NYPD Shooting Incident Project

12/06/2023

Import Library

```
library(tidyverse)
library(lubridate)
library(scales)
```

Importing NYPD Data

Read csv format of NYPD Shooting Incident Data form NYPD catalog Data

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

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

```
nypd_input
```

```
## # A tibble: 27,312 x 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
##  1    228798151 05/27/2021 21:30      QUEENS   <NA>                   105
##  2    137471050 06/27/2014 17:40      BRONX    <NA>                    40
##  3    147998800 11/21/2015 03:56      QUEENS   <NA>                   108
##  4    146837977 10/09/2015 18:30      BRONX    <NA>                    44
##  5     58921844 02/19/2009 22:58      BRONX    <NA>                    47
##  6    219559682 10/21/2020 21:36      BROOKLYN <NA>                    81
##  7     85295722 06/17/2012 22:47      QUEENS   <NA>                   114
##  8     71662474 03/08/2010 19:41      BROOKLYN <NA>                    81
##  9     83002139 02/05/2012 05:45      QUEENS   <NA>                   105
## 10     86437261 08/26/2012 01:10      QUEENS   <NA>                   101
## # i 27,302 more rows
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, 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>
```

Summary of NYPD data

```
summary(nypd_input)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms         Class :character
##  Median : 90372218   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE             X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

TIDY

Keeping fields that are needed for my analysis and removing others as first step and using mutate change OCCUR_DATE datatype from character to Date.

I am using below fields for my analysis. - INCIDENT_KEY - PERP_SEX - VIC_SEX - PERP_AGE_GROUP - VIC_AGE_GROUP - BORO - OCCUR_TIME - OCCUR_DATE - STATISTICAL_MURDER_FLAG

```
nypd_input <- nypd_input %>%
select(-c(LOCATION_DESC,PRECINCT,PERP_RACE,JURISDICTION_CODE,VIC_RACE,
          X_COORD_CD,Y_COORD_CD,Latitude,Longitude)) %>%
mutate(OCCUR_DATE = mdy(OCCUR_DATE))

nypd_input
```

```
## # A tibble: 27,312 x 12
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO  LOC_OF_OCCUR_DESC LOC_CLASSFCTN_DESC
##           <dbl> <date>     <time>     <chr> <chr>             <chr>
##  1    228798151 2021-05-27 21:30      QUEE~ <NA>              <NA>
##  2    137471050 2014-06-27 17:40      BRONX <NA>              <NA>
##  3    147998800 2015-11-21 03:56      QUEE~ <NA>              <NA>
##  4    146837977 2015-10-09 18:30      BRONX <NA>              <NA>
##  5     58921844 2009-02-19 22:58      BRONX <NA>              <NA>
##  6    219559682 2020-10-21 21:36      BROO~ <NA>              <NA>
##  7     85295722 2012-06-17 22:47      QUEE~ <NA>              <NA>
##  8     71662474 2010-03-08 19:41      BROO~ <NA>              <NA>
##  9     83002139 2012-02-05 05:45      QUEE~ <NA>              <NA>
## 10     86437261 2012-08-26 01:10      QUEE~ <NA>              <NA>
## # i 27,302 more rows
## # i 6 more variables: STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, Lon_Lat <chr>
```

summary of nypd data after above step

```
  summary(nypd_input)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:27312      Length:27312
##  1st Qu.: 63860880   1st Qu.:2009-07-18   Class1:hms        Class :character
##  Median : 90372218   Median :2013-04-29   Class2:difftime   Mode  :character
##  Mean   :120860536   Mean   :2014-01-06   Mode  :numeric
##  3rd Qu.:188810230   3rd Qu.:2018-10-15
##  Max.   :261190187   Max.   :2022-12-31
##  LOC_OF_OCCUR_DESC  LOC_CLASSFCTN_DESC STATISTICAL_MURDER_FLAG
##  Length:27312       Length:27312       Mode :logical
##  Class :character   Class :character   FALSE:22046
##  Mode  :character   Mode  :character   TRUE :5266
##
##
##
##  PERP_AGE_GROUP       PERP_SEX         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##    Lon_Lat
##  Length:27312
##  Class :character
##  Mode  :character
##
##
##
```

ANALYSIS

Total incidents Borough wise and year

```
Incidents_by_boro_year <- nypd_input %>%
  mutate(year = lubridate::year(OCCUR_DATE)) %>%
    group_by(BORO, year) %>%
    summarize(total_incidents_by_year = n()) %>%
ungroup()
```

```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```

```
Incidents_by_boro_year
```

```
## # A tibble: 85 x 3
##    BORO   year total_incidents_by_year
##    <chr> <dbl>                   <int>
##  1 BRONX  2006                     568
##  2 BRONX  2007                     533
##  3 BRONX  2008                     520
##  4 BRONX  2009                     529
##  5 BRONX  2010                     525
##  6 BRONX  2011                     571
##  7 BRONX  2012                     531
##  8 BRONX  2013                     371
##  9 BRONX  2014                     446
## 10 BRONX  2015                     409
## # i 75 more rows
```

filtering

```
filter_by_year <- Incidents_by_boro_year %>% filter(year == '2020')
filter_by_year
```

```
## # A tibble: 5 x 3
##   BORO           year total_incidents_by_year
##   <chr>         <dbl>                   <int>
## 1 BRONX          2020                     504
## 2 BROOKLYN       2020                     819
## 3 MANHATTAN      2020                     272
## 4 QUEENS         2020                     303
## 5 STATEN ISLAND  2020                      50
```

slicing

#Now, calculate Total Incidents by Borough and the year having maximum incidents

```
Incidents_by_boro <- nypd_input %>%
group_by(BORO,year = lubridate::year(OCCUR_DATE))  %>%
summarize(total_incidents = n()) %>%
mutate(max_year = year[which.max(total_incidents)]) %>%
summarize(total_incidents = sum(total_incidents),
year_with_max_incidents = first(max_year)) %>%
select(BORO, total_incidents, year_with_max_incidents) %>%
ungroup()
```

```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```

```
Incidents_by_boro
```

```
## # A tibble: 5 x 3
##   BORO          total_incidents year_with_max_incidents
##   <chr>                   <int>                   <dbl>
## 1 BRONX                    7937                    2021
## 2 BROOKLYN                10933                    2006
## 3 MANHATTAN                3572                    2021
## 4 QUEENS                   4094                    2008
## 5 STATEN ISLAND             776                    2008
```

joining

Now, I am joining this dataframe 'Incidents_by_boro' with 'Incidents_by_boro_year' to get result borough wise total incidents along with year that has maximum incidents and incident count in that year borough wise

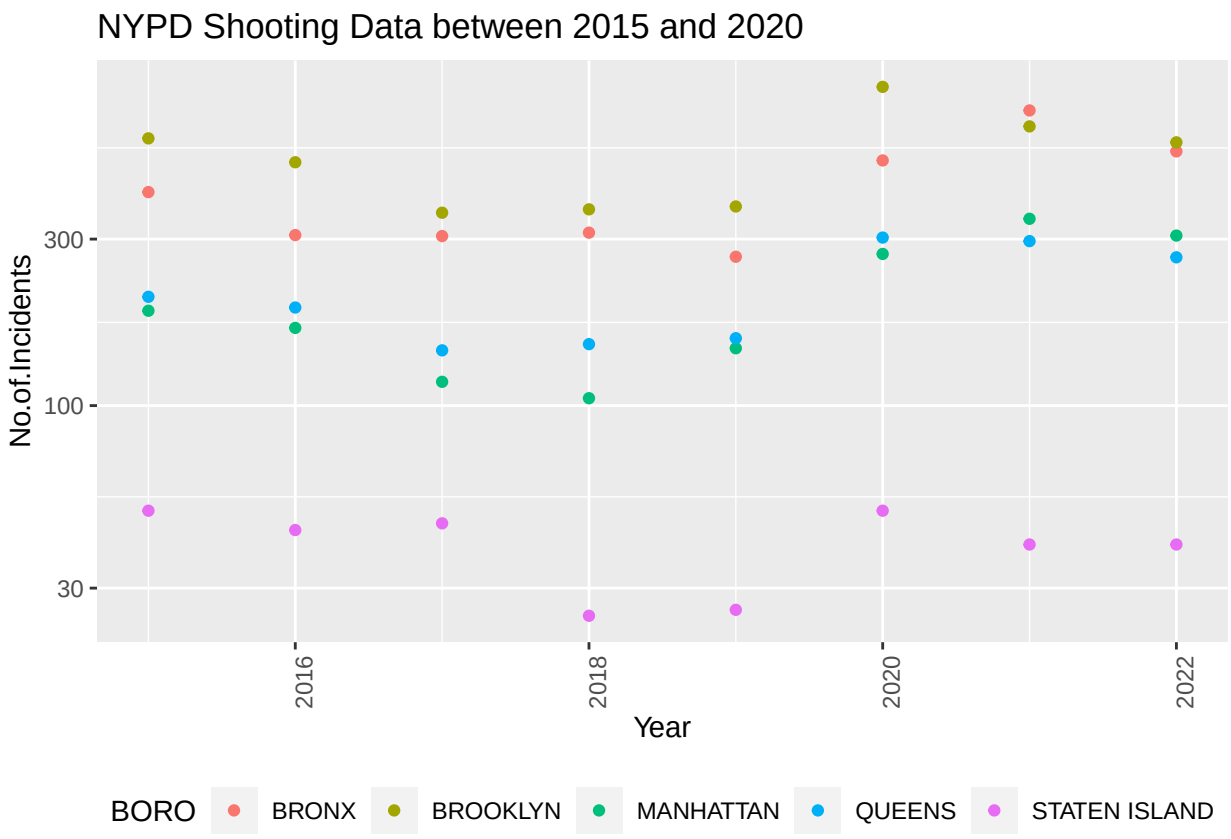
```
final_incidents_rep_boro <- Incidents_by_boro %>%
  left_join(Incidents_by_boro_year ,by = c("BORO", "year_with_max_incidents"="year"))
final_incidents_rep_boro
```

```
## # A tibble: 5 x 4
##   BORO          total_incidents year_with_max_incidents total_incidents_by_year
##   <chr>                   <int>                   <dbl>                    <int>
## 1 BRONX                    7937                    2021                      701
## 2 BROOKLYN                10933                    2006                      850
## 3 MANHATTAN                3572                    2021                      343
## 4 QUEENS                   4094                    2008                      326
## 5 STATEN ISLAND             776                    2008                       69
```

VISUALIZATION

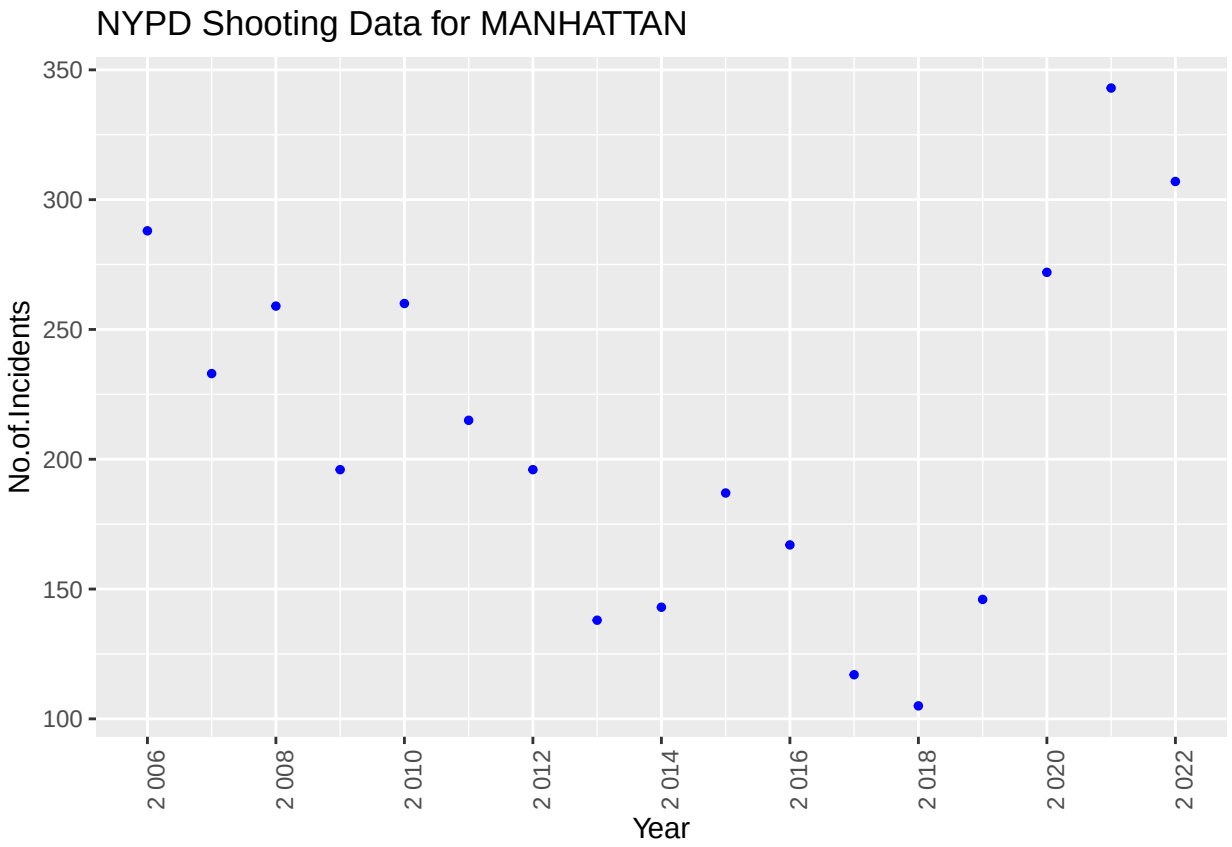
Now, lets start visualization data in different ways. First , lets visualize no.of incidents between year 2015 and 2020 by Borough wise For that, we already have Incidents_by_boro_year that was derived in previous steps

```
year_input <- 2014
  Incidents_by_boro_year %>% filter(year > year_input) %>%
           ggplot(aes(x=year,y= total_incidents_by_year, color = BORO)) +
           geom_point() +
           ggtitle("NYPD Shooting Data between 2015 and 2020") +
           xlab("Year") + ylab("No.of.Incidents") +
           scale_color_discrete(name = "BORO") +
  scale_y_log10() +
  theme(legend.position="bottom", axis.text.x = element_text(angle = 90))
```



incidents-year-Borough plot

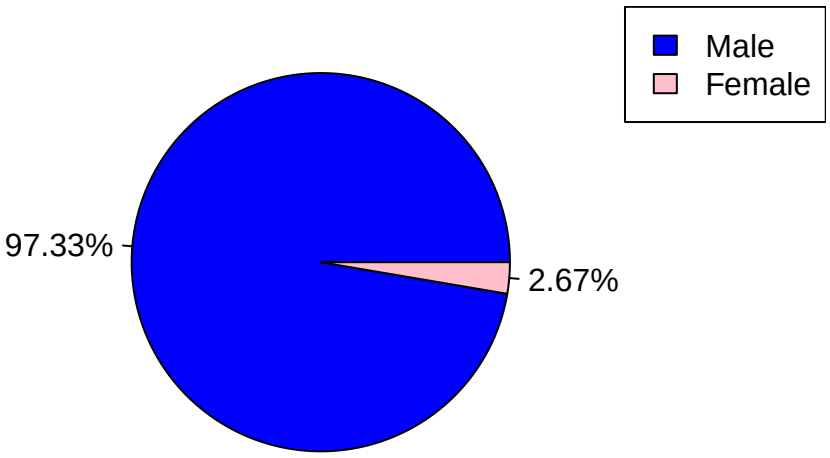
```
### plot no.of incidents borough wise and year


boro_input <- "MANHATTAN"
  Incidents_by_boro_year %>% filter(BORO == boro_input) %>%
           ggplot(aes(x=year,y= total_incidents_by_year)) +
           geom_point(na.rm=TRUE, color="blue", size=1) +
           ggtitle("NYPD Shooting Data for MANHATTAN") +
           xlab("Year") + ylab("No.of.Incidents") +
  theme(legend.position="bottom", axis.text.x = element_text(angle = 90)) +
  scale_x_continuous(breaks = seq(2000, 2025, by = 2),
                     labels = scales::number_format())
```



From above analysis, Observed that no.of shooting incidents have been increased from 2021 though the incidents number drop during 2017- 2018

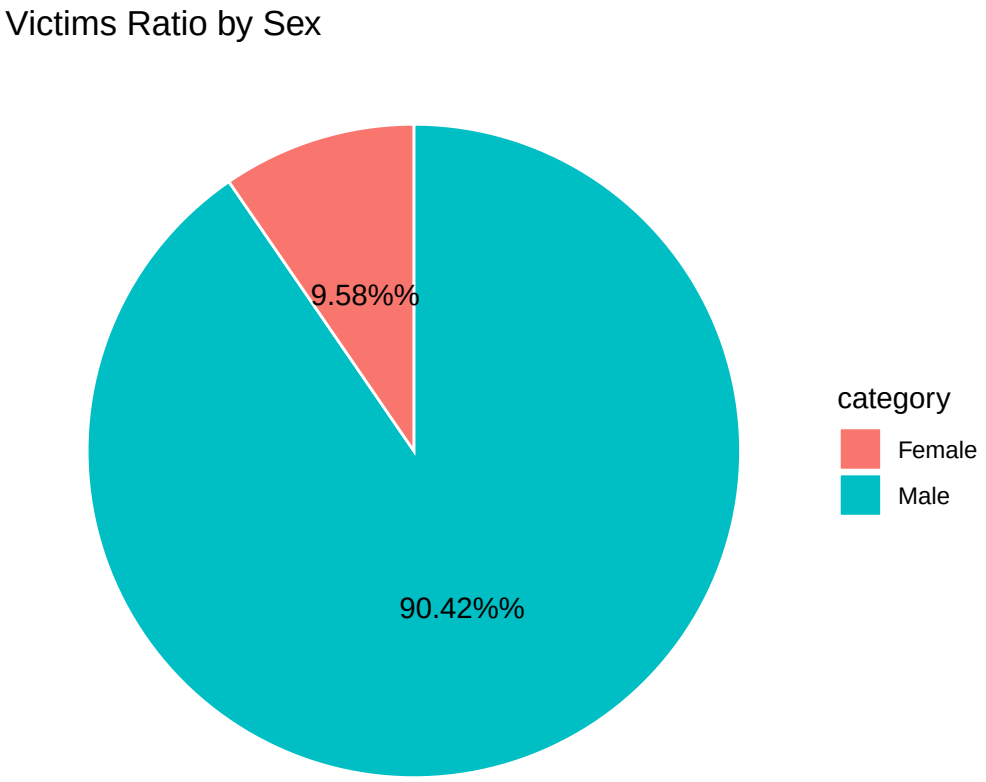
perpetrators ratio based on sex

```
x <- c("Male","Female")
y <- c(nrow(nypd_input %>%
    filter(PERP_SEX == 'M')),nrow(nypd_input %>% filter(PERP_SEX == 'F')))
perc <- paste0(round(100 * y/sum(y), 2), "%")
colors <- c('blue','pink')
pie(y,label = perc, main = "perpetrators Ratio sex based", col = colors)
legend("topright", x, fill = colors)
```



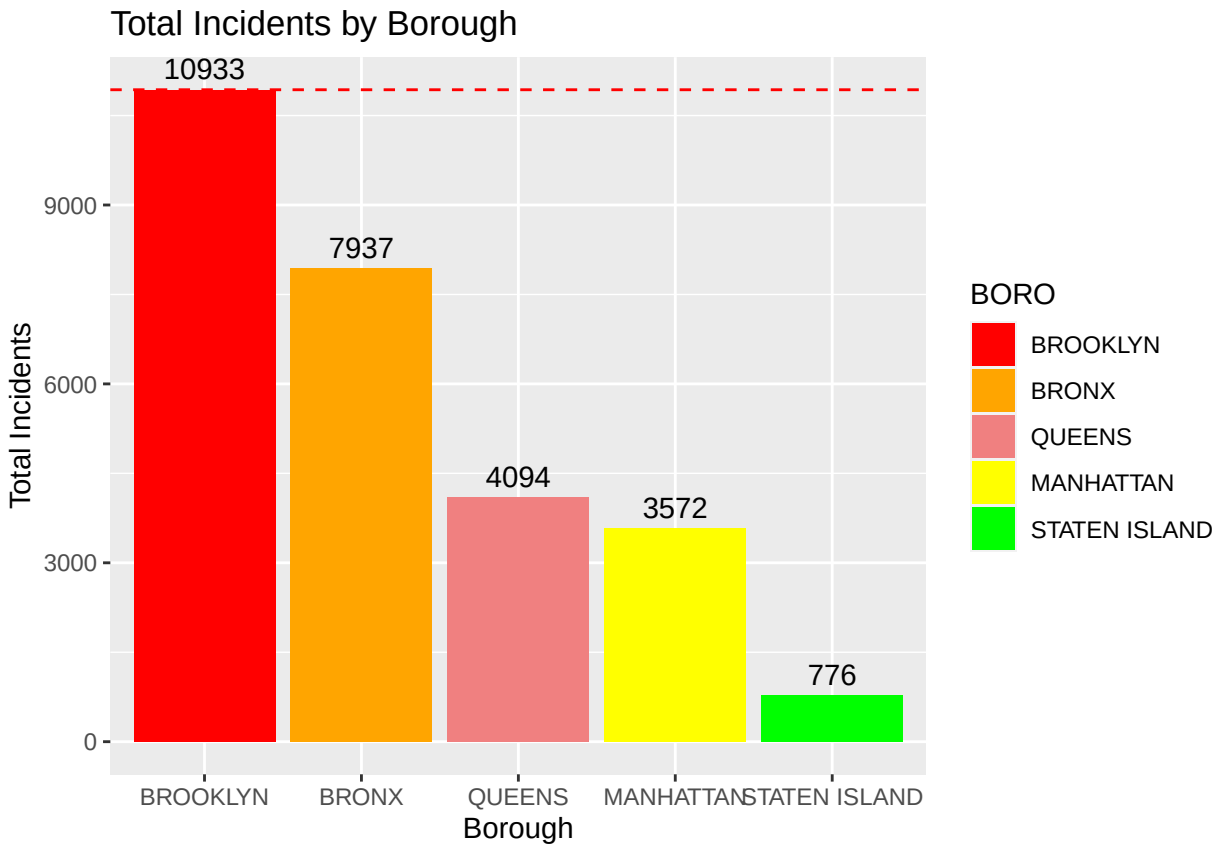
victims ratio based on sex

```
x <- c("Male","Female")
y <- c(nrow(nypd_input %>% filter(VIC_SEX == 'M')),nrow(nypd_input %>%
filter(VIC_SEX == 'F')))
perc <- paste0(round(100 * y/sum(y), 2), "%")
data <- data.frame(category = x, value = y, perc = perc)
ggplot(data, aes(x = "", y = value, fill = category)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y") +
  theme_void() +
  geom_text(aes(label = paste0(perc, "%")), position = position_stack(vjust = 0.5)) +
  ggtitle("Victims Ratio by Sex")
```



Total incidents by Borough wise ranking

```
Incidents_by_boro_ranked <- Incidents_by_boro %>%
  arrange(desc(total_incidents)) %>%
  mutate(BORO = factor(BORO, levels = BORO)) %>%
  mutate(boro_ranking = row_number())

# Create a color palette
colors <- c("red", "orange", "lightcoral", "yellow", "green")

# Plotting
ggplot(Incidents_by_boro_ranked, aes(x = BORO, y = total_incidents, fill = BORO)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = colors) +
  geom_text(aes(label = total_incidents), vjust = -0.5) +
  geom_hline(yintercept = max(Incidents_by_boro_ranked$total_incidents), linetype = "dashed", color = "red")
  ggtitle("Total Incidents by Borough") +
  xlab("Borough") +
  ylab("Total Incidents")
```



LINEAR MODEL based on gender to year

A Linear Mode which predicts the no.of incidents by victim gender=‘M’ by year. This model uses the existing data to predict the outcome, which has been compared with the real outcomes.

```
incident_count_vic_sex <- nypd_input %>% filter(STATISTICAL_MURDER_FLAG == 'TRUE') %>%
  group_by(BORO,year = lubridate::year(OCCUR_DATE)) %>%
  summarise(total_incidents = n(),
            total_incidents_M = sum(VIC_SEX == 'M', na.rm = TRUE),
            total_incidents_F = sum(VIC_SEX == 'F', na.rm = TRUE)) %>%
select(BORO, year, total_incidents_M, total_incidents_F) %>%
ungroup()
```

```
## ‘summarise()‘ has grouped output by ’BORO’. You can override using the
## ‘.groups‘ argument.
```

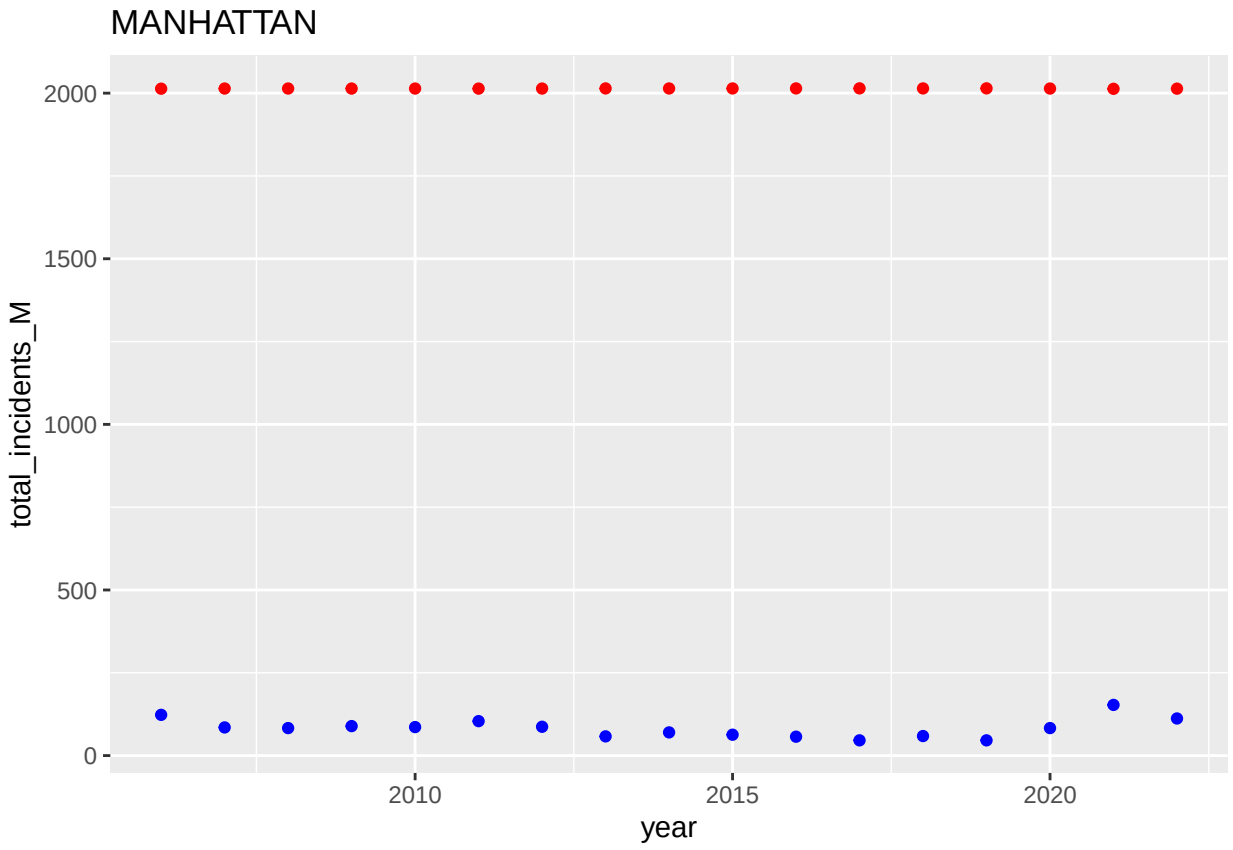
```
incident_count_vic_sex
```

```
## # A tibble: 85 x 4
##    BORO   year total_incidents_M total_incidents_F
##    <chr> <dbl>             <int>             <int>
##  1 BRONX  2006               123                14
##  2 BRONX  2007                85                 8
##  3 BRONX  2008                83                 6
##  4 BRONX  2009                89                 9
##  5 BRONX  2010                86                10
##  6 BRONX  2011               104                10
```

```
##  7 BRONX  2012                87                4
##  8 BRONX  2013                58                4
##  9 BRONX  2014                70                6
## 10 BRONX  2015                63                8
## # i 75 more rows
```

using the data above, below is the prediction for Borough BRONX

```
#prediction
mod_data <- incident_count_vic_sex %>% filter(BORO == "BRONX")
mod <- lm(year ~ total_incidents_M,data = mod_data)
pred_data <- mod_data %>% mutate(pred = predict(mod))
pred_data %>% ggplot() +
 ggtitle("MANHATTAN") +
 geom_point(aes(x = year,y = total_incidents_M),color = "blue") +
 geom_point(aes(x = year,y = pred),color = "red")
```



BIAS

Inaccuracies or missing data can introduce bias. certain incidents may be excluded based on certain criteria. With increasing data volume, differences in reporting rates over the years can lead to biased trends. Sometimes changes in law or policies might lead to variations in how incidents are recorded over time.

Project Conclusion

By analyzing the pie charts, graphs, and models I generated above, I am able to find the rise or fall in the number of incidents based on boroughs and then by states by year. I identified the borough with the highest number of incidents

overall and also visualized the ratio of each gender in both victims and perpetrators compared to the other.