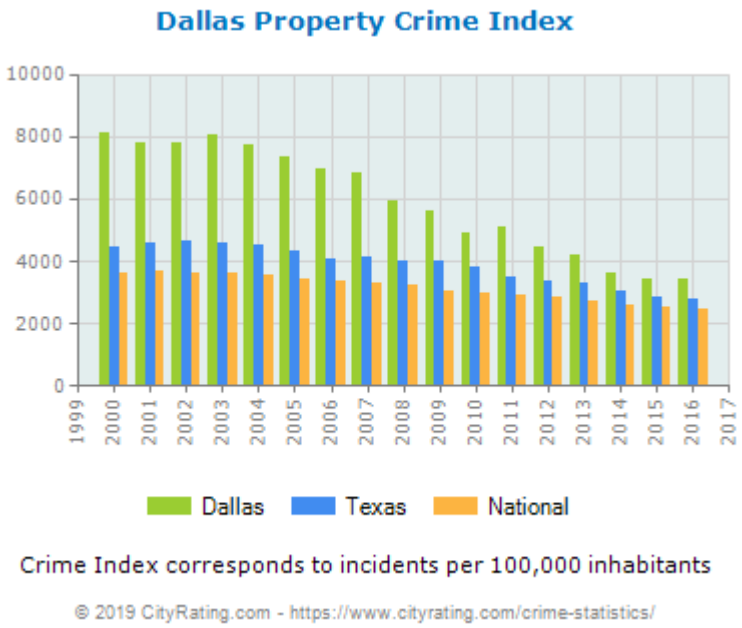
DALLAS CRIME DATA ANALYSIS

INTRODUCTION

Crime has been prevalent over the years and has become a huge concern to us. It should be taken on a serious note and analyzed properly to reduce the crime rate. Hundreds of crimes are reported daily across the country despite having strict rules and regulations. Crime analysis involves identification of patterns and trends in crime and disorder. It plays a major role in finding solutions to crime incidents and formulating crime prevention strategies. This information helps the law enforcement agencies to utitlize the resources in an effective way.

Although the crime statistics in Dallas report an overall downward trend in crime based on data from 18 years, crimes continue to happen every day.



We aim at further reducing the crime rate by analyzing the crime incident reports that had happened over the years. This analysis can help alert security officials to be prepared on how to deal with them in the future. For this purpose, we have data that the Dallas Crime department - Records Management System has recorded from June 2014 till current date.

PROBLEM DEFINITION

The objective of this project is to analyze the trend in crimes from the year June 2014 to current date (2019-12-10).

- Find the time of day and day of week when most crimes tend to occur
- Analyze the trend of different categories of crime
- Find the most used weapons in crime scenes
- Identify areas in the city where there are less crimes
- Relationship between crime types and victim's characteristics
- Year on year increase in overall crimes and also specific types of crimes
- Predict the crime rate (for each category of crime) for the next year

EXTRACTION, EXPLORATION AND PREPROCESSING OF DATA

For this project, we use the crime data reported in the city of Dallas, beginning June 2014 to current date (2019-12-10). The data is collected from www.dallasopendata.com. This information reflects the crimes reported to the

Dallas Police Department. There are about 100 columns and over 585K records.

We use the read.delim() function to read the crime data and store it in a dataframe. We can either read the data directly from the URL or download and read it from the file. Here, we are using the second approach to read the data in R.

```
library(dplyr)
library(tidyr)
library(lubridate)
```

```
crimedata <- read.delim(input_file, header = TRUE,sep = "\t", check.names = FALSE,
                        stringsAsFactors = FALSE)
```

Now that we have read the data, let us understand the data, its fields and its types. The summary() function provides the detailed summary of data. str() function displays the structure of the data in a compact way.

```
str(crimedata)
```

```
## 'data.frame': 628964 obs. of 100 variables:
## $ Incident Number w/year : chr "196700-2018" "248560-2018" "080509-2018" "081335-2018" ...
## $ Year of Incident : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
## $ Service Number ID : chr "196700-2018-01" "248560-2018-01" "080509-2018-01" "081335-2018-01"
...
## $ Watch : int 1 3 2 2 3 2 3 1 3 2 ...
## $ Call (911) Problem : chr "58 - ROUTINE INVESTIGATION" "41/20 - ROBBERY - IN PROGRESS" "58 -
ROUTINE INVESTIGATION" "58 - ROUTINE INVESTIGATION" ...
## $ Type of Incident : chr "LOST PROPERTY (NO OFFENSE)" "FOUND PROPERTY (NO OFFENSE)" "INJURED
PERSON- PUBLIC PROPERTY (OTHER THAN FIREARM) (NO OFFENSE)" "LOST PROPERTY (NO OFFENSE)" ...
## $ Type Location : chr "Single Family Residence - Occupied" "Highway, Street, Alley ETC" "Airport
- All Others" "Hotel/Motel/ETC" ...
## $ Type of Property : chr "N/A" "N/A" "N/A" "N/A" ...
## $ Incident Address : chr "133 LAGUNA DR" "933 SIX FLAS AVE" "10701 LAMBERT INTERNATIONAL BLDV."
"200 MAIN ST" ...
## $ Apartment Number : chr "" "" "" "" ...
## $ Reporting Area : int 4404 4074 NA 2123 NA 2123 4290 8811 2005 NA ...
## $ Beat : int 623 424 122 111 621 111 454 436 146 229 ...
## $ Division : chr "NORTH CENTRAL" "SOUTHWEST" "CENTRAL" "CENTRAL" ...
## $ Sector : int 620 420 120 110 620 110 450 430 140 220 ...
## $ Council District : chr "" "" "" "9" ...
## $ Target Area Action Grids : chr "" "" "" "" ...
## $ Community : chr "" "" "" "" ...
## $ Date1 of Occurrence : chr "09/03/2018" "11/16/2018" "04/17/2018" "04/15/2018" ...
## $ Year1 of Occurrence : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
## $ Month1 of Occurence : chr "September" "November" "April" "April" ...
## $ Day1 of the Week : chr "Mon" "Fri" "Tue" "Sun" ...
## $ Time1 of Occurrence : chr "14:00" "20:18" "09:00" "12:00" ...
## $ Day1 of the Year : int 246 320 107 105 61 256 110 193 152 89 ...
## $ Date2 of Occurrence : chr "09/04/2018" "11/16/2018" "04/17/2018" "04/15/2018" ...
## $ Year2 of Occurrence : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
## $ Month2 of Occurence : chr "September" "November" "April" "April" ...
## $ Day2 of the Week : chr "Tue" "Fri" "Tue" "Sun" ...
## $ Time2 of Occurrence : chr "13:00" "20:18" "09:01" "12:00" ...
## $ Day2 of the Year : int 247 320 107 105 66 282 110 193 181 89 ...
## $ Date of Report : chr "09/04/2018 06:19:00 PM" "11/16/2018 08:19:00 PM" "04/17/2018 10:15:00
AM" "04/17/2018 10:24:00 AM" ...
## $ Date incident created : chr "09/04/2018 06:22:55 PM" "11/16/2018 11:44:33 PM" "04/17/2018
```

```
10:25:02 AM" "04/18/2018 10:25:12 AM" ...
## $ Offense Entered Year : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
## $ Offense Entered Month : chr "September" "November" "April" "April" ...
## $ Offense Entered Day of the Week : chr "Tue" "Fri" "Tue" "Wed" ...
## $ Offense Entered Time : chr "18:22" "23:44" "10:25" "10:25" ...
## $ Offense Entered Date/Time : int 247 320 107 108 66 283 132 193 198 263 ...
## $ CFS Number : chr "18-1617276" "18-2086439" "18-0668631" "18-0676330" ...
## $ Call Received Date Time : chr "09/04/2018 06:19:05 PM" "11/16/2018 08:18:25 PM" "04/17/2018
06:00:49 AM" "04/18/2018 10:24:19 AM" ...
## $ Call Date Time : chr "09/04/2018 06:19:05 PM" "11/16/2018 08:18:25 PM" "04/17/2018 06:00:49
AM" "04/18/2018 10:24:19 AM" ...
## $ Call Cleared Date Time : chr "09/04/2018 07:37:30 PM" "11/16/2018 11:59:18 PM" "04/17/2018
01:40:53 PM" "04/18/2018 10:24:34 AM" ...
## $ Call Dispatch Date Time : chr "09/04/2018 06:19:05 PM" "11/16/2018 08:22:35 PM" "04/17/2018
06:00:51 AM" "04/18/2018 10:24:19 AM" ...
## $ Special Report (Pre-RMS) : chr "" "" "" "" ...
## $ Person Involvement Type : chr "Victim" "Victim" "Victim" "Victim" ...
## $ Victim Type : chr "Individual" "Society/Public" "Individual" "Individual" ...
## $ Victim Name : chr "LIN, HUAN" "CITY OF DALLAS" "ODEN, RUSSELL, DEAN" "NEIGHBOR, BENJAMIN,
JOSEPH" ...
## $ Victim Race : chr "Asian" "" "White" "White" ...
## $ Victim Ethnicity : chr "Non-Hispanic or Latino" "" "Non-Hispanic or Latino" "Non-Hispanic or
Latino" ...
## $ Victim Gender : chr "Female" "" "Male" "Male" ...
## $ Victim Age : int 25 NA 82 31 NA 56 34 NA 58 NA ...
## $ Victim Age at Offense : int 25 NA 82 31 NA 56 34 NA 58 NA ...
## $ Victim Home Address : chr "133 LAGUNA DR" "725 N JIM MILLER RD" "5SCHOOLHOUSE CT" "2660 N
HASKELL AVE" ...
## $ Victim Apartment : chr "" "" "" "1160" ...
## $ Victim Zip Code : chr "75252" "75217" "63368" "75204" ...
## $ Victim City : chr "DALLAS" "DALLAS" "O FALLON" "DALLAS" ...
## $ Victim State : chr "TX" "TX" "MO" "TX" ...
## $ Victim Business Name : chr "" "" "" "" ...
## $ Victim Business Address : chr "" "" "" "" ...
## $ Victim Business Phone : chr "" "" "" "" ...
## $ Responding Officer #1 Badge No : chr "9949" "10987" "5796" "7881" ...
## $ Responding Officer #1 Name : chr "MACIAS,OSCAR,IVAN" "KIM,DANIEL,K" "HERNANDEZ JR,BENITO,F"
"CONWAY,MICHAEL,SHANE" ...
## $ Responding Officer #2 Badge No : chr "" "10983" "" "" ...
## $ Responding Officer #2 Name : chr "" "RAMIREZ,DANIEL" "" "" ...
## $ Reporting Officer Badge No : chr "9949" "10987" "5796" "7881" ...
## $ Assisting Officer Badge No : chr "" "" "" "" ...
## $ Reviewing Officer Badge No : chr "77397" "118918" "105273" "106845" ...
## $ Element Number Assigned : chr "C691" "C325" "S324" "U155" ...
## $ Investigating Unit 1 : chr "" "" "" "" ...
## $ Investigating Unit 2 : chr "" "" "" "" ...
## $ Offense Status : chr "Suspended" "Suspended" "Suspended" "Suspended" ...
## $ UCR Disposition : chr "Suspended" "Suspended" "Suspended" "Suspended" ...
## $ Victim Injury Description : chr "" "" "" "" ...
## $ Victim Condition : chr "" "" "" "" ...
## $ Modus Operandi (MO) : chr "LOST PASSPORT" "ROS FOUND CRACK COCAINE IN A TRUCK" "WHILE ON PLANE
IN MO. VICTIM HIT FRONT OF FOREHEAD THEN BACK OF H" "LOST PROPERTY" ...
## $ Family Offense : chr "false" "false" "false" "false" ...
## $ Hate Crime : chr "" "" "" "" ...
## $ Hate Crime Description : chr "None" "None" "None" "None" ...
## $ Weapon Used : chr "" "" "" "" ...
```

```
## $ Gang Related Offense : chr "" "" "" "" ...
## $ Victim Package : logi NA NA NA NA NA NA ...
## $ Drug Related Istevencident : chr "No" "Yes" "No" "No" ...
## $ RMS Code : chr "NA-99999999-X1" "NA-99999999-X3" "NA-99999999-W1" "NA-99999999-X1" ...
## $ Criminal Justice Information Service Code: int 99999999 99999999 99999999 99999999 99999999
99999999 99999999 99999999 99999999 99999999 ...
## $ Penal Code : chr "No Offense" "No Offense" "UCR" "No Offense" ...
## $ UCR Offense Name : chr "" "" "INJURED PUBLIC" "LOST" ...
## $ UCR Offense Description : chr "" "" "ACCIDENTAL INJURY" "LOST PROPERTY" ...
## $ UCR Code : int NA NA 3300 4200 4300 NA 4200 NA NA NA ...
## $ Offense Type : chr "" "" "NOT CODED" "NOT CODED" ...
## $ NIBRS Crime : chr "MISCELLANEOUS" "MISCELLANEOUS" "MISCELLANEOUS" "MISCELLANEOUS" ...
## $ NIBRS Crime Category : chr "MISCELLANEOUS" "MISCELLANEOUS" "MISCELLANEOUS" "MISCELLANEOUS" ...
## $ NIBRS Crime Against : chr "MISCELLANEOUS" "MISCELLANEOUS" "MISCELLANEOUS" "MISCELLANEOUS" ...
## $ NIBRS Code : chr "999" "999" "999" "999" ...
## $ NIBRS Group : chr "D" "D" "D" "D" ...
## $ NIBRS Type : chr "Not Coded" "Not Coded" "Not Coded" "Not Coded" ...
## $ Update Date : chr "2018-09-06 09:27:31.0000000" "2018-11-17 23:20:26.0000000" "2018-06-11
10:03:26.0000000" "2018-06-11 10:03:27.0000000" ...
## $ X Coordinate : num NA NA NA 2541289 NA ...
## $ Y Cordinate : num NA NA NA 7020042 NA ...
## $ Zip Code : int 75252 75208 63145 75040 75252 75216 75224 75249 75204 75043 ...
## $ City : chr "DALLAS" "ARLINGTON" "STLOUIS" "GARLAND" ...
## $ State : chr "TX" "TX" "MO" "TX" ...
## [list output truncated]
```

Looking at the summary of data, we understand that there is one observation for each crime incident in the data frame. We have 628964 (rows) of 100 variables (columns) where each row is a crime incident reported to the Dallas Police Department. For the ease of data analysis, we select only the fields that are necessary. The fields are selected based on the problem definition.

```
crimedata <-crimedata %>% select(`Incident Number w/year`,`Year of Incident`,`Type of Incident`,
                  Beat,Division, `Type of Property`,`Date1 of Occurrence`,`Month1 of Occurence`,
                  `Time1 of Occurrence`,`Day1 of the Week`,`Day1 of the Year`,`Victim Age`,
                  `Victim Gender`, `Offense Status`,`UCR Offense Name`, `Victim Condition`,
                  `Weapon Used`)
```

Now with the selected fields, we can start cleaning the data and make it ready for analysis.

- Incident Number should be unique

Each incident has a unique identifier associated with it which is stored in the variable `Incident number w/ year`. However we have some instances where two or more rows have the same identifier. These duplicated instances should be removed. We use the duplicated() function to remove the duplicates.

```
crimedata <- crimedata[!duplicated(crimedata$`Incident Number w/year`), ]
```

- Date of Occurrence should be between June 2014 and 2019-12-10

Date of Occurrence is a field that stores the date when the incident was reported. Hence it should be of type *Date*. The values should be from June 2014 to 2019-12-10

```
crimedata$`Date1 of Occurrence`<-as.Date(crimedata$`Date1 of Occurrence`,format="%m/%d/%Y")
crimedata<-crimedata %>% filter(between(`Date1 of Occurrence`,as.Date("2014-06-01"),Sys.Date()))
```

- Year of Incident, Month of Occurrence, Day of the Week, Day of the Year should be derived from `Date of Occurrence`

It makes much sense to have the Year, Month, Day of week and Day of year values to be derived from the fields `Date of Occurrence`. This can be easily done with the *lubridate* package.

```
crimedata$`Year of Incident` <- with(crimedata, ifelse(`Year of Incident` ==
            year(`Date1 of Occurrence`),`Year of Incident` ,year(`Date1 of Occurrence`)))
crimedata$`Month1 of Occurence` <- with(crimedata, ifelse(`Month1 of Occurence` ==
            month(`Date1 of Occurrence`),`Month1 of Occurence`,month(`Date1 of Occurrence`)))
crimedata$`Day1 of the Year` <- with(crimedata, ifelse(`Day1 of the Year` ==
            yday(`Date1 of Occurrence`), `Day1 of the Year`, yday(`Date1 of Occurrence`)))
crimedata$`Day1 of the Week` <- with(crimedata, ifelse(`Day1 of the Week` ==
            wday(`Date1 of Occurrence`), `Day1 of the Week`, wday(`Date1 of Occurrence`)))
```

- Victim Age - Reset values greater than 125 to 125

```
crimedata$`Victim Age`[crimedata$`Victim Age` > 125 ] = 125
```

- Victim Age cannot be a negative value.

```
crimedata$`Victim Age` <- ifelse(crimedata$`Victim Age`< 0 , abs(crimedata$`Victim Age`) ,
                                 crimedata$`Victim Age`)
```

The variable `Division` tells us in which part of the Dallas city the crime incident took place. It is categorized into 7 divisions - CENTRAL, NORTHEAST, SOUTH CENTRAL, SOUTHWEST, NORTH CENTRAL, NORTHWEST, SOUTHEAST. Therefore, it has to be of type *Factor*

```
unique(crimedata$Division)
```

```
##  [1] "NORTH CENTRAL" "SOUTHWEST"     "CENTRAL"       "NORTHEAST"
##  [5] "NORTHWEST"     "SOUTHEAST"     "SOUTH CENTRAL" ""
##  [9] "SouthEast"     "NorthEast"     "Central"       "North Central"
## [13] "SouthWest"     "South Central" "NorthWest"
```

It's clear that the values here are not unique, hence converting them all to Upper case help make the analysis easy.

- Division - Change all the values to Upper case

```
tmp <- crimedata %>% mutate(new_division = toupper(Division))
crimedata$Division <- tmp$new_division
crimedata$Division <- as.factor(crimedata$Division)
```

In order to find out how the crime rate trends throughout the year, we create a new varaiable named `Season` based on the `Month of Occurence`. It takes the values - *Spring*, *Summer*, *Fall*, *Winter* and is of type *Factor*.

- Find the Season of the Year i.e., Spring/Summer/Fall/Winter

```
crimedata <- crimedata %>% mutate(Season = ifelse(`Month1 of Occurence` %in% c(3,4,5), "Spring",
                    ifelse(`Month1 of Occurence`%in% c(6,7,8), "Summer",
                          ifelse(`Month1 of Occurence` %in% c(9,10,11), "Fall",
                                 ifelse(`Month1 of Occurence` %in% c(12,1,2), "Winter", NA)))))
crimedata$Season <- as.factor(crimedata$Season)
```

- Columns - Victim Gender, Victim Condition, Offense Status should of type *Factor*

```
crimedata$`Victim Gender` <- as.factor(crimedata$`Victim Gender`)
crimedata$`Victim Condition` <- as.factor(crimedata$`Victim Condition`)
crimedata$`Offense Status` <- as.factor(crimedata$`Offense Status`)
```

We can also try to find out the time of day most crimes tend to happen. For this, we extract the hour of the day from the `Time of Occurrence` variable.

- Extract the hour of the day when the crime took place.

```
crimedata <- crimedata %>% mutate(`Hour of the Day` = sub(":.*", "",`Time1 of Occurrence`))
```

Variable `Type of Property` stores the target item of the incident. Example : Motor vehicle, Apartment. It cannot take numeric values.

- `Type of Property` cannot have numeric values

```
crimedata$`Type of Property` <- as.factor(crimedata$`Type of Property`)
crimedata$`Type of Property`<-droplevels(crimedata$`Type of Property`,exclude=c(910,920,932,510))
```

`UCR Offense Name` stores the type of crime incident that took place. For all the crimes that happened in the year 2019, there is no value for `UCR Offense Name`. This can be extracted from the column `Type of Incident`.

To do this, we create a dataframe which maps the unique `Type of Incident` to its corresponding `UCR Offense Name` called *offenseNames*. Based on this mapping, we find all the missing values for the column `UCR Offense Names`.

```
temp <- crimedata[!duplicated(crimedata$`Type of Incident`), ]
offenseNames <- temp %>% dplyr::select(`Type of Incident`, `UCR Offense Name`)

offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`, "ARSON"))] <-
  "ARSON"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^GRAFFITI","^CRUELTY TO","^CRIM MISCHIEF"),
          collapse="|"),offenseNames$`Type of Incident` ))] <- "VANDALISM & CRIM MISCHIEF"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^ASSAULT","^DEADLY CONDUCT"),collapse="|"),
                                            offenseNames$`Type of Incident` ))] <- "ASSAULT"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`, "BMV"))] <-
  "THEFT/BMV"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^CREDIT CARD", "^COMPUTER SECURITY",
 "^DECEPTIVE", "^FRAUD", "^THEFT OF SERVICE", "^FALSE STATEMENT", "^TAMPER W" , "^SECURE EXE",
 "^FAIL TO", "FALSE ALARM"), collapse="|"),offenseNames$`Type of Incident` ))] <- "FRAUD"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,
                              "CRIMINAL TRESPASS"))] <- "CRIMINAL TRESPASS"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^DELIVERY", "^MAN DEL", "^POSS CONT",
"^POSS MARIJUANA"), collapse="|"),offenseNames$`Type of Incident` ))] <- "NARCOTICS & DRUG"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^DISORDERLY", "^DISRUPT", "^ILLUMINA",
      "^ONLINE IMPRESS", "^STALKING", "^SEX OFFENDERS", "^HARASSMENT"), collapse="|"),
       offenseNames$`Type of Incident` ))] <- "DISORDERLY CONDUCT"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`, "DWI"))] <-
  "DWI"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`, "ESCAPE"))] <-
  "ESCAPE"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,"EVADING"))] <-
  "EVADING"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,"FORGERY"))] <-
```

```
  "FORGE & COUNTERFEIT"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,
                                                 "KIDNAPPING"))] <- "KIDNAPPING"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^ILLEGAL", "^INTERFERE", "^INTERFER",
  "^FLEEING", "^MISAPP", "^OTHER OFFENSES", "^WARRANT"), collapse="|"),
  offenseNames$`Type of Incident` ))] <- "OTHERS"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,
                        "TRAFFICKING"))]<- "HUMAN TRAFFICKING"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,
                                                 "UNAUTHORIZED USE OF"))] <- "UUMV"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^UNLAWFULLY", "^UNLAWFUL", "^PROHIBITED"),
                          collapse="|"),offenseNames$`Type of Incident` ))] <- "WEAPONS"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^VIO BOND", "^VIO PROTECT"), collapse="|"),
                            offenseNames$`Type of Incident` ))] <- "OFFENSE AGAINST CHILD"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^TRAFFIC VIO", "^TRAF VIO"), collapse="|"),
                            offenseNames$`Type of Incident` ))] <- "TRAFFIC VIOLATION"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,
                                  "TRADEMARK"))] <- "FORGE & COUNTERFEIT"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,
                                    "THEFT ORG"))] <- "THEFT ORG RETAIL"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,
                                    "TERRORISTIC THREAT"))] <- "TERRORISTIC THREAT"
offenseNames$`UCR Offense Name`[which(startsWith(offenseNames$`Type of Incident`,
                                    "RESIST ARREST"))] <- "RESIST ARREST"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^MURDER", "^MANSLAUGHTER"), collapse="|"),
                                            offenseNames$`Type of Incident` ))] <- "MURDER"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^PUBLIC INTOX", "^PURCHASE FURN"),
                   collapse="|"), offenseNames$`Type of Incident` ))] <- "DRUNK & DISORDERLY"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^ROBBERY *OF BUSINESS"), collapse="|"),
                                 offenseNames$`Type of Incident` ))] <- "ROBBERY-BUSINESS"
offenseNames$`UCR Offense Name`[which(grepl(paste(c("^ROBBERY *OF INDIVIDUAL"), collapse="|"),
                                 offenseNames$`Type of Incident` ))] <- "ROBBERY-INDIVIDUAL"
offenseNames$`UCR Offense Name` <- with(offenseNames, ifelse(`UCR Offense Name` == "" &
                     grepl("^THEFT", `Type of Incident`), "OTHER THEFTs" , `UCR Offense Name` ))
offenseNames$`UCR Offense Name`[which(offenseNames$`UCR Offense Name` == "")] <- "OTHERS"

crimedata$`UCR Offense Name` <- offenseNames[match(crimedata$`Type of Incident`,
                                                   offenseNames$`Type of Incident`),2]
```

```
length(unique(crimedata$`UCR Offense Name`))
```

```
## [1] 49
```

There are 49 different types of crime. We can group similar categories of crime into one and make this number smaller.

- Group similar offense types

```
crimedata$`Crime Type` <- crimedata$`UCR Offense Name`

crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("ASSAULT", "AGG ASSAULT - NFV"))] =
  "ASSAULT"
crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("BURGLARY-BUSINESS",
                                            "BURGLARY-RESIDENCE"))] = "BURGLARY"
```

```
crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("THEFT/BMV", "THEFT/SHOPLIFT",
                            "OTHER THEFTS", "THEFT ORG RETAIL", "EMBEZZLEMENT"))] = "THEFT"
crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("ROBBERY-BUSINESS",
                                                  "ROBBERY-INDIVIDUAL"))] = "ROBBERY"
crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("ACCIDENT MV",
                                               "MOTOR VEHICLE ACCIDENT"))] = "ACCIDENT"
crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("NARCOTICS & DRUGS",
  "NARCOTICS & DRUG", "DRUNK & DISORDERLY", "DWI", "LIQUOR OFFENSE",
  "INTOXICATION MANSLAUGHTER"))] = "DRUGS"
crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("TRAFFIC VIOLATION",
                                                    "TRAFFIC FATALITY"))] = "TRAFFIC"
crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("MURDER","SUDDEN DEATH&FOUND BODIES",
 "VANDALISM & CRIM MISCHIEF", "WEAPONS", "ARSON", "TERRORISTIC THREAT", "KIDNAPPING",
 "HUMAN TRAFFICKING", "OFFENSE AGAINST CHILD", "ORANIZED CRIME"))] = "VIOLENCE"
crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("DISORDERLY CONDUCT" ,
 "CRIMINAL TRESPASS", "EVADING", "RESIST ARREST", "FAIL TO ID", "GAMBLING", "ESCAPE",
 "FRAUD", "UUMV", "FORGE & COUNTERFEIT"))] = "NONVIOLENCE"
crimedata$`Crime Type`[which(crimedata$`Crime Type` %in% c("NOT CODED", "LOST", "ANIMAL BITE",
 "OTHERS", "FOUND", "INJURED FIREARM", "INJURED HOME", "INJURED OCCUPA", "INJURED PUBLIC"))] =
  "OTHERS"

crimedata$`Crime Type` <- as.factor(crimedata$`Crime Type`)
```

Similarly, let us group all the similar weapon categories and store it as *Factor*

```
crimedata$`Weapon Type` <- ""

crimedata$`Weapon Type`[which(grepl("gun",crimedata$`Weapon Used`, ignore.case = TRUE))] <-
  "Gun"
crimedata$`Weapon Type`[which(crimedata$`Weapon Used` %in% c("Rifle", "Missile/Rock"))] = "Gun"
crimedata$`Weapon Type`[which(crimedata$`Weapon Used` %in% c("Hands-Feet"))] = "Hands/Feet"
crimedata$`Weapon Type`[which(crimedata$`Weapon Used` %in% c("Vehicle", "MOTOR VEHICLE"))] =
  "Vehicle"
crimedata$`Weapon Type`[which(crimedata$`Weapon Used` %in% c("None"))] = "No Weapons"
crimedata$`Weapon Type`[which(crimedata$`Weapon Used` %in% c("Threats"))] = "Threat"
crimedata$`Weapon Type`[which(grepl("knife",crimedata$`Weapon Used`, ignore.case = TRUE))] =
  "Knife"
crimedata$`Weapon Type`[which(crimedata$`Weapon Used` %in% c("Other Cutting Stabbing Inst.",
                                        "SWITCHBLADE", "AXE", "ICE PICK"))] = "Knife"
crimedata$`Weapon Type`[which(grepl("fire",crimedata$`Weapon Used`, ignore.case = TRUE))] =
  "Fire"
crimedata$`Weapon Type`[which(crimedata$`Weapon Used` %in% c("Explosives","Gas/Carbon Monoxide",
                                                      "Burn/Scald"))] = "Knife"
crimedata$`Weapon Type`[which(grepl("drugs",crimedata$`Weapon Used`, ignore.case = TRUE))] =
  "Drugs"
crimedata$`Weapon Type`[which(crimedata$`Weapon Used` %in% c("Omission/Neglect",
"ANY WEAPON OF FORCE DEADLY DISEASE, ETC"))] = "Drugs"
crimedata$`Weapon Type`[which(crimedata$`Weapon Used` %in% c("Other", "Blunt", "Stangulation",
              "Assault", "Crowbar", "Asphixation", "BlackJack/Club", "Omission"))] <- "Others"
crimedata$`Weapon Type`[which(crimedata$`Weapon Type` == "")] <- "Others"

crimedata$`Weapon Type` <- as.factor(crimedata$`Weapon Type`)
```

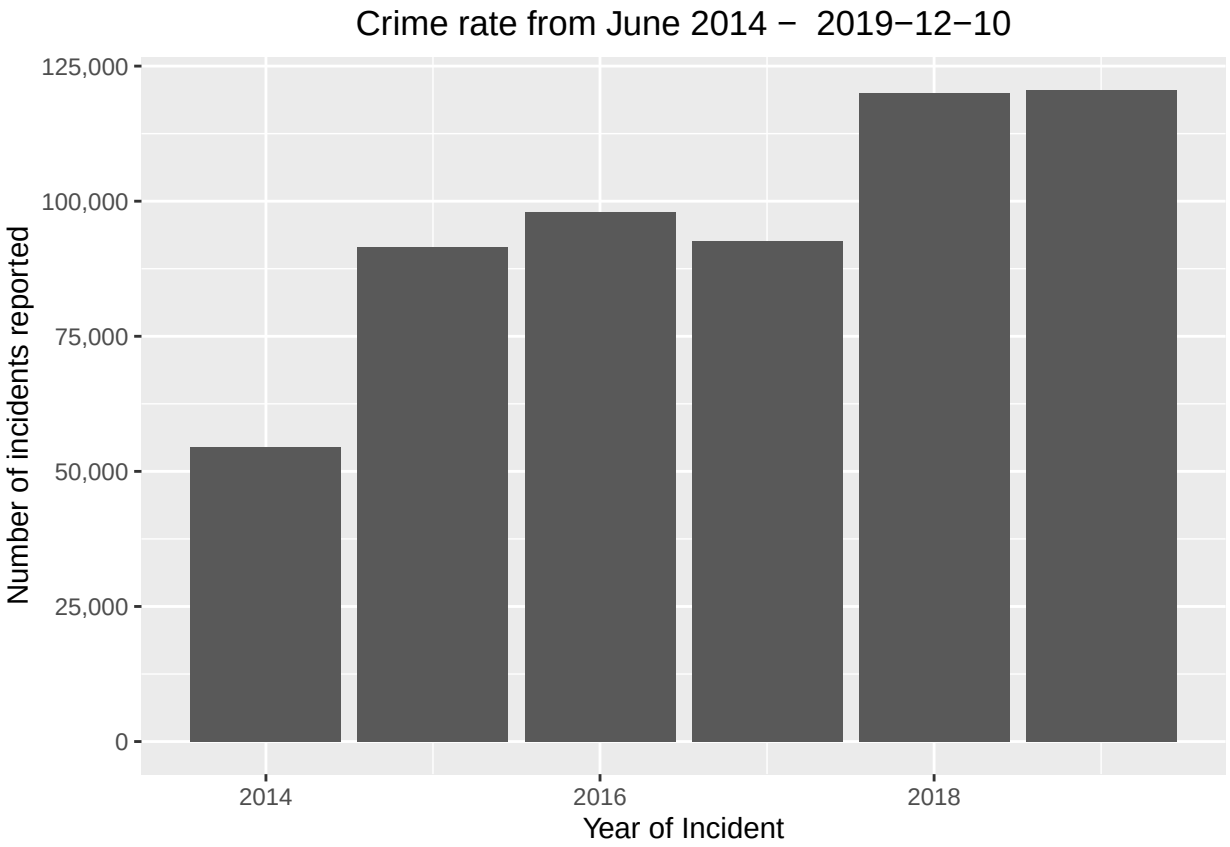
DATA VISUALIZATION

Data Visualization is a powerful way to understand the data and its underlying patterns. *ggplot2* is a package to draw graphics, which implements grammar of graphics. We will perform analysis on Dallas crime data using parameters like time, year, crime type, victim condition, weapons used etc. For this analysis we will use the powerful aggregation functions like *summarise*, *tally* from *dplyr* package.

```
library(ggplot2)
library(RColorBrewer)
library(scales)
```

Let us have a look at the crime rates in Dallas from June 2014 to 2019-12-10.

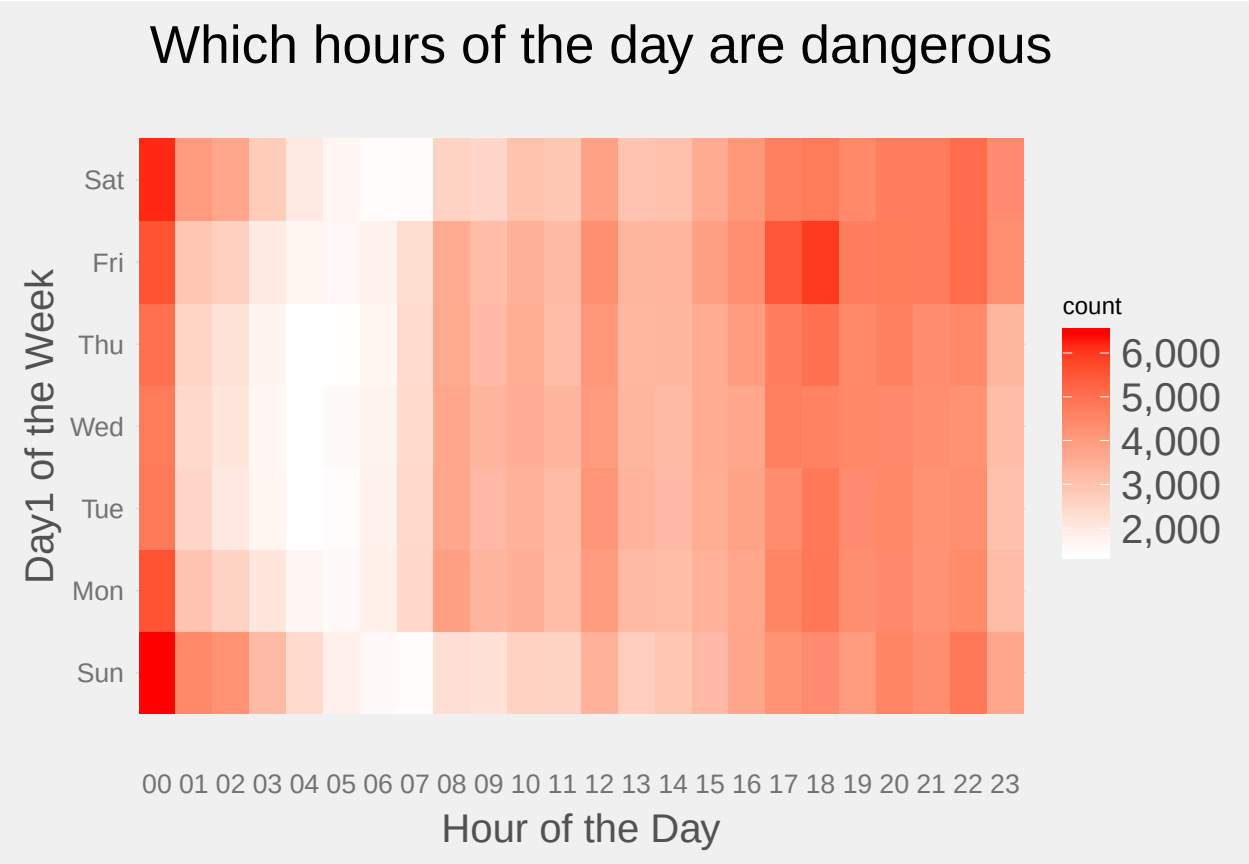
```
ggplot(crimedata, aes(`Year of Incident`)) + geom_bar() +
  ggtitle(paste("Crime rate from June 2014 - ", Sys.Date())) +
  labs( y = "Number of incidents reported", x = "Year of Incident") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(labels = comma)
```



The above plot shows the occurence of crime from June 2014 to 2019-12-10. Note that the crime rates has drastically increased in 2018 compared to the previous years. Let us also look at how crimes varied with respect to time of day, day of week and season.

```
crime_per_day <- crimedata %>%
                  group_by(`Hour of the Day`, `Day1 of the Week`) %>%
                  dplyr::summarise(count = n())
```

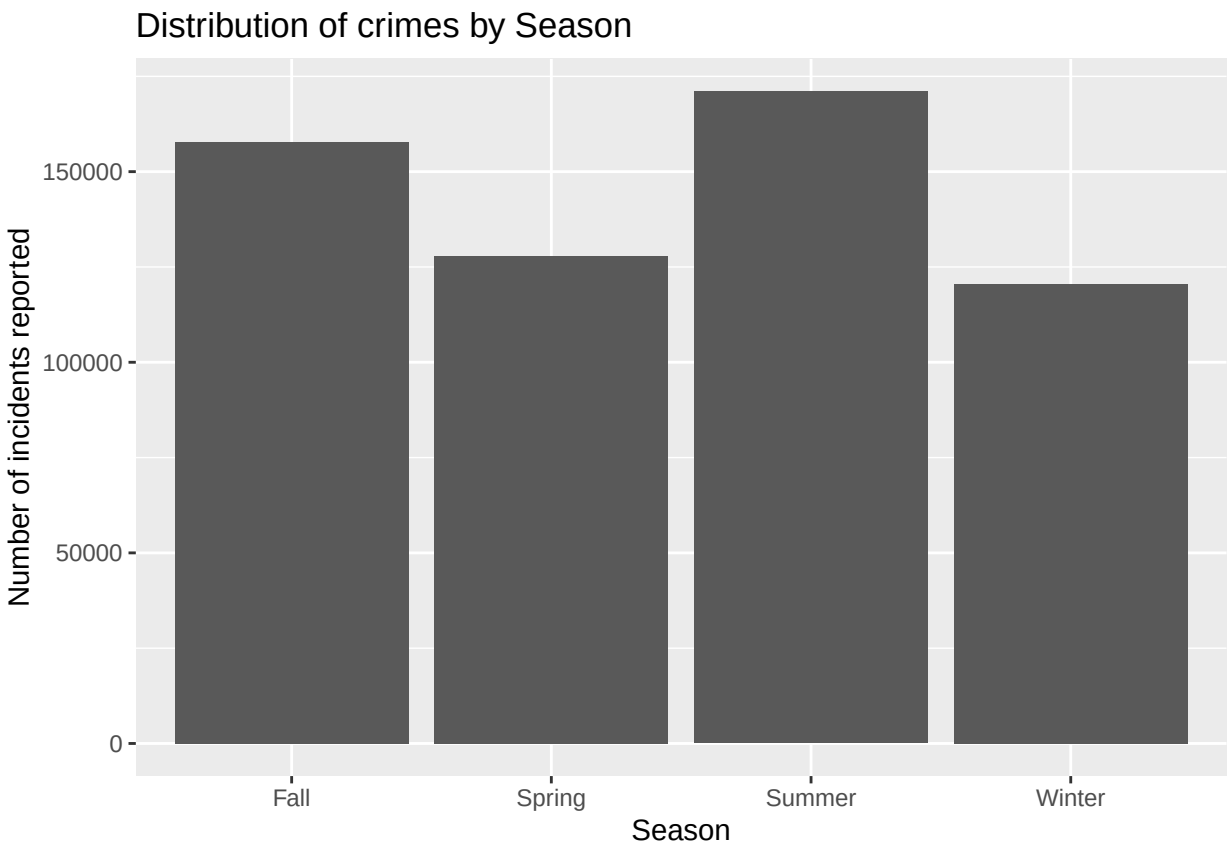
```
days <- c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat")
week.name <- factor(days, levels = days)
ggplot(crime_per_day, aes(x =`Hour of the Day`, y =`Day1 of the Week`, fill = count)) +
  geom_tile() +
  fte_theme() +
  scale_fill_gradient(low = "White", high = "Red", labels = comma) +
  scale_y_discrete(limits = week.name) +
  ggtitle(" Which hours of the day are dangerous")
```



A closer look at the heat map shows that most theft happens around midnight. Especially on the later part of the day on Fridays and weekends, there are significantly more number of crimes happening in the evening compared to the other days. The bar chart below witnesses more crimes during Summer months compared to others.

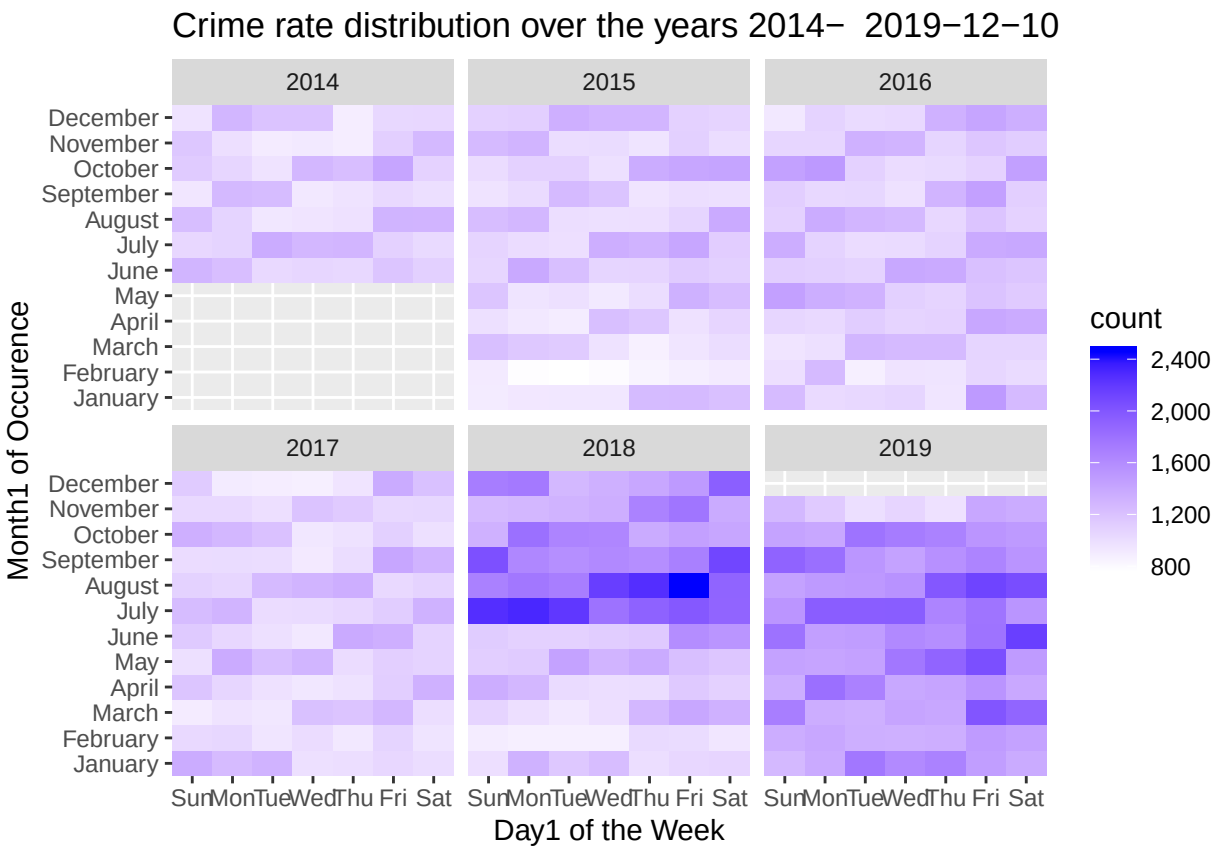
```
crime_month <- crimedata %>% group_by(Season) %>% summarise(n = n())
```

```
ggplot(crime_month, aes(x = Season , y = n)) + geom_bar(stat = "identity") +
  ggtitle("Distribution of crimes by Season") +
  labs( y = "Number of incidents reported")
```




```
crime_per_week <- crimedata %>%
                  group_by(`Day1 of the Week`, `Month1 of Occurence`, `Year of Incident`) %>%
                  dplyr::summarise(count = n())
```

```
days <- c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat")
week.name <- factor(days, levels = days)
ggplot(crime_per_week,aes(x=`Day1 of the Week`,y=`Month1 of Occurence`,fill=count,na.rm=TRUE)) +
  geom_tile() +
  scale_fill_gradient(low = "White", high = "Blue", labels = comma) +
  facet_wrap( ~ `Year of Incident`) +
  scale_x_discrete(limits = week.name, expand = c(0, 0)) +
  scale_y_discrete(limits = month.name, expand = c(0, 0)) +
  ggtitle(paste("Crime rate distribution over the years 2014- ", Sys.Date()))
```



The above heat map briefly shows us the number of crimes that occurred every day of the week for all months from June 2014 till 2019-12-10. This plot proves that crime rates are low during winter (Nov, Dec and Jan) compared to other seasons. It is more evident there is a sharp increase in crime incidents in late 2018 continuing through 2019, in comparison with other years. Let us take a look into this.

```
data <- crimedata %>% filter(`Month1 of Occurence` %in% c(7,8) & `Crime Type` != "ACCIDENT") %>%
  group_by(`Year of Incident`, `Month1 of Occurence`, `Crime Type`) %>% tally()
spread(data, `Crime Type`, n)
```

```
## # A tibble: 12 x 12
## # Groups:   Year of Incident, Month1 of Occurence [12]
##    `Year of Incide~ `Month1 of Occu~ ASSAULT BURGLARY DRUGS NONVIOLENCE
##               <dbl>            <dbl>   <int>    <int> <int>       <int>
##  1             2014                7     617     1007   157        1312
##  2             2014                8     553      971   195        1204
##  3             2015                7     613      939   246        1133
##  4             2015                8     649      951   224        1114
##  5             2016                7     640      807   439        1059
##  6             2016                8     604      869   372        1062
##  7             2017                7     663      834   308        1057
##  8             2017                8     698      881   285        1134
##  9             2018                7     724      740   896        2169
## 10             2018                8     667      745   896        2393
## 11             2019                7     771      697   818        2404
## 12             2019                8     849      726   670        2403
## # ... with 6 more variables: `OTHER THEFTs` <int>, OTHERS <int>,
## #   ROBBERY <int>, THEFT <int>, TRAFFIC <int>, VIOLENCE <int>
```

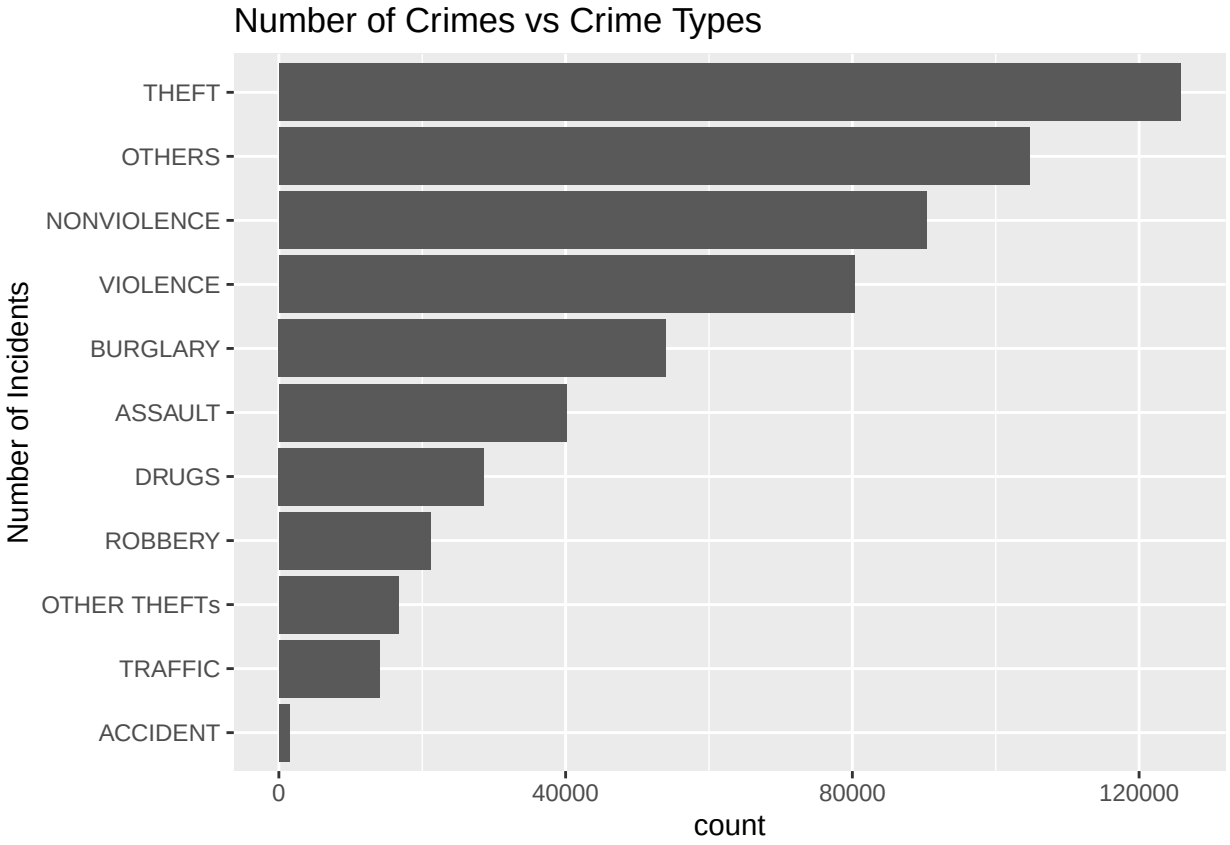
Looking at just July and August months, there is a significant increase in incidents under *DRUGS*, *NONVIOLENCE*,

OTHER THEFTs and *OTHERS* categories. This article also confirms there is an increase in crimes starting July 2018 but no proper reason could be identified.

Now let us try to identify what type of crime is more prevalent in the city.

```
crimedata_top <- within(crimedata,
                        `Crime Type` <- factor(`Crime Type`,
                                               levels=names(sort(table(`Crime Type`),
                                                                 decreasing=FALSE))))
```

```
ggplot(crimedata_top, aes(x=`Crime Type`, na.rm = TRUE)) + geom_histogram(stat = "count") +
  coord_flip() + ggtitle("Number of Crimes vs Crime Types") + labs(x = "Number of Incidents")
```

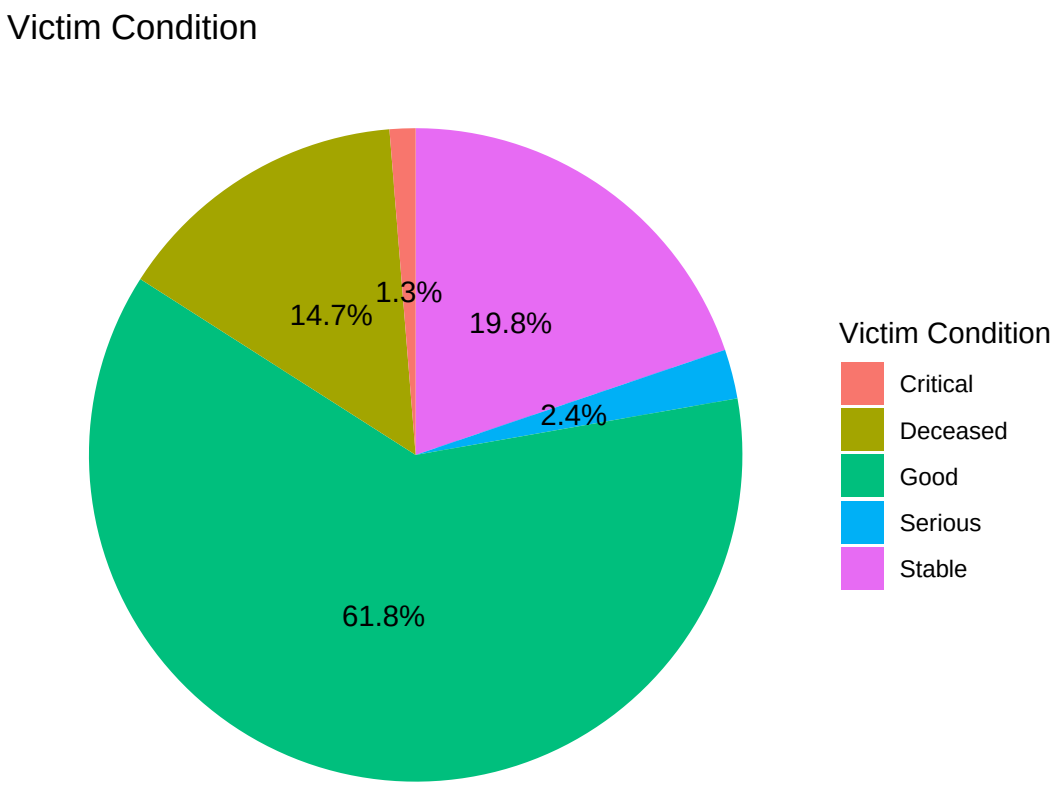


This graph tells that there are more number of thefts happening in the city rather than any other type of crime. We shall also see how bad the victims were affected by these crimes and what type of weapons were used in the crime scene.

```
crime_by_victim <- crimedata %>%
                    filter(`Victim Condition` != "") %>%
                    group_by(`Victim Condition`) %>%  dplyr::summarise(value = n()) %>%
                    ungroup() %>%
                    mutate(per=value/sum(value)) %>%
                    arrange(desc(`Victim Condition`))
crime_by_victim$label <- scales::percent(crime_by_victim$per)
```

```
ggplot(data=crime_by_victim)+
  geom_bar(aes(x="", y=per, fill=`Victim Condition`), stat="identity", width = 1)+
  coord_polar("y", start=0)+
```

```
  theme_void() +
  geom_text(aes(x=1, y = cumsum(per) - per/2, label=label)) +
  ggtitle("Victim Condition")
```



The pie chart on the victims' condition shows that about 61.6% of the victims are in *Good* condition and 21.1% victims are *Stable*. Although a major percentage of the victims are in good health, there are 1.3% and 13.4% victims in *Critical* and *Deceased* states respectively. We should be aiming at reducing this value. To have more idea on how the victims are affected by the crime incidents, let us do some analysis on the weapons used, what weapon is most likely used in a particular crime scene and how they affected the victims.

Recalling the new factor variable created `Weapon Type` which groups similar `Weapons Used`. Levels of `Weapon Type` are :

```
levels(crimedata$`Weapon Type`)
```

```
## [1] "Drugs"      "Fire"       "Gun"        "Hands/Feet" "Knife"
## [6] "No Weapons" "Others"     "Threat"     "Vehicle"
```

Let us also find how many values of each levels are present in our dataset.

```
crimedata %>% count(crimedata$`Weapon Type`, sort = TRUE)
```

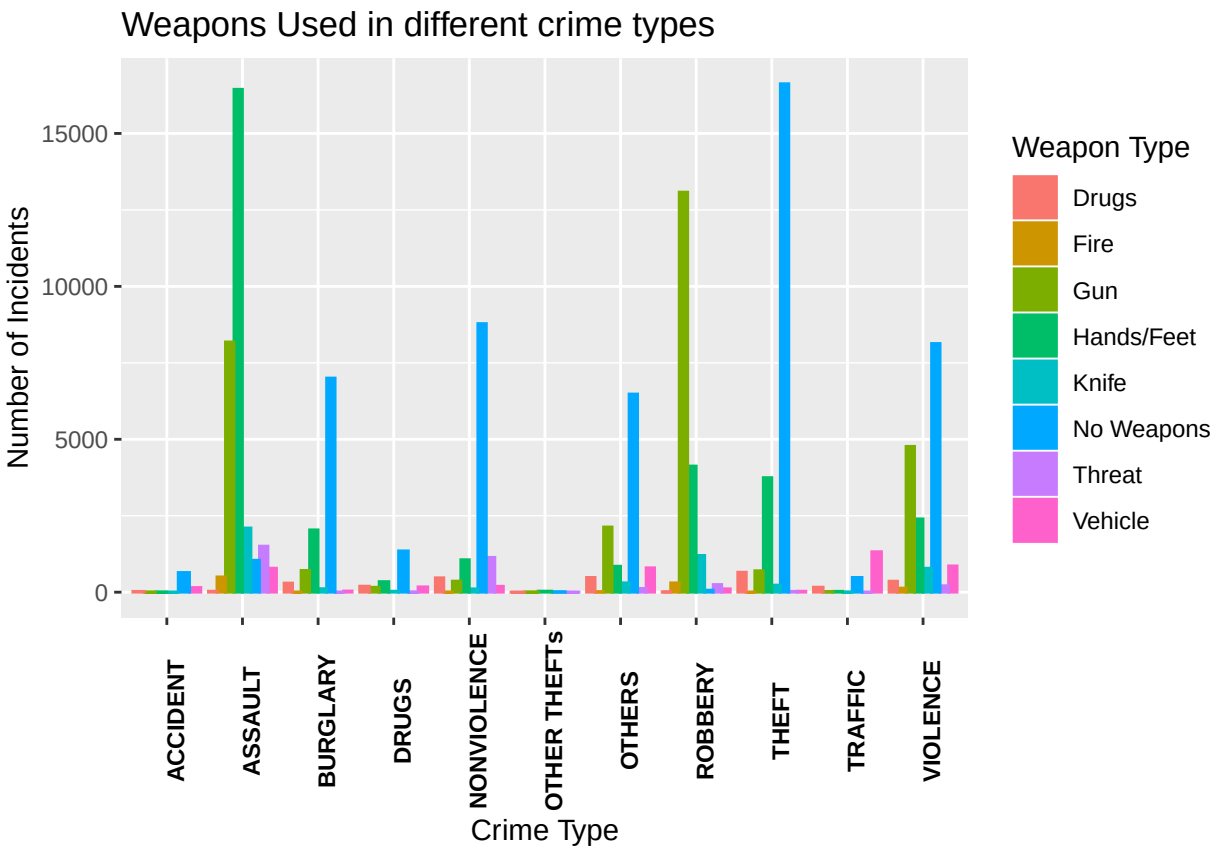
```
## # A tibble: 9 x 2
##   `crimedata$\`Weapon Type\``      n
##   <fct>                        <int>
## 1 Others                      449208
## 2 No Weapons                   50622
```

```
## 3 Hands/Feet                  31077
## 4 Gun                         30129
## 5 Knife                        4888
## 6 Vehicle                      4469
## 7 Threat                       3266
## 8 Drugs                        2699
## 9 Fire                          958
```

There are 9 levels in `Weapon Type` variable however more than 76% of the incidents fall under *Others* category which includes weapons like Blunt, Assault, Crowbar, Asphixation, BlackJack/Club, etc. For the purposes of this study to accurately analyze the weapon usage in various crime categories and not skew the results, we will be ignoring *Others* category from the dataset.

```
weapons_used <- crimedata %>% filter(`Weapon Type` != "Others") %>%
  group_by(`Crime Type`, `Weapon Type`) %>%
  summarise(count = n())
```

```
ggplot(weapons_used, aes(x =`Crime Type`,y = count,color=`Weapon Type`,fill=`Weapon Type`)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  theme(axis.text.x = element_text(angle=90, face="bold", colour="black")) +
  ggtitle("Weapons Used in different crime types") + labs(y = "Number of Incidents")
```



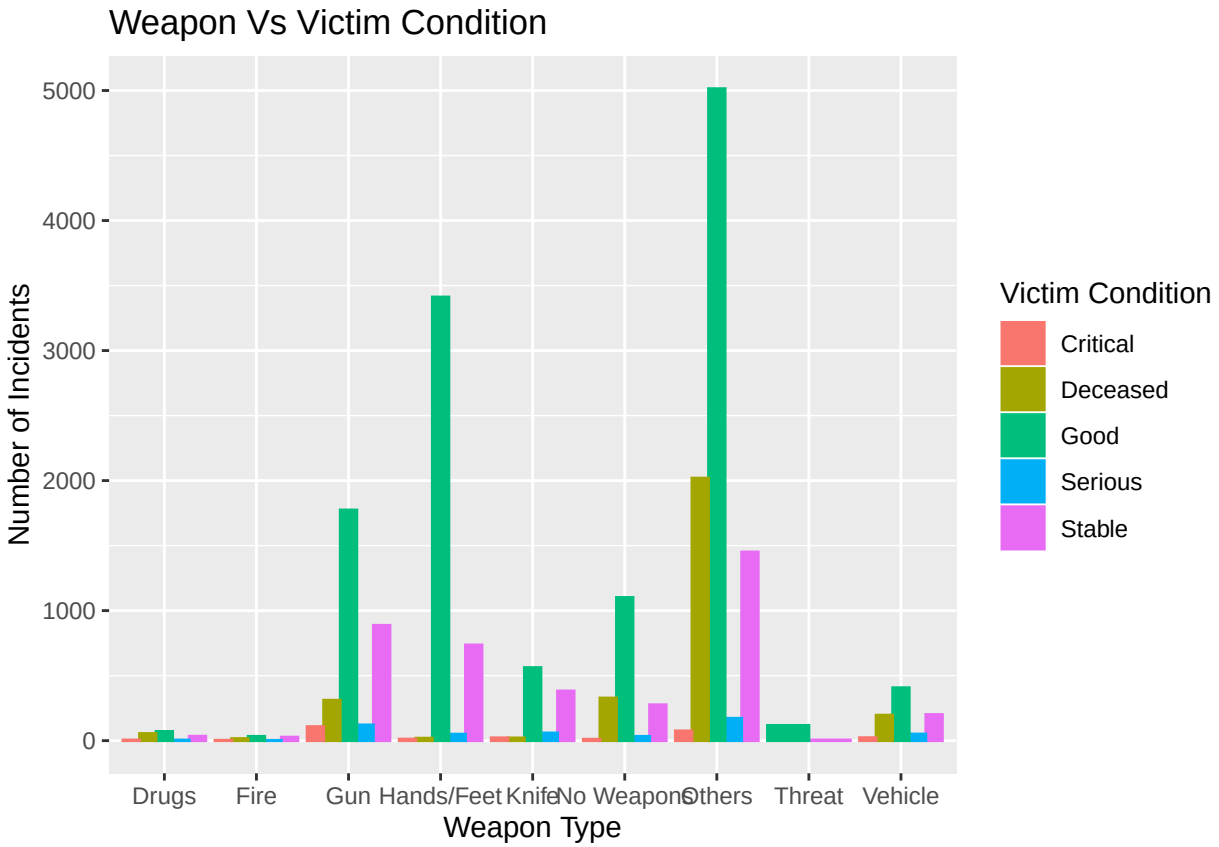
- As we have seen earlier, *Theft* is the most frequent crime in the city compared to other crime types
- Most of the *Theft* cases reported doesn't involve any weapon
- Hands/Feet were the mode of operation for most of the reported *Assault* cases
- Non-violent crimes that include Disorderly conduct, Gambling, Fraud, Failing to ID etc. and Burglary cases show no weapons used on most incidents

- Guns were primarily used in large numbers in crimes like Assault, Robbery and Violence. Gun violence results in thousands on deaths and injuries annually in the U.S

Let us also try to understand how the victims were affected by different types of weapons used in various crime scenes.

```
weapons_victim <- crimedata %>% filter(`Victim Condition` != "") %>%
  group_by(`Weapon Type`, `Victim Condition`) %>% tally(name = "cnt")
```

```
ggplot(weapons_victim,aes(x =`Weapon Type`,y=cnt,color=`Victim Condition`,fill=`Victim Condition`)) +
  geom_bar(stat = "identity", position =position_dodge()) +
   ggtitle("Weapon Vs Victim Condition") + labs(y = "Number of Incidents")
```

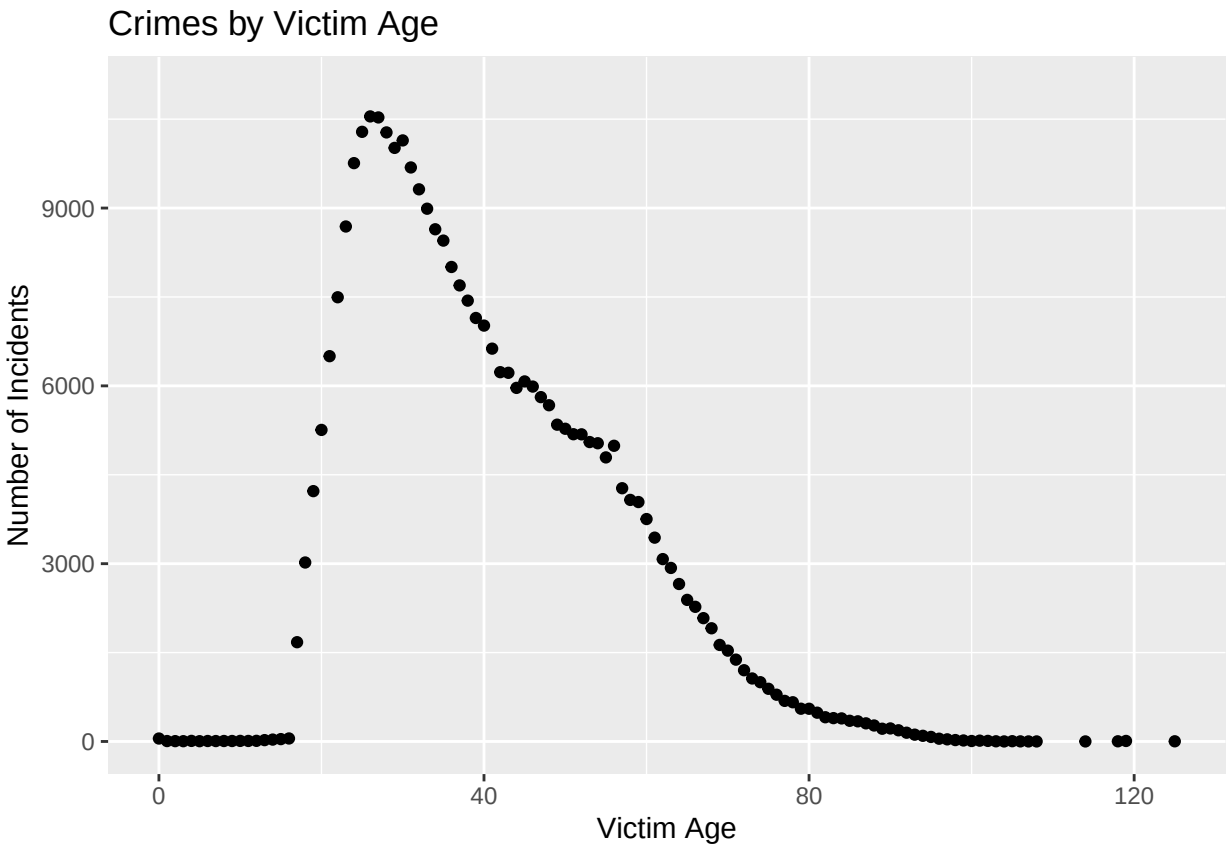


The Weapon type *Others* that includes Blunt, Assault, Crowbar, BlackJack/Club etc., affects more victims in comparison with other types. A majority of them are said to be in Good condition, but not to forget there are also some deaths recorded in this category. This also backs up the fact that gun violence has dropped dramatically in 3 states - New York, California and Texas because of its stringent gun laws. This drop approaches 74% in Dallas.

It is also very important to find out what age group people are most affected by the crime incidents happening in the city.

```
victim_affected <- crimedata %>% group_by(`Victim Age`) %>% summarize(n = n())
```

```
ggplot(victim_affected, aes(x=`Victim Age`, y = n)) + geom_point() + ylim(0,11000) +
  labs(y = "Number of Incidents") + ggtitle("Crimes by Victim Age")
```

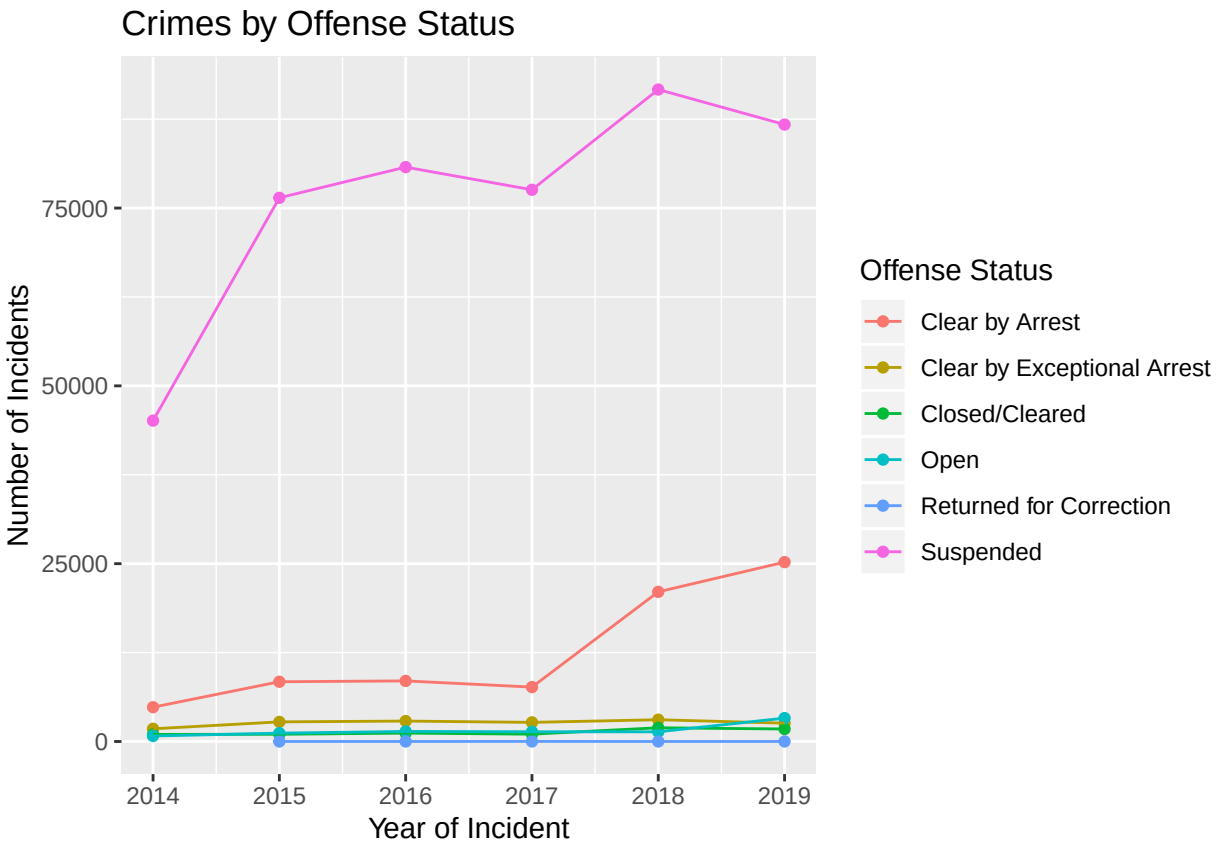


This plots shows us that people in age 20-40 are extremely affected by the crimes. We can also see some casualities in the age group 80-120.

On an average 285 crime events are reported to the Dallas Police Department each day. It will be interesting to find out the status of these crime events.

```
offense_trends <- crimedata %>% filter(`Offense Status` != "") %>%
  group_by(`Offense Status`,`Year of Incident`) %>% summarize(count = n())
```

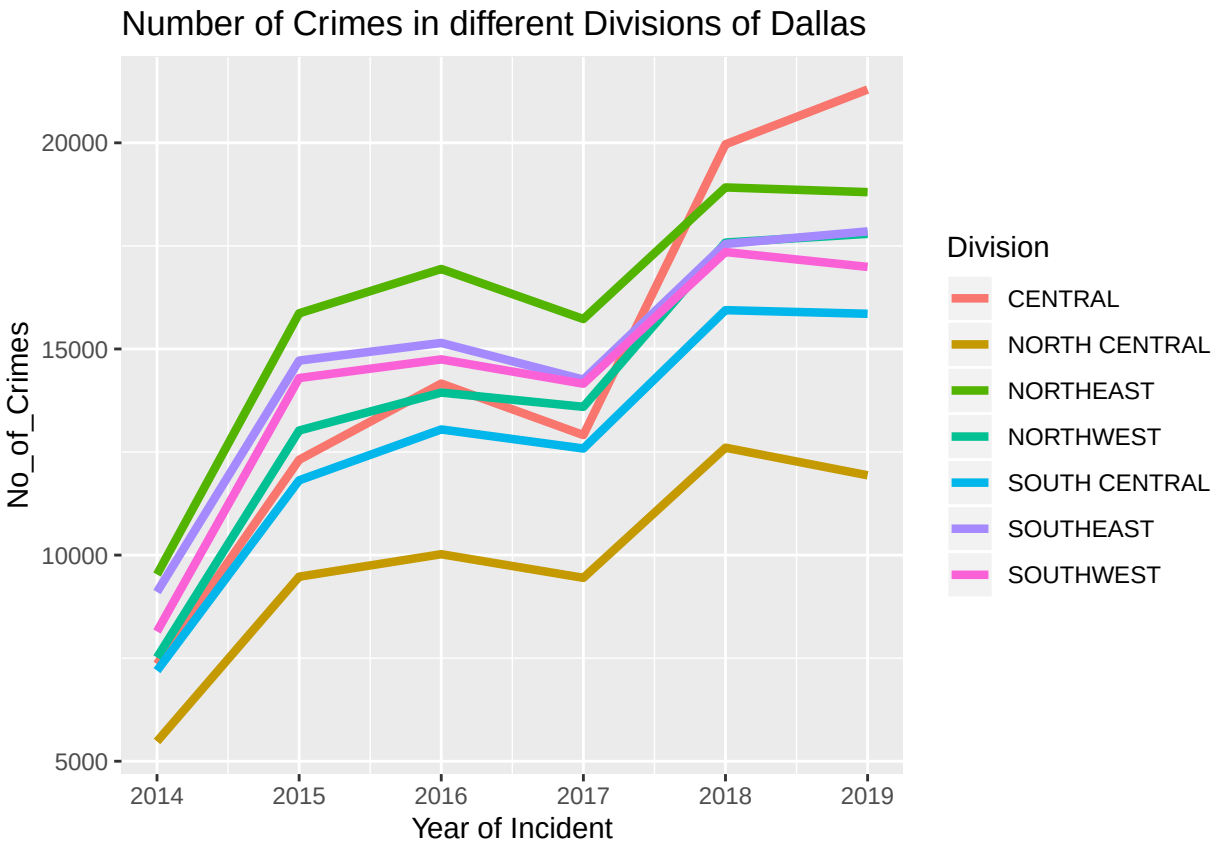
```
ggplot(offense_trends, aes(x=`Year of Incident`, y = count, color = `Offense Status`)) +
  geom_point() + geom_line() + labs(y = "Number of Incidents") +
  ggtitle("Crimes by Offense Status")
```



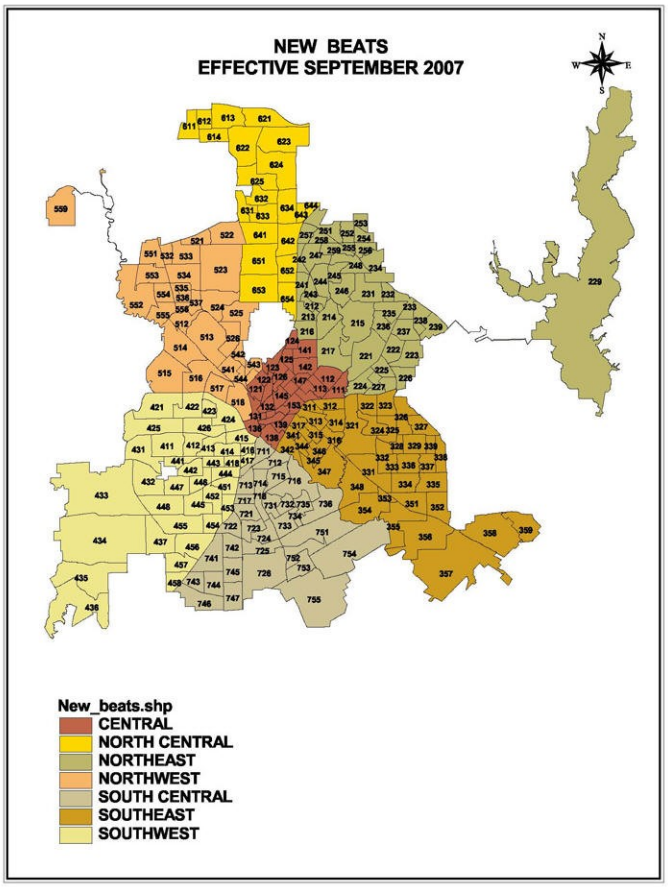
A large number of cases are in *Suspended* state - meaning they are not actively being worked, but not *Closed*. When new evidences are received, the cases may become active. After a period of time, they go cold and are then archived. It is also good to note that certain amount of crimes are *Cleared by arrests* and few are even *Closed*.

```
crime_trends<-crimedata %>% filter(Division != "") %>% group_by(Division,`Year of Incident`) %>%
  tally(name = "No_of_Crimes")
```

```
ggplot(crime_trends, aes(x=`Year of Incident`, y = No_of_Crimes , color = Division, size = 0)) +
  geom_line(size = 1.5) + ggtitle("Number of Crimes in different Divisions of Dallas")
```



This plot clearly explains how the crime rate has increased in various parts of the city over the years. *NORTH CENTRAL* Dallas has recorded lesser crime rates. *CENTRAL* and *NORTHEAST* Dallas have higher crime rates than the other parts of the city.

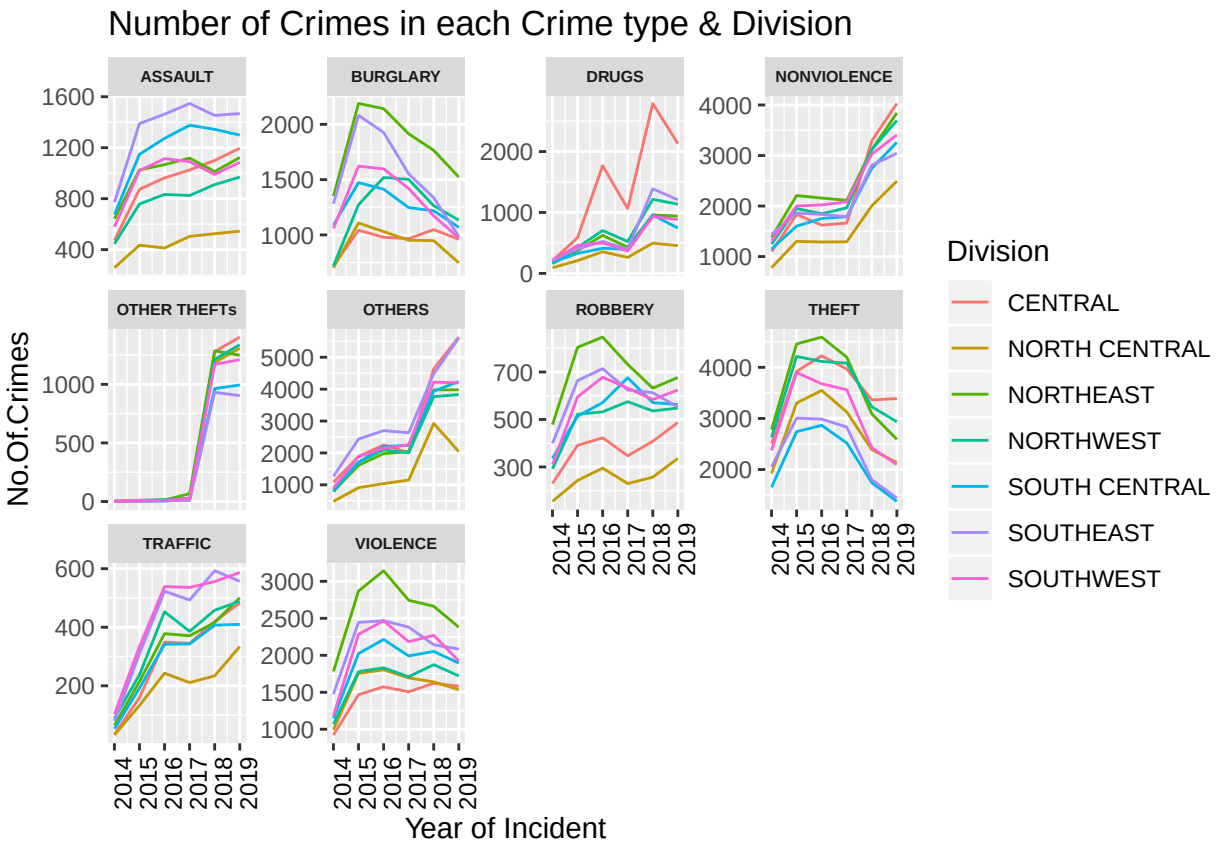


CENTRAL Dallas being the smallest division in terms of area, has recorded highest number of incidents. *NORTH CENTRAL* Dallas which stands next to *CENTRAL* Dallas, has the lowest crime rate.

It is also very important to identify what type of crime is recorded high in all the divisions of Dallas. For this, let us exclude `Crime type` *Accidents* and look at other crimes.

```
crime_div <- crimedata %>% filter(`Crime Type` != "ACCIDENT" & Division != "") %>%
  group_by(`Crime Type`, Division, `Year of Incident`) %>% tally(name = "No.Of.Crimes")
```

```
ggplot(crime_div, aes(x = `Year of Incident`, y = `No.Of.Crimes`, color = Division)) +
  geom_line() + facet_wrap( ~ `Crime Type`, scales = "free_y") +
  theme(axis.text.x = element_text(angle=90, colour="black")) +
  ggtitle("Number of Crimes in each Crime type & Division")  +
  theme(strip.text = element_text(size = 6, face = "bold"))
```



From our previous analysis we know that *NORTH CENTRAL* Dallas is the safest place in the city. But *Theft* in *NORTH CENTRAL* region is higher than the *SOUTH CENTRAL* region.

MODELING

The main objective of this project is to predict the crime rate for the future year. Our dataset fits right in the supervised learning technique.

Supervised learning is a learning where it takes a sample of input and desired outputs(training data), analyses them and effectively predicts output data. The correct output produced is entirely based on the training data. Supervised learning can be of two categories:

- Classification : When the output variable is a category
- Regression : When the output variable is a real or continuous value

Clearly, our problem falls under the regression category.

With our visualizations, we learnt that there is a significant increase in crime incidents with respect to time of the day. So, we try to build a model with respect to the time(Hour) of the day the crime incidents happened. We need to prepare our data for modeling. The predictor variables to solve our problems are : Year of Incident and Hour of the Day

```
crime_by_year <- crimedata %>% group_by(`Year of Incident`, `Hour of the Day`) %>%
  summarise(total = n())

crime_by_year <- as.data.frame(crime_by_year)

crime_by_year$`Hour of the Day` <- as.factor(crime_by_year$`Hour of the Day`)
```

- **ANOVA**

Analysis of Variance (ANOVA) is a statistical method used to test the differences between two or more means.Inferences about means are made by analysing the variance.

```
res.aov <- aov(crime_by_year$total ~  crime_by_year$`Hour of the Day` +
                 crime_by_year$`Year of Incident`)
summary(res.aov)
```

```
##                                 Df    Sum Sq   Mean Sq F value Pr(>F)
## crime_by_year$`Hour of the Day`   23 216496947   9412911   33.27 <2e-16
## crime_by_year$`Year of Incident`   1 100615747 100615747  355.61 <2e-16
## Residuals                        119  33669754    282939
##
## crime_by_year$`Hour of the Day`  ***
## crime_by_year$`Year of Incident` ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary of ANOVA shows that the P values of both predictor variables are less than 0.05, proving that they are *statistically significant*

- **LINEAR REGRESSION**

Linear Regression is a method for fitting a regression line, y = f(x). The typical use of the model is to predict y given a set of predictors x. The predictors can be continuous, categorical or a mix of both. Here the function to be called is *lm()*.

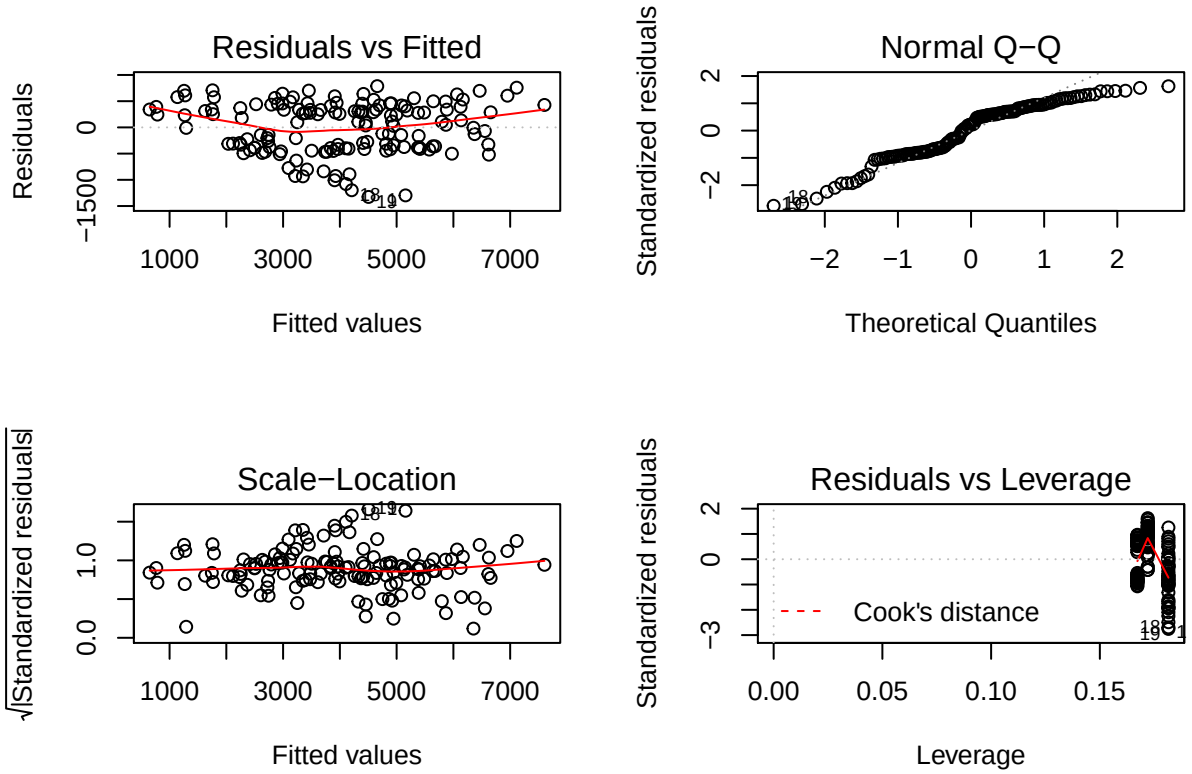
```
lm_mod <- lm(total ~`Hour of the Day` + `Year of Incident` , data = crime_by_year)
summary(lm_mod)
```

```
##
## Call:
## lm(formula = total ~ `Hour of the Day` + `Year of Incident`,
##     data = crime_by_year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1326.2  -386.3   106.8   395.1   786.5
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -980595.42   52338.81 -18.736  < 2e-16 ***
## `Hour of the Day`01   -2740.67     307.10  -8.924 6.55e-15 ***
## `Hour of the Day`02   -3121.67     307.10 -10.165  < 2e-16 ***
## `Hour of the Day`03   -3864.17     307.10 -12.583  < 2e-16 ***
## `Hour of the Day`04   -4370.83     307.10 -14.232  < 2e-16 ***
## `Hour of the Day`05   -4507.83     307.10 -14.679  < 2e-16 ***
## `Hour of the Day`06   -4391.00     307.10 -14.298  < 2e-16 ***
## `Hour of the Day`07   -3889.50     307.10 -12.665  < 2e-16 ***
## `Hour of the Day`08   -2473.00     307.10  -8.053 6.97e-13 ***
## `Hour of the Day`09   -2854.00     307.10  -9.293 8.84e-16 ***
## `Hour of the Day`10   -2524.17     307.10  -8.219 2.88e-13 ***
```

```
## `Hour of the Day`11   -2788.50     307.10  -9.080 2.82e-15 ***
## `Hour of the Day`12   -1735.83     307.10  -5.652 1.10e-07 ***
## `Hour of the Day`13   -2659.50     307.10  -8.660 2.73e-14 ***
## `Hour of the Day`14   -2657.17     307.10  -8.652 2.84e-14 ***
## `Hour of the Day`15   -2216.00     307.10  -7.216 5.44e-11 ***
## `Hour of the Day`16   -1808.00     307.10  -5.887 3.71e-08 ***
## `Hour of the Day`17    -947.83     307.10  -3.086 0.002521 **
## `Hour of the Day`18    -651.67     307.10  -2.122 0.035914 *
## `Hour of the Day`19   -1250.33     307.10  -4.071 8.45e-05 ***
## `Hour of the Day`20   -1049.83     307.10  -3.418 0.000863 ***
## `Hour of the Day`21   -1230.33     307.10  -4.006 0.000108 ***
## `Hour of the Day`22    -987.83     307.10  -3.217 0.001671 **
## `Hour of the Day`23   -2193.00     307.10  -7.141 7.97e-11 ***
## `Year of Incident`      489.45      25.96  18.858  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 531.9 on 119 degrees of freedom
## Multiple R-squared:  0.904,  Adjusted R-squared:  0.8847
## F-statistic:  46.7 on 24 and 119 DF,  p-value: < 2.2e-16
```

On evaluating the Estimate coefficients in the summary, we predict that there will be around 115,884 crime incidents to happen in the year 2020.

```
par(mfrow=c(2, 2))
plot(lm_mod)
```



Looking at the Residuals vs Fitted plot, the prediction made by the model is in x-axis and the accuracy of prediction is on the y-axis. The distance from the line 0 is how bad the prediction was for that value.

Residual = Actual - Predicted

Positive values for the residual means the prediction was too low, negative values means the prediction was too high and 0 means the guess was exactly correct. The Residual vs Fitted values indicates that as we get to the edges of regression, we are less accurate in predicting the crime rate. Values for the year 2020 is more difficult to fit into the regression. Hence there is a room for improvement in our model.

- **POISSON REGRESSION**

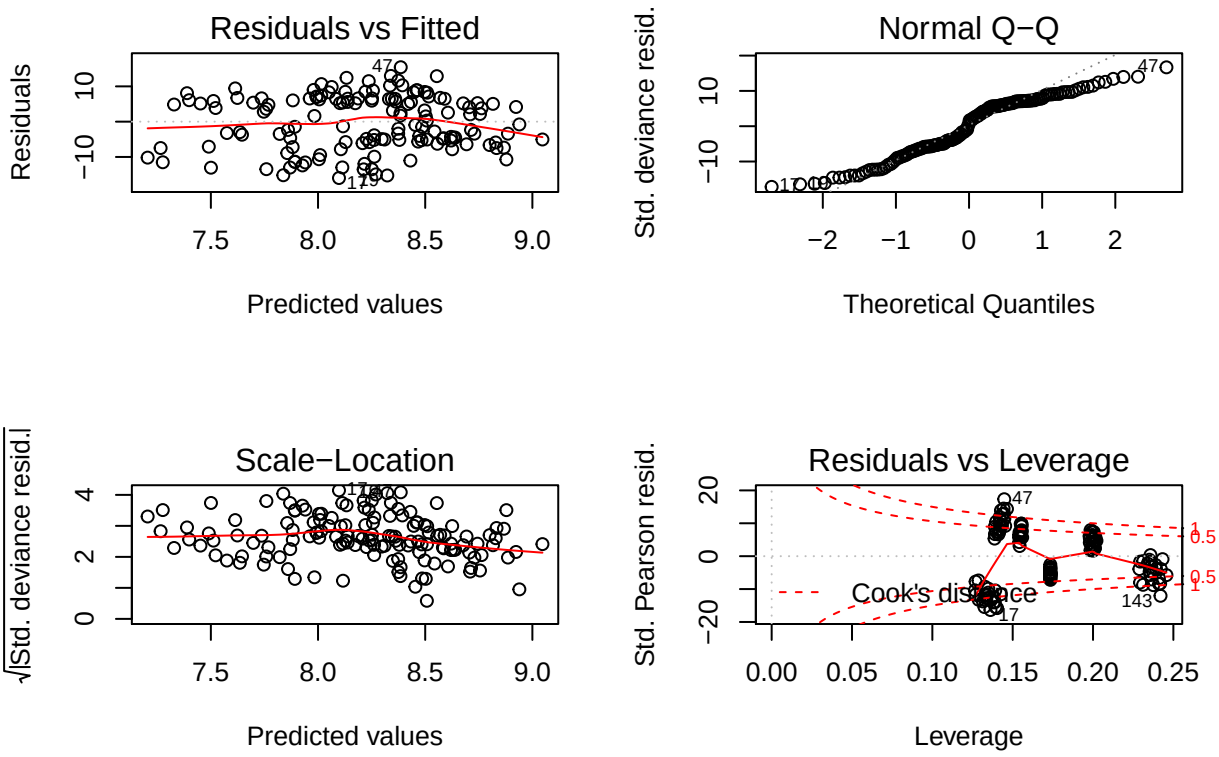
Poisson regression is used to model response variables (Y-values) that are counts. It tells which explanatory variables have a statistically significant effect on the response variables. We perform poisson model analysis with glm() where *family = poisson*

```
poisson.glm <- glm(total ~`Hour of the Day` +`Year of Incident`, data = crime_by_year,
                   family = "poisson")
summary(poisson.glm)
```

```
##
## Call:
## glm(formula = total ~ `Hour of the Day` + `Year of Incident`,
##     family = "poisson", data = crime_by_year)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -15.9689   -5.7929    0.8951    6.4559   15.4215
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.397e+02  1.576e+00 -152.11   <2e-16 ***
## `Hour of the Day`01 -5.613e-01  8.480e-03  -66.19   <2e-16 ***
## `Hour of the Day`02 -6.719e-01  8.790e-03  -76.44   <2e-16 ***
## `Hour of the Day`03 -9.304e-01  9.610e-03  -96.82   <2e-16 ***
## `Hour of the Day`04 -1.155e+00  1.044e-02 -110.63   <2e-16 ***
## `Hour of the Day`05 -1.226e+00  1.073e-02 -114.25   <2e-16 ***
## `Hour of the Day`06 -1.165e+00  1.048e-02 -111.17   <2e-16 ***
## `Hour of the Day`07 -9.406e-01  9.645e-03  -97.52   <2e-16 ***
## `Hour of the Day`08 -4.903e-01  8.293e-03  -59.13   <2e-16 ***
## `Hour of the Day`09 -5.929e-01  8.566e-03  -69.22   <2e-16 ***
## `Hour of the Day`10 -5.035e-01  8.327e-03  -60.47   <2e-16 ***
## `Hour of the Day`11 -5.745e-01  8.516e-03  -67.47   <2e-16 ***
## `Hour of the Day`12 -3.175e-01  7.874e-03  -40.32   <2e-16 ***
## `Hour of the Day`13 -5.393e-01  8.421e-03  -64.04   <2e-16 ***
## `Hour of the Day`14 -5.386e-01  8.419e-03  -63.98   <2e-16 ***
## `Hour of the Day`15 -4.267e-01  8.133e-03  -52.46   <2e-16 ***
## `Hour of the Day`16 -3.332e-01  7.910e-03  -42.12   <2e-16 ***
## `Hour of the Day`17 -1.608e-01  7.537e-03  -21.34   <2e-16 ***
## `Hour of the Day`18 -1.077e-01  7.431e-03  -14.50   <2e-16 ***
## `Hour of the Day`19 -2.181e-01  7.656e-03  -28.49   <2e-16 ***
## `Hour of the Day`20 -1.798e-01  7.575e-03  -23.73   <2e-16 ***
## `Hour of the Day`21 -2.142e-01  7.647e-03  -28.01   <2e-16 ***
## `Hour of the Day`22 -1.682e-01  7.552e-03  -22.27   <2e-16 ***
## `Hour of the Day`23 -4.211e-01  8.119e-03  -51.87   <2e-16 ***
## `Year of Incident`   1.232e-01  7.815e-04  157.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 90256.4  on 143  degrees of freedom
## Residual deviance:  8453.1  on 119  degrees of freedom
## AIC: 9949.9
##
## Number of Fisher Scoring iterations: 4
```

```
par(mfrow=c(2,2))
plot(poisson.glm)
```



To perform the goodness of fit for this model, we need to look at overdispersion. Overdispersion is a situation where the residual deviance of the glm is large relative to the residual degrees of freedom. If the ratio of the residual deviance to the residual degrees of freedom exceeds 1.5, then the model is said to be overdispersed.

```
poisson.glm$deviance/poisson.glm$df.residual
```

```
## [1] 71.03412
```

One potential solution to avoid overdispersion is to use a quasi model.

- **QUASSIPOISSON REGRESSION**

Over-dispersion is a problem if the conditional variance (residual variance) is larger than the conditional mean. One way to check for and deal with over-dispersion is to run a quasi-poisson model.

```r
qpoisson.glm <- glm(total ~`Hour of the Day` +`Year of Incident`, data = crime_by_year,
                    family = "quasipoisson")
summary(qpoisson.glm)
```

```
##
## Call:
## glm(formula = total ~ `Hour of the Day` + `Year of Incident`,
##     family = "quasipoisson", data = crime_by_year)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -15.9689   -5.7929    0.8951    6.4559   15.4215
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.397e+02  1.318e+01 -18.190  < 2e-16 ***
## `Hour of the Day`01 -5.613e-01  7.091e-02  -7.915 1.44e-12 ***
## `Hour of the Day`02 -6.719e-01  7.351e-02  -9.141 2.03e-15 ***
## `Hour of the Day`03 -9.304e-01  8.036e-02 -11.578  < 2e-16 ***
## `Hour of the Day`04 -1.155e+00  8.733e-02 -13.229  < 2e-16 ***
## `Hour of the Day`05 -1.226e+00  8.972e-02 -13.663  < 2e-16 ***
## `Hour of the Day`06 -1.165e+00  8.766e-02 -13.294  < 2e-16 ***
## `Hour of the Day`07 -9.406e-01  8.066e-02 -11.661  < 2e-16 ***
## `Hour of the Day`08 -4.903e-01  6.935e-02  -7.071 1.14e-10 ***
## `Hour of the Day`09 -5.929e-01  7.164e-02  -8.277 2.12e-13 ***
## `Hour of the Day`10 -5.035e-01  6.963e-02  -7.231 5.03e-11 ***
## `Hour of the Day`11 -5.745e-01  7.121e-02  -8.068 6.43e-13 ***
## `Hour of the Day`12 -3.175e-01  6.585e-02  -4.822 4.24e-06 ***
## `Hour of the Day`13 -5.393e-01  7.042e-02  -7.658 5.55e-12 ***
## `Hour of the Day`14 -5.386e-01  7.040e-02  -7.650 5.76e-12 ***
## `Hour of the Day`15 -4.266e-01  6.801e-02  -6.273 5.92e-09 ***
## `Hour of the Day`16 -3.332e-01  6.615e-02  -5.037 1.70e-06 ***
## `Hour of the Day`17 -1.608e-01  6.302e-02  -2.552 0.011987 *
## `Hour of the Day`18 -1.077e-01  6.214e-02  -1.734 0.085542 .
## `Hour of the Day`19 -2.181e-01  6.402e-02  -3.407 0.000897 ***
## `Hour of the Day`20 -1.798e-01  6.335e-02  -2.838 0.005342 **
## `Hour of the Day`21 -2.142e-01  6.395e-02  -3.350 0.001084 **
## `Hour of the Day`22 -1.682e-01  6.315e-02  -2.664 0.008802 **
## `Hour of the Day`23 -4.211e-01  6.790e-02  -6.203 8.32e-09 ***
## `Year of Incident`   1.232e-01  6.535e-03  18.856  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 69.92958)
##
##     Null deviance: 90256.4  on 143  degrees of freedom
## Residual deviance:  8453.1  on 119  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

The summary shows that the residual deviance has not changed. The dispersion parameter, which was forced to be 1 in our last model, is allowed to be estimated here. This model again has overdispersion.

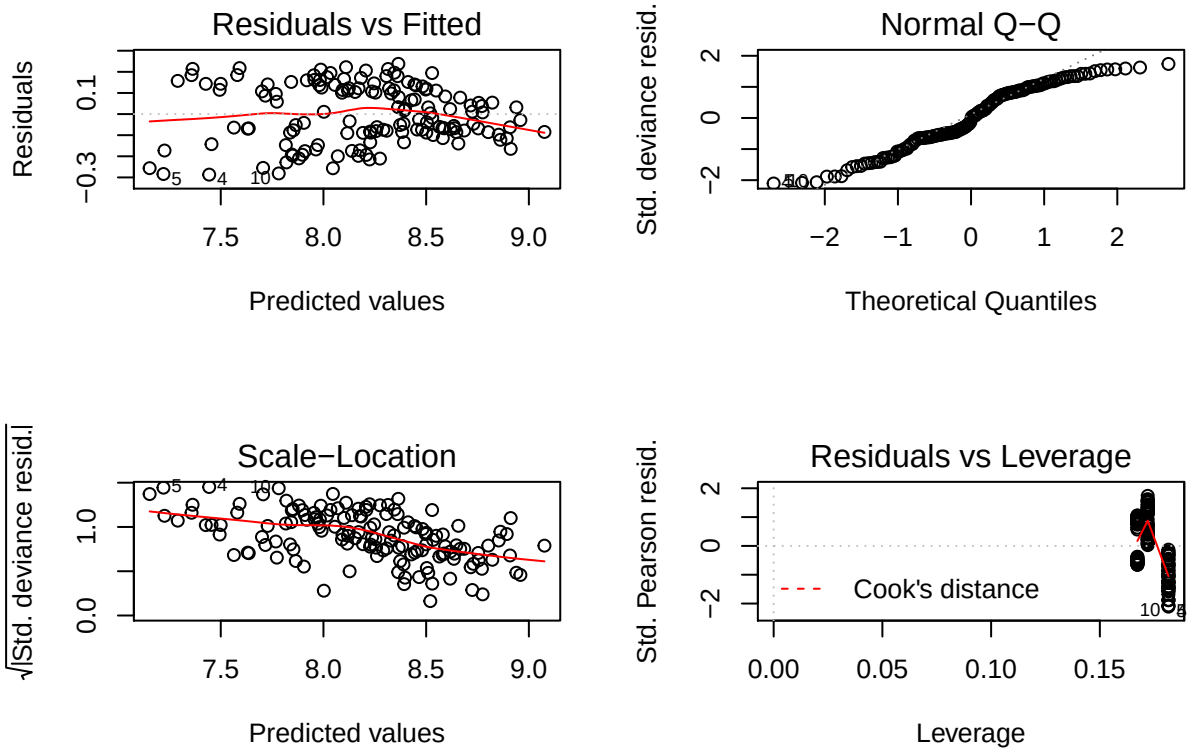
We can transform our data, by performing a mathematical operation and use these transformed values in our statistical tests. We can choose from an infinite number of transformations, but the most common ones are *log* and *sqrt* transformations.

- **LOG TRANSFORMATION OF LM**

```
log_lm <- glm(log(total) ~ `Hour of the Day` + `Year of Incident`, data = crime_by_year)
summary(log_lm)
```

```
##
## Call:
## glm(formula = log(total) ~ `Hour of the Day` + `Year of Incident`,
##     data = crime_by_year)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.286770  -0.087879  -0.007004   0.121708   0.237821
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.670e+02  1.480e+01 -18.035  < 2e-16 ***
## `Hour of the Day`01 -5.726e-01  8.686e-02  -6.591 1.25e-09 ***
## `Hour of the Day`02 -6.861e-01  8.686e-02  -7.899 1.57e-12 ***
## `Hour of the Day`03 -9.485e-01  8.686e-02 -10.919  < 2e-16 ***
## `Hour of the Day`04 -1.171e+00  8.686e-02 -13.482  < 2e-16 ***
## `Hour of the Day`05 -1.238e+00  8.686e-02 -14.257  < 2e-16 ***
## `Hour of the Day`06 -1.166e+00  8.686e-02 -13.423  < 2e-16 ***
## `Hour of the Day`07 -9.372e-01  8.686e-02 -10.789  < 2e-16 ***
## `Hour of the Day`08 -4.948e-01  8.686e-02  -5.696 9.02e-08 ***
## `Hour of the Day`09 -6.099e-01  8.686e-02  -7.022 1.46e-10 ***
## `Hour of the Day`10 -5.114e-01  8.686e-02  -5.887 3.72e-08 ***
## `Hour of the Day`11 -5.753e-01  8.686e-02  -6.624 1.07e-09 ***
## `Hour of the Day`12 -3.224e-01  8.686e-02  -3.711 0.000315 ***
## `Hour of the Day`13 -5.463e-01  8.686e-02  -6.289 5.49e-09 ***
## `Hour of the Day`14 -5.419e-01  8.686e-02  -6.239 6.99e-09 ***
## `Hour of the Day`15 -4.290e-01  8.686e-02  -4.938 2.60e-06 ***
## `Hour of the Day`16 -3.462e-01  8.686e-02  -3.985 0.000117 ***
## `Hour of the Day`17 -1.677e-01  8.686e-02  -1.930 0.055938 .
## `Hour of the Day`18 -1.171e-01  8.686e-02  -1.348 0.180260
## `Hour of the Day`19 -2.225e-01  8.686e-02  -2.561 0.011683 *
## `Hour of the Day`20 -1.835e-01  8.686e-02  -2.113 0.036718 *
## `Hour of the Day`21 -2.145e-01  8.686e-02  -2.469 0.014959 *
## `Hour of the Day`22 -1.643e-01  8.686e-02  -1.892 0.060955 .
## `Hour of the Day`23 -4.190e-01  8.686e-02  -4.824 4.22e-06 ***
## `Year of Incident`   1.367e-01  7.341e-03  18.625  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.02263592)
##
##     Null deviance: 27.4902  on 143  degrees of freedom
## Residual deviance:  2.6937  on 119  degrees of freedom
## AIC: -112.31
##
## Number of Fisher Scoring iterations: 2
```

```
par(mfrow=c(2, 2))
plot(log_lm)
```



There are several other predicting techniques that can be built to fit a better model and predict crime rates for future years. Such techniques can be used to further investigate and identify crime prone localities, help deploy security systems and make use of them effectively. In this way we can try to reduce future crime rates and bring them under control.