EDA of Crimes in Toronto - Phase 1 Capstone Project
Submission by Vidyasankar Sundar

Key Findings from Exploratory Data analysis

Below are the key insights gleaned from the exploratory analysis conducted on the toronto open dataset containing the crimes that occurred between 2014 and 2019.The supporting R code and graphs are presented following the findings

1.In Toronto, the average number of crime incidents is 31179 per year, 2647 per month, and 86 per day. Overall, crime has seen an increasing trend since 2014.

2.Crime occurs the most during summer and fall (May-October). The first and the last week of any month is seen to have the most incidents. Notably, the peak observed is on the first day of every month. During weekdays, peak crime is observed at noon and then starts to gradually build from 3PM onwards and steadily increases into the night. During weekends, peak is observed at midnight. In addition, there are more crimes on Fridays and weekends than any other day in the week.

3.Most number of crimes happen outside followed by apartments and commercial establishments.

4.Assault is the most prevalent form of crime in Toronto followed by home/commercial break and enter and auto theft.

5.All crime types have seen an increasing trend since 2014 with the exception of robbery

6.Most assault incidents happen between May and August,auto theft between July and November while other types are fairly consistent across all months. Maximum number of Assault incidents occur usually on weekends, while other types increasingly occur on Fridays. Assault, Break and Enter, Theftover peaks are observed at both midnight and noon, auto theft at 10PM, and robbery tops at 9PM.

7.Auto theft happens mostly outside and at houses while other crime types are common across all premise types

8.Top offences within the assault category include - assault with weapon, bodily harm, and assault peace officer while top offences within robbery include - mugging, other, robbery with weapons, and robbery-business.

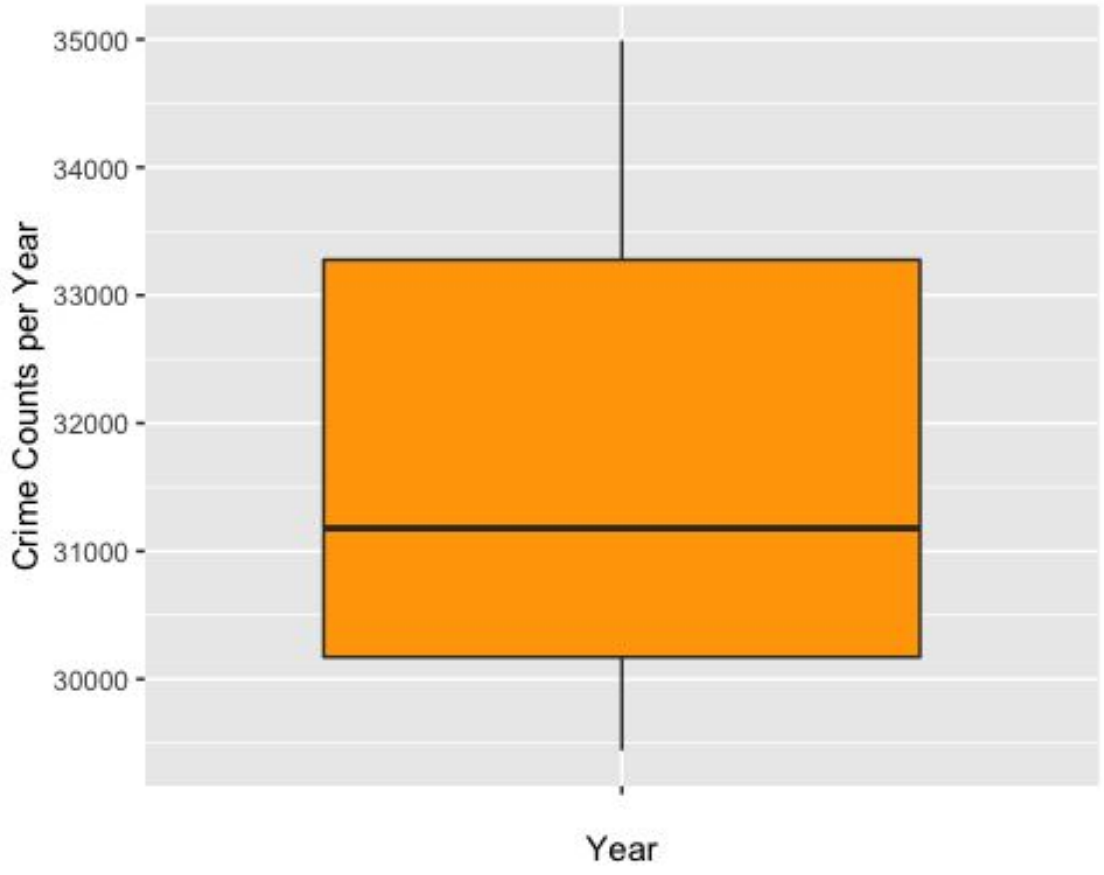
9.The most dangerous neighbourhood is the Waterfront. The other two most dangerous hoods are the Bay Street Corridor and Yonge-Church Corridor. The most safest hoods are Lambton Baby Point, Woodbine-Lumsden, and Guildwood

10.Besides assaults, Bay Street Corridor, Church-Yonge Corridor and Waterfront had the most break and enter crimes while West Humber-Clairville had the most vehicle stolen crimes. Assaults, and Break and Enter occur all over the city, with a concentration in the Waterfront areas. Other crimes, such as Auto Theft, have more points on the west side than the east side. Robbery and Theft Over are primarily in the Waterfront areas.

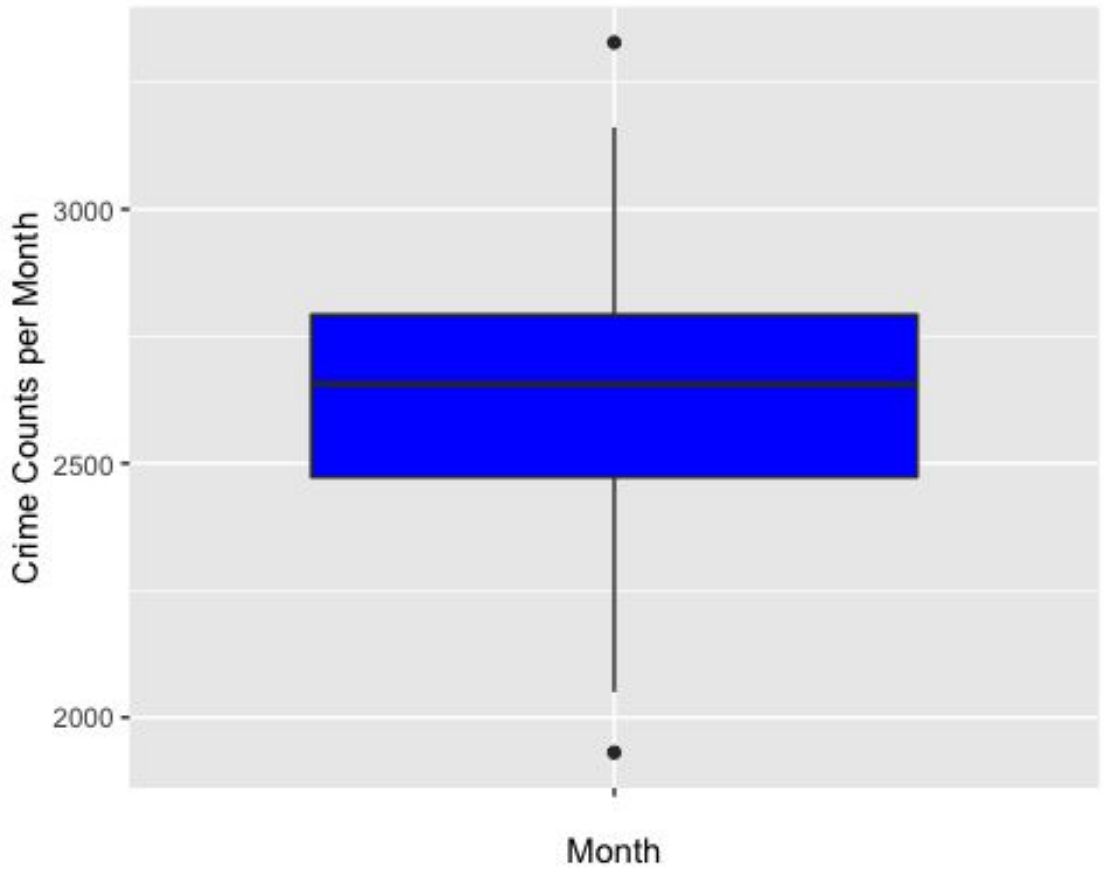
1.Box Plot Crime Distribution

In Toronto, the average number of crime incidents is 31179 per year, 2647 per month, and 86 per day.

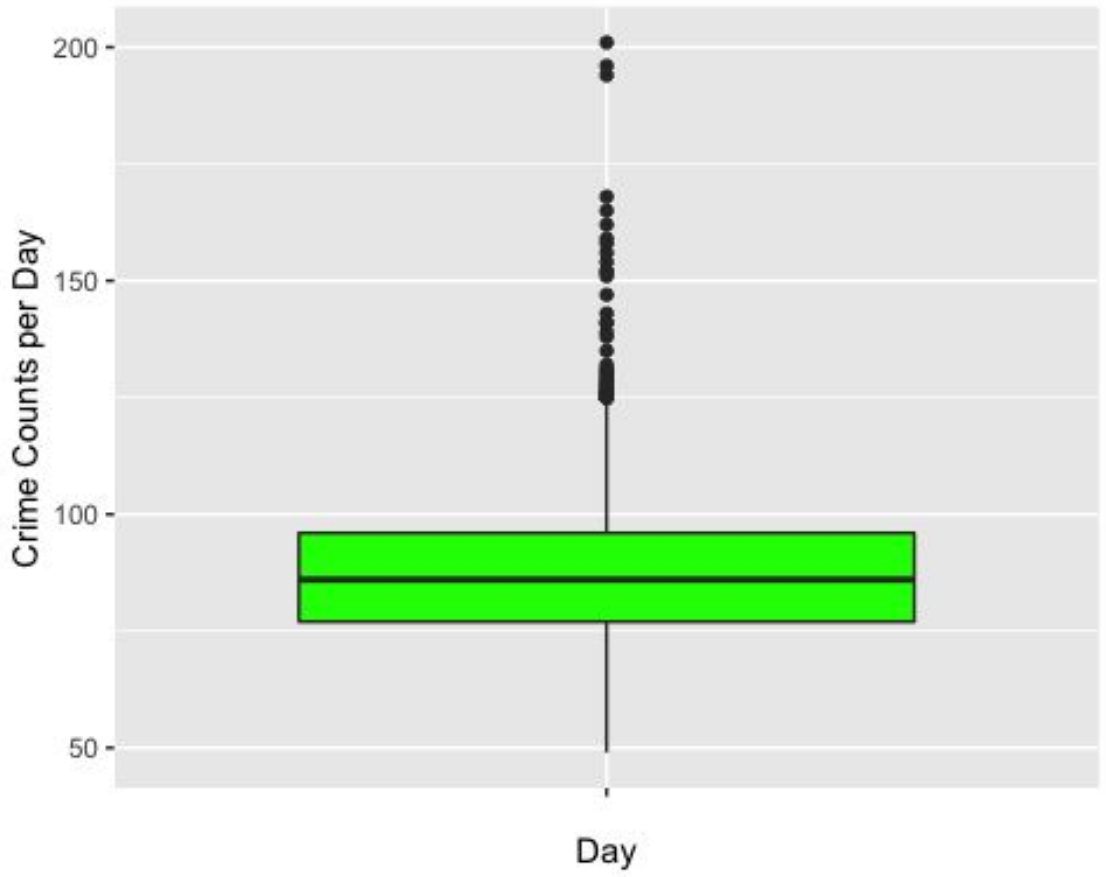
```
by_date <- EDAfilter %>% group_by(occurrenceyear) %>% dplyr::summarise(Total
= n())
ggplot(by_date, aes(x = "",Total, fill = Total)) +
geom_boxplot(fill="orange") + xlab("Year") + ylab("Crime Counts per Year")
```




```
by_month <- EDAfilter %>% group_by(occurrencemonth,occurrenceyear) %>%
dplyr::summarise(Total = n())
ggplot(by_month, aes(x = "",Total, fill = Total)) + geom_boxplot(fill="blue")
+ xlab("Month") + ylab("Crime Counts per Month")
```



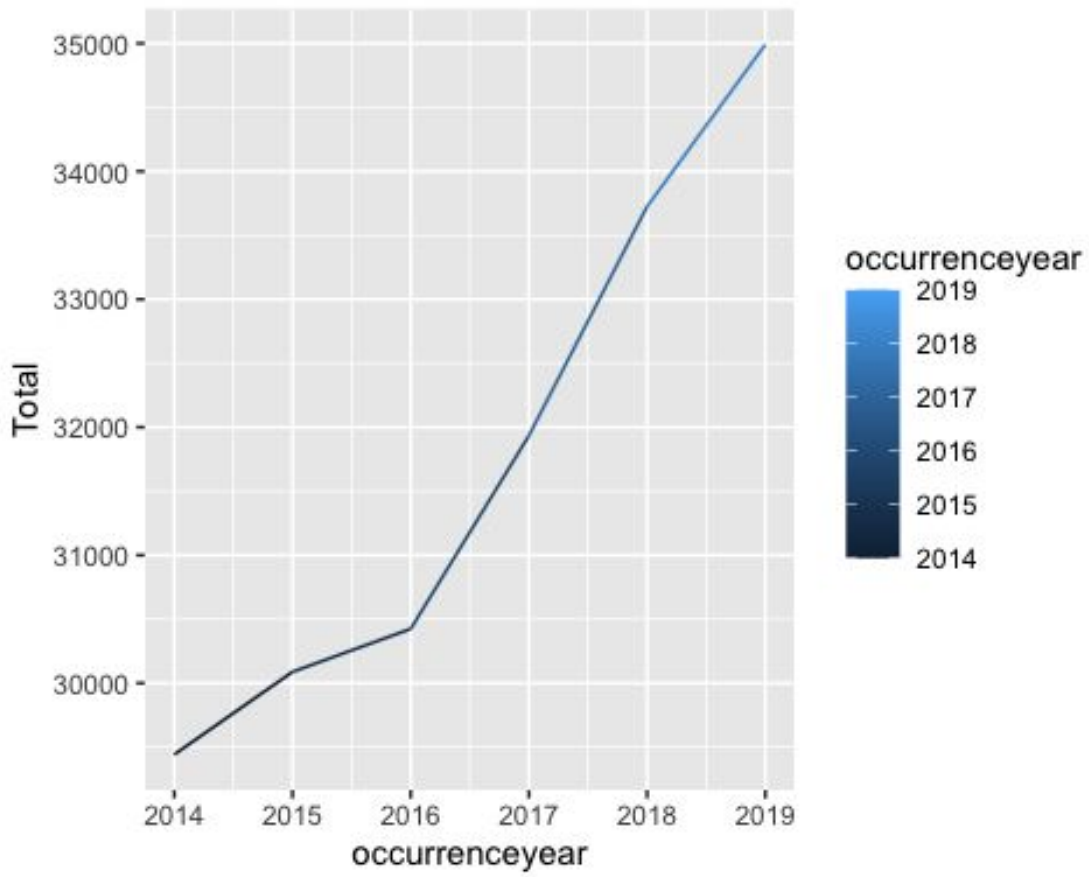

```
by_day <- EDAfilter %>% group_by(occurrenceyear,occurrencedayofyear) %>%
dplyr::summarise(Total = n())
ggplot(by_day, aes(x = "",Total, fill = Total)) + geom_boxplot(fill="green")
+ xlab("Day") + ylab("Crime Counts per Day")
```



2.Crime Trends by Year

Overall, crime has seen an increasing trend since 2014.

```
ggplot(by_date, aes(occurrenceyear, Total, color = occurrenceyear)) +
geom_line()
```

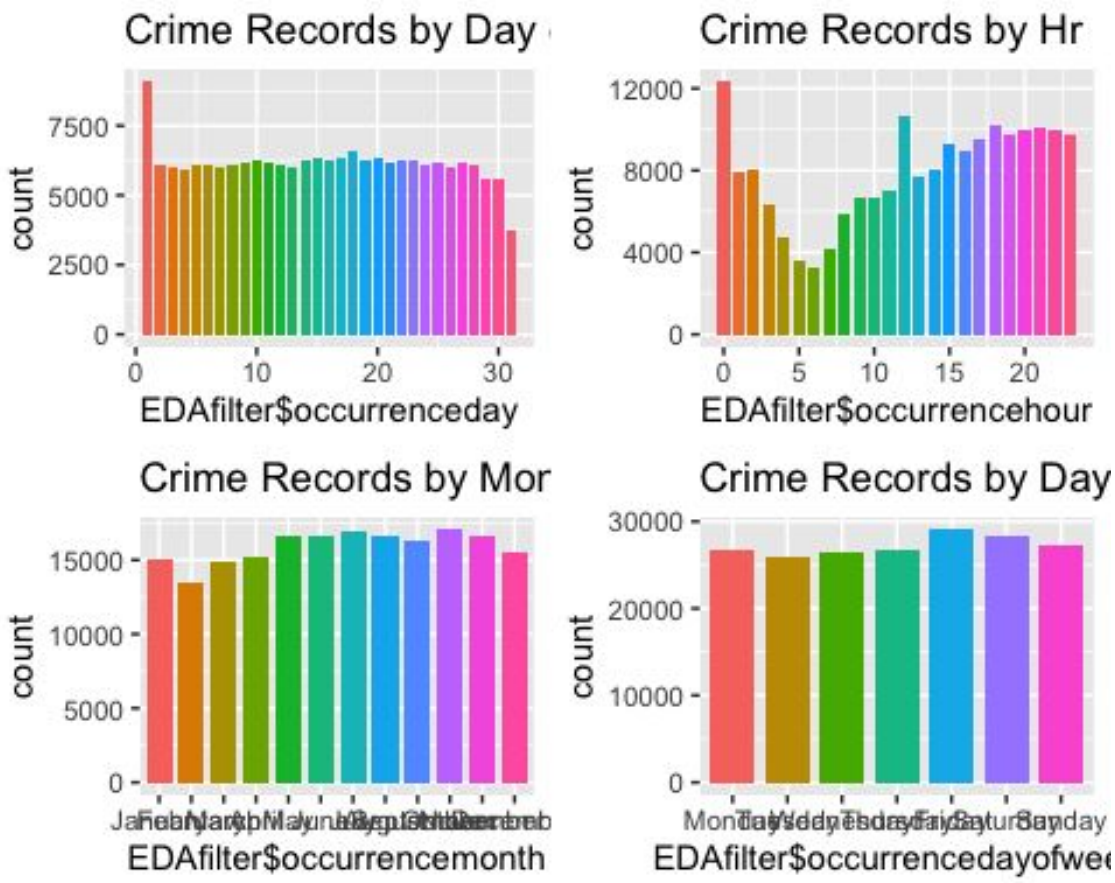


3.Understand crime trends by month, day, hour and day of the week and plot them side by side

Crime incidents are most during summer and fall (May-October), with frequencies peaking on the first of every month.Crimes are also most around Fridays and weekends.

```
mon.bp <- ggplot(EDAfilter, aes(x = EDAfilter$occurrencemonth,
fill=as.factor(EDAfilter$occurrencemonth))) +
                    geom_bar(width=0.8, stat="count") +
theme(legend.position="none") +
                ggtitle("Crime Records by Month of Year")
sdom.bp <- ggplot(EDAfilter, aes(x = EDAfilter$occurrenceday,
fill=as.factor(EDAfilter$occurrenceday))) +
  geom_bar(width=0.8, stat="count") + theme(legend.position="none") +
  ggtitle("Crime Records by Day of Month")
shour.bp <- ggplot(EDAfilter, aes(x = EDAfilter$occurrencehour,
fill=as.factor(EDAfilter$occurrencehour))) +
  geom_bar(width=0.8, stat="count") + theme(legend.position="none") +
  ggtitle("Crime Records by Hr")
```

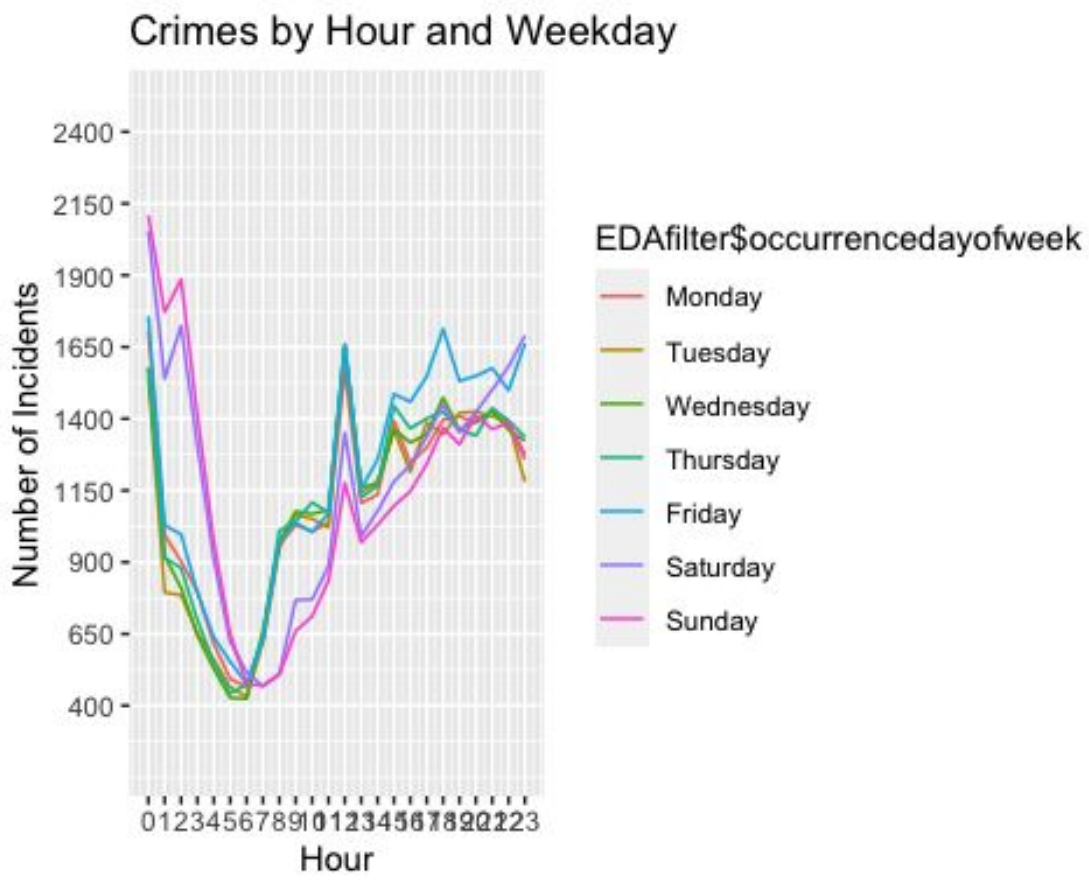
```
dow.bp <- ggplot(EDAfilter, aes(x = EDAfilter$occurrencedayofweek,
fill=as.factor(EDAfilter$occurrencedayofweek))) +
  geom_bar(width=0.8, stat="count") + theme(legend.position="none") +
  ggtitle("Crime Records by Day of Week")
grid.arrange(sdom.bp, shour.bp, mon.bp, dow.bp)
```



4.Crimes by Hour and Weekday, Trends

Crime peaks at noon, tapers off and gradually picks up steam to continously progress through evening and late into night.

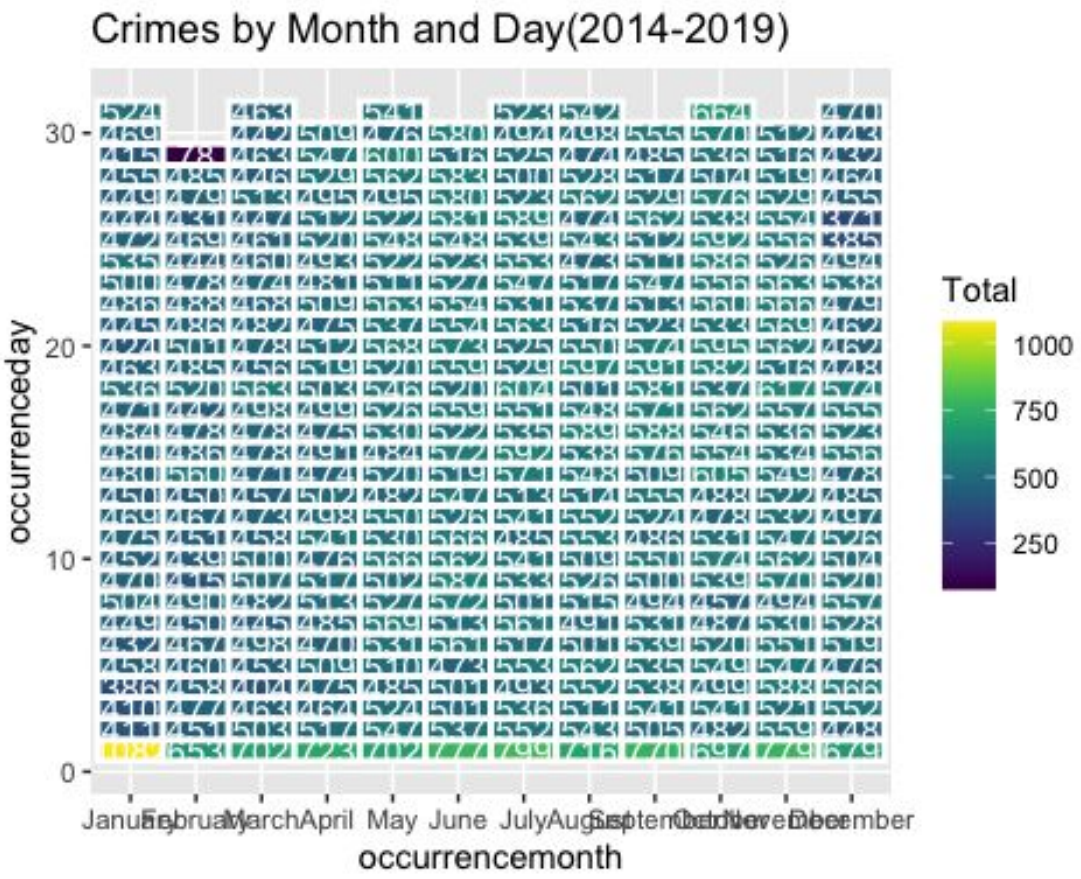
```
ggplot(EDAfilter)+
  aes(x=EDAfilter$occurrencehour, colour=EDAfilter$occurrencedayofweek)+
  geom_line(stat="count")+
  scale_x_continuous(breaks = seq(0,23,1),limit=c(0,23))+
  scale_y_continuous(breaks = seq(400,3000,250),limit=c(200,2500))+
  labs(title="Crimes by Hour and Weekday",x="Hour",y="Number of Incidents")
```



5.Heatmap of Crimes by Month and Day

The first and the last week of any month is seen to have the most incidents. The peak observed is on the first day of every month.

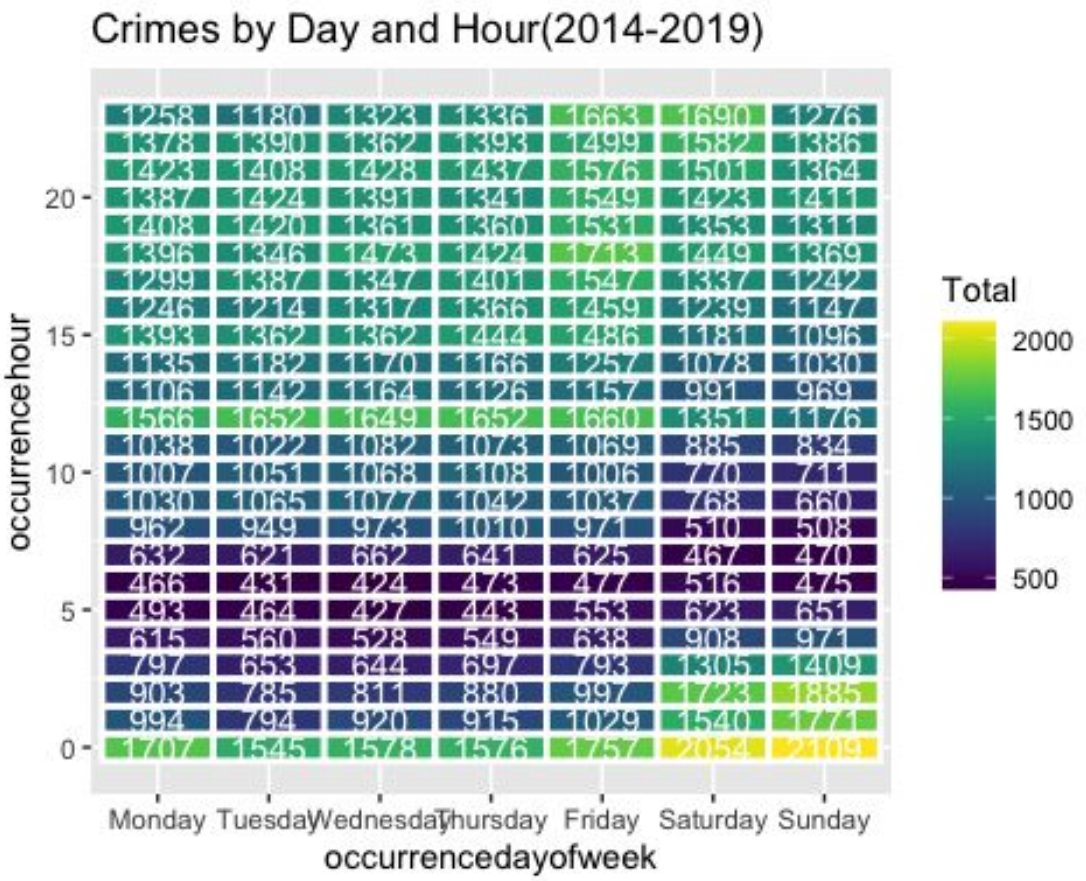
```
crime_heatmap <- EDAfilter %>% group_by(occurrencemonth, occurrenceday) %>%
dplyr::summarise(Total = n())
ggplot(crime_heatmap, aes(occurrencemonth, occurrenceday, fill = Total)) +
  geom_tile(size = 1, color = "white") +
  scale_fill_viridis()  +
  geom_text(aes(label=Total), color='white') +
  ggtitle("Crimes by Month and Day(2014-2019)")
```



6.Heatmap of Crimes by Day and Hour

During weekdays, peak is observed at noon and then starts to gradually build from 3PM. During weekends, peak is observed at midnight.

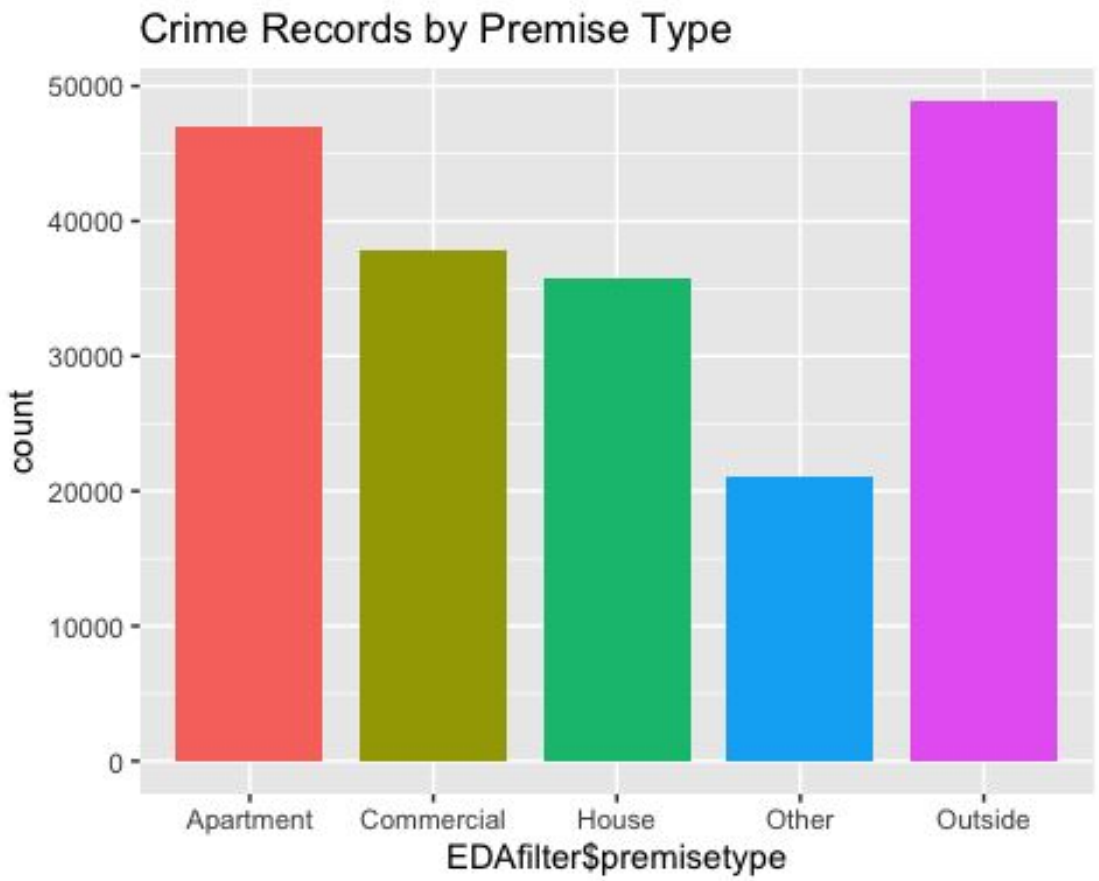
```
hour_heatmap <- EDAfilter %>% group_by(occurrencedayofweek, occurrencehour)
%>% dplyr::summarise(Total = n())
ggplot(hour_heatmap, aes(occurrencedayofweek, occurrencehour, fill = Total))
+
  geom_tile(size = 1, color = "white") +
  scale_fill_viridis()  +
  geom_text(aes(label=Total), color='white') +
  ggtitle("Crimes by Day and Hour(2014-2019)")
```



7.Crimes by Premise Type

Most number of crimes happen outside followed by apartments and commercial establishments.

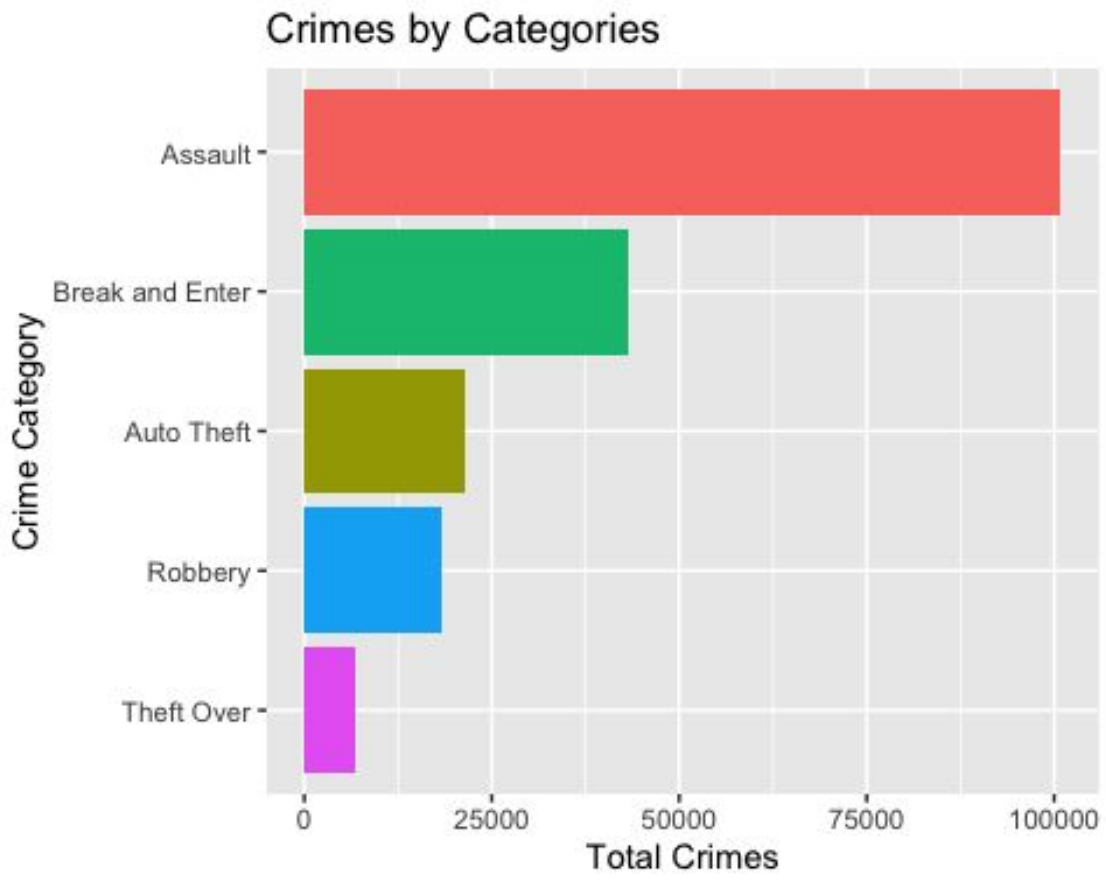
```
ggplot(EDAfilter, aes(x = EDAfilter$premisetype,
fill=as.factor(EDAfilter$premisetype))) +
  geom_bar(width=0.8, stat="count") + theme(legend.position="none") +
  ggtitle("Crime Records by Premise Type")
```



8.Major Crime Incident Categories

Assault is the most prevalent form of crime in Toronto followed by home/commercial break and enter and auto theft.

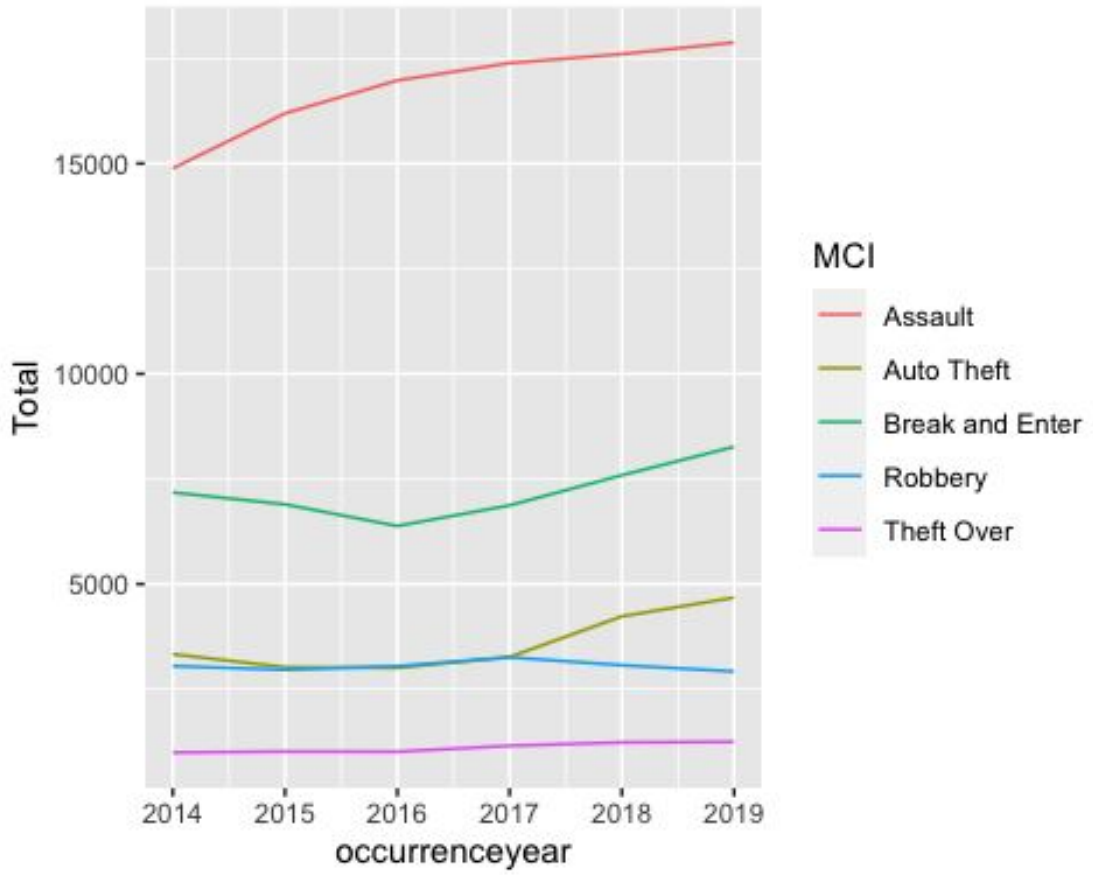
```
mci<-EDAfilter %>% group_by(MCI) %>% dplyr::summarise(Total = n())
ggplot(mci, aes(reorder(MCI, Total), Total, fill = MCI)) +
  geom_bar(stat = "identity") +
  coord_flip() + ggtitle("Crimes by Categories") +
theme(legend.position="none") +
  xlab("Crime Category") +
  ylab("Total Crimes")
```



9.MCI Trends by Year

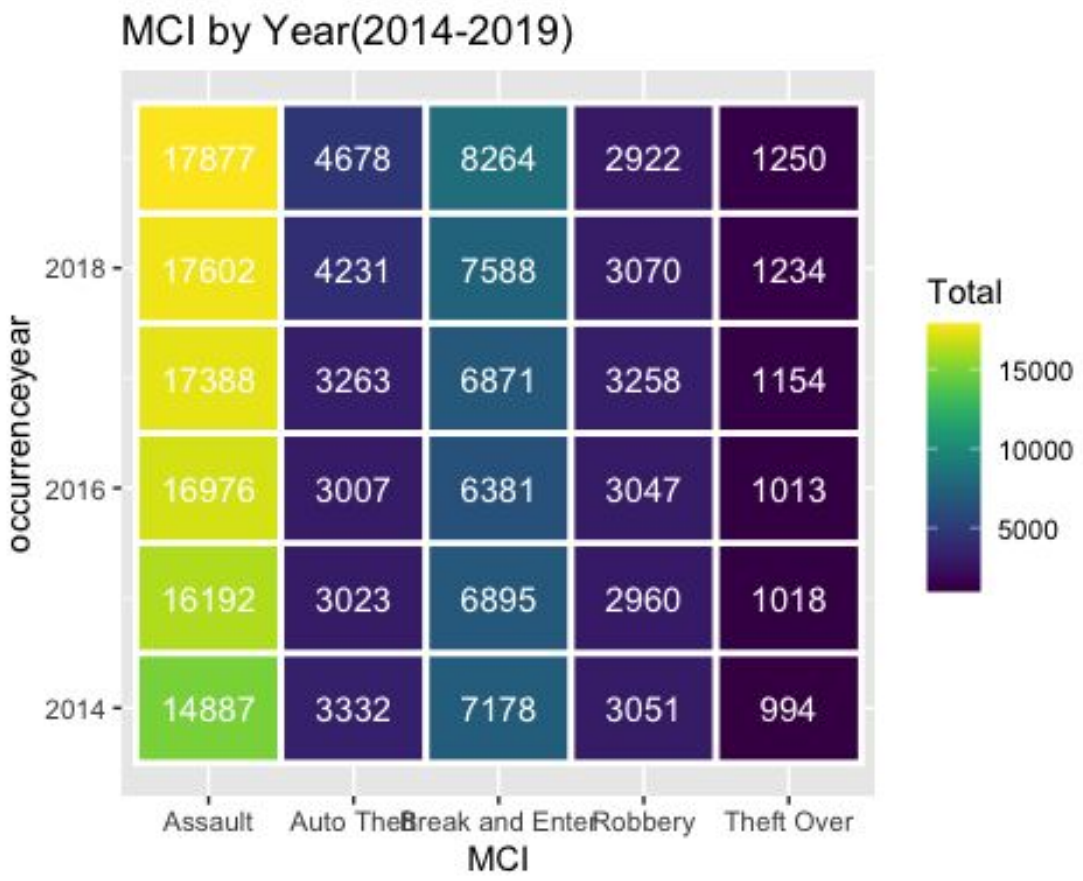
All crime types have seen an increasing trend since 2014 with the exception of robbery

```
mciyear<-EDAfilter %>% group_by(MCI,occurrenceyear) %>%
dplyr::summarise(Total = n())
ggplot(mciyear, aes(occurrenceyear, Total, color = MCI)) + geom_line()
```



10.MCI Heatmap by Year

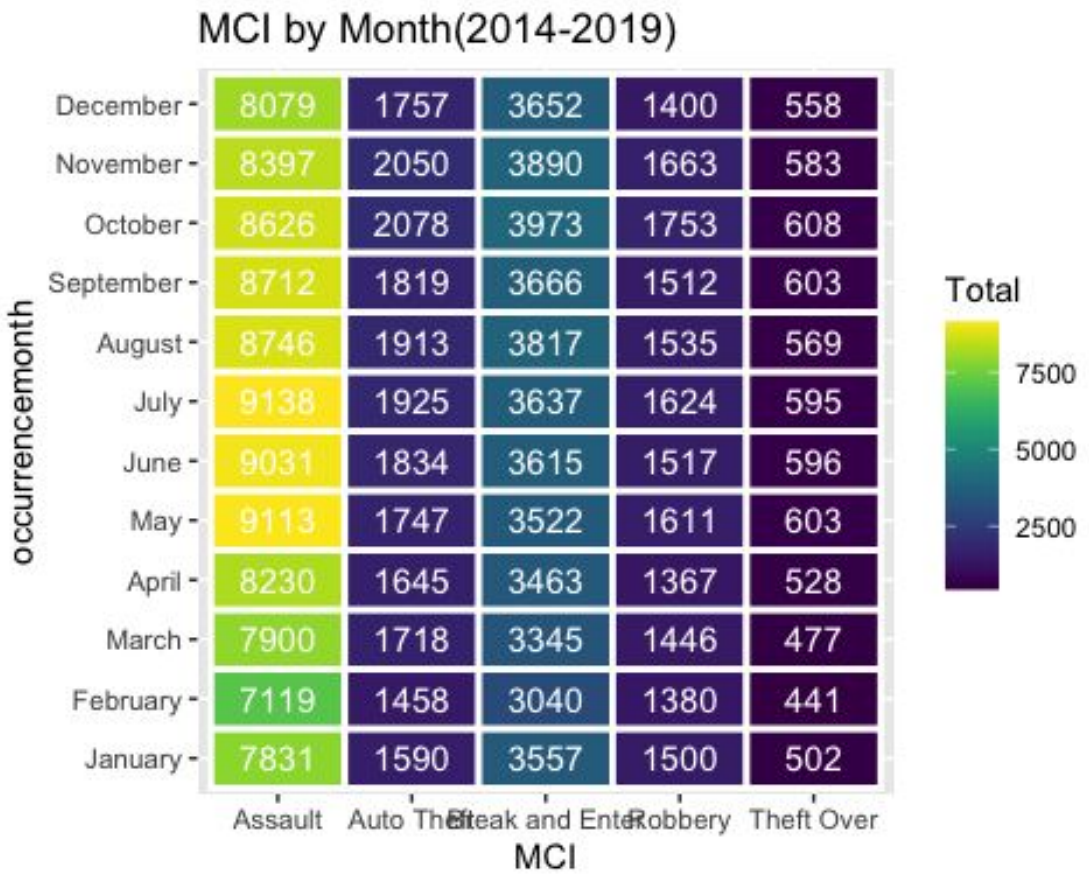
```
ggplot(mciyear, aes(MCI, occurrenceyear, fill = Total)) +
  geom_tile(size = 1, color = "white") +
  scale_fill_viridis()  +
  geom_text(aes(label=Total), color='white') +
  ggtitle("MCI by Year(2014-2019)")
```



11.MCI Heatmap by Month

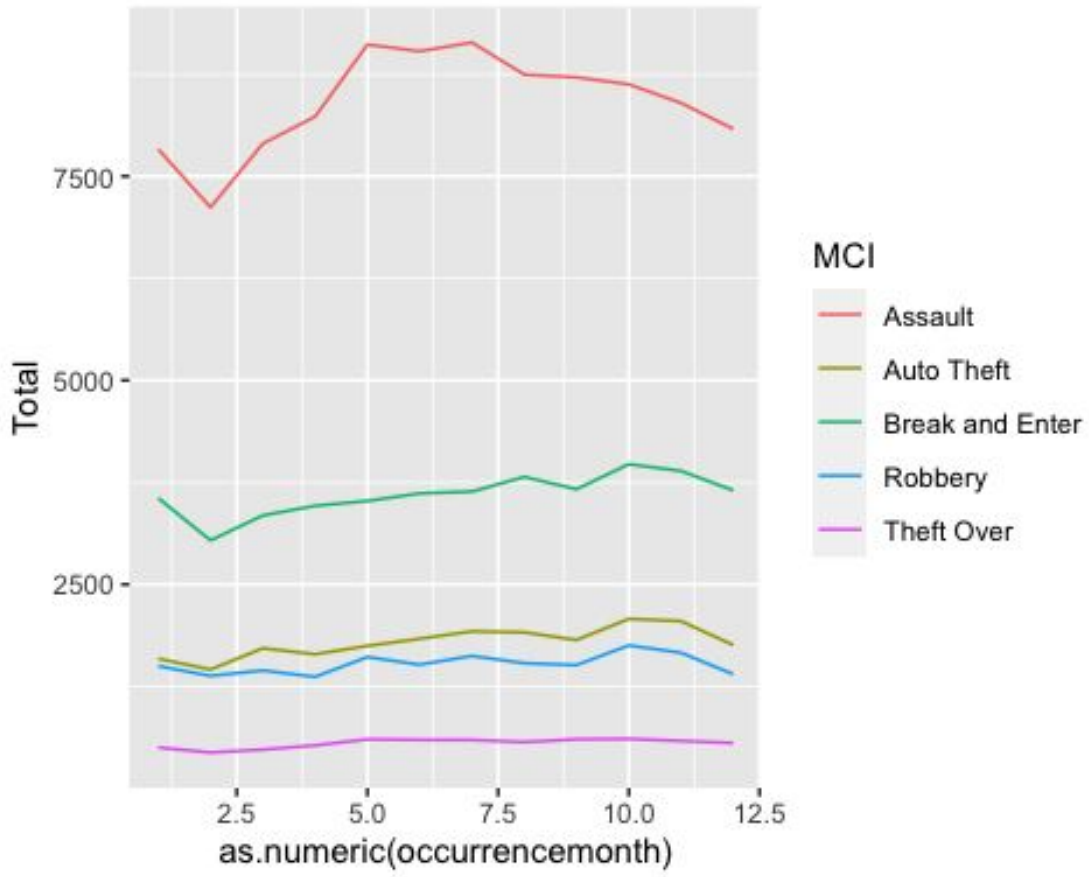
Most assault incidents happen between May and August,autotheft between July and November while other types are fairly consistent across all months.

```
mcimonth<-EDAfilter %>% group_by(MCI,occurrencemonth) %>%
dplyr::summarise(Total = n())
ggplot(mcimonth, aes(MCI, occurrencemonth, fill = Total)) +
  geom_tile(size = 1, color = "white") +
  scale_fill_viridis()  +
  geom_text(aes(label=Total), color='white') +
  ggtitle("MCI by Month(2014-2019)")
```



12.Trend MCI by Month

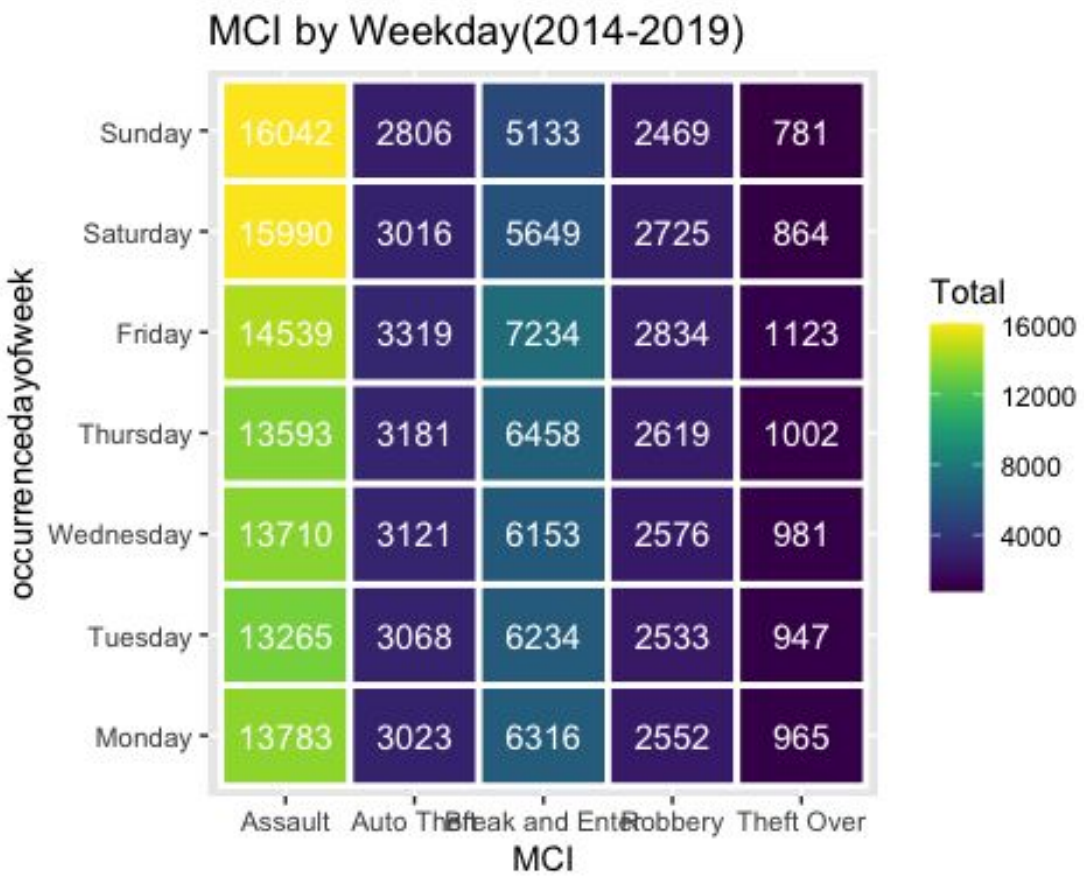
```
ggplot(mcimonth, aes(as.numeric(occurrencemonth), Total, color = MCI)) +
geom_line()
```



13.MCI Heatmap by Day of Week

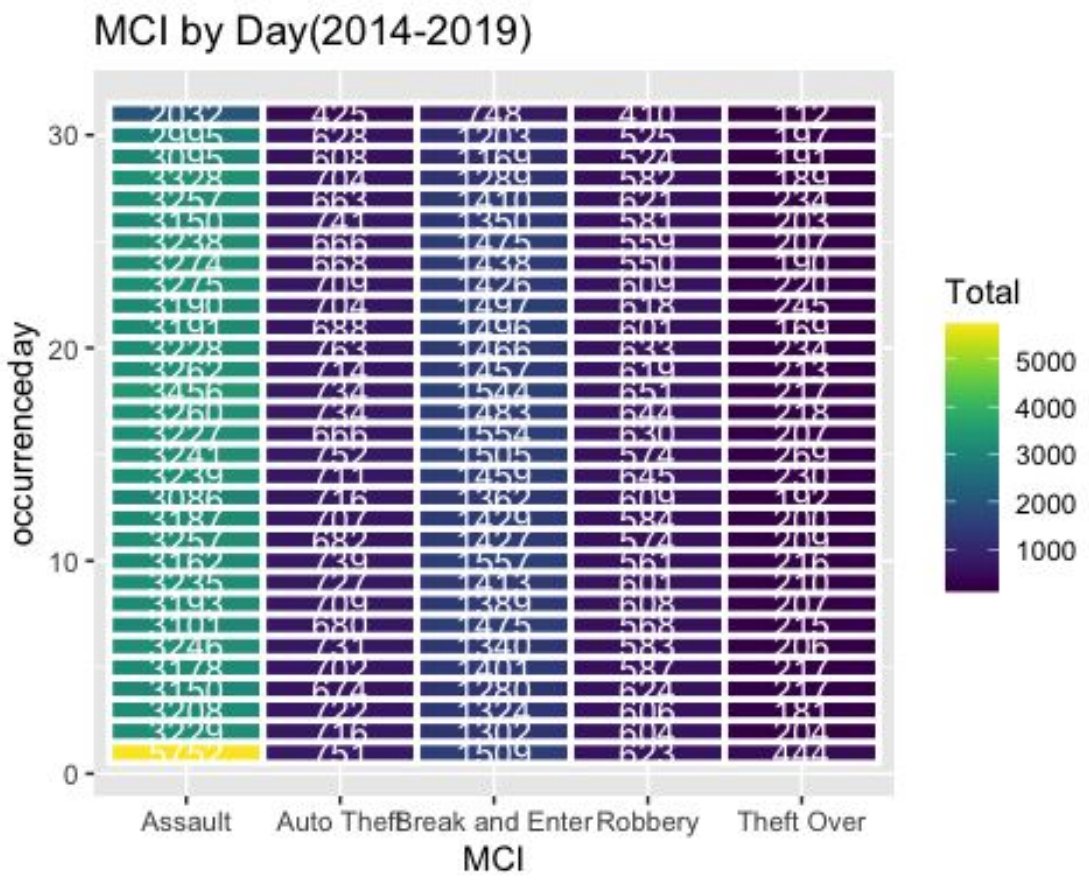
Assault peak is observed usually on weekends, while other types peak on Fridays.

```
mciweek <- EDAfilter %>% group_by(MCI, occurrencedayofweek) %>%
dplyr::summarise(Total = n())
ggplot(mciweek, aes(MCI, occurrencedayofweek, fill = Total)) +
  geom_tile(size = 1, color = "white") +
  scale_fill_viridis()  +
  geom_text(aes(label=Total), color='white') +
  ggtitle("MCI by Weekday(2014-2019)")
```



14.MCI Heatmap by Day of Month

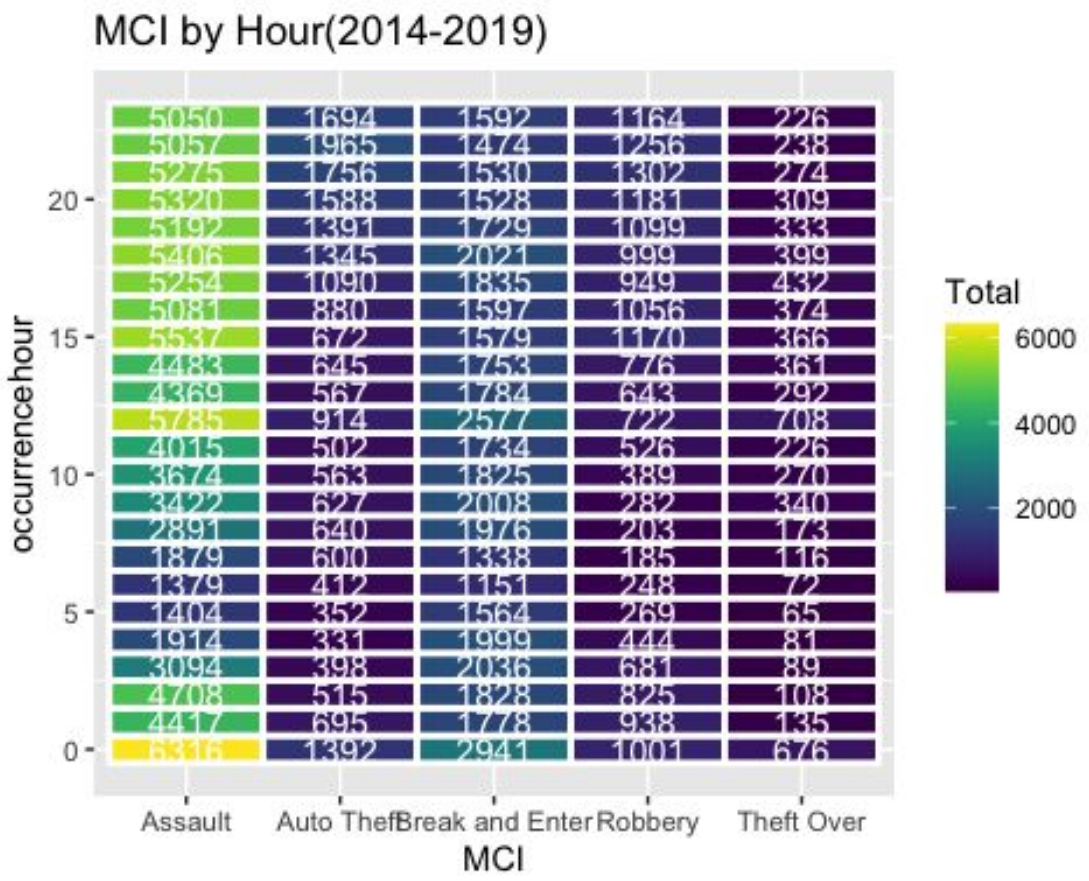
```
mciday <- EDAfilter %>% group_by(MCI, occurrenceday) %>%
dplyr::summarise(Total = n())
ggplot(mciday, aes(MCI, occurrenceday, fill = Total)) +
  geom_tile(size = 1, color = "white") +
  scale_fill_viridis()  +
  geom_text(aes(label=Total), color='white') +
  ggtitle("MCI by Day(2014-2019)")
```



15.MCI Heatmap by Hour

Assault, Break and Enter, Theftover peaks are observed at both midnight and noon, autotheft at 10PM, robbery at 9PM.

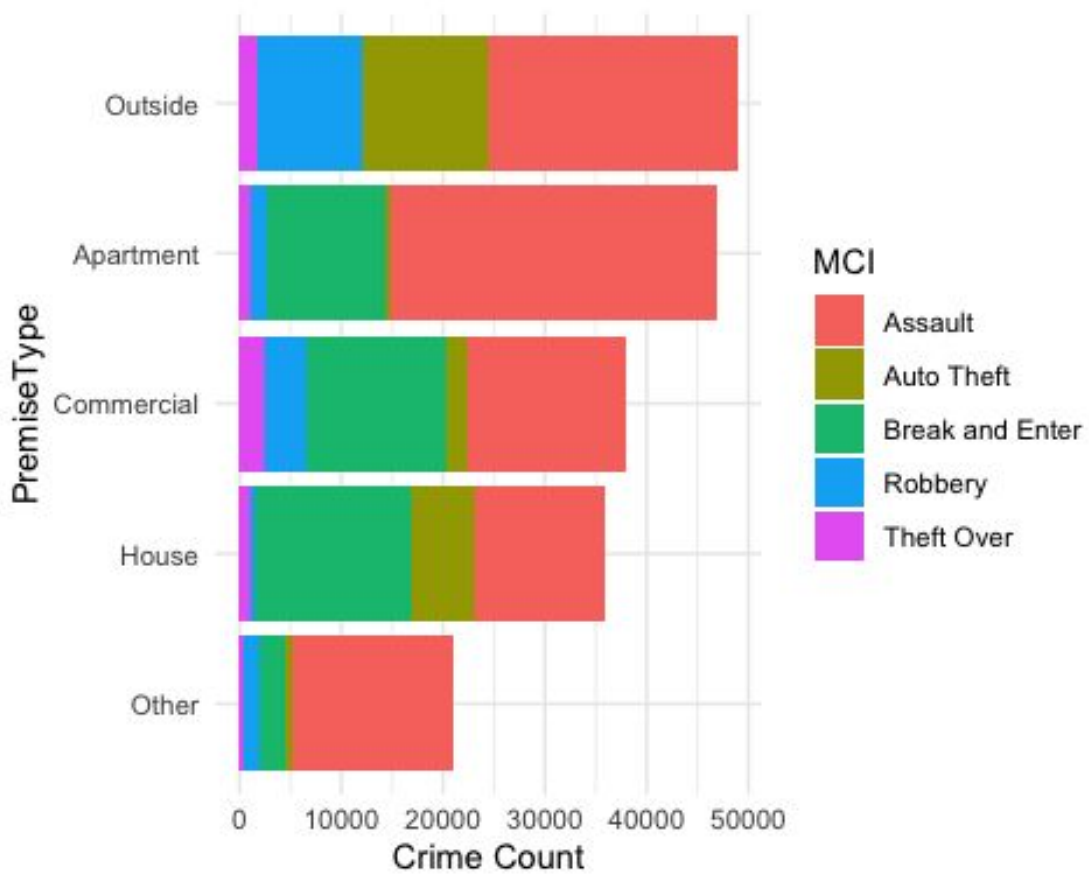
```
mcihour <- EDAfilter %>% group_by(MCI, occurrencehour) %>%
dplyr::summarise(Total = n())
ggplot(mcihour, aes(MCI, occurrencehour, fill = Total)) +
  geom_tile(size = 1, color = "white") +
  scale_fill_viridis()  +
  geom_text(aes(label=Total), color='white') +
  ggtitle("MCI by Hour(2014-2019)")
```



16.Crime Category by Premise Type

Auto theft happens mostly outside and at houses while other crime types are common across all premise types.

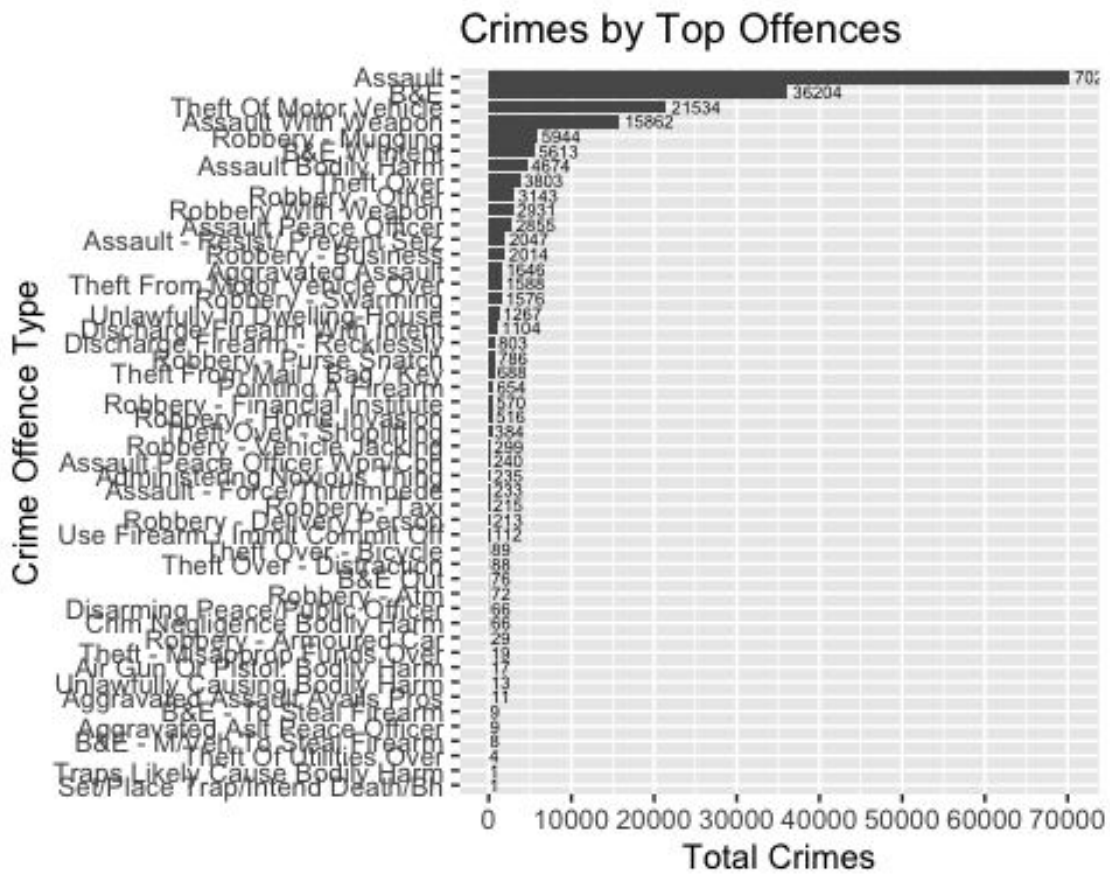
```
premci<-EDAfilter %>% group_by(MCI,premisetype) %>% dplyr::summarise(Total =
n())
ggplot(premci, aes(reorder(premisetype,Total),Total,fill=MCI))+
  geom_bar(stat="identity")+
  coord_flip()+
  xlab("PremiseType") +
  ylab("Crime Count")+
  theme_minimal()
```



17.Top Offences

Top offences within the assault category include - assault with weapon, bodily harm, and assault peace officer while top offences within robbery include - mugging, other, robbery with weapons, and robbery-business.

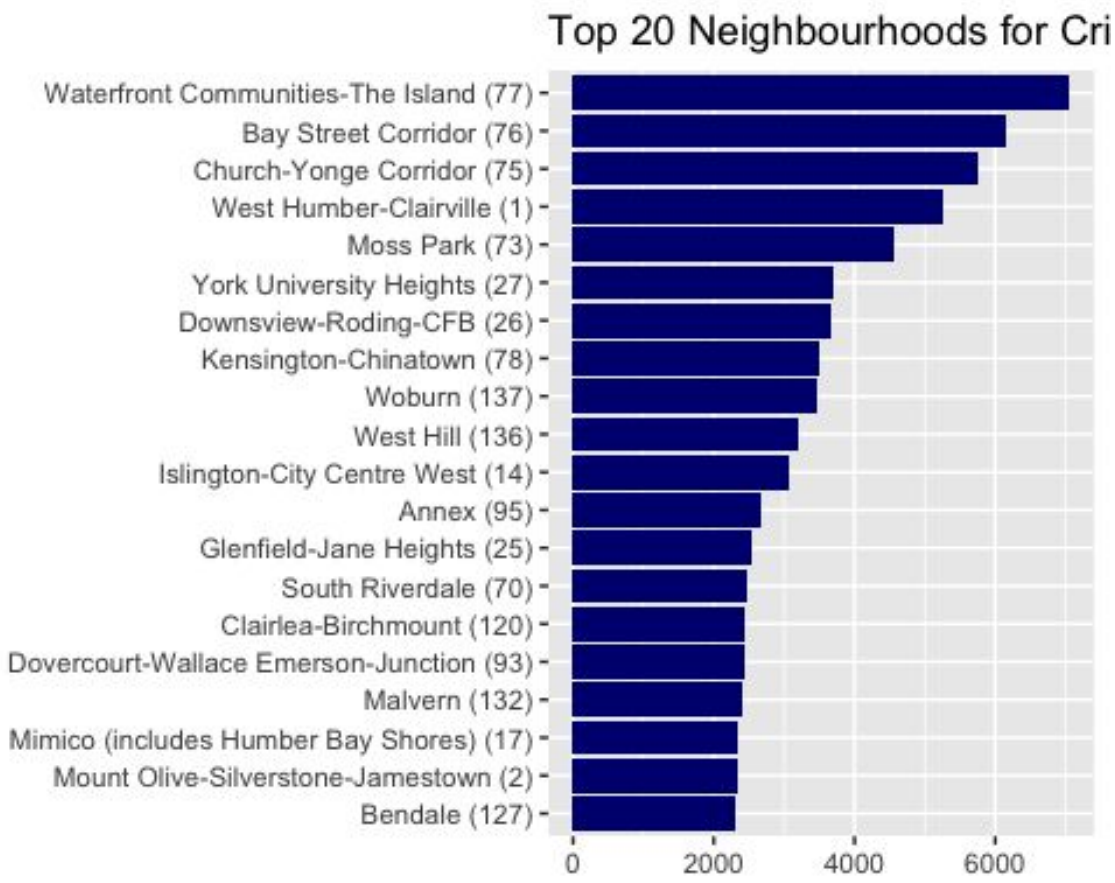
```
type<-EDAfilter %>% group_by(offence) %>% dplyr::summarise(Total = n())
ggplot(type, aes(reorder(offence, Total), Total)) +
  geom_bar(stat = "identity") + geom_text(aes(label = Total), stat =
'identity', data = type, hjust = -0.1, size = 2) +
  scale_y_continuous(breaks = seq(0,80000,10000)) +
  coord_flip() + ggtitle("Crimes by Top Offences") +
  xlab("Crime Offence Type") +
  ylab("Total Crimes")
```



18.Top 20 Neighbourhoods for Crime

The most dangerous neighbourhood is the Waterfront. The other two most dangerous hoods are the Bay Street Corridor and Yonge-Church Corridor.

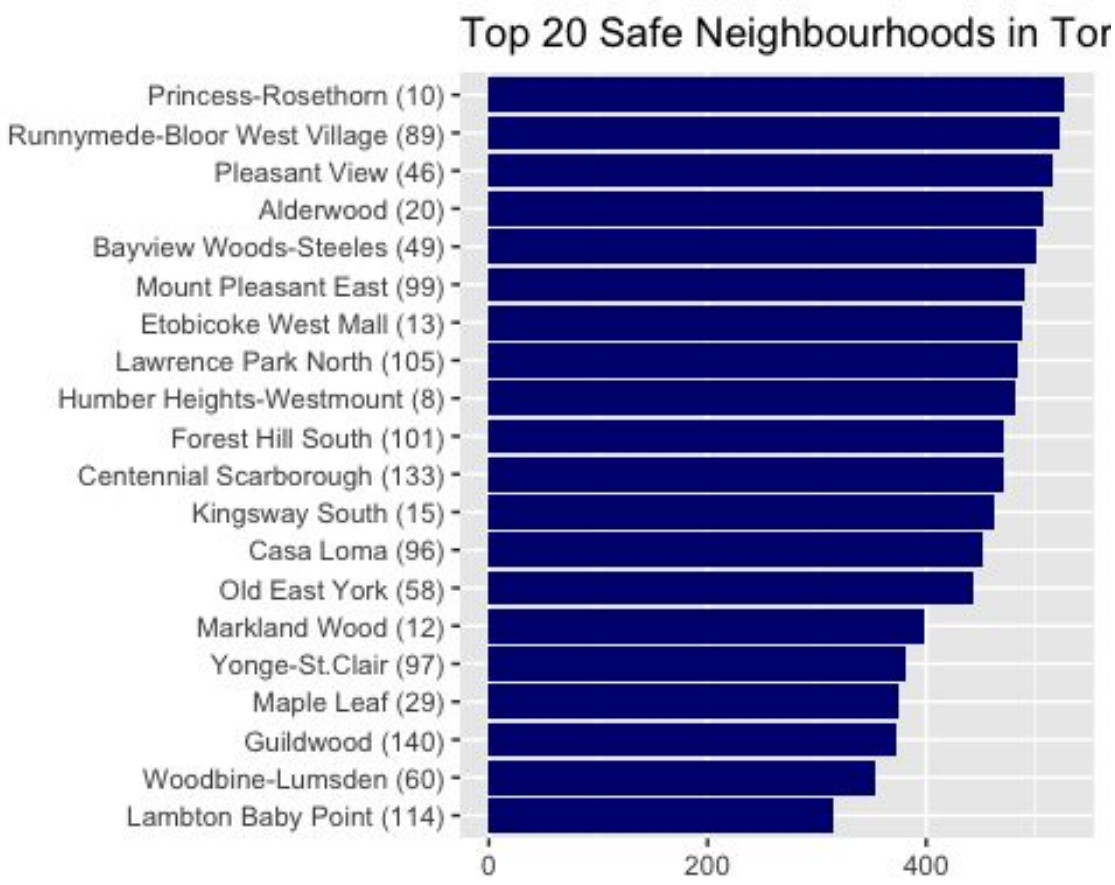
```
df<-EDAfilter
df$Neighbourhood<-as.factor(df$Neighbourhood)
dfhood<-df %>% group_by(Neighbourhood) %>% dplyr::summarise(Total = n()) %>%
dplyr::top_n(20) %>% dplyr::arrange(desc(Total))

## Selecting by Total

ggplot(dfhood,aes(reorder(Neighbourhood,Total), Total))+
  geom_col(fill = "navy")+
  coord_flip()+
  ggtitle("Top 20 Neighbourhoods for Crime") +
  labs(x = NULL, y = NULL)
```



19.Top 20 Safe Neighbourhoods in Toronto

The most safest hoods are Lambton Baby Point, Woodbine-Lumsden, and Guildwood.

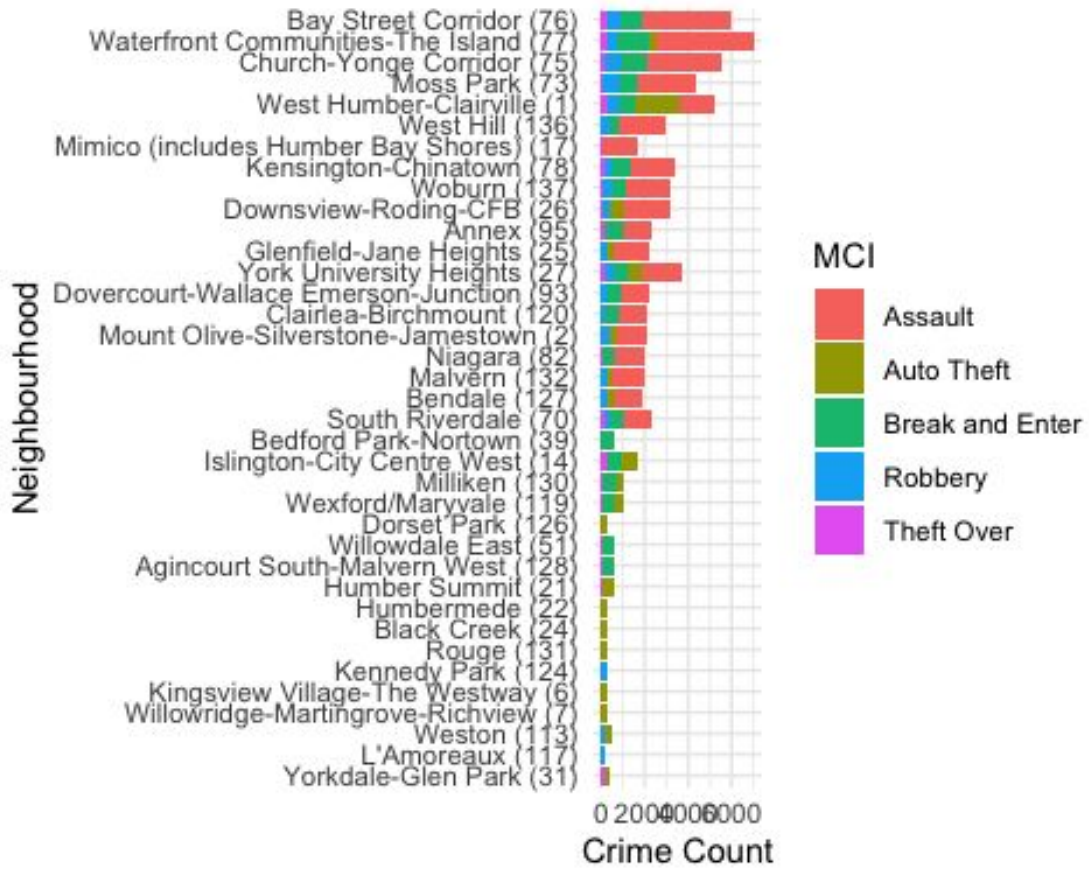
```
dfsafe<-df %>% group_by(Neighbourhood) %>% dplyr::summarise(Total = n()) %>%
dplyr::top_n(-20) %>% dplyr::arrange(desc(Total))

## Selecting by Total

ggplot(dfsafe,aes(reorder(Neighbourhood,Total), Total))+
  geom_col(fill = "navy")+
  coord_flip()+
  ggtitle("Top 20 Safe Neighbourhoods in Toronto")+
  labs(x = NULL, y = NULL)
```



20.Top 20 Neighbourhoods by Crime Categories

Besides assaults, Bay Street Corridor, Church-Yonge Corridor and Waterfront had the most break and enter crimes while West Humber-Clairville had the most vehicle stolen crimes.

```
dfmci<-df %>% group_by(MCI,Neighbourhood) %>% dplyr::summarise(Total = n())
%>% dplyr::top_n(20) %>% dplyr::arrange(desc(Total))

## Selecting by Total

ggplot(dfmci, aes(reorder(Neighbourhood,Total),Total,fill=MCI))+
  geom_bar(stat="identity")+
  coord_flip()+
  xlab("Neighbourhood") +
  ylab("Crime Count")+
  theme_minimal()
```



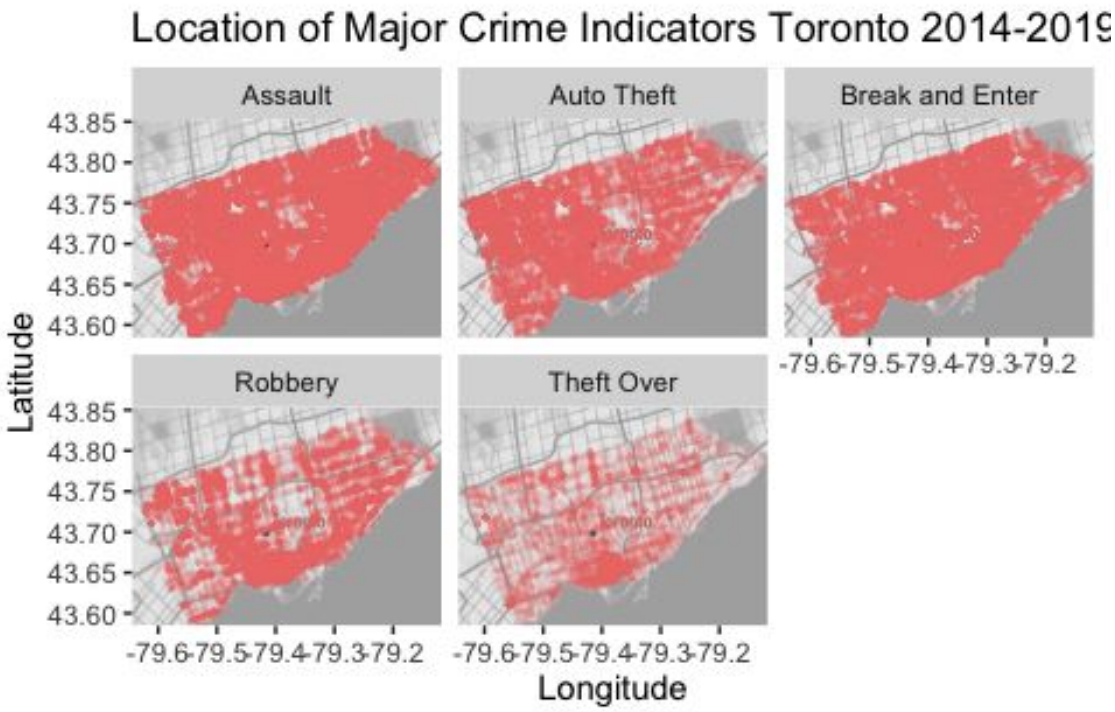
21.Map of Toronto's crimes, simple visualization

Assaults, and Break and Enter occur all over the city, with a concentration in the Waterfront areas. Other crimes, such as Auto Theft has more points on the west side than the east side. Robbery and Theft Over are primarily in the Waterfront areas.

```
lat <- df$Lat
lon <- df$Long
crimes <- df$MCI
to_map <- data.frame(crimes, lat, lon)
colnames(to_map) <- c('crimes', 'lat', 'lon')
sbbox <- make_bbox(lon = df$Long, lat = df$Lat, f = 0.01)
my_map <- get_map(location = sbbox, maptype = "roadmap", scale = 2,
color="bw", zoom = 10)

## Source : http://tile.stamen.com/terrain/10/285/372.png

## Source : http://tile.stamen.com/terrain/10/286/372.png

## Source : http://tile.stamen.com/terrain/10/285/373.png
```

```
## Source : http://tile.stamen.com/terrain/10/286/373.png

ggmap(my_map) +
  geom_point(data=to_map, aes(x = lon, y = lat, color = "#27AE60"),
             size = 0.5, alpha = 0.03) +
  xlab('Longitude') +
  ylab('Latitude') +
  ggtitle('Location of Major Crime Indicators Toronto 2014-2019') +
  guides(color=FALSE)
```


```
ggmap(my_map) +
  geom_point(data=to_map, aes(x = lon, y = lat, color = "#27AE60"),
             size = 0.5, alpha = 0.05) +
  xlab('Longitude') +
  ylab('Latitude') +
  ggtitle('Location of Major Crime Indicators Toronto 2014-2019') +
  guides(color=FALSE) +
  facet_wrap(~ crimes, nrow = 2)
```


Location of Major Crime Indicators Toronto 2014-2019
Assault
Auto Theft
Break and Enter
Robbery
Theft Over
Latitude
43.85
43.80
43.75
43.70
43.65
43.60
-79.6
-79.5
-79.4
-79.3
-79.2
Longitude