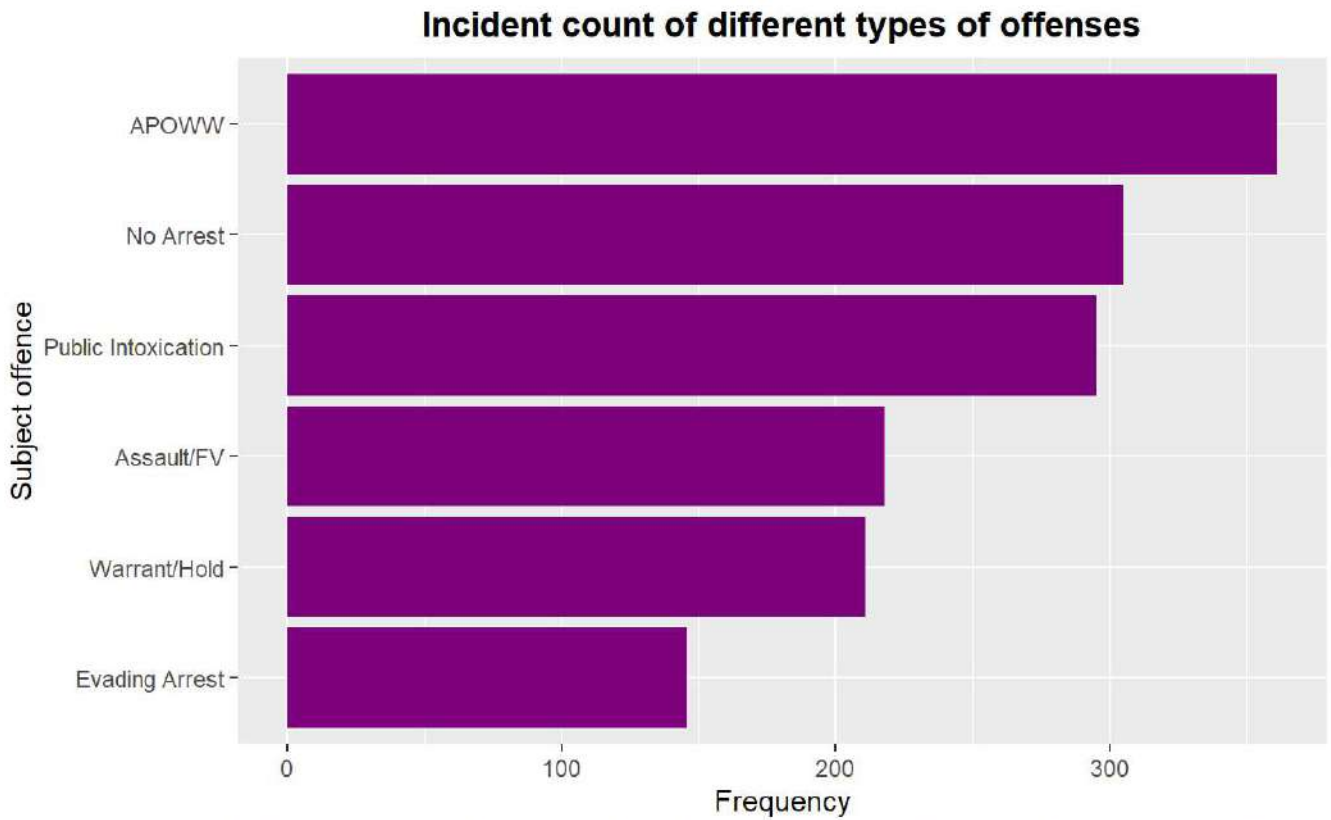


Figure 8:Column plot showing the incident count of different types of offenses, with the X-axis representing different types of subject offenses and the Y-axis representing the frequency count of each offense.The plot highlights the most frequent types of subject offenses based on the available data.

The data represents the count of incidents of different types of offences committed by the subjects in Dallas. It shows the top 6 offenses of the subjects. Most of the crimes were reported as APOWW(Assault public official while intoxicated) followed by the no Arrest, public intoxication, Assault/FV(family violence). This data helps to identify the common root cause of the crimes. By analyzing this data government can implement certain schemes which works provides the counseling sessions, enhancing awareness about the society and further more.

```
# Computing the monthly crime rate
month_record<- dallas_crime_data %>%
  group_by(month(INCIDENT_DATE)) %>%                                    # grouping by the factor of month of incident date
  count()                                                               # counting number of incidents recorded in each month
names(month_record)<-c("months","n")                                    # naming the columns of month_record

#plotting the number of crimes occured in each month over the year
month_plot<-ggplot(month_record,aes(x=months,y=n))+                     # defining the aesthetics of the ggplot months
  geom_line()+                                                          # using geom_line of the trend analysis in the records
of the
  geom_point()+
  scale_x_continuous(breaks = seq(1:12))+
  xlab("Month")+
  ylab("Number of incidents occured")+
  ggtitle("Incident count per month \nover a year")+
  labs(caption=str_wrap("Figure 9:Line plot showing the incident count per month  over the year, with the X-axis representin
g the month of the year and the Y-axis representing the total count of incidents. It displays a trend line of the incident c
ount over the year.",width=100))+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust=0.5,face="bold"))

# evaluating the total crimes in an year
montly_crime_rate<-dallas_crime_data %>%
  count(factor(month(INCIDENT_DATE)),sort = T)

# calculating the average crime rate over an year
average<-sum(montly_crime_rate$n)/12

# calculating number of crimes recorded each day of a week
crime_dataset <- dallas_crime_data %>%
         mutate(day_of_week = factor(format(INCIDENT_DATE, "%a"), levels = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sa
t")),
           day=wday(INCIDENT_DATE,label=TRUE,abbr=TRUE),
           month=month(INCIDENT_DATE,label=TRUE,abbr=TRUE)
         )

# tabulating the crime rate per day of a week
day_record <- crime_dataset %>%
  group_by(day_of_week) %>%
```

```
  count()
day_record<-as.data.frame(day_record)

#plotting the number of crimes occurred in each day of a week over the year
day_plot<-ggplot(day_record,aes(x=day_of_week,y=n,group=1))+        # defining the aesthetics of the ggplot day
  geom_line()+                                                      # using geom_line of the trend analysis in the records o
f the
  geom_point()+
  xlab("Day of the week")+
  ylab(" Number fo incidents occured ")+
  ggtitle("Incident count per week of the day \nover the year")+
  labs(caption=str_wrap("Figure 10:Line plot showing the incident count per week of the day over the year, with the X-axis r
epresenting the day of the week and the Y-axis representing the total count of incidents. It displays a trend line of the in
cident count over a year.",width=100))+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust=0.5,face="bold"))


grid.arrange(month_plot,day_plot)
```

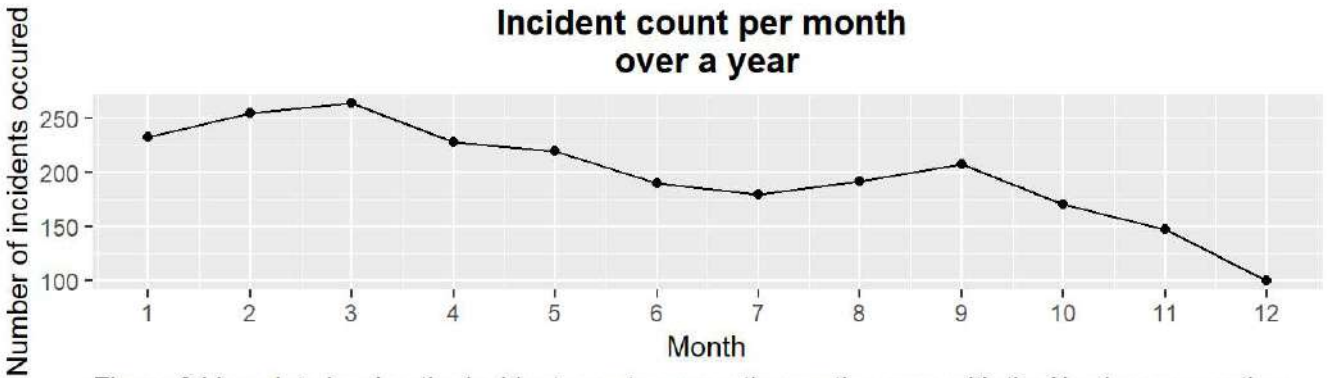


Figure 9:Line plot showing the incident count per month over the year, with the X-axis representing the month of the year and the Y-axis representing the total count of incidents. It displays a trend line of the incident count over the year.

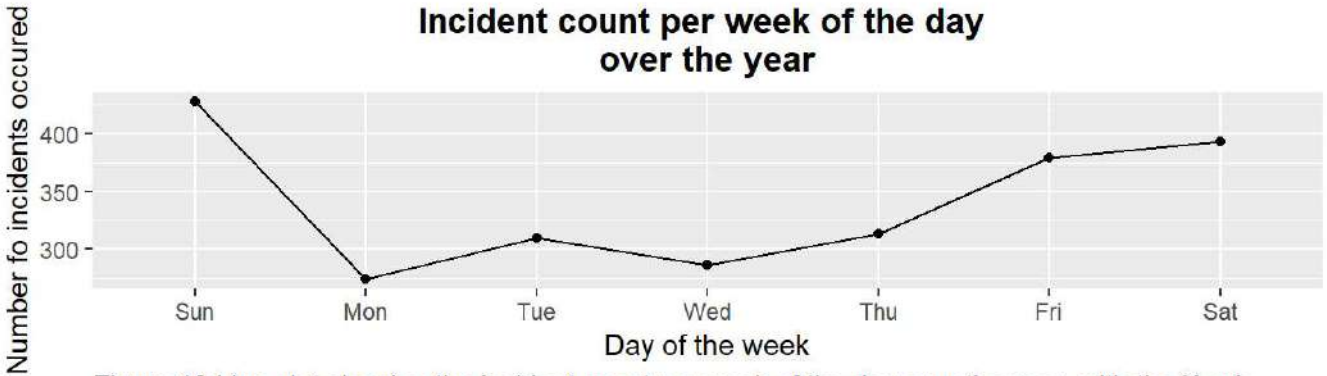


Figure 10:Line plot showing the incident count per week of the day over the year, with the X-axis representing the day of the week and the Y-axis representing the total count of incidents. It displays a trend line of the incident count over a year.

The line graph shows the variation of the recorded crimes over the year per month and per day receptively.It shows that till march the crime rate has increased linearly since march the crime rate started to decrease till July. Since July the crime rate was gradually increasing till September. We can see a minor peak at September and after September the crime rate was plummeted. The line graph of day of the week shows that the crime count is more during weekends and it has sharply plummeted during start of the week and gradually increased through out the week with a small peak on Tuesday. This may be due to the high social gathering during the weekends.This information is useful in identifying the patterns or trends in crime incidents across the different months and week days of the year.

```r
# time interval plot

times_dt <- strptime(dallas_crime_data$INCIDENT_TIME, format="%I:%M:%S %p")     # converting the 12 hrs time format to a 24 hrs time format

# Convert to 24-hour format
times_24h <- format(times_dt, format="%H:%M:%S")
hours <- substr(times_24h, start = 1, stop = 2)                                 # extracting only hour
mino<-substr(times_24h,start=4,stop=5)                                          # extracting only minutes
dallas_data<-dallas_crime_data
dallas_data$hours<-hours                                                        # mutating the hour column
dallas_data$times_24h<-times_24h                                                # mutating the 24 hrs column
dallas_clean<-drop_na(dallas_data)                                              # dropping na's


ggplot(dallas_clean, aes(x =factor(hours))) +                                   # defining the aesthetics of the ggplot of Dallas crime data
  geom_histogram(stat='count',fill="#008080") +                                                 # using histogram to visualize the data
  labs(x = "Time of the day", y = "Total Incidents")+                           # labeling the x axis and y axis
  ggtitle("Crime rate per hour")+                                                # labeling the graph
  labs(caption=str_wrap("Figure 11(a):Line plot showing the incident count per hour of the day over the year, with the X-axis representing the hour of the hour and the Y-axis representing the total count of incidents.",width=90))+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust=0.5,face="bold"))
```

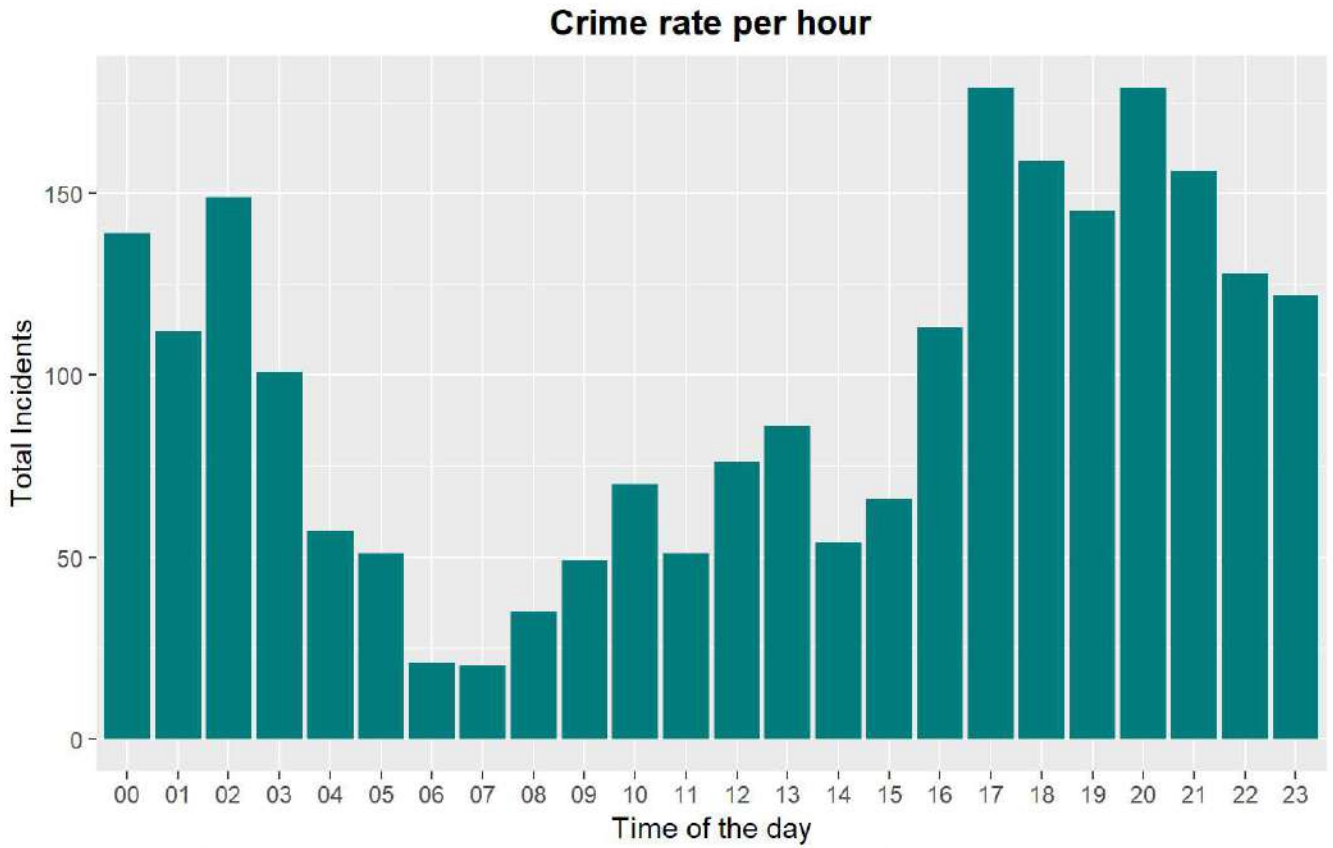


Figure 11(a):Line plot showing the incident count per hour of the day over the year, with the X-axis representing the hour of the hour and the Y-axis representing the total count of incidents.

```r
# adding a  column containing the time intervals
crime_dataset<-dallas_clean %>%
  mutate(time=as.numeric(substr(times_24h, start = 1, stop = 2))*60+
           as.numeric(substr(times_24h,start=4,stop=5)),
         time_group=cut(as.numeric(time),
                        breaks=c(0,6*60,12*60,18*60,23*60+59),
                        labels=c("00-06","06-12","12-18","18-00"),
                        include.lowest = TRUE))


# tabulating the crime count grouped by the time intervals
crime_time<-crime_dataset %>%
  group_by(time_group) %>%                                  # grouping by the time intervals
  count()                                                   # counting number of crimes recorded over the time interval

# plotting the graph of the incident records over the time interval
crime_dataset %>%
  ggplot(aes(x=time_group))+                                # defining the aesthetics of the ggplot of crime dataset
  geom_bar(fill="orange")+                                  # using geom bar to visualize the data
  xlab("Time intervals")+                                   # labeling x axis
  ylab("Number of crime")+                                  # labeling y axis
  ggtitle("Crimes count in time intervals of day")+                        # labeling the graph
  labs(caption=str_wrap("Figure 11(b):Bar plot displaying the frequency count of crimes across different time intervals in a
day, with the X-axis representing the time intervals and the Y-axis representing the total count of crimes.",width=100))+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust=0.5,face = "bold"))
```

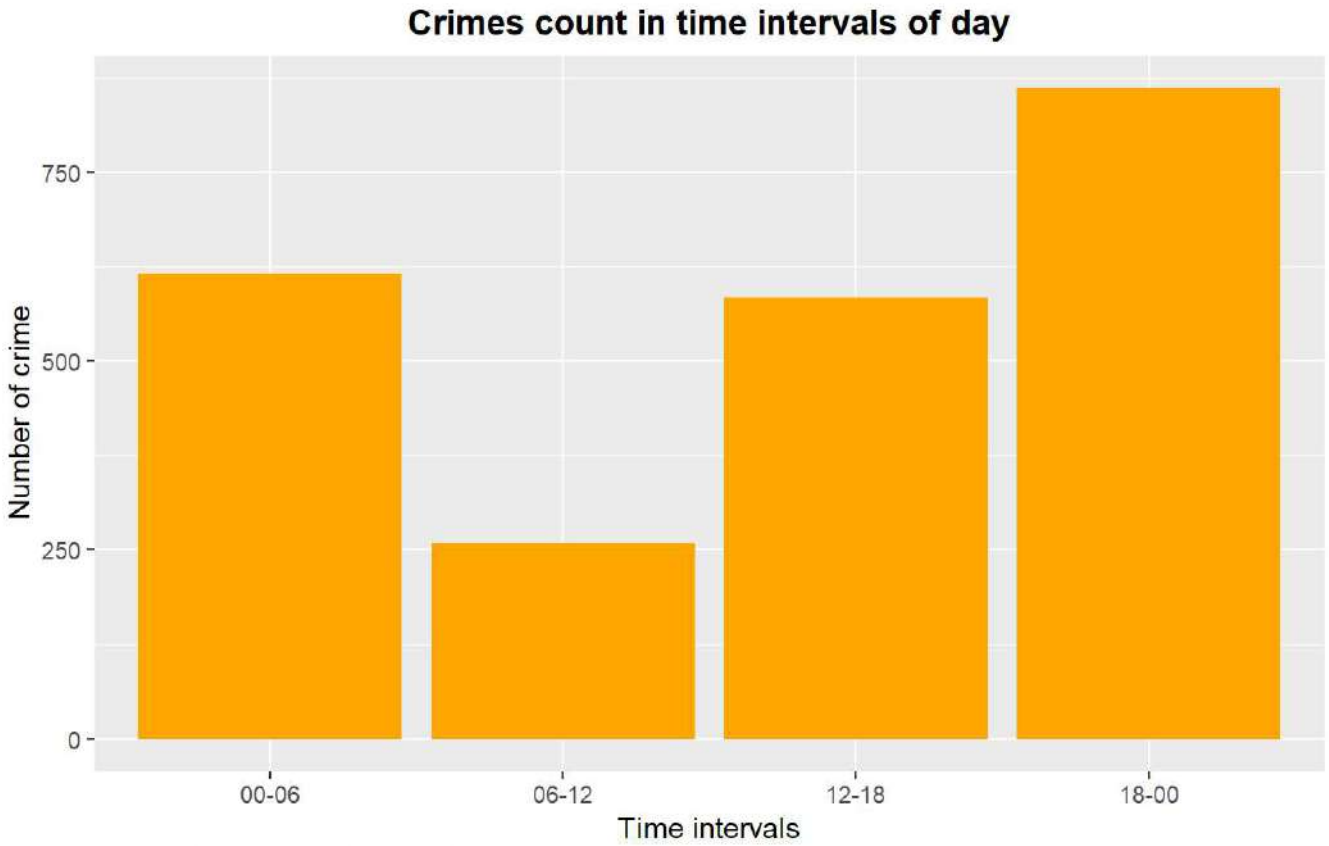


Figure 11(b):Bar plot displaying the frequency count of crimes across different time intervals in a day, with the X-axis representing the time intervals and the Y-axis representing the total count of crimes.

By graphical interpretation of crime rate per hour(Figure 11a) across over all year,it can be observed that the crime rate is very minimum from 6am to 7am in morning while it is very high from 17pm to 20pm evening.It appears that mostly crimes are committed between after the sunset and before the sunrise. During the day time the number of crimes was moderate. It can be observed that analyzing the crime through the hour exactly doesn't give the actual results, because the incident to happen may take sometime to happen so it would be better for analysis to use the time interval rather than the time hour alone. 24 hours of the time is divided into 4 groups (mid-night to 6AM, 6AM to mid afternoon, 12PM to 6PM and 6PM to mid night).[2]From above graph(figure 11(b)) it can be interpreted that most on the crimes took place in between 6pm to 12am, this could be due to reduced visibility, increased opportunity,followed by the time interval from 12 am to 6am, this may be due to the fewer witnesses and less human activity. Due to the higher visibility and activity ,the crime rate has shown a deeper decline in between 6am and 12pm. This information could be utilized by the law enforcement agencies to enhance the security during these hours.

```
# computing the percentage of officers were hospitalized over the year
No_of_off_hos<- dallas_crime_data %>%
  count(OFFICER_HOSPITALIZATION)                                    # counting number of officers hospitalized

names(No_of_off_hos)<-c("OFFICER_HOSPITALIZATION","Count_of_officers")
No_of_off_hos$percent<-No_of_off_hos$Count_of_officers/sum(No_of_off_hos$Count_of_officers) *100        # computing the perc
entage

kable(No_of_off_hos,caption = "Table 5:Table showing the count and percentage of officers by hospitalization status. ")%>%
    kable_styling(position="center",font_size = 12)
```

Table 5:Table showing the count and percentage of officers by hospitalization status.

OFFICER_HOSPITALIZATION	Count_of_officers	percent
No	2335	97.985732
Yes	48	2.014268

The data indicates that the vast majority of the officers (approximately 98%), were not hospitalized,whereas a small percentage of officers(around 2%) undergone treatment, highlighting that they experienced injuries or health issues during the course of their duties.Overall, this implicates that the police department were provided with an effective safety measures, equipment and training.However, the injures of a small percent of officers indicates the danger associated with the law enforcement work. Taking this into consideration government officials may understand the risk in the area and treat to the police. Discussing this would help both police and public.

```
#Barplot of the criminal description categorized by the status of arrest
sub_arr<- dallas_clean %>%
  count(SUBJECT_WAS_ARRESTED)                              # count number of subjects arrested or not

sub<-dallas_clean %>%
  count(SUBJECT_DESCRIPTION,SUBJECT_WAS_ARRESTED,sort=T) %>%
  arrange(desc(n)) %>% filter(n>10)

ggplot(sub,aes(x=reorder(SUBJECT_DESCRIPTION,n),y=n,fill=factor(SUBJECT_WAS_ARRESTED)))+ # defining the aesthetics of the gg
plot
  geom_bar(stat='identity')+                                                            # using geom_bar for the visual inte
rpretation
  scale_fill_discrete(name = "Arrest Status", labels = c("Not Arrested", "Arrested"))+
  coord_flip()+ylab("count")+
  xlab("Subject description")+
  labs(caption=str_wrap("Figure 12:Bar plot showing the count of subjects categorized by description, with the X-axis repres
enting different subject descriptions and the Y-axis representing the count of subjects. The plot is differentiated by the s
ubject's arrest status",width=90))+
  ggtitle("Subject description categorized by the status of arrest")+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust=0.5,face="bold"))
```

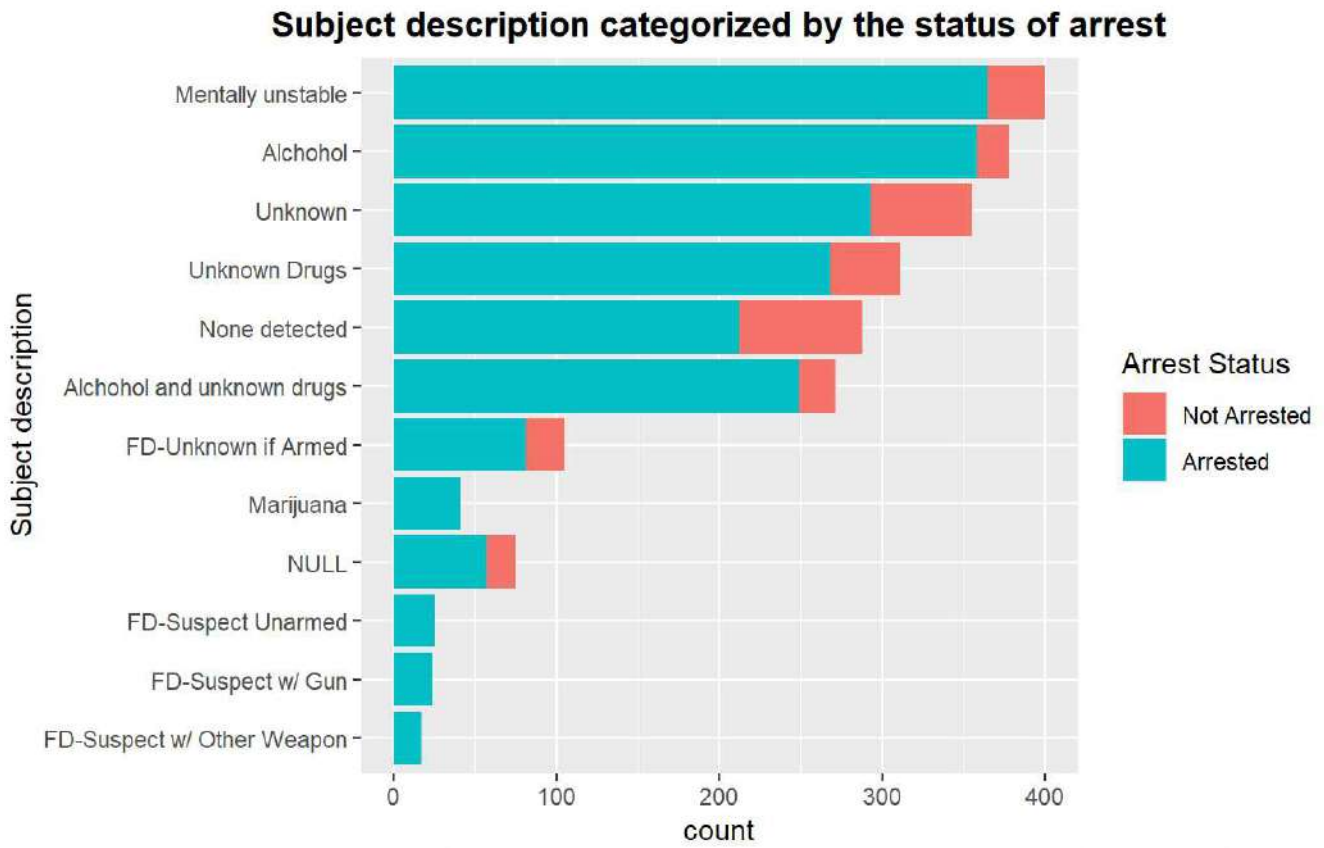


Figure 12:Bar plot showing the count of subjects categorized by description, with X-axis representing different subject descriptions and the Y-axis representing the of subjects. The plot is differentiated by the subject's arrest status

The data reveals that subjects who have committed crime are mostly mentally unstable ,alcoholic and due to unknown drugs and unknown reasons. People who committed crime by being unarmed,suspect with gun or other weapon and marijuana are fully arrested leaving none spared.Whereas it comes to other factors, people who are addicted to drugs and alcohol are arrested with around 3% of people are not arrested. This helps the policymakers to address various contributing factors leading to arrests and develop targeted inventions. This information helps to identify the common factors that contribute to the police incidents and help the police department in reducing the similar incidents by taking a precautionary measure.

```
# Plotting density curve for the analysis of the density of officers with significant number of years on force
ggplot(dallas_clean, aes(x = OFFICER_YEARS_ON_FORCE)) +          # defining the aesthetics of the ggplot of dallas crime dat
a over officer years of experience
  geom_density(color = "darkblue", fill = "lightblue") +         # using geom_density for plotting the density curves
  ggtitle("Officer experience Distribution") +                   # labeling graph
  xlab("Officer experience") +                                   # labeling x axis
  ylab("Density") +                                              # labeling y axis
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5,face="bold"))+
  labs(caption=str_wrap("Figure 13:Density plot showing the distribution of officer experience in the Dallas crime dataset,
with the X-axis representing the years of experience and the Y-axis representing the density of officers at each experience
level. ",width=110))+
  theme(plot.caption = element_text(hjust=0,size=10))
```

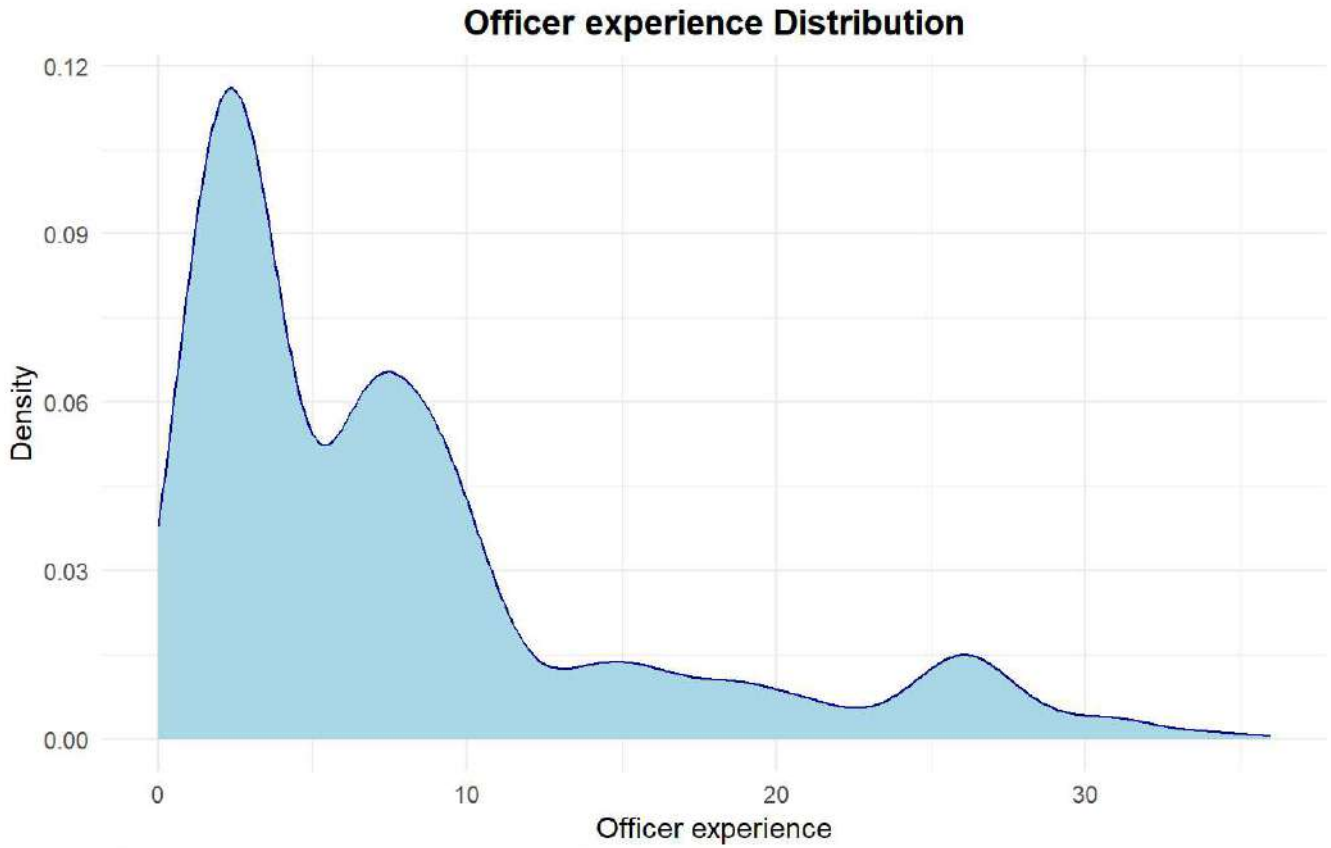


Figure 13:Density plot showing the distribution of officer experience in the Dallas crime dataset, with the X-axis representing the years of experience and the Y-axis representing the density of officers at ea experience level.

The above plot shows the density distribution years of officers experience. It indicates that there are huge number of officers who have been working for less than 5 years in Dallas during 2016. There very few highly experienced people(in force more than 30 years).The density of the officers decreases with the increase in the years of experience. The density of officers ranging experience from 5 to 10 years is half the density of officers having experience ranging from 0 to 5 years. It can be said that there is a sharp decline in the density of officers after 8 years of experience.This pattern could be of many reasons, such as hiring trends, attrition rates, promotions or retirements.

```
# selecting the variables having numeric values for correlation plot
selected_variables<- dallas_clean %>%
  select_if(is.numeric)

#computing the correlation matrix
correlation_matrix<-cor(na.omit(selected_variables))   # omitting the null values
colnames(correlation_matrix)<- c("Officer_ID","Years_Force","Subject_ID","Area","Beat","Sector","Street_number","Latitud
e","Longitude")                                                    # redefining the column names of the correlation matrix
rownames(correlation_matrix)<- c("Officer_ID","Years_Force","Subject_ID","Area","Beat","Sector","Street_number","Latitud
e","Longitude")                                                    # redefining the row names of the correlation matrix
kable( correlation_matrix,caption ="Table 6: Correlation matrix containing values rangng from -1 to 1" )%>%
    kable_styling(position="center",font_size = 12)
```

Table 6: Correlation matrix containing values rangng from -1 to 1

	Officer_ID	Years_Force	Subject_ID	Area	Beat	Sector	Street_number	Latitude	Longitude
Officer_ID	1.0000000	-0.9143921	-0.0105707	-0.0170891	-0.0205871	-0.0204534	-0.0083556	-0.0654106	0.0904271
Years_Force	-0.9143921	1.0000000	0.0161761	0.0116090	0.0102904	0.0102709	0.0321063	0.0814586	-0.0798826
Subject_ID	-0.0105707	0.0161761	1.0000000	-0.0634632	-0.0251385	-0.0249143	-0.0269621	0.0112660	-0.0493642
Area	-0.0170891	0.0116090	-0.0634632	1.0000000	0.4052187	0.4059910	0.0954091	0.0882964	-0.2907411
Beat	-0.0205871	0.0102904	-0.0251385	0.4052187	1.0000000	0.9999538	0.1662100	0.0007688	-0.3445452
Sector	-0.0204534	0.0102709	-0.0249143	0.4059910	0.9999538	1.0000000	0.1657822	0.0020711	-0.3473841
Street_number	-0.0083556	0.0321063	-0.0269621	0.0954091	0.1662100	0.1657822	1.0000000	0.5671245	0.2092809
Latitude	-0.0654106	0.0814586	0.0112660	0.0882964	0.0007688	0.0020711	0.5671245	1.0000000	-0.0094215
Longitude	0.0904271	-0.0798826	-0.0493642	-0.2907411	-0.3445452	-0.3473841	0.2092809	-0.0094215	1.0000000

```
ggcorrplot(correlation_matrix,hc.order = T,type="lower",lab=T)+     # plotting the correlation matrix values
  labs(caption = str_wrap("Figure 14:Lower triangle correlation plot showing the correlation coefficients between selected v
ariables, with the size and color of each cell representing the strength of the correlation",width=70))+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust=0.5,face="bold"))+
  ggtitle("Correlation plot of Dallas dataset")
```

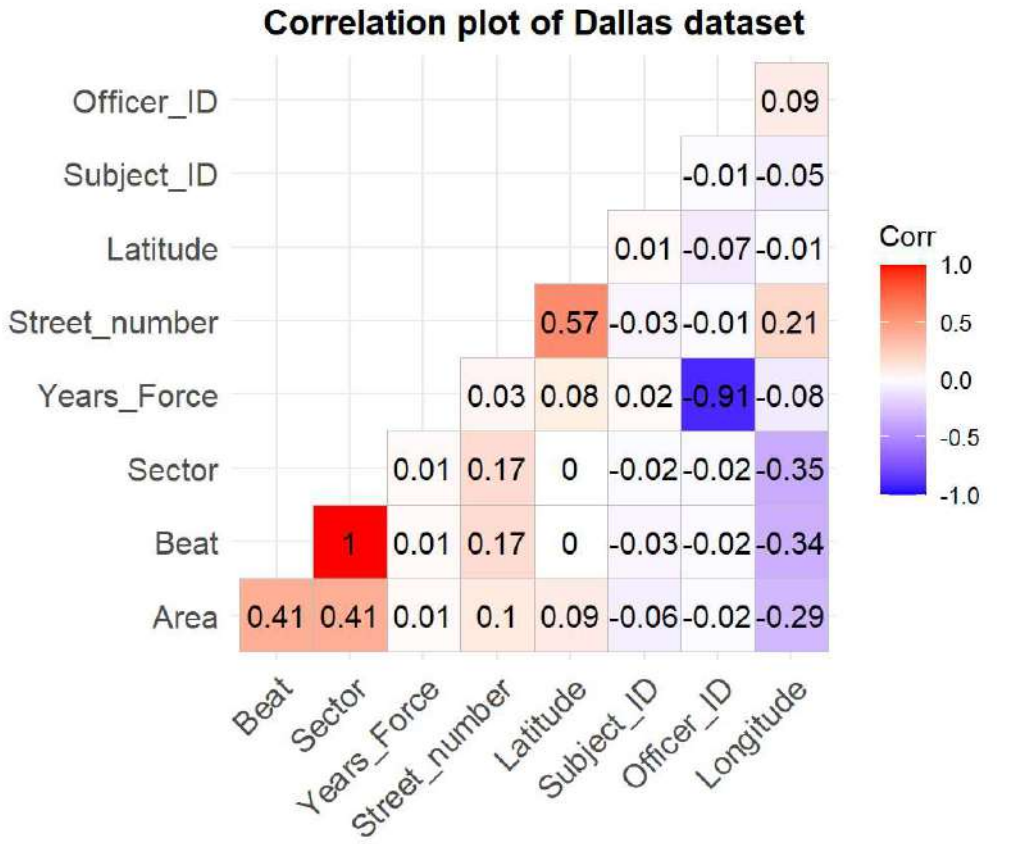


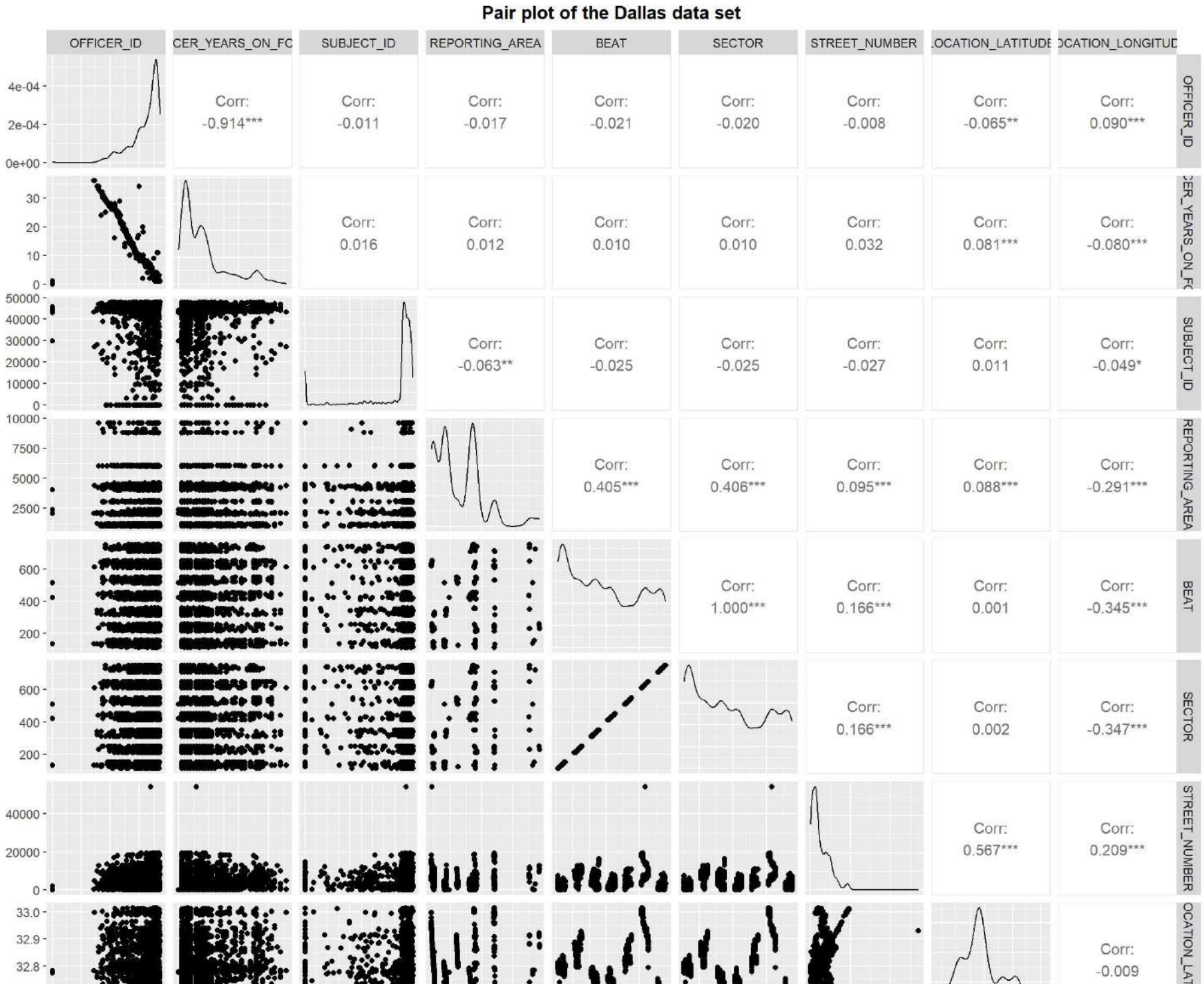
Figure 14:Lower triangle correlation plot showing the correlation coefficients between selected variables, with the size and color of each cell representing the strength of the correlation

The above graph shows the correlation matrix plot. If the correlation of two variables is positive then they are said to be positively correlated , both the variables vary similarly in a same direction, whereas if two variables are negative correlated then they are said to be varying similarly in opposite direction. If correlation value is near to 1 or -1 then they are said to be significant variables, if the correlation co-efficient is '0' then those two variables are not correlated. The correlation matrix shows the relationship between two variables.

Officer id and officer years on force shows a strong negative correlation(-0.914), indicating that as the officers id decreases with the increase in the number of years they have serviced. Beat and sector shows a highly positive correlation(0.99), indicates that the two variables are almost perfectly related, i.e they may trying to convey similar information. The legend of color scale indicates the correlation measure ,color towards red indicates positively correlated and color towards blue indicates negatively correlated,white color indicates no correlation.

Reporting area, beat and sector and street number and location latitude shows a moderate positive correlation indicating that the street number increases with the increases in latitude. The correlation of location longitude with reporting area , beat and sector is moderate. The other correlation coefficients have value near to 0 indicating a week correlation and not significant. Through this analysis one can analyze the pattern in variation between variables. This helps the police department to check and make important decision on stabilizing the law in that areas.

```
#plotting the pair plot containing the correlation values, distribution of the variables and scatter plot
ggpairs(selected_variables,cardinality_threshold = 600)+                        # selecting numeric variables
labs(caption=str_wrap("Figure 15: Scatterplot matrix showing pairwise scatterplots and correlation coefficients for selected
variables, with each variable represented in both the X and Y axes. ",width=100))+
  theme(plot.caption = element_text(hjust=0,size=13),
        plot.title = element_text(hjust=0.5,face="bold"))+
  ggtitle("Pair plot of the Dallas data set")
```


Pair plot of the Dallas data set
OFFICER_ID
CER_YEARS_ON_FC
SUBJECT_ID
REPORTING_AREA
BEAT
SECTOR
STREET_NUMBER
.OCATION_LATITUDE
OCATION_LONGITUD
Corr: -0.914***
Corr: -0.011
Corr: -0.017
Corr: -0.021
Corr: -0.020
Corr: -0.008
Corr: -0.065**
Corr: 0.090***
Corr: 0.016
Corr: 0.012
Corr: 0.010
Corr: 0.010
Corr: 0.032
Corr: 0.081***
Corr: -0.080***
Corr: -0.063**
Corr: -0.025
Corr: -0.025
Corr: -0.027
Corr: 0.011
Corr: -0.049*
Corr: 0.405***
Corr: 0.406***
Corr: 0.095***
Corr: 0.088***
Corr: -0.291***
Corr: 1.000***
Corr: 0.166***
Corr: 0.001
Corr: -0.345***
Corr: 0.166***
Corr: 0.002
Corr: -0.347***
Corr: 0.567***
Corr: 0.209***
Corr: -0.009

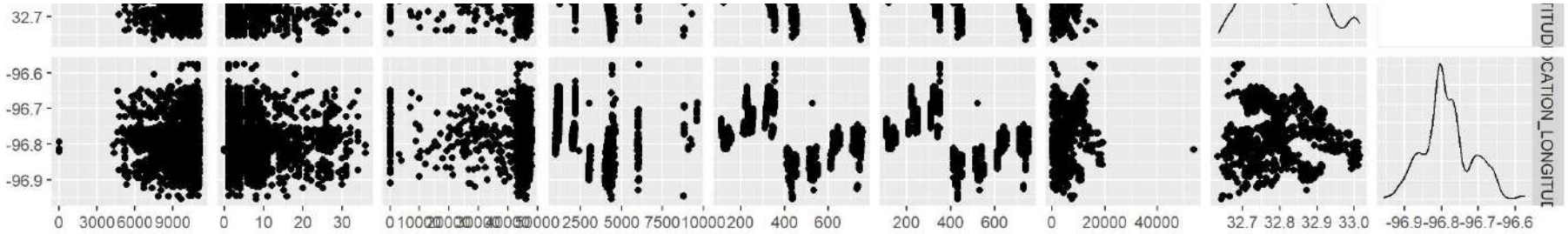


Figure 15: Scatterplot matrix showing pairwise scatterplots and correlation coefficients for selected variables, with each variable represented in both the X and Y axes.

The graph shows the correlation and the graphical representation of the each variable. It shows the scatter plot of multi variables. It provides the visual summary of the relationships between multiple variable of a data set and helps to identify the potential patterns and outliers.It can be said that the beat and sector are highly correlated. The street number provides a similar pattern and outlier across all variables.It also shows the relation between two variables. The plot in the diagonal shows the distribution of each variable. It provides the range of values of each variable.It can be said that the officer id , subject id,location of latitude and location of longitude shows a left skewness while the officer years on force , street number shows high right skewness. The scatter plot of beat nd sector shows a straight line indicating high correlation. while the pattern of subject id , operating areas with all other variables shows a similar pattern.A single outlier can be seen in street number across any variable.

```
ggplot(dallas_crime_data , aes(y=BEAT,x=OFFICER_YEARS_ON_FORCE))+    # defining the aesthetics for the ggplot
  geom_point()+                                                      # using geom_point for the plotting
  geom_smooth(method="lm",color="orange")+                           # linear regression model
  labs(caption = str_wrap("Figure 16: The smoothend graph of relationship between Officer Experience and Beats in Dallas: Sc
atter plot with Regression Line",width=100))+
  xlab("Experience of Officers")+
  ylab("Beat")+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust = 0.5,face="bold"))
```

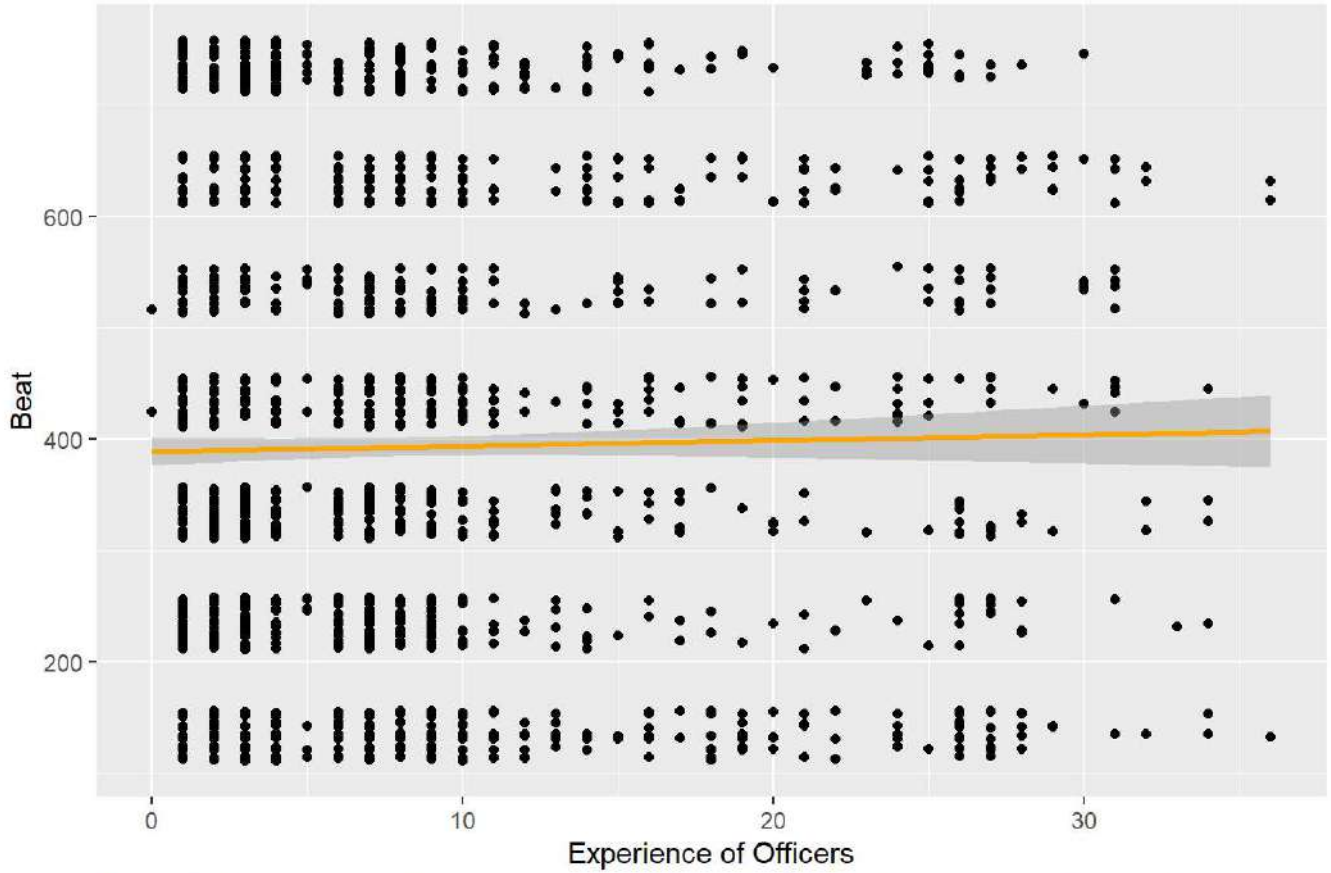
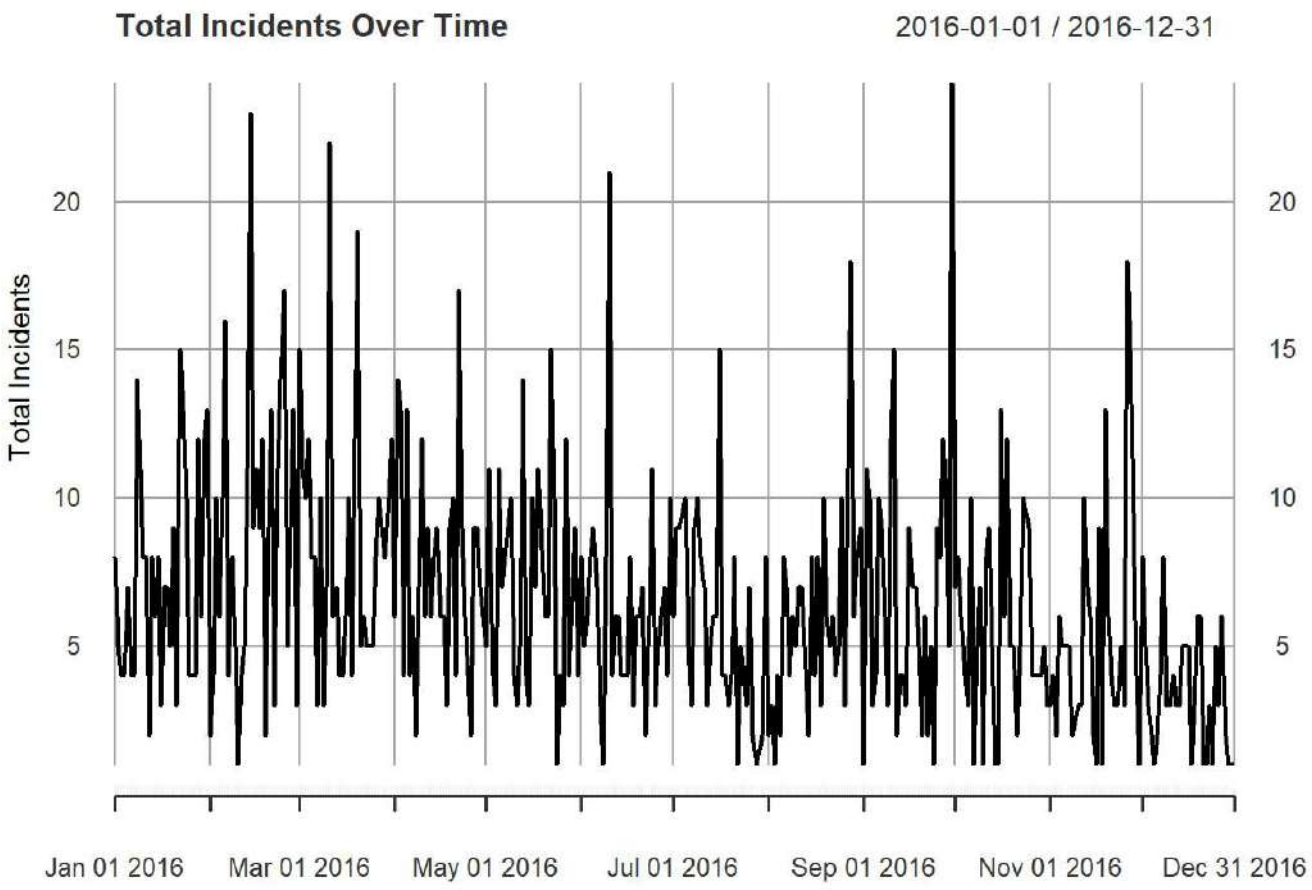


Figure 16: The smoothend graph of relationship between Officer Experience and Beats in Dallas: Scatter plot with Regression Line

The Graph shows the relationship between the number of years an officer has served on the force and the beats they are assigned to in Dallas. The scatter plot shows the distribution of the officers across different beats based on their experience.Officers between 0 to 10 years of experience were assigned mostly to the large number of beats . this may be due to the large population density of the officers who fall under this experience range. There are only few beats allocated to the officers how have experience more than 25 years of experience.The linear regression line shows the overall trend between the two variables, it represents the best-fit straight line that summarizes the relationship between the two variables. The line suggests that there is a positive relation associated between these variables, which implies that the officer with more number of beats were more likely to be assigned to the beats with higher crime rate.

```r
# Group by date and count incidents per day
daily_incidents <- dallas_clean %>%
  group_by(INCIDENT_DATE) %>%        # grouping by the incident date
  count()                             # counting the number of incidents took place on that date

dallas_xts <- xts(daily_incidents$n, order.by = daily_incidents$INCIDENT_DATE)  # using the xts package function for time series graph

# Plot the time series
plot(dallas_xts, main = "Total Incidents Over Time", xlab = "Date", ylab = "Total Incidents") # plotting the time series graph
```



The above graph indicates the time series analyses of the total number of crime incidents over the given period , i.e from 1 Jan,2016 to 31st Dec,2016. It can be interpreted that the crime rate is increased by the end of the year. xts library package is used to visualize the time series pattern over the entire period[5]. It is bit harder to interpret as there are many strokes and bit noisy . so we use the smoothed version of the time series graph to analyse the pattern of crimes over the time period.

```
quaterly<-ma(daily_incidents$n,12)
quater_xts<-xts(data.frame(daily_incidents=daily_incidents$n,quaterly_incidents=quaterly),order.by = daily_incidents$INCIDEN
T_DATE)
autoplot(quater_xts,facet= NULL)+
  geom_line(size=1.1)+
  scale_color_manual(values=c("darkgrey","blue"))+
  theme_bw()+
  xlab("Time intervals")+
  ylab("Number of crimes")+
  ggtitle("Total incidents over time")+
  labs(caption=str_wrap("Figure 17:Line plot showing the total number of crimes over time, with each line representing a dif
ferent quarter. The X-axis represents the time intervals, and the Y-axis represents the total count of crimes. The plot prov
ides a visual representation of the trend in the total number of crimes over time and helps identify any seasonal patterns o
r anomalies in the data",width=100))+
  theme(plot.caption = element_text(hjust=0,size=10),
        plot.title = element_text(hjust = 0.5,face="bold"))
```

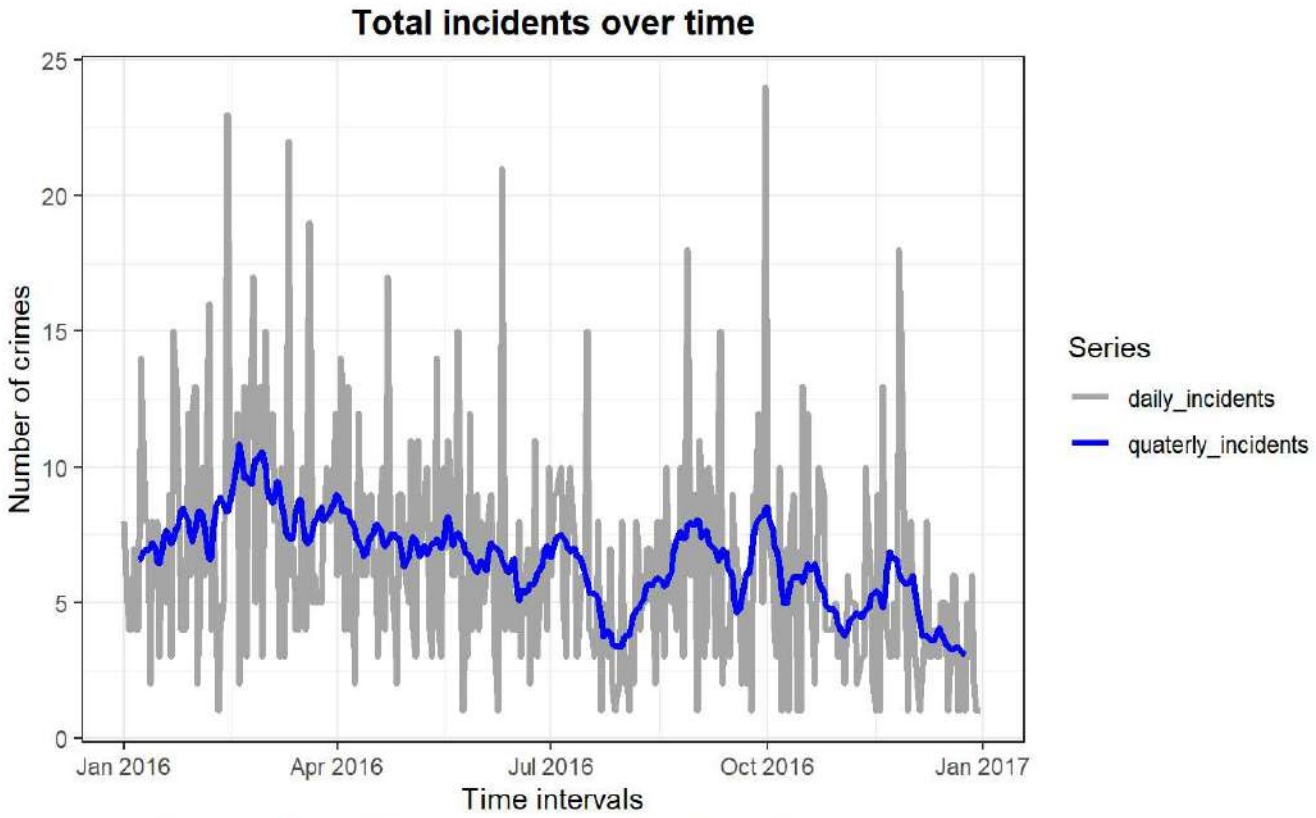


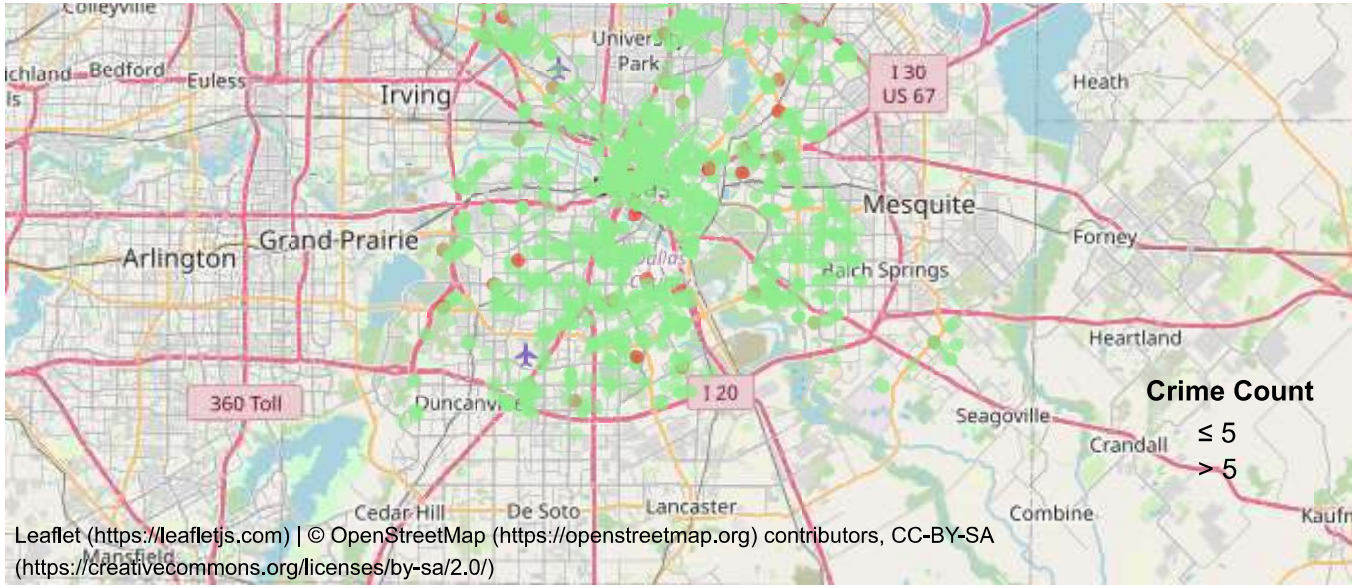
Figure 17:Line plot showing the total number of crimes over time, with each line representing a different quarter. The X-axis represents the time intervals, and the Y-axis represents the total count of crimes. The plot provides a visual representation of the trend in the total number of crimes over time and helps identify any seasonal patterns or anomalies in the data

The above graph indicates the smoothed version of the daily time series graph plotted previously. a smooth curve for a quarterly crime count is drawn. It indicates that the distribution of crimes quarterly rather than daily.It indicates that the crime rate is plummeted during the mid of third quarter while crime rate is high during the start of the year. The crime rate was decreased over the first three quarters, however the crime rate is again increased over the last quarter with minor fluctuations.

Interactive Maps

Interactive map of Dallas crimes

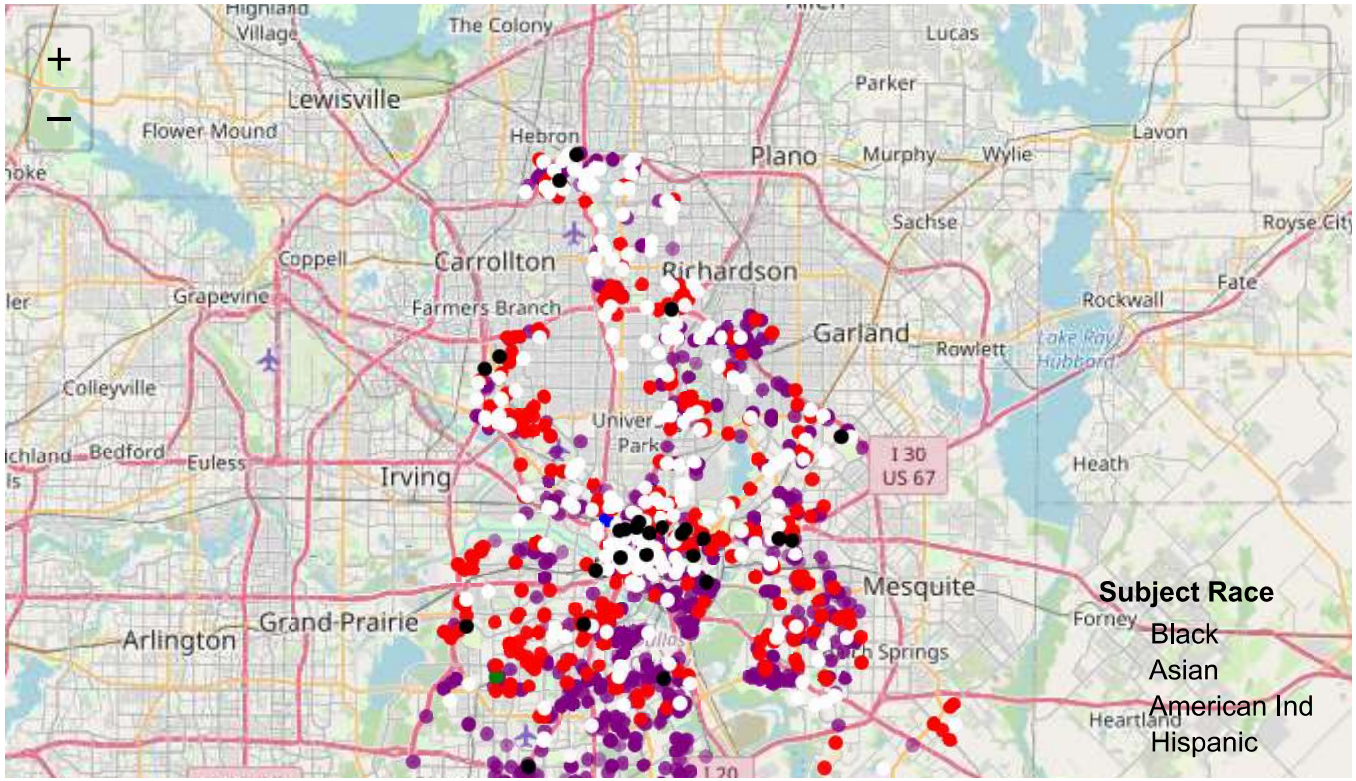
```
# crime count in each area grouped by latitude and longitude
crimesrate <- dallas_clean %>%
  group_by(LOCATION_LONGITUDE,LOCATION_LATITUDE) %>%    # grouping by the latitude and longitude
  count() %>%             # counting crimes
  arrange(desc(n)) %>%    # arranging in the descending order of count
  drop_na()              # dropping null values

names(crimesrate) <- c("longitude","latitude","n")   # renaming columns

maplf <- crimesrate %>%
  leaflet() %>%                    # leaflet package function is used
  addTiles() %>%                   # adding the tiles to map
  addCircleMarkers(lng=dallas_clean$LOCATION_LONGITUDE,lat=dallas_clean$LOCATION_LATITUDE,radius=1,    # adding circle marke
rs as symbol of crimes
                   popup = paste("division:", dallas_clean$DIVISION,
                                 "<br>",
                                 "street Name:",dallas_clean$LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION,
                                 "<br>",
                                 "Number of crimes recorded: ",crimesrate$n),
                   color= ifelse(crimesrate$n >5 ,"red","lightgreen")) %>%
  leaflet::addLegend(
    position = "bottomright",
    title = "Crime Count",
    colors = c("lightgreen", "red"),
    labels = c("≤ 5", "> 5"),
    opacity = 1
  )
maplf
```

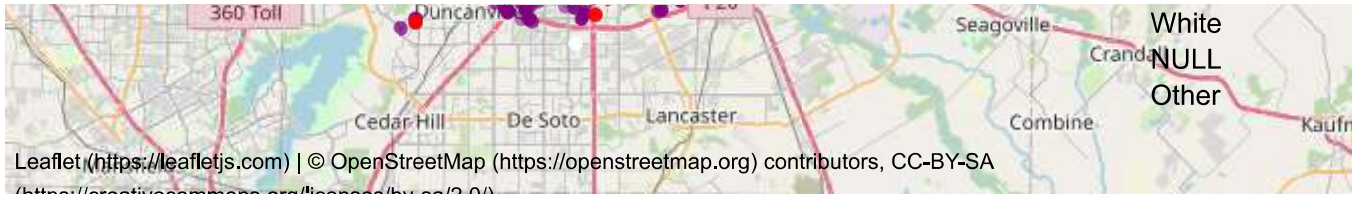


The above map shows the view of geographic view[4] of Dallas city and its crimes in 2016. Red dots shows the ares having crimes more than 5 in that year, whereas light green dots shows the areas having crimes less than 5. Clicking on a circle marker reveals additional information about the crime, including the division and street name associated with the location. Using this maps , one can easily identify and circle the areas having more crimes such that a proper action and strategies can be drawn by the law enforcement agencies for the better safety and securities in areas. through map it can be said that mostly the crime rate is high in D14 district with 313 crimes over the year followed by D2(310) and D7(231).Through this one can easily identify the crime hot spots (areas where crime is high) and can develop needed strategies for the safety of the people over that area. This helps to understand the relationship between the crime and demographic factors, such as race, gender, age and income. This information can be used to develop targeted interventions aimed at reducing crime in specific communities.

```
# create a leaflet map object
map <- leaflet(data=dallas_clean) %>%
  addTiles()

# add circles for each race
m1<-map %>%
  addCircles(data = dallas_clean[dallas_clean$SUBJECT_RACE == "Black",], group = "Black",
             col = "purple", lng = ~LOCATION_LONGITUDE, lat = ~LOCATION_LATITUDE,
                                popup = paste("division:", dallas_clean$DIVISION,
                                 "<br>",
                                 "street Name:",dallas_clean$LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION)) %>%
  addCircles(data = dallas_clean[dallas_clean$SUBJECT_RACE == "Asian",], group = "Asian",
             col = "blue", lng = ~LOCATION_LONGITUDE, lat = ~LOCATION_LATITUDE,
                                popup = paste("division:", dallas_clean$DIVISION,
                                 "<br>",
                                 "street Name:",dallas_clean$LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION)) %>%
  addCircles(data = dallas_clean[dallas_clean$SUBJECT_RACE == "American Ind",], group = "American Ind",
             col = "green", lng = ~LOCATION_LONGITUDE, lat = ~LOCATION_LATITUDE,
                                popup = paste("division:", dallas_clean$DIVISION,
                                 "<br>",
                                 "street Name:",dallas_clean$LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION)) %>%
  addCircles(data = dallas_clean[dallas_clean$SUBJECT_RACE == "Hispanic",], group = "Hispanic",
             col = "red", lng = ~LOCATION_LONGITUDE, lat = ~LOCATION_LATITUDE,
                                popup = paste("division:", dallas_clean$DIVISION,
                                 "<br>",
                                 "street Name:",dallas_clean$LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION)) %>%
  addCircles(data = dallas_clean[dallas_clean$SUBJECT_RACE == "White",], group = "White",
             col = "white", lng = ~LOCATION_LONGITUDE, lat = ~LOCATION_LATITUDE,
                                popup = paste("division:", dallas_clean$DIVISION,
                                 "<br>",
                                 "street Name:",dallas_clean$LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION)) %>%
  addCircles(data = dallas_clean[dallas_clean$SUBJECT_RACE == "NULL",], group = "NULL",
             col = "black", lng = ~LOCATION_LONGITUDE, lat = ~LOCATION_LATITUDE,
                                popup = paste("division:", dallas_clean$DIVISION,
                                 "<br>",
                                 "street Name:",dallas_clean$LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION)) %>%
  addCircles(data = dallas_clean[dallas_clean$SUBJECT_RACE == "Other,",], group = "Other",
             col = "orange", lng = ~LOCATION_LONGITUDE, lat = ~LOCATION_LATITUDE,
                                popup = paste("division:", dallas_clean$DIVISION,
```

```
                                "<br>",
                                "street Name:",dallas_clean$LOCATION_FULL_STREET_ADDRESS_OR_INTERSECTION))



m1 %>%
addLayersControl(
baseGroups = c("OSM (default)", "Toner Lite"),
overlayGroups = c("American Ind", "Asian", "Black", "Hispanic", "NULL", "Other", "White"),
options = layersControlOptions(collapsed = TRUE)
) %>%
leaflet::addLegend(
position = "bottomright",
title = "Subject Race",
colors = c("purple", "blue", "green", "red", "white", "black", "orange"),
labels = c("Black", "Asian", "American Ind", "Hispanic", "White", "NULL", "Other"),
opacity = 1
)
```





The above map shows the interactive plot of the crime data of Dallas ,it is layered by the subject race, to find number of subjects,who have committed crime, belong to a particular race or present in that particular area .Race plays a prominent role in united states of america in each and every state. While we filter the map by the races we can identify the patterns and the root cause of the crime.Thus by observing two maps one can be able to construct a strategy what need to be done to get the crime rate under control by taking race into consideration. It shows the crimes conducted by each race across different regions. This helps in identifying the subjects demographically so that based on the count of the subjects government could take an action to prevent the crimes in that area.

Discussion:

The crime rate of the Dallas over the year 2016,has been fluctuating maximum at the first quarter of the year. The number of crimes committed by each subject has also been decreased.It is recorded that most of the crimes where .There are some missing values which indicates that the data is not recorded properly(which includes missing values in latitude, longitude,subject id,some other variables).The crime rate is more in central in D6 district. There are different types of forces used by the police to control subjects from committing crimes.The most common type of reason for the arrest is the verbal command and holding an wepon. This says that the weapons are being used by the public which is not safe for the society.

Conclusion:

Based on the Dallas crime data in 2016, we can conclude that the crime rates were higher in central division. Crimes where more likely to occurred during the late afternoon ans the evening hours, mostly occurred after the sunset and midnight.There are more black male subjects who are arrested and involved in crimes. There appears a greater gender and race disparities in the Dallas during 2016. The most common reason for the offense is the APOWW and intoxication. The interactive maps provide a deeper insight on the demo graphical distribution of the crimes. This information helps the policy makers to identify the crime patterns and analyze the root cause for the crimes. This could aid the law enforcement agencies to allocate their resources more effectively and efficiently. It also helps the policy makers and community leaders to understand the crime trends and take necessary actions to address the root causes of crime in the city. Thus, the visualization of Dallas crime in 2016 provides a valuable insights for crime prevention and public safety efforts in the city. But it can be observed that the subject injury type is not noticed mostly,and mostly subjects were arrested. There are more number of black race male people as the subjects while the government jobs were mostly occupied by the white race people.Thus we can see a significant racism.This findings using the data visualization aids in analyzing the data and can understand the reason for the crime, which in turn helps in reducing the crimes across the zone by introducing essential strategies. This enhances the trust of people on government and helps to maintain balance between law and order.

Utilizing this data government can get a deeper insights on the race disparities and tries to reduce this conflicts by checking the root cause of racism and could take a instant charge if required. It helps in introducing new schemes for the minority categories, it also increases the transparency of the crime rate and people could be safe. Thus, it shows that data visualization plays a prominent role in every department in making effective decisions. One can easily analyze the data through visualizations than a tabulated excel sheet with large number of observations. This tells a story where we can predict the future outcomes based on the past records. It tells us the correlations of the variables.

References

1.Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. Hoboken, N.J.: Wiley

2.Meloncon, L., & Warner, E. (2017). Data visualizations: A literature review and opportunities for technical and professional communication. Paper presented at the Professional Communication Conference (ProComm), 2017 IEEE International, Madison, WI, USA.

3.Chen, C. H., Härdle, W. K., & Unwin, A. (Eds.). (2007). Handbook of data visualization. Springer Science & Business Media.

4.Lu, W., Ai, T., Zhang, X., & He, Y. (2017). An interactive web mapping visualization of urban air quality monitoring data of China. Atmosphere, 8(8), 148.

5.Krispin, R. (2019). Hands-On Time Series Analysis with R: Perform time series analysis and forecasting using R. Packt Publishing Ltd.