Project1: NYC Crime Analysis

2022-07-31

Aim

The aim of this markdown is to analyse dataset for NYPD Shooting Incident Data (Historic)

The dataset captures all shooting incidents reported In New York City starting from the year 2006. The details of the incident such as location, time, victim and perpetrator are also included in the dataset.

Publisher : City of New York

Prepare for Analyses

Import Library

```
set.seed(1234)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library('lubridate')
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

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

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

Get URL

Copy the URL and Store into a variable so that it can be read as a csv:

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
```

Read Dataset from URL

Using read_csv function, store the dataset into a tibble

```
df = read_csv(url_in)
```

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Print Dataset

Let us print the dataset stored in the tibble called df to see what its contents are(rows and columns):

```
head(df)
```

```
## # A tibble: 6 x 19
##   INCIDE~1 OCCUR~2 OCCUR~3 BORO  PRECI~4 JURIS~5 LOCAT~6 STATI~7 PERP_~8 PERP_~9
##      <dbl> <chr>   <time>  <chr>   <dbl>   <dbl> <chr>   <lgl>   <chr>   <chr>
## 1   2.41e7 08/27/~ 05:35   BRONX      52       0 <NA>    TRUE    <NA>    <NA>
## 2   7.77e7 03/11/~ 12:03   QUEE~     106       0 <NA>    FALSE   <NA>    <NA>
## 3   2.27e8 04/14/~ 21:08   BRONX      42       0 COMMER~ TRUE    <NA>    <NA>
## 4   2.38e8 12/10/~ 19:30   BRONX      52       0 <NA>    FALSE   <NA>    <NA>
## 5   2.25e8 02/22/~ 00:18   MANH~      34       0 <NA>    FALSE   <NA>    <NA>
## 6   2.25e8 03/07/~ 06:15   BROO~      75       0 <NA>    TRUE    25-44   M
## # ... with 9 more variables: PERP_RACE <chr>, VIC_AGE_GROUP <chr>,
## #   VIC_SEX <chr>, VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>,
## #   Latitude <dbl>, Longitude <dbl>, Lon_Lat <chr>, and abbreviated variable
## #   names 1: INCIDENT_KEY, 2: OCCUR_DATE, 3: OCCUR_TIME, 4: PRECINCT,
## #   5: JURISDICTION_CODE, 6: LOCATION_DESC, 7: STATISTICAL_MURDER_FLAG,
## #   8: PERP_AGE_GROUP, 9: PERP_SEX
## # i Use `colnames()` to see all variable names
```

Tidy and Transform Data

Lets get a list of all the columns in the dataset:

```
colnames(df)
```

```
##  [1] "INCIDENT_KEY"            "OCCUR_DATE"
##  [3] "OCCUR_TIME"              "BORO"
##  [5] "PRECINCT"                "JURISDICTION_CODE"
##  [7] "LOCATION_DESC"           "STATISTICAL_MURDER_FLAG"
##  [9] "PERP_AGE_GROUP"          "PERP_SEX"
## [11] "PERP_RACE"               "VIC_AGE_GROUP"
```

```
## [13] "VIC_SEX"                 "VIC_RACE"
## [15] "X_COORD_CD"              "Y_COORD_CD"
## [17] "Latitude"                "Longitude"
## [19] "Lon_Lat"
```

Reorganize Columns : Lets remove the columns we don't want in this dataset: We will remove the following - **JURISDICTION_CODE**, **X_COORD_CD**, **Y_COORD_CD**, **Lon_Lat**

```
df <- df %>% select(INCIDENT_KEY,
                    OCCUR_DATE,
                    OCCUR_TIME,
                    BORO,
                    PRECINCT,
                    LOCATION_DESC,
                    STATISTICAL_MURDER_FLAG,
                    PERP_AGE_GROUP,
                    PERP_SEX,
                    PERP_RACE,
                    VIC_AGE_GROUP,
                    VIC_SEX,
                    VIC_RACE,
                    Latitude,
                    Longitude)
```

Handle NA/Blank values : Next, we want handles blank values in the data. For various reasons such as case or investigation currently active, or lack of proof, certain columns would have incomplete or blank data. Lets check how many columns have blank data:

```
lapply(df, function(x) sum(is.na(x)))
```

```
## $INCIDENT_KEY
## [1] 0
##
## $OCCUR_DATE
## [1] 0
##
## $OCCUR_TIME
## [1] 0
##
## $BORO
## [1] 0
##
## $PRECINCT
## [1] 0
##
## $LOCATION_DESC
## [1] 14977
##
## $STATISTICAL_MURDER_FLAG
## [1] 0
##
## $PERP_AGE_GROUP
## [1] 9344
##
## $PERP_SEX
## [1] 9310
```

```
##
## $PERP_RACE
## [1] 9310
##
## $VIC_AGE_GROUP
## [1] 0
##
## $VIC_SEX
## [1] 0
##
## $VIC_RACE
## [1] 0
##
## $Latitude
## [1] 0
##
## $Longitude
## [1] 0
```

We see that 4 columns i.e LOCATION_DESC,PERP_AGE_GROUP,PERP_SEX and PERP_RACE have blanks. Lets replace these to 'UNKNOWN'

```
df <- df %>% replace_na(list(LOCATION_DESC = "UNKNOWN",
                        PERP_AGE_GROUP = "UNKNOWN",
                        PERP_SEX = "UNKNOWN",
                        PERP_RACE = "UNKNOWN"))
```

Column Data Type Conversion : We will convert OCCUR_DATE to a date field(it is a String currently).
We will also create a factor for the fields - BORO,PERP_AGE_GROUP,VIC_AGE_GROUP,LOCATION_DESC,STATISTICA

```
df$BORO = as.factor(df$BORO)
df$PERP_AGE_GROUP = as.factor(df$PERP_AGE_GROUP)
df$VIC_AGE_GROUP = as.factor(df$VIC_AGE_GROUP)
df$LOCATION_DESC = as.factor(df$LOCATION_DESC)
df$STATISTICAL_MURDER_FLAG = as.factor(df$STATISTICAL_MURDER_FLAG)
df$PERP_SEX = as.factor(df$PERP_SEX)
df$PERP_RACE = as.factor(df$PERP_RACE)
df$VIC_SEX = as.factor(df$VIC_SEX)
df$VIC_RACE = as.factor(df$VIC_RACE)
df$OCCUR_DATE <- as.Date(df$OCCUR_DATE , format = "%m/%d/%Y")
```

Tracking Hour Column : We will create a new column which will store the hour of the day for each incident. This will be useful to analyze and easily group incidents by Hour.

```
df$HOUR = hour(df$OCCUR_TIME)
```

Summary of the Dataset

```
summary(df)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:25596
##  1st Qu.: 61593633   1st Qu.:2009-05-10   Class1:hms
##  Median : 86437258   Median :2012-08-26   Class2:difftime
##  Mean   :112382648   Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:166660833   3rd Qu.:2017-07-01
```

```
##  Max.   :238490103   Max.   :2021-12-31
##
##             BORO          PRECINCT                        LOCATION_DESC
##  BRONX        : 7402   Min.   :  1.00   UNKNOWN                   :14977
##  BROOKLYN     :10365   1st Qu.: 44.00   MULTI DWELL - PUBLIC HOUS: 4559
##  MANHATTAN    : 3265   Median : 69.00   MULTI DWELL - APT BUILD  : 2664
##  QUEENS       : 3828   Mean   : 65.87   PVT HOUSE                :  893
##  STATEN ISLAND:  736   3rd Qu.: 81.00   GROCERY/BODEGA           :  622
##                        Max.   :123.00   BAR/NIGHT CLUB           :  588
##                                         (Other)                  : 1293
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  FALSE:20668             UNKNOWN:12492   F      :  371
##  TRUE : 4928             18-24  : 5844   M      :14416
##                          25-44  : 5202   U      : 1499
##                          <18    : 1463   UNKNOWN: 9310
##                          45-64  :  535
##                          65+    :   57
##                          (Other):    3
##                           PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2681   F: 2403
##  ASIAN / PACIFIC ISLANDER      :  141   18-24  : 9604   M:23182
##  BLACK                         :10668   25-44  :11386   U:   11
##  BLACK HISPANIC                : 1203   45-64  : 1698
##  UNKNOWN                       :11146   65+    :  167
##  WHITE                         :  272   UNKNOWN:   60
##  WHITE HISPANIC                : 2164
##                             VIC_RACE        Latitude       Longitude
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Min.   :40.51   Min.   :-74.25
##  ASIAN / PACIFIC ISLANDER      :  354   1st Qu.:40.67   1st Qu.:-73.94
##  BLACK                         :18281   Median :40.70   Median :-73.92
##  BLACK HISPANIC                : 2485   Mean   :40.74   Mean   :-73.91
##  UNKNOWN                       :   65   3rd Qu.:40.82   3rd Qu.:-73.88
##  WHITE                         :  660   Max.   :40.91   Max.   :-73.70
##  WHITE HISPANIC                : 3742
##       HOUR
##  Min.   : 0.00
##  1st Qu.: 3.00
##  Median :15.00
##  Mean   :12.19
##  3rd Qu.:20.00
##  Max.   :23.00
##
```

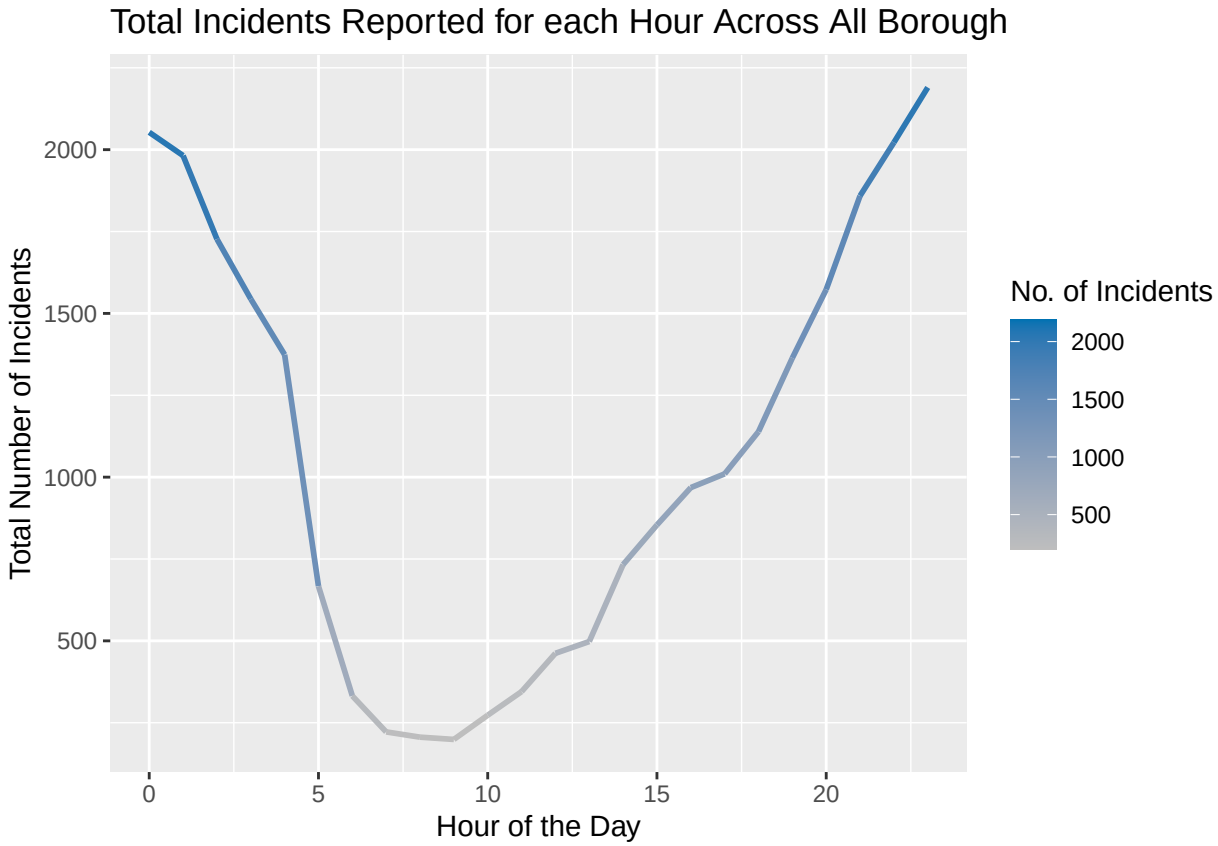
Add Visualizations and Analysis

1. What time of the day is crime most prevelant? Is this trend consistent across all neighbourhoods?

```
df_h = df %>%
  group_by(HOUR) %>%
  count()
```

```
ggplot(df_h, aes(x = HOUR, y = n)) +
  geom_line(aes(color = n), size = 1) +

    scale_colour_gradient(low = "gray", high = "#0072B2") +
  labs(
    x = "Hour of the Day",
    y = "Total Number of Incidents",
    colour = "No. of Incidents",
    title = "Total Incidents Reported for each Hour Across All Borough "
  )  + scale_fill_brewer(palette='Accent')
```



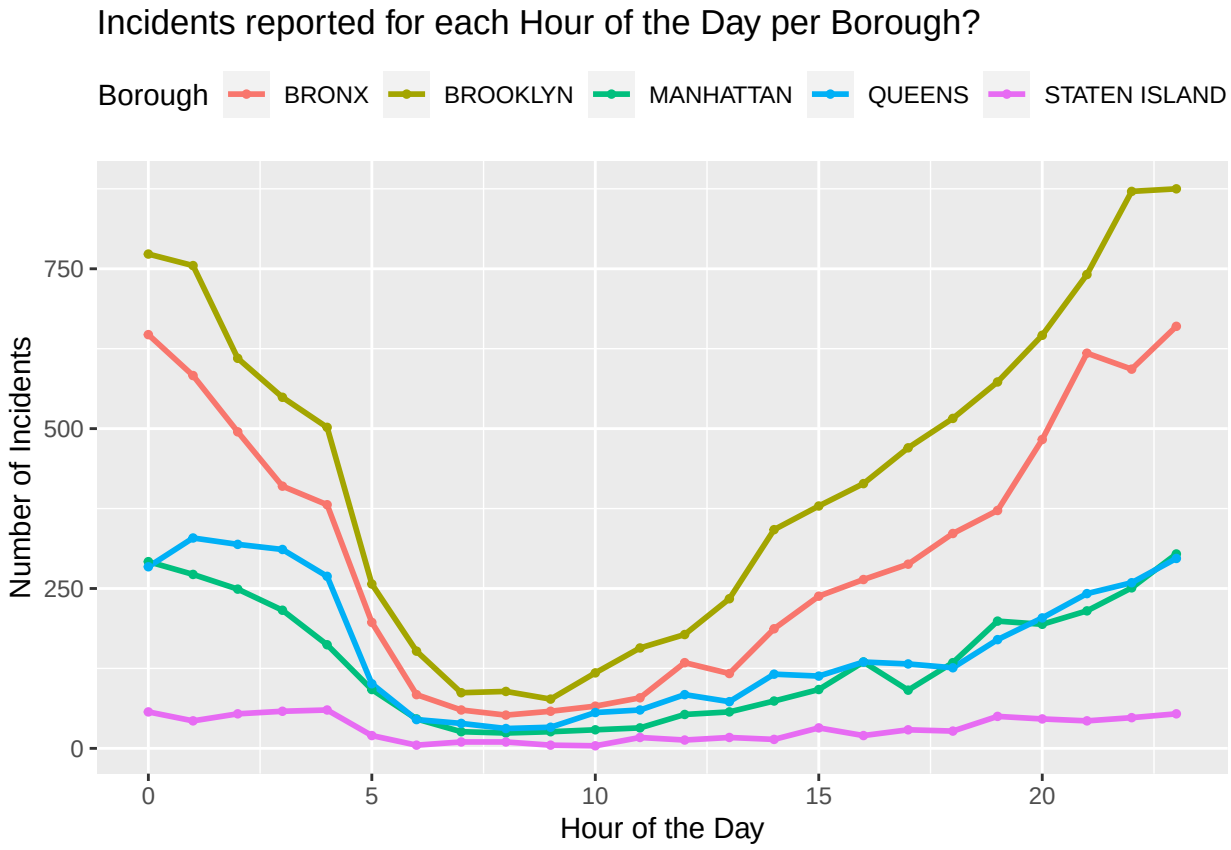
Observations : We observe, that the number of crime incidents reported drastically increase after dark and drop down once it is daylight outside. On average, we see a big spike in crime post 6PM and it reaches its highest point around midnight(200+ reported incidents at this hour). Post which, it slowly starts to drop down, while still maintaining a high mark and finally drops down by sunrise. The lowest number of incidents reported is as the workday starts around 8AM(less than 250 incidents).

Question Raised : Do various other BOROughs follow the same pattern that is observed citywide i.e crime drastically increase after sunset and dips during work hours? Or is this something mainly observed by less residential and more business areas ?

To look into this question, we will plot the number of incidents reported per hour for each Bourough -

```
df_hb = df %>%
  group_by(HOUR,BORO) %>%
  count()
```

```
ggplot(df_hb, aes(x = HOUR, y = n, group = BORO)) +
  geom_line(aes(color = BORO), size = 1) +
  geom_point(aes(color = BORO), size = 1) +
  theme(legend.position = "top") +
  labs(
    x = "Hour of the Day",
    y = "Number of Incidents",
    colour = "Borough",
    title = "Incidents reported for each Hour of the Day per Borough? "
  )
```



Observation : We can say that The citywide trend of crime increase after dark is followed by almost all the Borough of NYC. We see an uptick in incidents reported for all, except Staten island which still has its highest number of incidents reported early morning and drops down post 5AM. This could also be since Staten Island is further away from the city, and people had a 1hr+ commute to the city, hence, most people in Staten island tend to leave for the city due to work, and due to the commute, have to leave early. Hence, less people tend to go to Staten Island during the day as compared to people migrating from Staten island.

We also observe that among all the Borough, Brooklyn,Bronx and Queens register the highest number of incidents reported for each hour, closely followed by Manhattan. All the Borough follow a citywide pattern of crime increasing after work hours and decreasing during the day.

Question raised: We have not accounted for population so far. This raises the question, that in any given hour, how many people per 10,000 people(or per capita) are affected by crime. And how many of those incidents are Murders. Does a neighborhood with highest number of incidents reported also have the highest number of murders? Can we have a scenario where we have an uptick in incidents reported but a decrease in

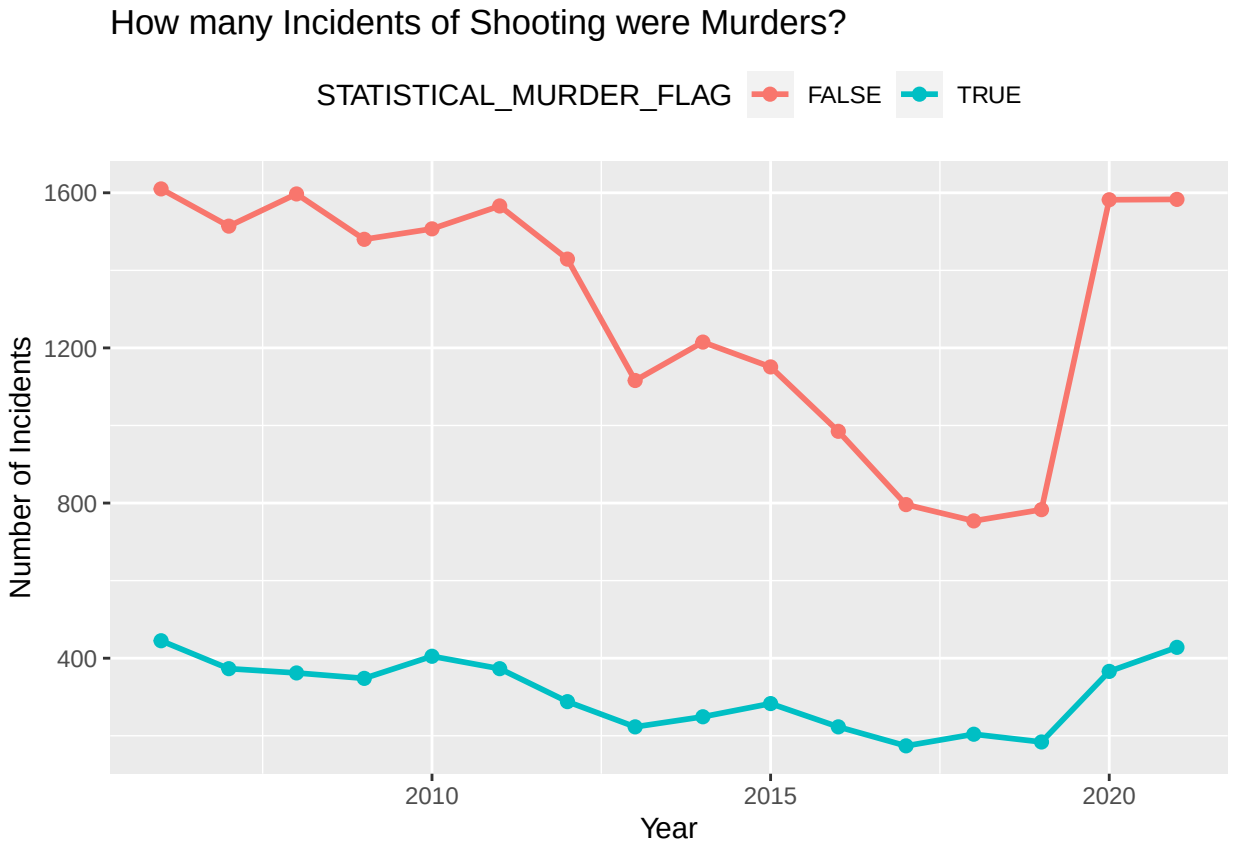
murders?

2. What is the proportion of Murders for total reported shooting incidents?

```
df_OCCUR_YEAR <- as.Date(df$OCCUR_DATE)
df$OCCUR_YEAR <- as.numeric(format(df_OCCUR_YEAR, "%Y"))

df_YEAR = df %>%
  group_by(OCCUR_YEAR,STATISTICAL_MURDER_FLAG) %>%
  count()

ggplot(df_YEAR, aes(x = OCCUR_YEAR, y = n,group=STATISTICAL_MURDER_FLAG)) +
  geom_line(aes(color = STATISTICAL_MURDER_FLAG), size = 1) +
  geom_point(aes(color = STATISTICAL_MURDER_FLAG), size = 2) +
  theme(legend.position = "top") +
  labs(
    x = "Year",
    y = "Number of Incidents",
    title = "How many Incidents of Shooting were Murders? "
  )
```

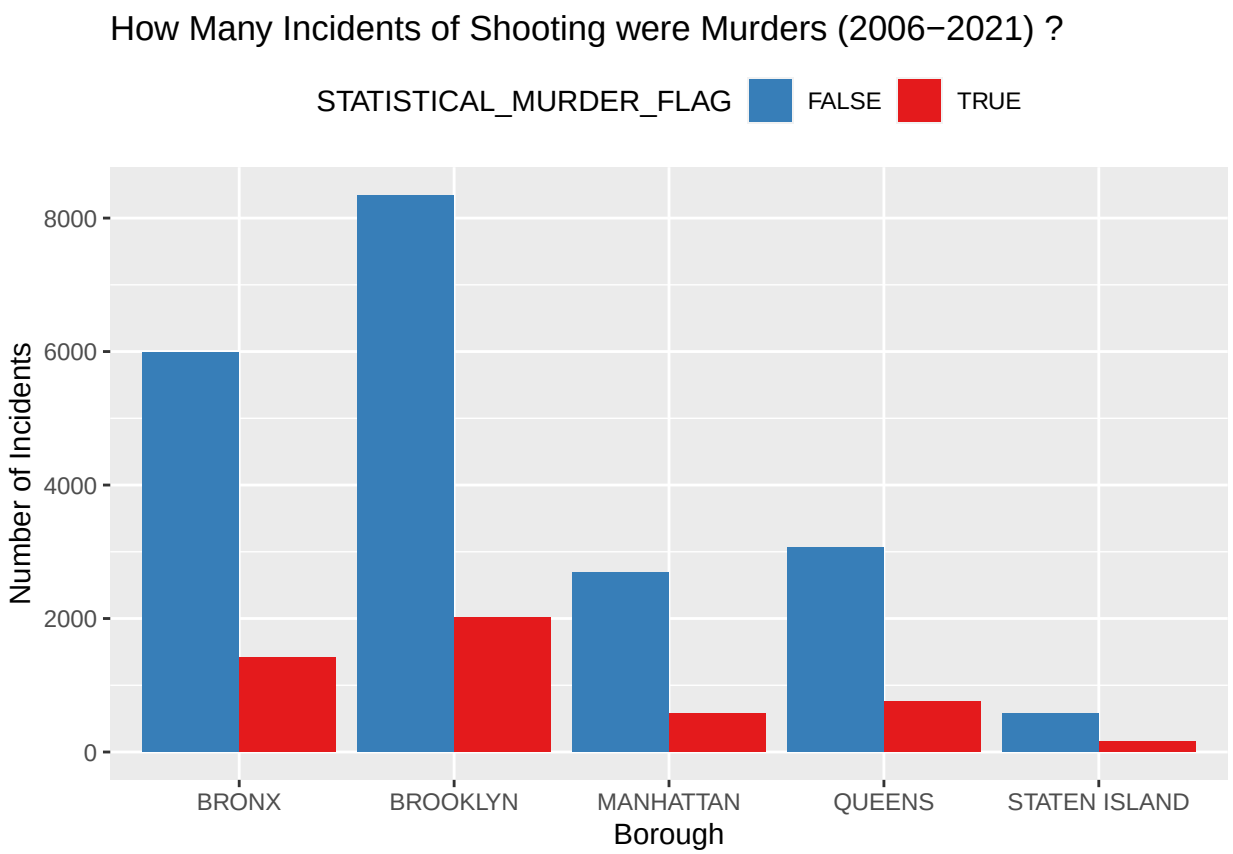


Observation About 20% of total shooting reported are Murders. In the graph above, we see that approx 400 incidents were Murders, while approx 1600 incidents were non-murder related shootings. Total 20% of reported shootings are murder.

Which borough has the highest number of reported murders?

```
df_YEAR_m = df %>%
  group_by(BORO,STATISTICAL_MURDER_FLAG) %>%
  count()

  # creating plot using the above data
ggplot(df_YEAR_m, aes(BORO,n, fill=STATISTICAL_MURDER_FLAG))  +
  geom_bar(stat="identity", position = "dodge") +
  scale_fill_brewer(palette = "Set1",direction=-1) +
    theme(legend.position = "top") +
  labs(
    x = "Borough",
    y = "Number of Incidents",
    title = "How Many Incidents of Shooting were Murders (2006-2021) ? "
  )
```



Observation : We see that just like the total number of non-murder shootings, murders are also the highest in Brooklyn, Bronx, Queens and closely followed by Manhattan. While Staten Island has the lowest amount of shootings, it has the highest proportion of murders when compared to its total reported shootings. Brooklyn, Bronx and Manhattan are lower than citywide trend of Murder - non murder proportion i.e around 18 percent (NYC City proportion for murder vs total shootings is 20%). Staten island has the highest proportion with 22% while queens is at 21%. overall, we can conclude that each Borough has 1/4 of total reported incidents classified as Murder.

Overall Data for Reported Shootings for Each Borough -

```
df_YEAR_m
```

```
## # A tibble: 10 x 3
```

```
## # Groups:   BORO, STATISTICAL_MURDER_FLAG [10]
##    BORO          STATISTICAL_MURDER_FLAG     n
##    <fct>         <fct>                   <int>
##  1 BRONX         FALSE                    5985
##  2 BRONX         TRUE                     1417
##  3 BROOKLYN      FALSE                    8345
##  4 BROOKLYN      TRUE                     2020
##  5 MANHATTAN     FALSE                    2691
##  6 MANHATTAN     TRUE                      574
##  7 QUEENS        FALSE                    3066
##  8 QUEENS        TRUE                      762
##  9 STATEN ISLAND FALSE                     581
## 10 STATEN ISLAND TRUE                      155
```

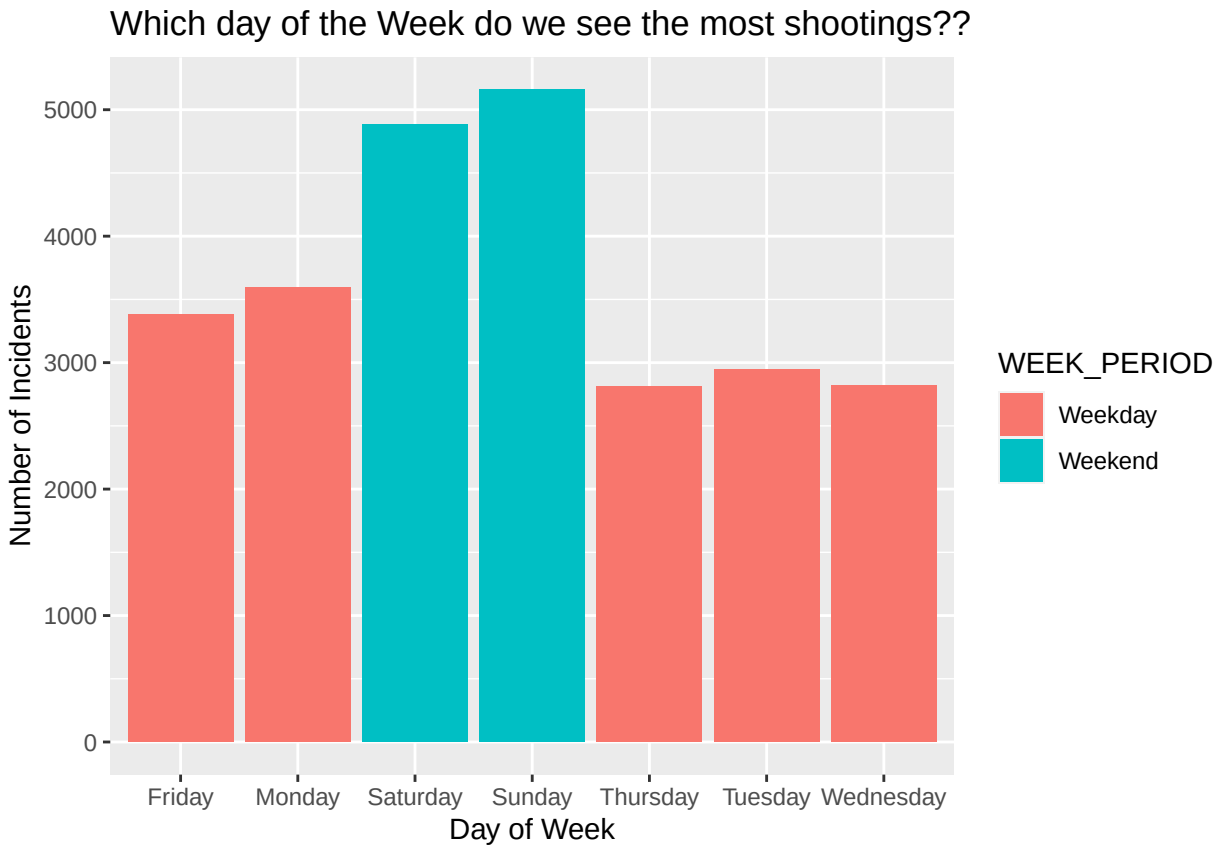
Questions Raised : We observe that there was a decline in total shootings for the year 2018-19 and then an uptick post 2020. What were the events that may have contributed to this? Also, post Covid wave of 2020, with property value changing, and people working remote permanently, along with a change is homelessness, how had that impacted the number of shootings in the city?

Which Day do we get the most shootings? Is it more on Weekdays or Weekends?

```
df$OCCUR_DAY <- weekdays(as.Date(df$OCCUR_DATE))
df_Day = df %>%
  group_by(OCCUR_DAY) %>%
  count()


df_Day$WEEK_PERIOD <- with(
  df_Day, ifelse(OCCUR_DAY == "Saturday" | OCCUR_DAY == "Sunday","Weekend", "Weekday"))

ggplot(df_Day, aes(x = OCCUR_DAY, y = n, fill=WEEK_PERIOD)) +
  geom_bar(stat="identity") +
  labs(title = "Which day of the Week do we see the most shootings??",
       x = "Day of Week",
       y = "Number of Incidents")
```



Observation : Shootings increase significantly(approx 25 percent) on weekends and then fall back on weekdays. Monday being the highest of the weekdays.

4. Model : Is there a correlation between murder and time of the day? Are there more reported murders for a particular time of the day?

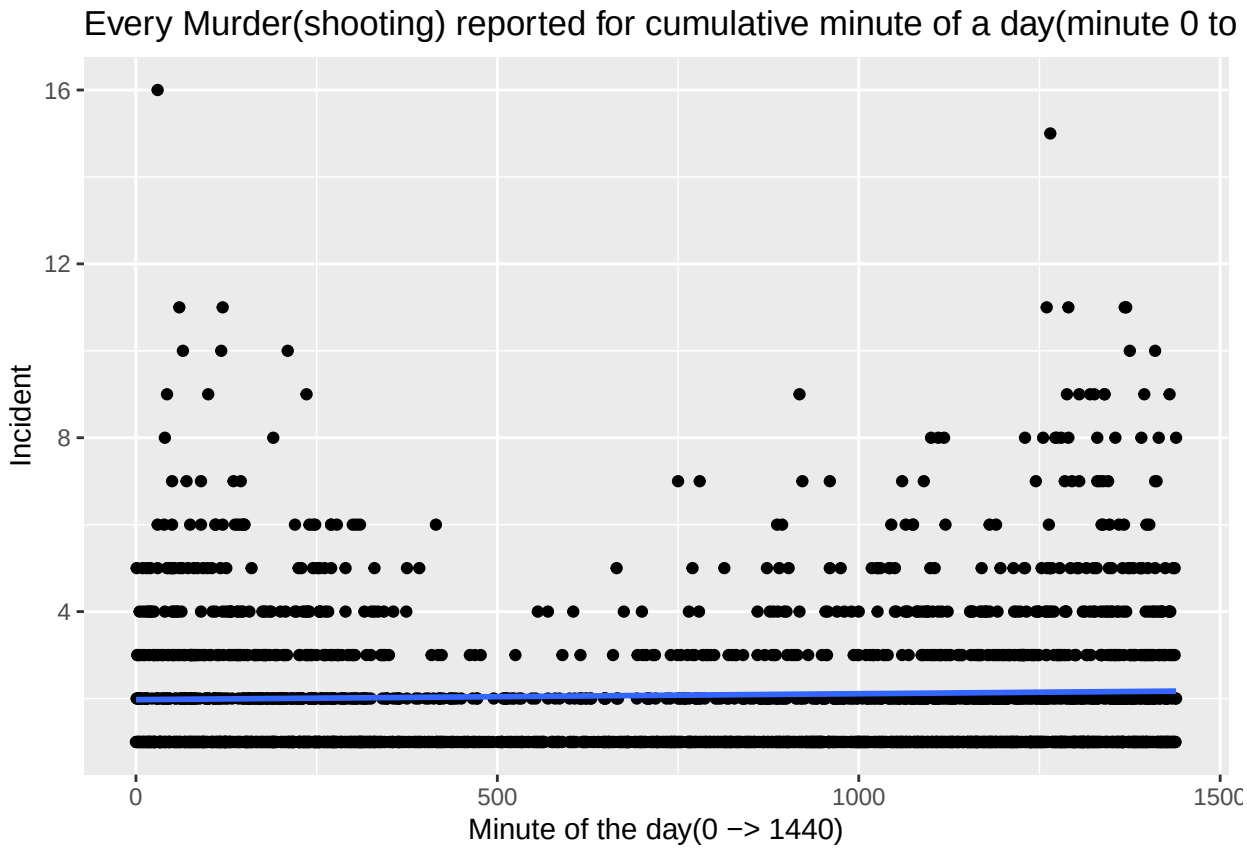
We will create a linear regression model for time of the day(represented by the minute of the day) and number of incidents reported on that minute for the city. The goal is to see a trend of how many reported incidents for a particular minute of the day(00:00 means minute 0 of the day and 23:59 means minute 1440 for a given day of 24 hours). The data will be split into number of incidents for a particular minute of the day. We will then find the p value for the model and check if our hypothesis is statistically significant.

```
df_mod = df%>%
     group_by(BORO,OCCUR_TIME,STATISTICAL_MURDER_FLAG) %>%
     count()

df_mod$hour <- hour(df_mod$OCCUR_TIME)
df_mod$minute_of_day <- hour(df_mod$OCCUR_TIME)*60 + minute(df_mod$OCCUR_TIME)
df_mod_m = filter(df_mod, STATISTICAL_MURDER_FLAG == TRUE )

 ggplot(df_mod_m, aes(x=minute_of_day, y=n)) +
     geom_point() +
     geom_smooth(method=lm, se=FALSE) +
  labs(title = "Every Murder(shooting) reported for cumulative minute of a day(minute 0 to 1440)",
       x = "Minute of the day(0 -> 1440)",
       y = "Incident")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Model Observation

We observe that the p-value is 0.0526 i.e p-value is greater tha 0.05. hence we cannot properly justify our hypothesis about the co-relation between time and reported murders. The model is dependent of factors such as day of the week, time of the year, neighbourhood and socio-political factors such as pandemic, social climate of the region etc and just time of the day is not significant for predicting crime.

```
mod <- lm(minute_of_day ~ n, data=df_mod_m)

mod
```

```
##
## Call:
## lm(formula = minute_of_day ~ n, data = df_mod_m)
##
## Coefficients:
## (Intercept)            n
##      721.06        11.62
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = minute_of_day ~ n, data = df_mod_m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -876.93 -494.85   91.32  446.32  705.32
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  721.061     15.830  45.551   <2e-16 ***
## n             11.617      5.991   1.939   0.0526 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 477.1 on 2372 degrees of freedom
## Multiple R-squared:  0.001582,   Adjusted R-squared:  0.001162
## F-statistic:  3.76 on 1 and 2372 DF,  p-value: 0.05262
```

Conclusion : Addressing Bias

It was unexpected to not see a direct correlation between time of the day and reported murders. It was my intuition that we would see a stronger relation between the two and the model would imply that both are directly relation i.e more murders likely to be reported during night time. It shows that this a a complicated idea with way too many dependent factors such as time of the year/socio-political climate in the country for that particular year, wealth inequality of the neighborhood instead of just looking at these two fields and building a citywide model.

As a personal observation, I would have also expected Manhattan to have the most crime due to it being the financial hub and less residential than Brooklyn or Bronx. For the victims, there are a lot more Male victims than Women which is something I also would not have predicted.

I find it a slippery slope to go into the race/gender of perpetrator because without knowledge of sociology and context, forming patterns for racial identity can be misunderstood/weaponized and that is something I would like to understand more not just by looking at this data but alos by being more informed about various factors around hopw to understand racial data.