# Crimes in NYC

***Karima_tajin***

*12/9/2019*

In this project, I am analyzing the Crimes in NYC during the first 6 month of 2019. the dataset is available in NYC Open Data website: https://data.cityofnewyork.us/browse?tags=crime, with this data I would like to know the answers to the following questions:

1. What parts of the city have the most crime complaints during 2019?

2. Has the crimes complaint improved over months and days during 2019?

3. What type of offenses or offenses categories is higher over the month and day?

4. Which Jurisdiction, premises have the highest crime rate?

5. Which victim gender to be more attacked by suspicious?

6. Which geographical area has the more crimes complaints not considered safe?

## I. Data Set Basic Statistics:

The first step is to start structuring the dataset by checking how many observations and variables, and looking for what informations could be interesting.

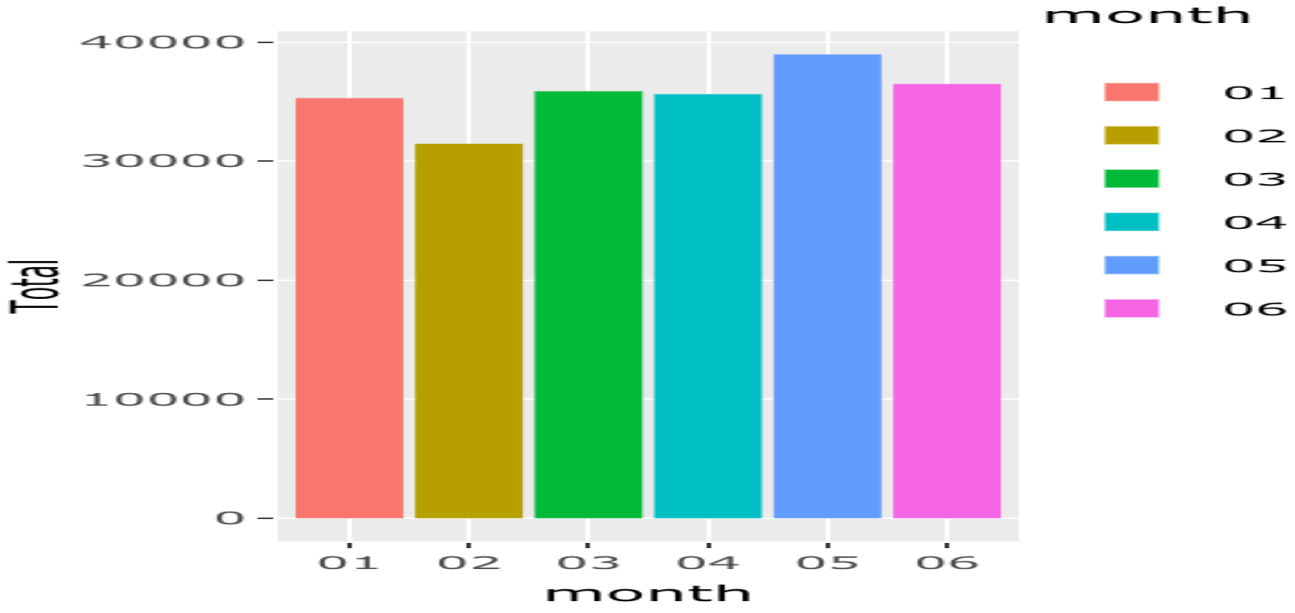
```
# set the working directory
setwd("/Users/karimaidrissi/Desktop/DSSA 5101")

# loading necessary libraries :

library(tidyverse)

library(lubridate)
```

```
library(plotly)

# read the file
crimes <- read_csv("complaint.csv")

# using object size to see the size in memory.
object.size(crimes)

# glimpse the data:
glimpse(crimes)

# basic statistics for all columns in the dataset:
summary(crimes)

# the missing values in our dataset:
sum(is.na(crimes))
rowSums(is.na(crimes))
colSums(is.na(crimes))
table(is.na(crimes))

# Extracting specific columns from our data:
crimes$CMPLNT_FR_DT <- as.Date(as.character(crimes$CMPLNT_FR_DT),
format = "%m/%d/%y")
crimes <- crimes %>%
  select(CMPLNT_NUM, BORO_NM, CMPLNT_FR_DT, CMPLNT_FR_TM,
JURIS_DESC, KY_CD,LAW_CAT_CD, OFNS_DESC,PREM_TYP_DESC,
SUSP_AGE_GROUP, SUSP_RACE, SUSP_SEX, VIC_AGE_GROUP,
VIC_RACE, VIC_SEX, Latitude, Longitude) %>% filter( CMPLNT_FR_DT >=
as.Date("2019-01-01"))
# renaming columns of our data:
names(crimes) <- c("ID","Borough","Date","Time","Jurisdiction","Code","Level
of offense", "Offense", "Premise" , " Suspicious age", " Suspicious race",
"Suspicious sex", "Victim age", " Victim race", " Victim sex","Latitude",
"Longtitude")

# Separate the date column into day, month, year:
crimes <- separate(crimes, col = Date, into = c("year","month","day"), sep ="-")
```

```
# Type of month, day and year:
class(crimes$month)
class(crimes$day)
class(crimes$year)
# change the month,day and year into factors:
crimes$month <- as.factor(crimes$month)
crimes$day <- as.factor(crimes$day)
crimes$year <- as.factor(crimes$year)
# checking the type of month, day and year:
typeof(crimes$month)
typeof(crimes$day)
typeof(crimes$year)
```

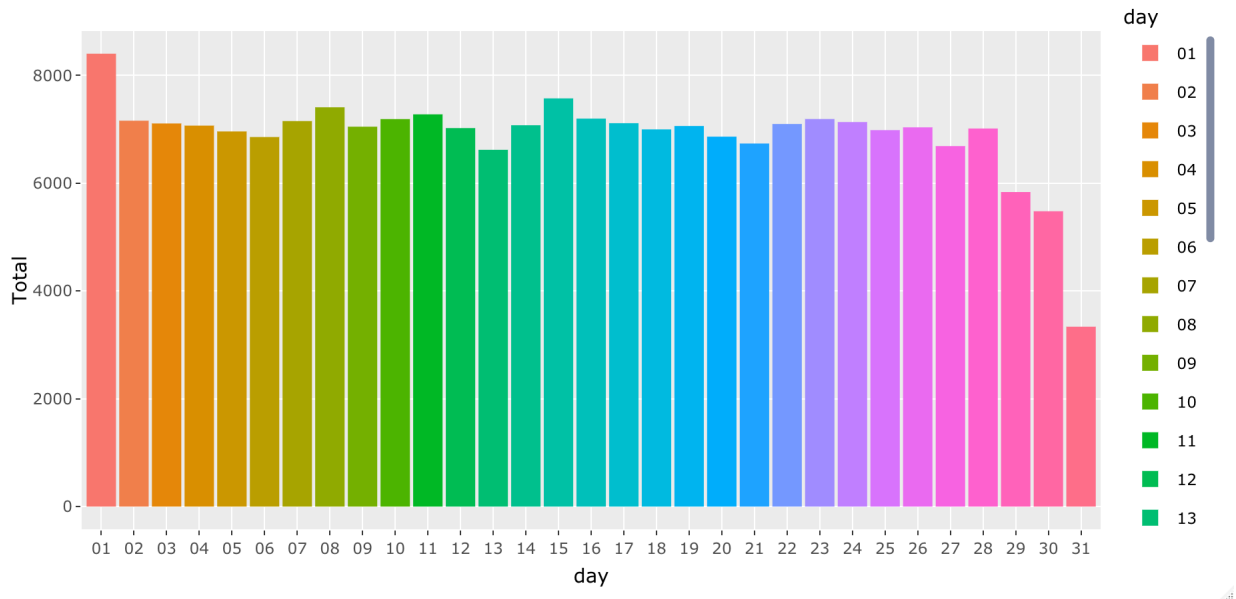
## II. Plotting the graphs:

```
# Crimes by Month:
by_month <- crimes %>% group_by(month) %>% dplyr::summarise(Total = n())
plot1 <- ggplot(subset(by_month), aes(x= month, y= Total, fill = month)) +
geom_bar(stat="identity")
ggplotly(plot1)
```



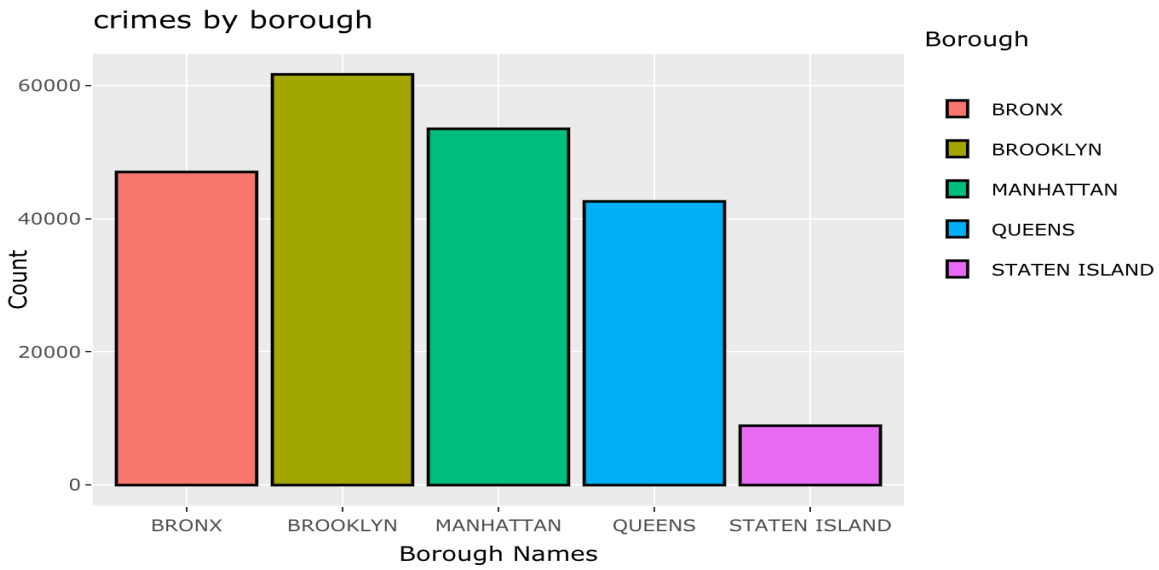
*# Crime by day:*

```
by_day <- crimes %>% group_b y(day) %>% dplyr::summarise(Total = n())
plot2 <- ggplot(by_day, aes(x=day, y=Total, fill =day)) +
geom_bar(stat="identity")
ggplotly(plot2)
```



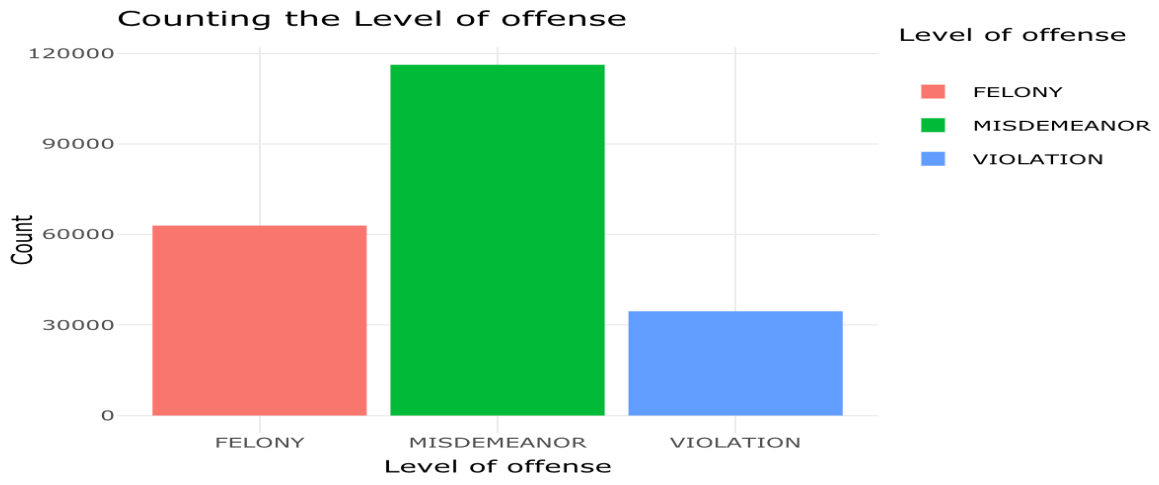

*# Which borough is the most dangerous:*

```
library(tidyverse)
crimes %>% drop_na("Borough")
crimes[!is.na(crimes$Borouh), ]
is.na(crimes$Borough)
plot3 <- ggplot(data = crimes, aes(x= crimes$Borough,fill = Borough))+
          geom_bar(colour = "black",stat = "count")+
          labs(x = "Borough Names", y="Count",
               title = "crimes by borough")
ggplotly(plot3)
```

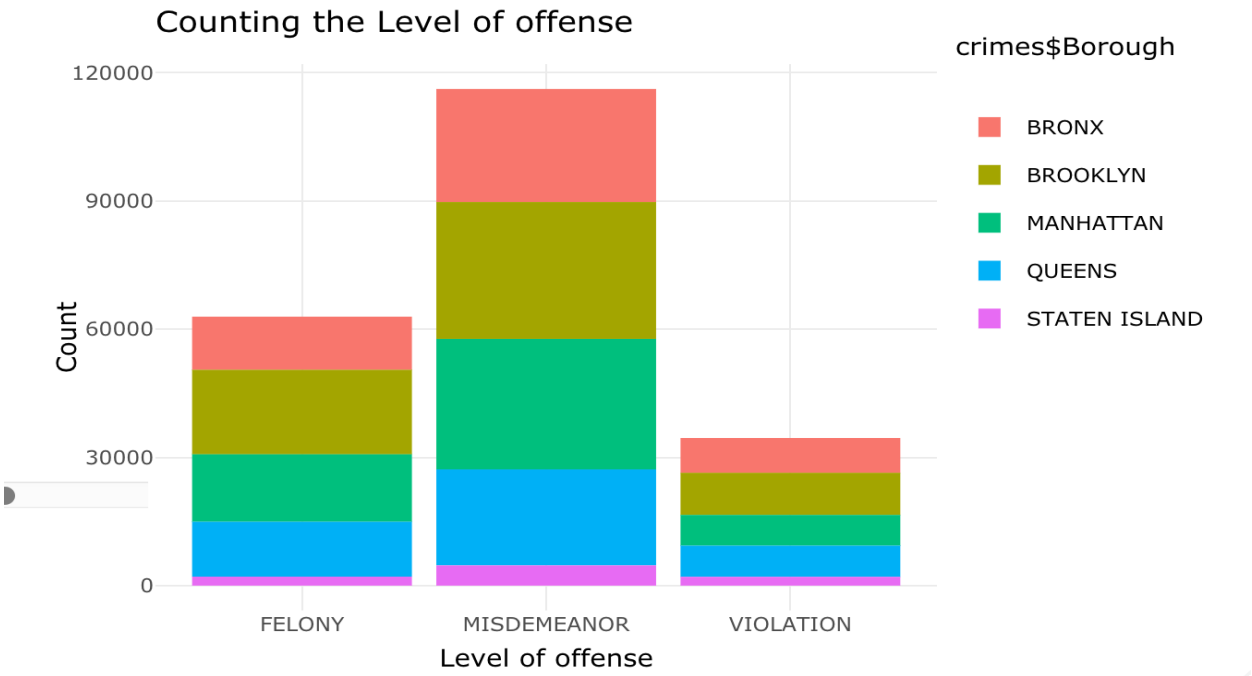



*# Level of offenses:*

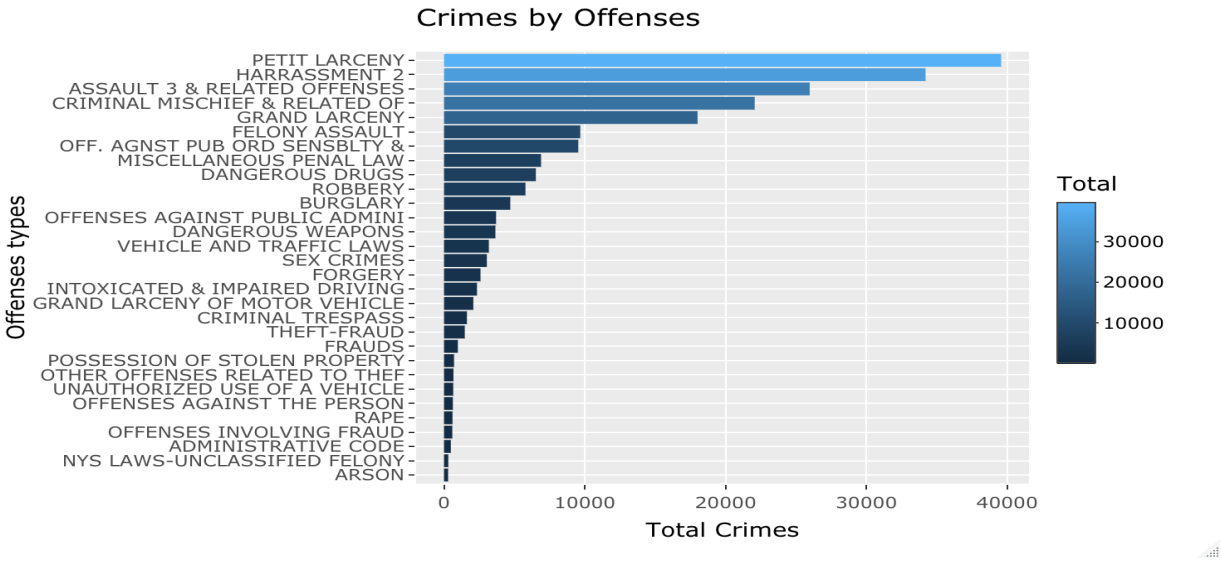
```
p <- ggplot(data = crimes, aes(x= `Level of offense`, fill = `Level of offense`))+
      geom_bar(stat = "count") + theme_minimal() +
      labs(x = "Level of offense",
           y= " Count",
           title = "Counting the Level of offense")

ggplotly(p)
```



```
# level of offense by borough
plot4 <- ggplot(data = crimes, aes(x= crimes$`Level of offense`, fill =
crimes$Borough))+
  geom_bar(stat = "count") + theme_minimal() +
  labs(x = "Level of offense",
       y= " Count",
       title = "Counting the Level of offense")

ggplotly(plot4)
```

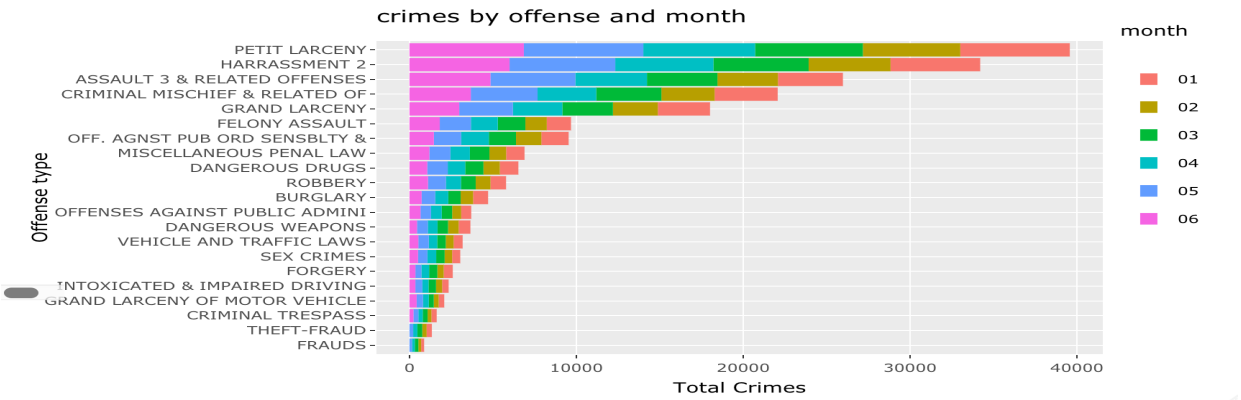



```
# The frequency of the offenses category

by_offenses <- crimes %>% group_by(Offense) %>% dplyr :: summarise(Total =
n()) %>% subset(Total > 200)
plot5 <- ggplot(by_offenses, aes(reorder(Offense, Total), Total, fill = Total)) +
geom_bar(stat = "identity") + coord_flip() +
    labs(y = "Total Crimes", x = "Offenses types", title = "Crimes by Offenses")


ggplotly(plot5)
```



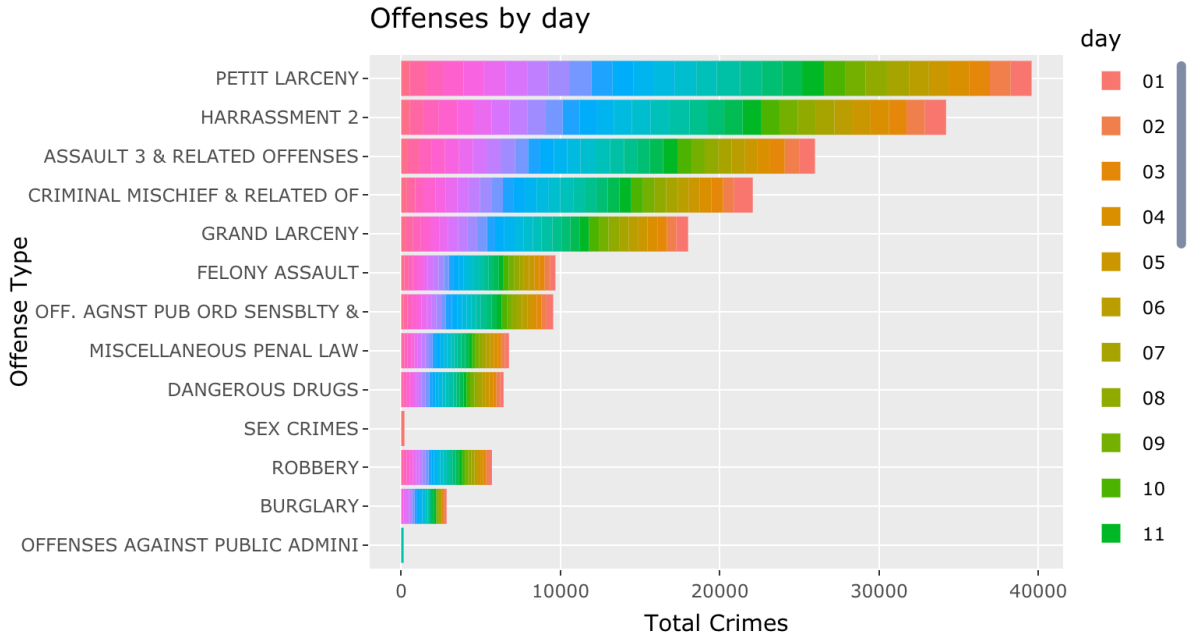

*# The offenses by month:*

```
by_offense_month <- crimes %>% group_by(Offense, month) %>% dplyr ::
summarise(Total = n()) %>% subset(Total > 150)
plot6 <- ggplot(by_offense_month, aes(reorder(Offense, Total), Total,fill =month))
+ geom_bar(stat= "identity") + coord_flip()+
        labs(y = "Total Crimes",x = "Offense type", title = " crimes by offense and
month ")

ggplotly(plot6)
```



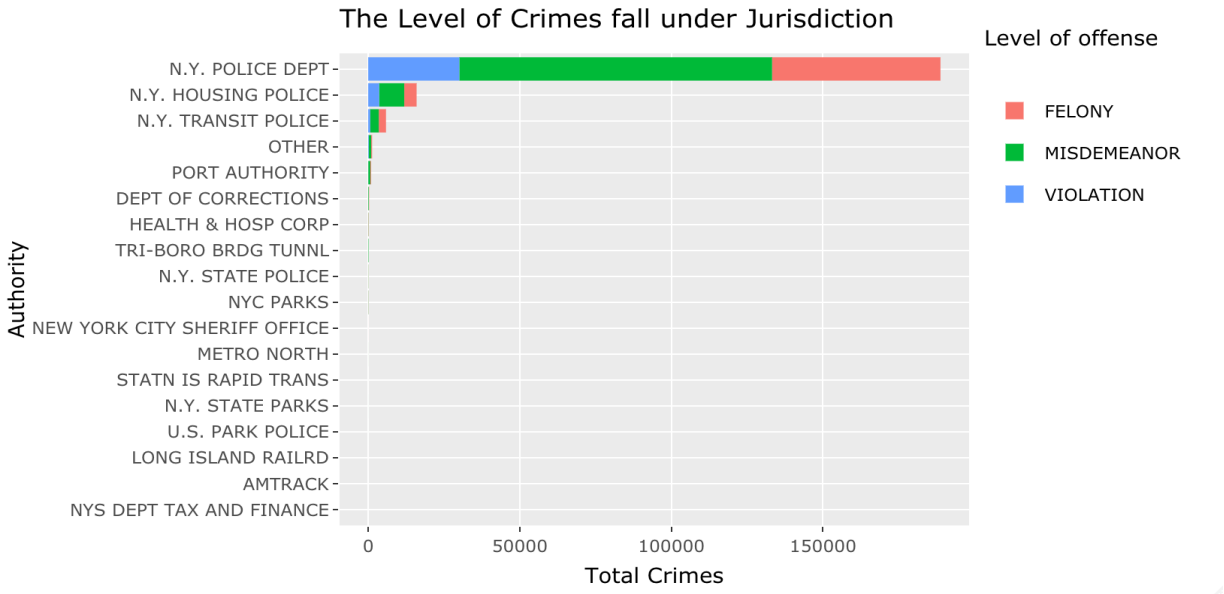
*# The offenses by day:*

```
by_offense_day <- crimes %>% group_by(Offense, day) %>% dplyr :: summarise(Total = n()) %>% subset(Total > 15)
plot7 <- ggplot(by_offense_day, aes(reorder(Offense, Total), Total, fill = day)) + geom_bar(stat="identity") + coord_flip()+
      labs(y = "Total Crimes", x = "Offense Type", title = "Offenses by day")

ggplotly(plot7)
```



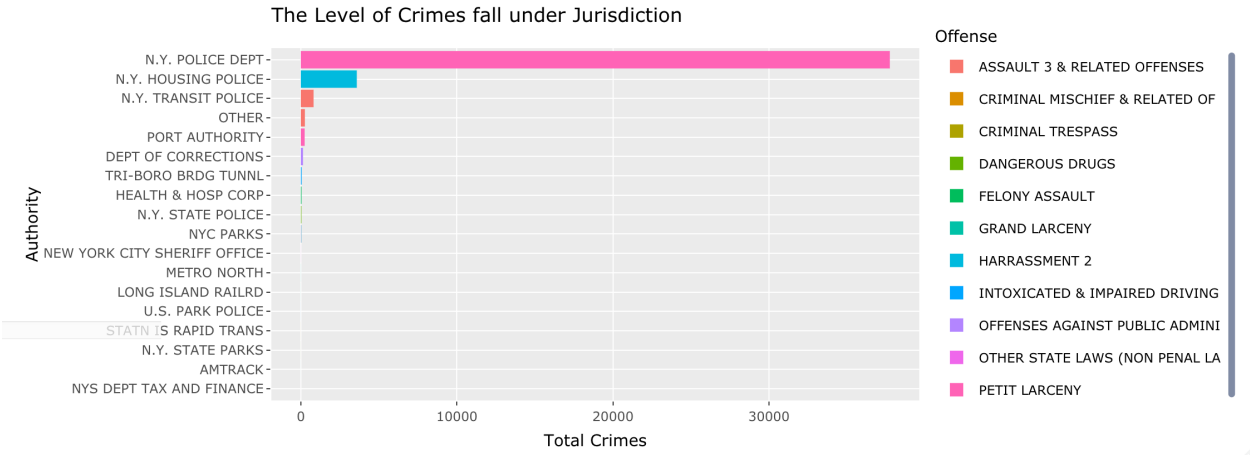

*# the top crimes in every jurisdiction:*

```
by_juri <- crimes %>% group_by(Jurisdiction, `Level of offense`) %>% dplyr::summarise(Total = n())

plot8<- ggplot(by_juri, aes(reorder(Jurisdiction,Total), Total, fill = `Level of offense`)) + geom_bar(stat= "identity") + coord_flip()+
    labs( y= "Total Crimes", x = "Authority", title= "The Level of Crimes fall under Jurisdiction")

ggplotly(plot8)
```



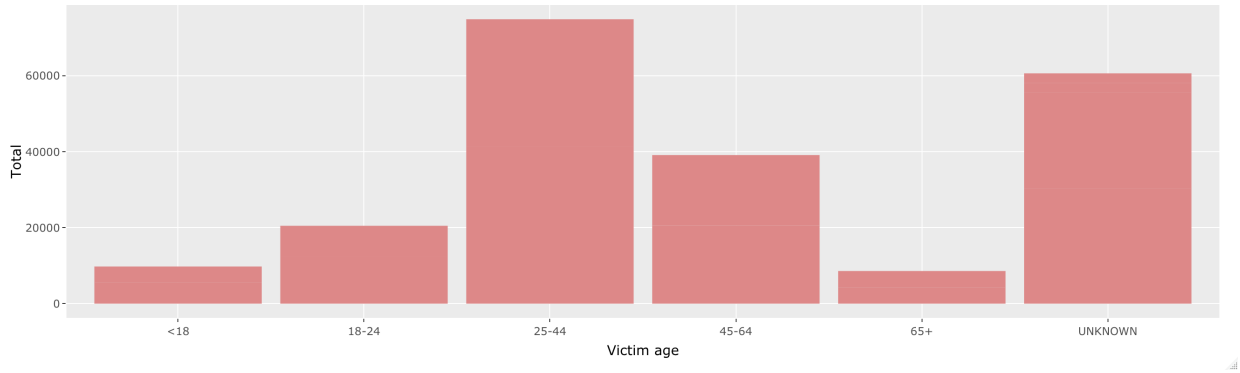

*# the top offenses in every jurisdiction:*

```
by_jur_offense <- crimes %>% group_by(Jurisdiction, Offense) %>% dplyr ::
summarise(Total = n()) %>% arrange(desc(Total)) %>% top_n(n=1)

plot9 <- ggplot(by_jur_offense, aes(reorder(Jurisdiction,Total), Total, fill =
Offense)) + geom_bar(stat= "identity") + coord_flip()+
        labs( y= "Total Crimes", x = "Authority", title= "The Level of Crimes fall
under Jurisdiction")

ggplotly(plot9)
```



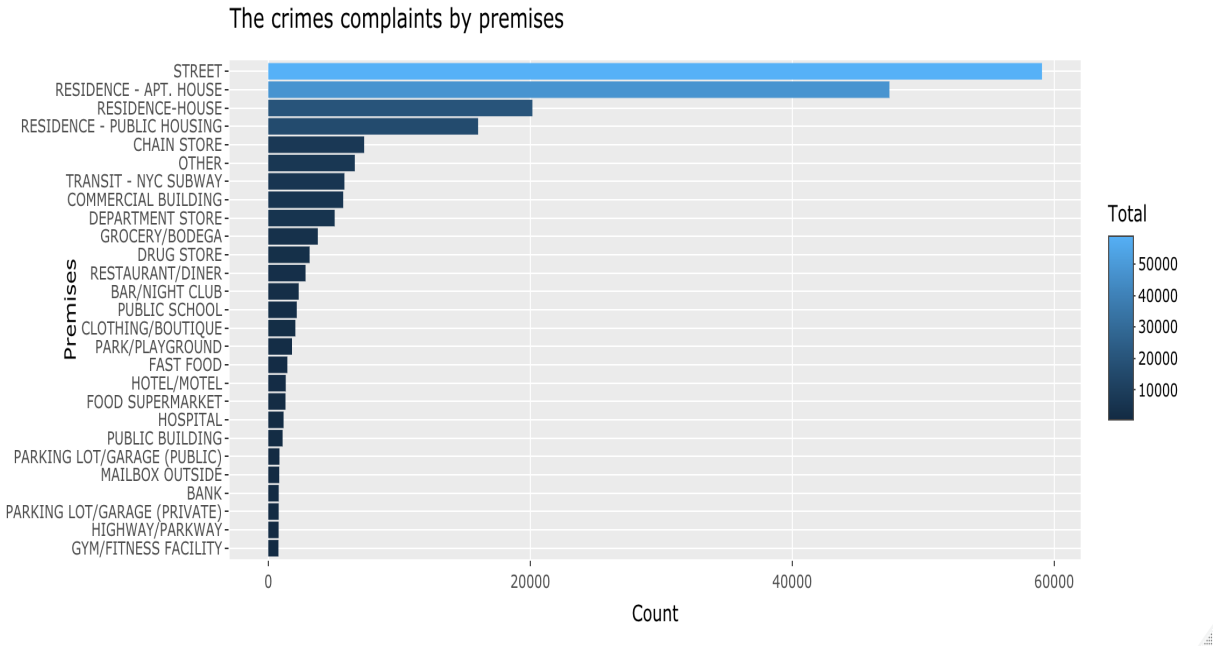
*# the victim age to be more attacked:*

```
by_victim_age <- crimes %>% group_by(`Victim age`,` Victim sex`) %>%
dplyr::summarise(Total = n()) %>% arrange(desc(Total)) %>% subset(Total>10)

plot11 <- ggplot(by_victim_age, aes(x=`Victim age`,y=Total))+
geom_bar(stat="identity", fill= "#DD8888")

ggplotly(plot11)
```

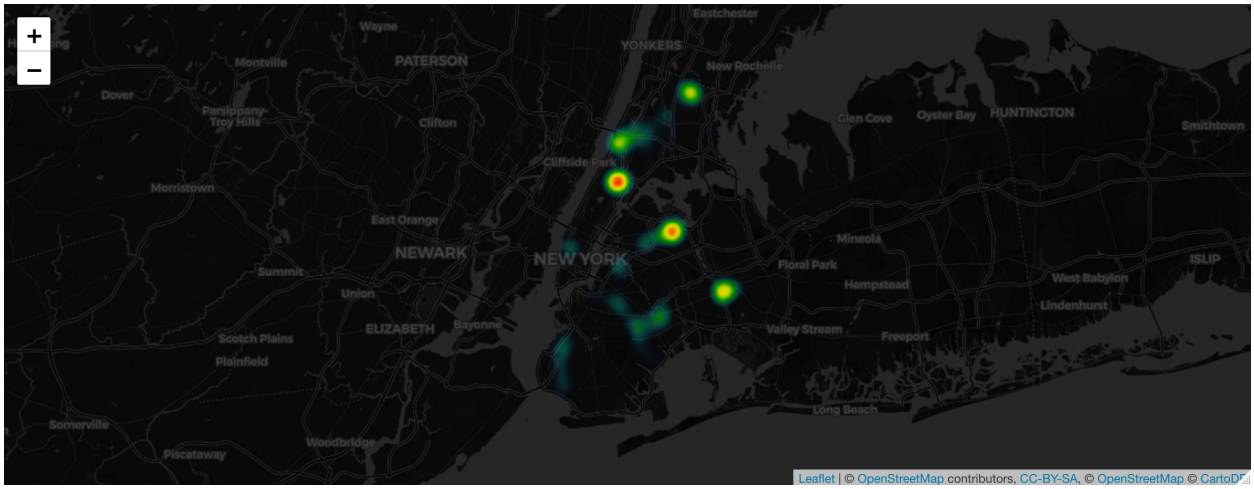



*# The crimes by premises:*

```
by_premises <- crimes %>% group_by(Premise) %>%
dplyr::summarise(Total=n()) %>% arrange(desc(Total)) %>% subset(Total >760)
plot12 <- ggplot(by_premises, aes(reorder(Premise, Total),Total,fill = Total)) +
geom_bar(stat="identity") + coord_flip() + labs( y= "Count", x = "Premises", title=
"The crimes complaints by premises")
ggplotly(plot12)
```

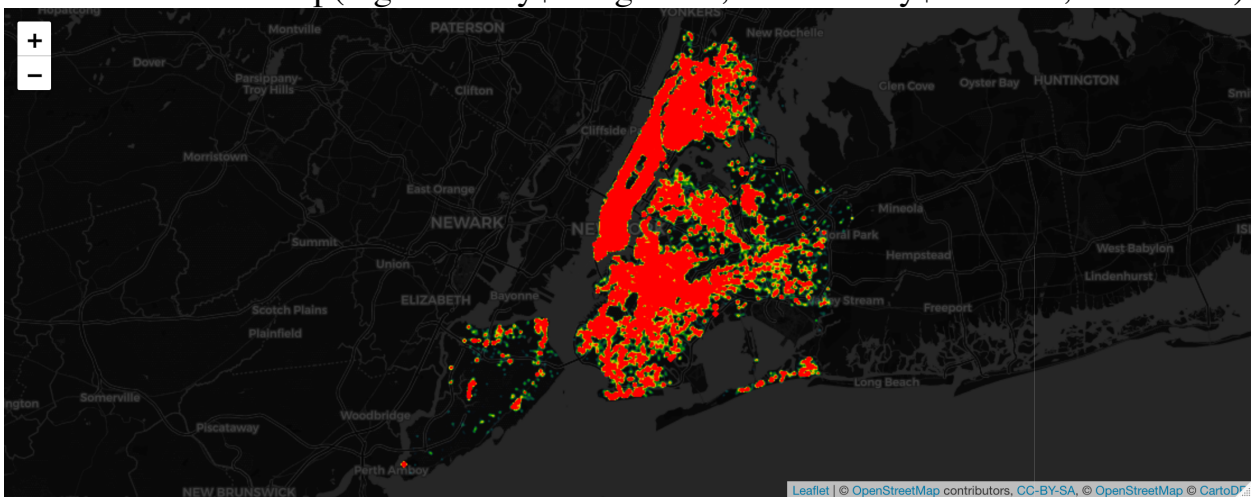



```
# Loading Leaflet for creating maps with latitude and longitude data we have:
library(leaflet)
library(leaflet.extras)

# we need to clean the data and make sure we don't have NA's:
sum(is.na(crimes))
crimes1 <- na.omit(crimes)
sum(is.na(crimes1))

# Mapping the Kidnapping dangerous place in NYC by using DarkMatter theme:
kidnapping <- crimes1[crimes1$Offense=="KIDNAPPING & RELATED
OFFENSES",]
kidnapping %>% leaflet() %>% addTiles() %>%
  addProviderTiles(providers$CartoDB.DarkMatter) %>%
  addWebGLHeatmap(lng=kidnapping$Longtitude, lat=kidnapping$Latitude, size
=5000)
```



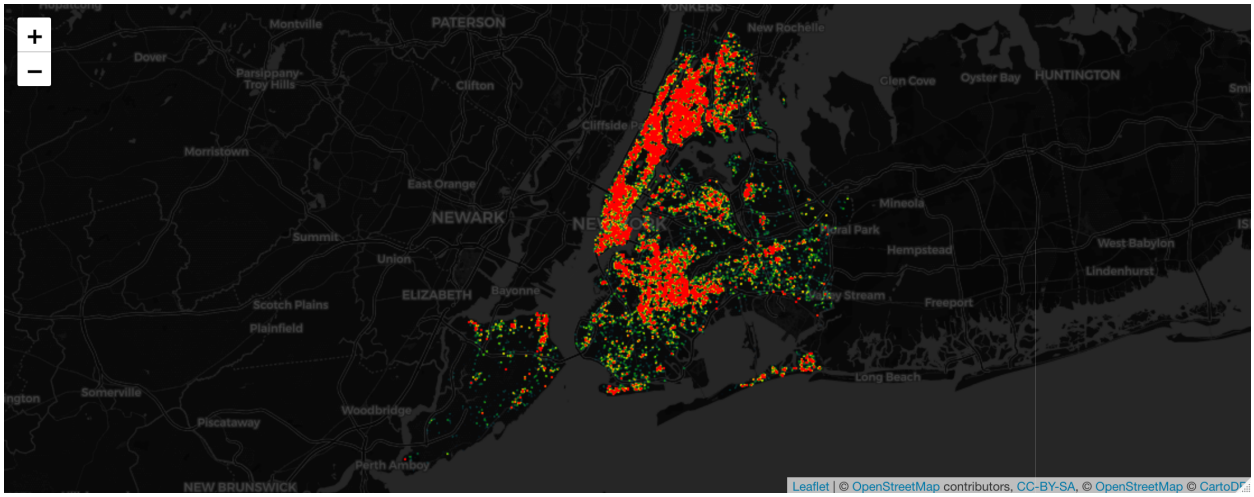

*# Mapping the PETIT LARCENY Offense in NYC:*

```
Larceny <- crimes1[crimes1$Offense == "PETIT LARCENY",]
Larceny %>%  leaflet() %>% addTiles() %>%
  addProviderTiles(providers$CartoDB.DarkMatter) %>%
  addWebGLHeatmap(lng=Larceny$Longtitude, lat=Larceny$Latitude, size = 700)
```




*# Mapping the Harrassment 2 offense in NYC:*

```
Harssment <- crimes1[crimes1$Offense == "HARRASSMENT 2",]
Harssment %>% leaflet() %>% addTiles() %>%
  addProviderTiles(providers$CartoDB.DarkMatter) %>%
  addWebGLHeatmap(lng=Harssment$Longtitude, lat=Harssment$Latitude, size =
400)
```




## Conclusion:

This is a review answers of the questions asked in introduction:

1. What parts of the city have the most crime complaints during 2019?

- Brooklyn has the highest crime complaints during the first six month of 2019

2. Has the crimes complaint improved over months and days during 2019?

- May has the most crimes rate during the first 6 month of 2019 maybe because of the summer season.
- the first day of each month is where the most crimes happened during 2019

3. what type of offenses or offenses categories to be considered high over the month and day?

- the MISDEMEANOR is the highest level of offense
- Petit larceny and harassment 2 considered to be high during may in 2019
- Petit Larceny, Harassment 2 and Assault & related offenses are the highest offenses during first day of week.

4. Which Jurisdiction, premises have the highest crime rate?

- The NYC Police Department received more crimes complaints during 2019
- Street and residence considered to have the most offenses complaints during the

year

6. Which victim gender to be more attacked by suspicious?

- The victim age group is between 25-44.

7. Which geographical area has the more crimes complaints not considered safe?

- The maps show how Larceny and Harassment are spreader more than kidnapping and related offenses in Manhattan area up to Bronx and relatively rarely in Staten island.

## III. Appendix:

```
# set the working directory
setwd("/Users/karimaidrissi/Desktop/DSSA 5101")
# loading necessary libraries:
library(tidyverse)
library(lubridate)
library(gganimate)
library(plotly)
# read the file
crimes <- read.csv("complaint.csv")
# using object size to see the size in memory.
object.size(crimes)
# glimpse the dataset:
glimpse(crimes)
# basic statistics for all columns in the dataset:
summary(crimes)
# structure of the dataset:
str(crimes)

#dimension of the data:
dim(crimes)
# the missing values in our dataset:
sum(is.na(crimes))
rowSums(is.na(crimes))
```

```
colSums(is.na(crimes))
table(is.na(crimes))
# extracting specific columns from our data:
crimes$CMPLNT_FR_DT <- as.Date(as.character(crimes$CMPLNT_FR_DT),
format = "%m/%d/%y")
crimes <- crimes %>%
  select(CMPLNT_NUM, BORO_NM, CMPLNT_FR_DT, CMPLNT_FR_TM,
JURIS_DESC, KY_CD,LAW_CAT_CD, OFNS_DESC,PREM_TYP_DESC,
SUSP_AGE_GROUP, SUSP_RACE, SUSP_SEX, VIC_AGE_GROUP,
VIC_RACE, VIC_SEX, Latitude, Longitude) %>% filter( CMPLNT_FR_DT >=
as.Date("2019-01-01"))
# renaming columns of our data:
names(crimes) <- c("ID","Borough","Date","Time","Jurisdiction","Code","Level
of offense", "Offense", "Premise" , " Suspicious age", " Suspicious race",
"Suspicious sex", "Victim age", " Victim race", " Victim sex","Latitude",
"Longtitude")


# need to separte the date column into day, month, year:
crimes <- separate(crimes, col = Date, into = c("year","month","day"), sep ="-")
# type of month, day and year:
class(crimes$month)
class(crimes$day)
class(crimes$year)
# change the month,day and year into factors:
crimes$month <- as.factor(crimes$month)
crimes$day <- as.factor(crimes$day)
crimes$year <- as.factor(crimes$year)
# checking the type of month, day and year:
typeof(crimes$month)
typeof(crimes$day)
typeof(crimes$year)

# crimes by Month:
by_month <- crimes %>% group_by(month) %>% dplyr::summarise(Total = n())
```

```
plot1 <- ggplot(subset(by_month), aes(x= month, y= Total, fill = month)) +
geom_bar(stat="identity")

ggplotly(plot1)
# by plotting the graph it's shows that May has the highest crimes during 2019
# crime by day:
by_day <- crimes %>% group_by(day) %>% dplyr::summarise(Total = n())
plot2 <- ggplot(by_day, aes(x=day, y=Total, fill =day)) +
geom_bar(stat="identity")


ggplotly(plot2)
# noticed that the first day of each month has the most frequency of crimes during
2019

#Which burough is the most dangerous one:
library(tidyverse)
crimes %>% drop_na("Borough")
crimes[!is.na(crimes$Borough), ]
is.na(crimes$Borough)
plot3 <- ggplot(data = crimes, aes(x= crimes$Borough,fill = Borough))+
          geom_bar(colour = "black",stat = "count")+
          labs(x = "Borough Names", y="Count",
               title = "crimes by borough")

ggplotly(plot3)
# by looking at the graph I have noticed that BROOKLYN has the most crime rate
#5 Which offense is the highest one:
p <- ggplot(data = crimes, aes(x= `Level of offense`, fill = `Level of offense`))+
     geom_bar(stat = "count") + theme_minimal() +
     labs(x = "Level of offense",
          y= " Count",
          title = "Counting the Level of offense")

ggplotly(p)
```

```
# the MISDEMEANOR is the most highest level of offense

# level of offense by borough
plot4 <- ggplot(data = crimes, aes(x= crimes$`Level of offense` , fill =
crimes$Borough))+
  geom_bar(stat = "count") + theme_minimal() +
  labs(x = "Level of offense",
       y= " Count",
       title = "Counting the Level of offense")

ggplotly(plot4)


# The frequency of the offenses category

by_offenses <- crimes %>% group_by(Offense) %>% dplyr :: summarise(Total =
n()) %>% subset(Total > 200)
plot5 <- ggplot(by_offenses, aes(reorder(Offense, Total), Total, fill = Total)) +
geom_bar(stat = "identity") + coord_flip() +
    labs(y = "Total Crimes", x = "Offenses types", title = "Crimes by Offenses")


ggplotly(plot5)

# the offense by month:

by_offense_month <- crimes %>% group_by(Offense, month) %>% dplyr ::
summarise(Total = n()) %>% subset(Total > 150)
plot6 <- ggplot(by_offense_month, aes(reorder(Offense, Total), Total,fill =month))
+ geom_bar(stat= "identity") + coord_flip()+
        labs(y = "Total Crimes",x = "Offense type", title = " crimes by offense and
month ")

ggplotly(plot6)
```

## petit larceny and harrasement considred one of the highest offenses crimes during may, april and march

```
by_offense_day <- crimes %>% group_by(Offense, day) %>% dplyr :: summarise(Total = n()) %>% subset(Total > 150)
plot7 <- ggplot(by_offense_day, aes(reorder(Offense, Total), Total, fill = day)) + geom_bar(stat="identity") + coord_flip()+
        labs(y = "Total Crimes", x = "Offense Type", title = "Offenses by day ")

ggplotly(plot7)
```

## Petit Larceny,Harrassment 2, Assualt & related offenses are the highest offenses during first day of week.

```
# the top crimes in every jurisdiction:

by_juri <- crimes %>% group_by(Jurisdiction, `Level of offense`) %>% dplyr::summarise(Total = n())

plot8<- ggplot(by_juri, aes(reorder(Jurisdiction,Total), Total, fill = `Level of offense`)) + geom_bar(stat= "identity") + coord_flip()+
     labs( y= "Total Crimes", x = "Authority", title= "The Level of Crimes fall under Jurisdiction")

ggplotly(plot8)
```

## by looking we can see clearly that NYC Police Department has the most crime during 2019.

```
# the top offenses in every jurisdiction:
by_jur_offense <- crimes %>% group_by(Jurisdiction, Offense) %>% dplyr :: summarise(Total = n()) %>% arrange(desc(Total)) %>% top_n(n=1)

plot9 <- ggplot(by_jur_offense, aes(reorder(Jurisdiction,Total), Total, fill = Offense)) + geom_bar(stat= "identity") + coord_flip()+
        labs( y= "Total Crimes", x = "Authority", title= "The Level of Crimes fall under Jurisdiction")
```

```
ggplotly(plot9)
## the Petit larceny is the highest offense in NY Police then the harrassment2 in
NY housing police departement.

# the victim age to be more attacked :

by_victim_age <- crimes %>% group_by(`Victim age`, ` Victim sex`) %>%
dplyr::summarise(Total = n()) %>% arrange(desc(Total)) %>% subset(Total>10)

plot11 <- ggplot(by_victim_age, aes(x=`Victim age`,y=Total))+
geom_bar(stat="identity", fill= "#DD8888")

ggplotly(plot11)

# the victim age is between 25-44

# The crimes by premises

by_premises <- crimes %>% group_by(Premise) %>%
dplyr::summarise(Total=n()) %>% arrange(desc(Total)) %>% subset(Total >760)
plot12 <- ggplot(by_premises, aes(reorder(Premise, Total),Total,fill = Total)) +
geom_bar(stat="identity") + coord_flip() + labs( y= "Count", x = "Premises", title=
"The crimes complaints by premises")
ggplotly(plot12)
## Street and residence considered to have the most level of offense during the
year.

# Loading Leaflet for creating maps with latitude and longtitude data we have:
library(leaflet)
library(leaflet.extras)

# we need to clean the data and make sure we don't have NA's:
sum(is.na(crimes))
crimes1 <- na.omit(crimes)
sum(is.na(crimes1))
```

```
# Mapping the Kidnapping dangerous place in NYC by using DarkMatter theme:
kidnapping <- crimes1[crimes1$Offense=="KIDNAPPING & RELATED
OFFENSES",]
kidnapping %>% leaflet() %>% addTiles() %>%
  addProviderTiles(providers$CartoDB.DarkMatter) %>%
  addWebGLHeatmap(lng=kidnapping$Longtitude, lat=kidnapping$Latitude, size
=5000)

# Mapping the PETIT LARCENY Offense in NYC:
Larceny <- crimes1[crimes1$Offense == "PETIT LARCENY",]
Larceny %>%  leaflet() %>% addTiles() %>%
  addProviderTiles(providers$CartoDB.DarkMatter) %>%
  addWebGLHeatmap(lng=Larceny$Longtitude, lat=Larceny$Latitude, size = 700)

# Mapping the Harrassment 2 offense in NYC:
Harssment <- crimes1[crimes1$Offense == "HARRASSMENT 2",]
Harssment %>% leaflet() %>% addTiles() %>%
  addProviderTiles(providers$CartoDB.DarkMatter) %>%
  addWebGLHeatmap(lng=Harssment$Longtitude, lat=Harssment$Latitude, size =
400)

# By looking on the maps we can see clearly how Larceny and Harassment crime
complaints are extended more than kidnapping and related offenses in Manhattan
area up to Bronx and relatively rarely in Staten island.
```