Midterm-Report-MCI

Chirag Bellara & Soumya Velamuri

2023-10-30

library(ggplot2)  
library(ggthemes)  
library(dplyr)  
library(viridis)  
library(tidyr)  
library(cluster)  
library(ggmap)  
library(maps)

#### Load the Dataset and Check for duplicated records

toronto <- read.csv('data/Major\_Crime\_Indicators\_Open\_Data.csv')  
head(toronto)

## X Y OBJECTID EVENT\_UNIQUE\_ID REPORT\_DATE  
## 1 -79.42590 43.75735 1 GO-20141262074 2014/01/01 05:00:00+00  
## 2 -79.35023 43.64629 2 GO-20141260701 2014/01/01 05:00:00+00  
## 3 -79.37650 43.66642 3 GO-20141260889 2014/01/01 05:00:00+00  
## 4 -85.48874 0.00000 4 GO-20141260973 2014/01/01 05:00:00+00  
## 5 -79.34484 43.67895 5 GO-20141261050 2014/01/01 05:00:00+00  
## 6 -79.39184 43.64664 6 GO-20141259344 2014/01/01 05:00:00+00  
## OCC\_DATE REPORT\_YEAR REPORT\_MONTH REPORT\_DAY REPORT\_DOY  
## 1 1998/06/01 04:00:00+00 2014 January 1 1  
## 2 2014/01/01 05:00:00+00 2014 January 1 1  
## 3 2014/01/01 05:00:00+00 2014 January 1 1  
## 4 2014/01/01 05:00:00+00 2014 January 1 1  
## 5 2014/01/01 05:00:00+00 2014 January 1 1  
## 6 2014/01/01 05:00:00+00 2014 January 1 1  
## REPORT\_DOW REPORT\_HOUR OCC\_YEAR OCC\_MONTH OCC\_DAY OCC\_DOY OCC\_DOW OCC\_HOUR  
## 1 Wednesday 12 NA NA NA 12  
## 2 Wednesday 3 2014 January 1 1 Wednesday 3  
## 3 Wednesday 4 2014 January 1 1 Wednesday 4  
## 4 Wednesday 4 2014 January 1 1 Wednesday 4  
## 5 Wednesday 4 2014 January 1 1 Wednesday 4  
## 6 Wednesday 2 2014 January 1 1 Wednesday 2  
## DIVISION LOCATION\_TYPE  
## 1 D32 Apartment (Rooming House, Condo)  
## 2 D51 Commercial Dwelling Unit (Hotel, Motel, B & B, Short Term Rental)  
## 3 D51 Apartment (Rooming House, Condo)  
## 4 NSA Streets, Roads, Highways (Bicycle Path, Private Road)  
## 5 D54 Streets, Roads, Highways (Bicycle Path, Private Road)  
## 6 D52 Bar / Restaurant  
## PREMISES\_TYPE UCR\_CODE UCR\_EXT OFFENCE MCI\_CATEGORY  
## 1 Apartment 1480 110 Administering Noxious Thing Assault  
## 2 Commercial 2120 200 B&E Break and Enter  
## 3 Apartment 1430 100 Assault Assault  
## 4 Outside 2130 210 Theft Over Theft Over  
## 5 Outside 1430 100 Assault Assault  
## 6 Commercial 1420 110 Assault Bodily Harm Assault  
## HOOD\_158 NEIGHBOURHOOD\_158 HOOD\_140 NEIGHBOURHOOD\_140  
## 1 38 Lansing-Westgate 38 Lansing-Westgate (38)  
## 2 70 South Riverdale 70 South Riverdale (70)  
## 3 74 North St.James Town 74 North St.James Town (74)  
## 4 NSA NSA NSA NSA  
## 5 69 Blake-Jones 66 Danforth (66)  
## 6 164 Wellington Place 77 Waterfront Communities-The Island (77)  
## LONG\_WGS84 LAT\_WGS84  
## 1 -79.42590 43.75735  
## 2 -79.35023 43.64629  
## 3 -79.37650 43.66642  
## 4 -85.48874 0.00000  
## 5 -79.34484 43.67895  
## 6 -79.39184 43.64664

sum(is.na(toronto))

## [1] 315

sapply(toronto, function(x) sum(is.na(x)))

## X Y OBJECTID EVENT\_UNIQUE\_ID   
## 0 0 0 0   
## REPORT\_DATE OCC\_DATE REPORT\_YEAR REPORT\_MONTH   
## 0 0 0 0   
## REPORT\_DAY REPORT\_DOY REPORT\_DOW REPORT\_HOUR   
## 0 0 0 0   
## OCC\_YEAR OCC\_MONTH OCC\_DAY OCC\_DOY   
## 105 0 105 105   
## OCC\_DOW OCC\_HOUR DIVISION LOCATION\_TYPE   
## 0 0 0 0   
## PREMISES\_TYPE UCR\_CODE UCR\_EXT OFFENCE   
## 0 0 0 0   
## MCI\_CATEGORY HOOD\_158 NEIGHBOURHOOD\_158 HOOD\_140   
## 0 0 0 0   
## NEIGHBOURHOOD\_140 LONG\_WGS84 LAT\_WGS84   
## 0 0 0

missingCols <- select(toronto, OCC\_YEAR, OCC\_MONTH, OCC\_DAY, OCC\_DOY, OCC\_DOW)  
toronto <- na.omit(toronto)  
(colMeans(is.na(missingCols)))\*100

## OCC\_YEAR OCC\_MONTH OCC\_DAY OCC\_DOY OCC\_DOW   
## 0.03247798 0.00000000 0.03247798 0.03247798 0.00000000

sum(duplicated(toronto$EVENT\_UNIQUE\_ID))

## [1] 41575

toronto <- subset(toronto, !duplicated(toronto$EVENT\_UNIQUE\_ID))  
unique(toronto$OCC\_YEAR)

## [1] 2014 2013 2012 2003 2011 2004 2010 2009 2008 2006 2000 2005 2002 2001 2015  
## [16] 2007 2016 2017 2018 2019 2020 2021 2022

unique(toronto$REPORT\_YEAR)

## [1] 2014 2015 2016 2017 2018 2019 2020 2021 2022

year\_group <- group\_by(toronto, OCC\_YEAR)  
crime\_by\_year <- summarise(year\_group,  
 n = n())  
crime\_by\_year

## # A tibble: 23 × 2  
## OCC\_YEAR n  
## <int> <int>  
## 1 2000 24  
## 2 2001 17  
## 3 2002 17  
## 4 2003 12  
## 5 2004 22  
## 6 2005 19  
## 7 2006 12  
## 8 2007 25  
## 9 2008 36  
## 10 2009 54  
## # ℹ 13 more rows

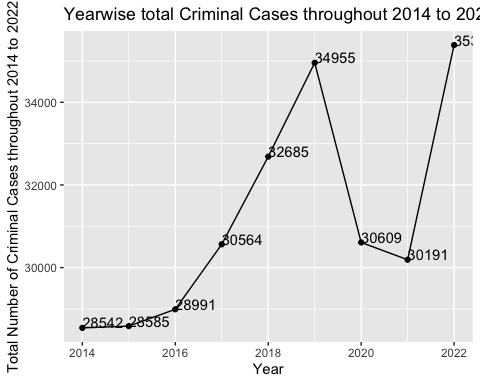
# Getting the Count of Each category of Crime  
table(unlist(toronto$MCI\_CATEGORY))

##   
## Assault Auto Theft Break and Enter Robbery Theft Over   
## 144916 41642 61089 23543 10426

drops <- c("X", "Y", "OBJECTID", "UCR\_CODE", "UCR\_EXT", "REPORT\_DATE", "REPORT\_MONTH", "REPORT\_DAY", "REPORT\_DOY", "REPORT\_DOW", "REPORT\_HOUR", "OCC\_DOY", "REPORT\_YEAR", "DIVISION", "HOOD\_158", "HOOD\_140")  
toronto <- toronto[, !(names(toronto) %in% drops)]

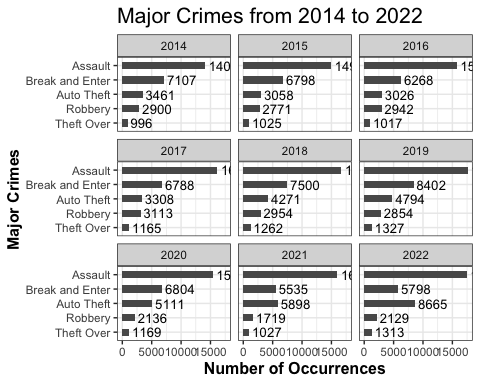
##### What are the total number of crimes in each year.

cases\_count <- toronto %>%   
 filter(OCC\_YEAR >= 2014) %>%  
 count(OCC\_YEAR) %>%  
 group\_by(OCC\_YEAR)  
  
# Getting the Count of Number of Crimes each year  
ggplot(cases\_count, aes(x = OCC\_YEAR, y = n, label=n)) +  
 geom\_line() +  
 geom\_point() +  
 geom\_text(hjust=0, vjust=0) +  
 labs(x = "Year",   
 y = "Total Number of Criminal Cases throughout 2014 to 2022",   
 title = "Yearwise total Criminal Cases throughout 2014 to 2022")



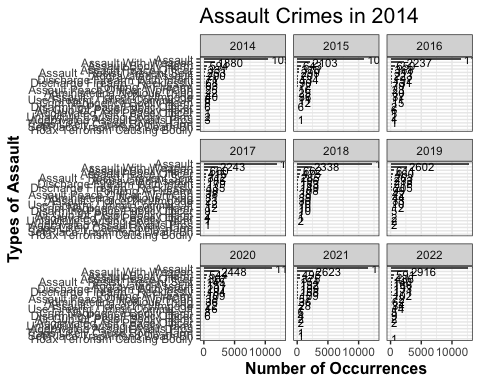
#### What are the Major Crimes in each year.

# Grouping Major Crimes from 2014 to 2022  
indicator\_group <- group\_by(filter(toronto, OCC\_YEAR >= 2014, OCC\_YEAR != "NA"), MCI\_CATEGORY, OCC\_YEAR)  
crime\_by\_indicator <- summarise(indicator\_group, n=n())  
crime\_by\_indicator <- crime\_by\_indicator[order(crime\_by\_indicator$n, decreasing = TRUE),]  
  
ggplot(aes(x = reorder(MCI\_CATEGORY, n), y = n), data = crime\_by\_indicator) +  
 geom\_bar(stat = 'identity', width = 0.5) +  
 geom\_text(aes(label = n), stat = 'identity', data = crime\_by\_indicator, hjust = -0.1, size = 3.5) +  
 coord\_flip() +  
 xlab('Major Crimes') +  
 ylab('Number of Occurrences') +  
 ggtitle('Major Crimes from 2014 to 2022') +  
 theme\_bw() +  
 theme(plot.title = element\_text(size = 16),  
 axis.title = element\_text(size = 12, face = "bold")) +  
 facet\_wrap(vars(OCC\_YEAR))



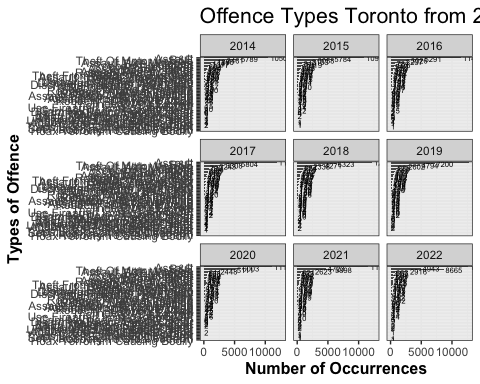
#### What are the different types of Assault and which one is the worst.

# Getting the crimes from Assault Category for 2014  
assault <- filter(toronto[toronto$MCI\_CATEGORY == 'Assault', ], OCC\_YEAR >= 2014, OCC\_YEAR != "NA")  
assault\_group <- group\_by(assault, OFFENCE, OCC\_YEAR)  
assault\_by\_offence <- summarise(assault\_group, n=n())  
assault\_by\_offence <- assault\_by\_offence[order(assault\_by\_offence$n, decreasing = TRUE), ]  
ggplot(aes(x = reorder(OFFENCE, n), y = n), data = assault\_by\_offence) +  
 geom\_bar(stat = 'identity', width = 0.6) +  
 geom\_text(aes(label = n), stat = 'identity', data = assault\_by\_offence, hjust = -0.1, size = 3) +  
 coord\_flip() +  
 xlab('Types of Assault') +  
 ylab('Number of Occurrences') +  
 ggtitle('Assault Crimes in 2014') +  
 theme\_bw() +  
 theme(plot.title = element\_text(size = 16),  
 axis.title = element\_text(size = 12, face = "bold")) +  
 facet\_wrap(vars(OCC\_YEAR))



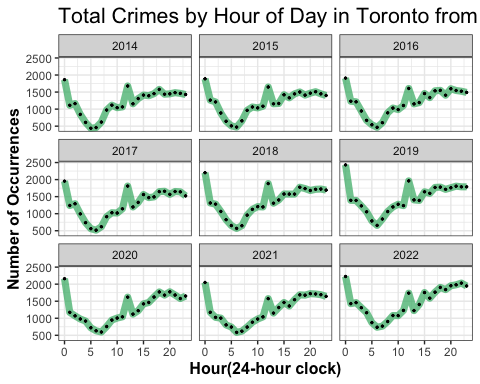
#### Let’s look at the offences instead.

# Getting the Offences from 2014 to 2022  
offence\_group <- group\_by(filter(toronto, OCC\_YEAR >= 2014, OCC\_YEAR != "NA"), OFFENCE, OCC\_YEAR)  
crime\_by\_offence <- summarise(offence\_group, n=n())  
crime\_by\_offence <- crime\_by\_offence[order(crime\_by\_offence$n, decreasing = TRUE), ]  
ggplot(aes(x = reorder(OFFENCE, n), y = n), data = crime\_by\_offence) +  
 geom\_bar(stat = 'identity', width = 0.7) +  
 geom\_text(aes(label = n), stat = 'identity', data = crime\_by\_offence, hjust = -0.1, size = 2) +  
 coord\_flip() +  
 xlab('Types of Offence') +  
 ylab('Number of Occurrences') +  
 ggtitle('Offence Types Toronto from 2014 to 2022') +  
 theme\_bw() +  
 theme(plot.title = element\_text(size = 16),  
 axis.title = element\_text(size = 12, face = "bold")) +  
 facet\_wrap(vars(OCC\_YEAR))



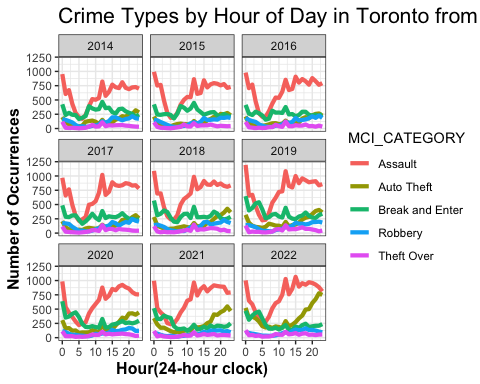
#### Checking for crimes wrt time of the day.

# Crimes WRT Time of Day from 2014 to 2022  
hour\_group <- group\_by(filter(toronto, OCC\_YEAR >= 2014, OCC\_YEAR != "NA"), OCC\_HOUR, OCC\_YEAR)  
crime\_hour <- summarise(hour\_group, n=n())  
ggplot(aes(x=OCC\_HOUR, y=n), data = crime\_hour) + geom\_line(size = 2.5, alpha = 0.7, color = "mediumseagreen", group=1) +   
 geom\_point(size = 0.5) +   
 ggtitle('Total Crimes by Hour of Day in Toronto from 2014 to 2022') +  
 ylab('Number of Occurrences') +  
 xlab('Hour(24-hour clock)') +  
 theme\_bw() +  
 theme(plot.title = element\_text(size = 16),  
 axis.title = element\_text(size = 12, face = "bold")) +  
 facet\_wrap(vars(OCC\_YEAR))



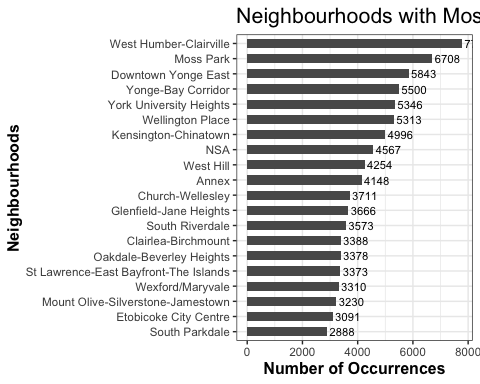
#### What types of Crimes are most frequent wrt the time of the day?

# Crime Types WRT Hour of the day from 2014 to 2022  
hour\_crime\_group <- group\_by(filter(toronto, OCC\_YEAR >= 2014), OCC\_HOUR, OCC\_YEAR, MCI\_CATEGORY)  
hour\_crime <- summarise(hour\_crime\_group, n=n())  
ggplot(aes(x=OCC\_HOUR, y=n, color=MCI\_CATEGORY), data = hour\_crime) +   
 geom\_line(size=1.5) +   
 ggtitle('Crime Types by Hour of Day in Toronto from 2014 to 2022') +  
 ylab('Number of Occurrences') +  
 xlab('Hour(24-hour clock)') +  
 theme\_bw() +  
 theme(plot.title = element\_text(size = 16),  
 axis.title = element\_text(size = 12, face = "bold")) +  
 facet\_wrap(vars(OCC\_YEAR))



#### **Where in Toronto were those crimes most likely to occur**

# Crimes WRT location from 2014 to 2022  
location\_group <- group\_by(filter(toronto, OCC\_YEAR >= 2014), NEIGHBOURHOOD\_158)  
crime\_by\_location <- summarise(location\_group, n=n())  
crime\_by\_location <- crime\_by\_location[order(crime\_by\_location$n, decreasing = TRUE), ]  
crime\_by\_location\_top20 <- head(crime\_by\_location, 20)  
ggplot(aes(x = reorder(NEIGHBOURHOOD\_158, n), y = n), data = crime\_by\_location\_top20) +  
 geom\_bar(stat = 'identity', width = 0.6) +  
 geom\_text(aes(label = n), stat = 'identity', data = crime\_by\_location\_top20, hjust = -0.1, size = 3) +  
 coord\_flip() +  
 xlab('Neighbourhoods') +  
 ylab('Number of Occurrences') +  
 ggtitle('Neighbourhoods with Most Crimes - Top 20 from 2014 to 2022') +  
 theme\_bw() +  
 theme(plot.title = element\_text(size = 16),  
 axis.title = element\_text(size = 12, face = "bold"))



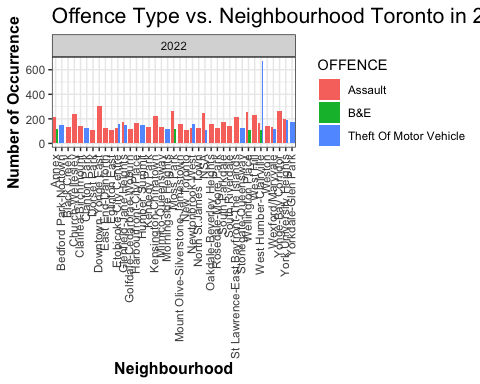
#### Which are the Safest Neighborhoods to Live in 2022

location\_group <- group\_by(filter(toronto, OCC\_YEAR == 2022), NEIGHBOURHOOD\_158)  
crime\_by\_location <- summarise(location\_group, n=n())  
crime\_by\_location <- crime\_by\_location[order(crime\_by\_location$n, decreasing = TRUE), ]  
tail(crime\_by\_location, 5)

## # A tibble: 5 × 2  
## NEIGHBOURHOOD\_158 n  
## <chr> <int>  
## 1 Bendale South 71  
## 2 Humber Heights-Westmount 71  
## 3 Woodbine-Lumsden 66  
## 4 Bayview Woods-Steeles 61  
## 5 Guildwood 50

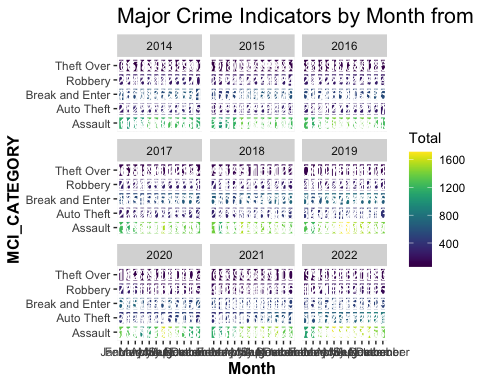
#### Comparing Neighborhoods with top offence types

# Top Offences in Each Neighborhood from in 2022  
offence\_location\_group <- group\_by(filter(toronto, OCC\_YEAR == 2022), NEIGHBOURHOOD\_158, OFFENCE, OCC\_YEAR)  
offence\_type\_by\_location <- summarise(offence\_location\_group, n=n())  
offence\_type\_by\_location <- offence\_type\_by\_location[order(offence\_type\_by\_location$n, decreasing = TRUE), ]  
offence\_type\_by\_location\_top20 <- head(offence\_type\_by\_location, 50)  
ggplot(aes(x = NEIGHBOURHOOD\_158, y=n, fill = OFFENCE), data=offence\_type\_by\_location\_top20) +  
 geom\_bar(stat = 'identity', position = position\_dodge(), width = 0.8) +  
 xlab('Neighbourhood') +  
 ylab('Number of Occurrence') +  
 ggtitle('Offence Type vs. Neighbourhood Toronto in 2022') +   
 theme\_bw() +  
 theme(plot.title = element\_text(size = 16),  
 axis.title = element\_text(size = 12, face = "bold"),  
 axis.text.x = element\_text(angle = 90, hjust = 1, vjust = .4)) +  
 facet\_wrap(vars(OCC\_YEAR))



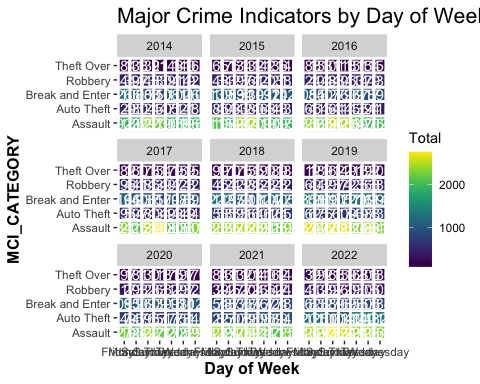
#### Month Wise Crime Rates

# Major Crimes for every month from 2014 to 2022  
crime\_count <- filter(toronto, OCC\_YEAR >= 2014) %>%   
 group\_by(OCC\_MONTH, MCI\_CATEGORY, OCC\_YEAR) %>%   
 summarise(Total = n())  
crime\_count$OCC\_MONTH <- ordered(crime\_count$OCC\_MONTH, levels = c('January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'December'))  
ggplot(crime\_count, aes(OCC\_MONTH, MCI\_CATEGORY, fill = Total)) +  
 geom\_tile(size = 1, color = "white") +  
 scale\_fill\_viridis() +  
 geom\_text(aes(label=Total), color='white') +  
 ggtitle("Major Crime Indicators by Month from 2014 to 2022") +  
 xlab('Month') +  
 theme(plot.title = element\_text(size = 16),   
 axis.title = element\_text(size = 12, face = "bold")) +  
 facet\_wrap(vars(OCC\_YEAR))



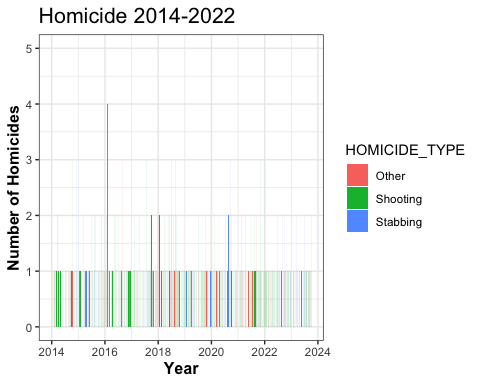
#### Day of the Week Wise Crime Rates

# Major Crimes for every DOW from 2014 to 2022  
day\_count <- filter(toronto, OCC\_YEAR >= 2014) %>%   
 group\_by(OCC\_DOW, MCI\_CATEGORY, OCC\_YEAR) %>%   
 summarise(Total = n())  
ggplot(day\_count, aes(OCC\_DOW, MCI\_CATEGORY, fill = Total)) +  
 geom\_tile(size = 1, color = "white") +  
 scale\_fill\_viridis() +  
 geom\_text(aes(label=Total), color='white') +  
 ggtitle("Major Crime Indicators by Day of Week from 2014 to 2022") +  
 xlab('Day of Week') +  
 theme(plot.title = element\_text(size = 16),   
 axis.title = element\_text(size = 12, face = "bold")) +  
 facet\_wrap(vars(OCC\_YEAR))

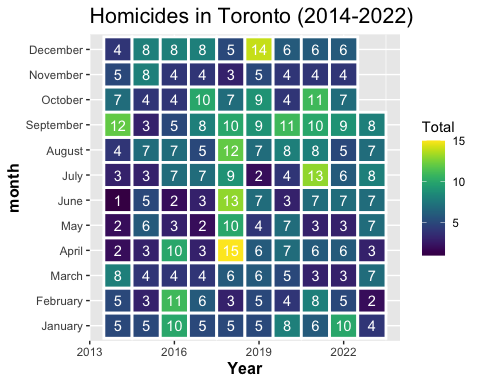


### Homicide Rates

homicide <- read.csv('data/Homicides.csv', stringsAsFactors = F)  
homicide <- filter(homicide, OCC\_DATE >= 2014)  
homicide$OCC\_DATE <- as.Date(homicide$OCC\_DATE)  
year\_group <- group\_by(homicide, OCC\_DATE, HOMICIDE\_TYPE)  
homicide\_by\_year <- summarise(year\_group, n=n())  
ggplot(aes(x = OCC\_DATE, y=n, fill = HOMICIDE\_TYPE), data = homicide\_by\_year) +  
 geom\_bar(stat = 'identity', position = position\_dodge(), width = 1) +  
 xlab('Year') +  
 ylab('Number of Homicides') +  
 ggtitle('Homicide 2014-2022') +   
 ylim(0, 5) +  
 theme\_bw() +  
 theme(plot.title = element\_text(size = 16),  
 axis.title = element\_text(size = 12, face = "bold"))



homicide$month <- format(as.Date(homicide$OCC\_DATE) , "%B")  
homicide\_count <- homicide %>% group\_by(OCC\_YEAR, month) %>% summarise(Total = n())  
homicide\_count$month <- ordered(homicide\_count$month, levels = c('January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'December'))  
ggplot(homicide\_count, aes(OCC\_YEAR, month, fill = Total)) +  
 geom\_tile(size = 1, color = "white") +  
 scale\_fill\_viridis() +  
 geom\_text(aes(label=Total), color='white') +  
 ggtitle("Homicides in Toronto (2014-2022)") +  
 xlab('Year') +  
 theme(plot.title = element\_text(size = 16),   
 axis.title = element\_text(size = 12, face = "bold"))



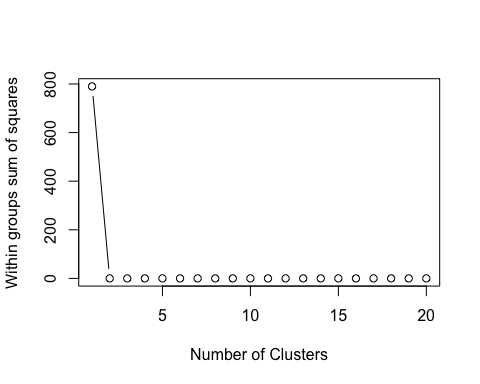
by\_groups <- group\_by(toronto, MCI\_CATEGORY, NEIGHBOURHOOD\_158)  
groups <- summarise(by\_groups, n=n())  
groups <- groups[c("NEIGHBOURHOOD\_158", "MCI\_CATEGORY", "n")]  
groups\_wide <- spread(groups, key = MCI\_CATEGORY, value = n)  
groups\_wide

## # A tibble: 159 × 6  
## NEIGHBOURHOOD\_158 Assault `Auto Theft` `Break and Enter` Robbery `Theft Over`  
## <chr> <int> <int> <int> <int> <int>  
## 1 Agincourt North 568 314 374 182 40  
## 2 Agincourt South-… 947 351 648 173 103  
## 3 Alderwood 277 193 241 49 57  
## 4 Annex 1993 246 1363 288 271  
## 5 Avondale 355 119 98 44 39  
## 6 Banbury-Don Mills 656 236 543 80 88  
## 7 Bathurst Manor 445 311 254 80 34  
## 8 Bay-Cloverhill 975 81 488 156 100  
## 9 Bayview Village 609 217 329 53 75  
## 10 Bayview Woods-St… 289 133 223 26 21  
## # ℹ 149 more rows

z <- groups\_wide[, -c(1,1)]  
z <- z[complete.cases(z), ]

m <- apply(z, 2, mean)  
s <- apply(z, 2, sd)  
z <- scale(z, m, s)

wss <- (nrow(z)-1) \* sum(apply(z, 2, var))  
for (i in 2:20) wss[i] <- sum(kmeans(z, centers=i)$withiness)  
plot(1:20, wss, type='b', xlab='Number of Clusters', ylab='Within groups sum of squares')



kc <- kmeans(z, 2)  
kc

## K-means clustering with 2 clusters of sizes 142, 17  
##   
## Cluster means:  
## Assault Auto Theft Break and Enter Robbery Theft Over  
## 1 -0.2583932 -0.1284152 -0.2604747 -0.2398382 -0.2728258  
## 2 2.1583428 1.0726446 2.1757297 2.0033547 2.2788979  
##   
## Clustering vector:  
## [1] 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2  
## [38] 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [75] 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1  
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 2 1  
## [149] 1 1 1 1 1 2 1 1 1 2 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 184.7964 229.2655  
## (between\_SS / total\_SS = 47.6 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

z1 <- data.frame(z, kc$cluster)  
clusplot(z1, kc$cluster, color=TRUE, shade=F, labels=0, lines=0, main='k-Means Cluster Analysis')

