NYPD Shooting Incident

N. Shrestha

2023-12-07

Introduction

This document provides an analysis of the shooting incident data from data.gov. This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the 2022. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included.

We will start by understanding the scope of the problem, followed by the methods we used to dissect the data, employing robust data visualization. We'll uncover key findings, including temporal and spatial patterns of shooting incidents, the demographics of victims, and highlight the critical issue of missing data. Without clear insights into the demographics of victims, including age and gender, as well as the geographical distribution of these incidents, policymakers and law enforcement are at a disadvantage when it comes to crafting effective crime prevention strategies.

```r
## Importing the tidyverse package
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Importing NYPD Shooting Incident (Historic) Data from DATA.GOV

We will begin by importing the dataset

```r
## Get the url
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
```

Let's read in the data to see what we have

```r
shooting_data <- read_csv(url[1])
```

Initial Data Exploration

```
## viewing the first few rows of the data
head(shooting_data)
```

```
## # A tibble: 6 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
## # 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>
```

Data Summary

Before cleaning the data, let’s add a summary to it to understand its structure

```
summary(shooting_data)
```

```
##   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.:1000028   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
```

```r
glimpse(shooting_data)
```

```
## Rows: 27,312
## Columns: 21
## $ INCIDENT_KEY            <dbl> 228798151, 137471050, 147998800, 146837977, 58~
## $ OCCUR_DATE              <chr> "05/27/2021", "06/27/2014", "11/21/2015", "10/~
## $ OCCUR_TIME              <time> 21:30:00, 17:40:00, 03:56:00, 18:30:00, 22:58~
## $ BORO                    <chr> "QUEENS", "BRONX", "QUEENS", "BRONX", "BRONX",~
## $ LOC_OF_OCCUR_DESC       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ PRECINCT                <dbl> 105, 40, 108, 44, 47, 81, 114, 81, 105, 101, 2~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 2~
## $ LOC_CLASSFCTN_DESC      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ LOCATION_DESC           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "MULTI DWE~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, ~
## $ PERP_AGE_GROUP          <chr> NA, NA, NA, NA, "25-44", NA, NA, NA, NA, "25-4~
## $ PERP_SEX                <chr> NA, NA, NA, NA, "M", NA, NA, NA, NA, "M", NA, ~
## $ PERP_RACE               <chr> NA, NA, NA, NA, "BLACK", NA, NA, NA, NA, "BLAC~
## $ VIC_AGE_GROUP           <chr> "18-24", "18-24", "25-44", "<18", "45-64", "25~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "WHITE", "WHITE HISPANIC", "~
## $ X_COORD_CD              <dbl> 1058925.0, 1005028.0, 1007667.9, 1006537.4, 10~
## $ Y_COORD_CD              <dbl> 180924.0, 234516.0, 209836.5, 244511.1, 262189~
## $ Latitude                <dbl> 40.66296, 40.81035, 40.74261, 40.83778, 40.886~
## $ Longitude               <dbl> -73.73084, -73.92494, -73.91549, -73.91946, -7~
## $ Lon_Lat                 <chr> "POINT (-73.73083868899994 40.662964620000025)~
```

Cleaning Data

Let's delete the columns we don't want and also convert the OCCUR_DATE to `<date>` type since it is originally in `<chr>` type.

```
shooting_data <- shooting_data %>%
  select(-c(LOC_OF_OCCUR_DESC:LOCATION_DESC, X_COORD_CD:Lon_Lat)) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))

## Viewing the cleaned data
head(shooting_data)
```

```
## # A tibble: 6 x 11
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO  STATISTICAL_MURDER_F~1 PERP_AGE_GROUP
##          <dbl> <date>     <time>     <chr> <lgl>                  <chr>
## 1    228798151 2021-05-27 21:30      QUEE~ FALSE                  <NA>
## 2    137471050 2014-06-27 17:40      BRONX FALSE                  <NA>
## 3    147998800 2015-11-21 03:56      QUEE~ TRUE                   <NA>
## 4    146837977 2015-10-09 18:30      BRONX FALSE                  <NA>
## 5     58921844 2009-02-19 22:58      BRONX TRUE                   25-44
## 6    219559682 2020-10-21 21:36      BROO~ TRUE                   <NA>
## # i abbreviated name: 1: STATISTICAL_MURDER_FLAG
## # i 5 more variables: PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>,
## #   VIC_SEX <chr>, VIC_RACE <chr>
```

Summary after Cleaning the Data

```
summary(shooting_data)
```

```
##   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
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Mode :logical           Length:27312       Length:27312
##  FALSE:22046             Class :character   Class :character
##  TRUE :5266              Mode  :character   Mode  :character
##
##
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```r
#just to see how the incidents are divided across the different boroughs
borough_count <- shooting_data %>%
  group_by(BORO) %>%
  summarise(count = n())
borough_count
```

```
## # A tibble: 5 x 2
##   BORO          count
##   <chr>         <int>
## 1 BRONX          7937
## 2 BROOKLYN      10933
## 3 MANHATTAN      3572
## 4 QUEENS         4094
## 5 STATEN ISLAND   776
```

```r
# Checking to see if the total number of data matches with the sum of borogh_count
total_incidents <- sum(borough_count$count)
total_incidents
```

```
## [1] 27312
```

Handling Missing Data

First we want to see the missing values in our data and from there we will decide what to do with it

```r
## Find the number of missing data in each category in our dataset
missing_data_summary <- sapply(shooting_data, function(x) sum(is.na(x)))
## Display the number of missing data
missing_data_summary
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                       0                       0                    9344
##                PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##                    9310                    9310                       0
##                 VIC_SEX                VIC_RACE
##                       0                       0
```

Most of the missing data is concentrated in specific fields, notably in details pertaining to the perpetrator, including their age group, sex, and race. This absence of information could stem from various factors, with one of the plausible explanations being that the perpetrator has not yet been apprehended, thereby limiting the availability of these details. To handle this absence, I could segment the data into two subsets; one with known perpetrator details and the other one with the unknown ones.

```r
#Creating indicator variables for the missing data
shooting_data$missing_age_group <- ifelse(is.na(shooting_data$PERP_AGE_GROUP), 1, 0)
shooting_data$missing_sex <- ifelse(is.na(shooting_data$PERP_SEX), 1, 0)
shooting_data$missing_race <- ifelse(is.na(shooting_data$PERP_RACE), 1, 0)
```

```
#Checking the amount of missing information for each borough
missing_by_location <- shooting_data %>%
  group_by(BORO)%>%
  summarize(missing_age_count = sum(missing_age_group),
            missing_sex_count = sum(missing_sex),
            missing_race_count = sum(missing_race)) %>%
  pivot_longer(cols = starts_with("missing"),
               names_to = "missing_data_type",
               values_to = "count")

missing_by_location
```

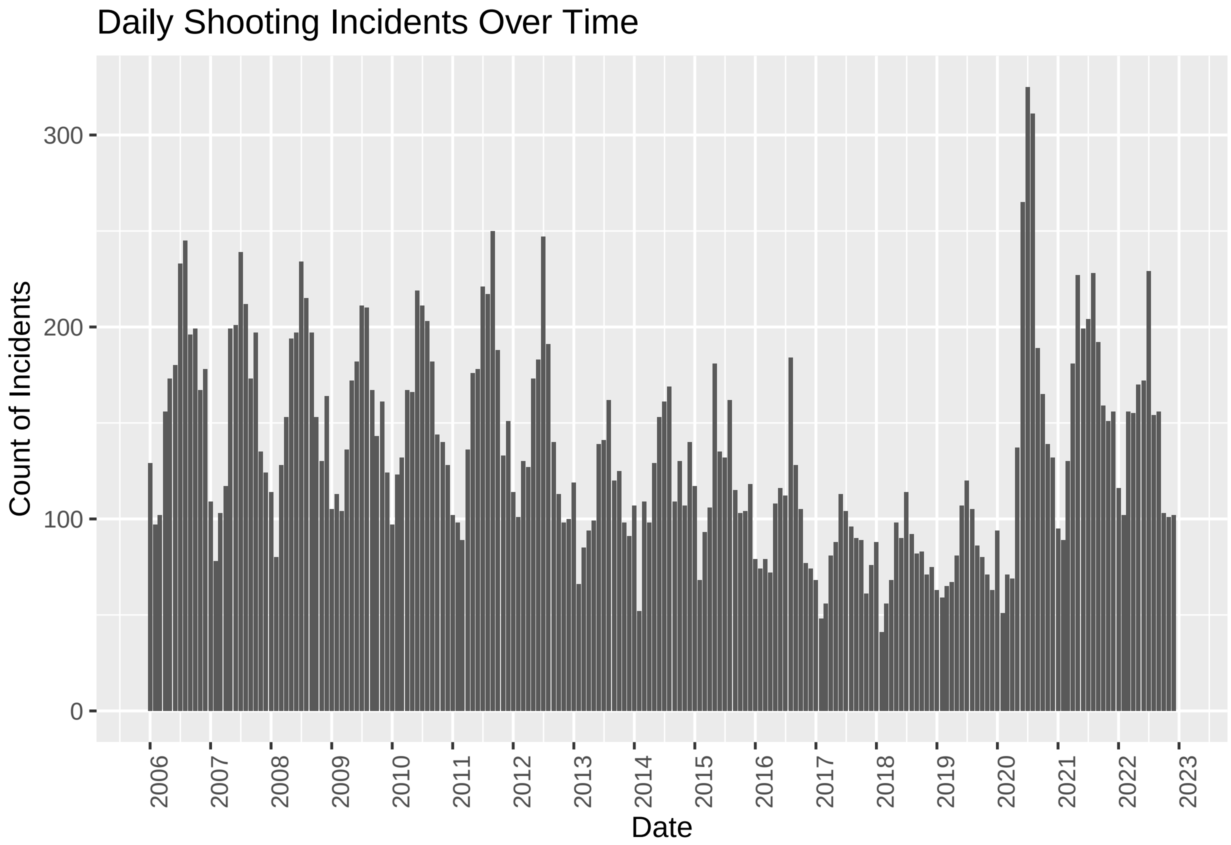
```
## # A tibble: 15 x 3
##    BORO          missing_data_type  count
##    <chr>         <chr>              <dbl>
##  1 BRONX         missing_age_count   2512
##  2 BRONX         missing_sex_count   2506
##  3 BRONX         missing_race_count  2506
##  4 BROOKLYN      missing_age_count   4291
##  5 BROOKLYN      missing_sex_count   4281
##  6 BROOKLYN      missing_race_count  4281
##  7 MANHATTAN     missing_age_count   1030
##  8 MANHATTAN     missing_sex_count   1024
##  9 MANHATTAN     missing_race_count  1024
## 10 QUEENS        missing_age_count   1366
## 11 QUEENS        missing_sex_count   1357
## 12 QUEENS        missing_race_count  1357
## 13 STATEN ISLAND missing_age_count    145
## 14 STATEN ISLAND missing_sex_count    142
## 15 STATEN ISLAND missing_race_count   142
```

```
#Summarizing the total count of missing and not missing data
borough_incident <- shooting_data %>%
  group_by(BORO) %>%
  summarise(total_incidents = n(),
    Missing_Count = sum(missing_sex),
    Non_Missing_Count = total_incidents - Missing_Count
  ) %>%
  pivot_longer(cols = c(Missing_Count, Non_Missing_Count),
               names_to = "Detail_Type",
               values_to = "Count")
```

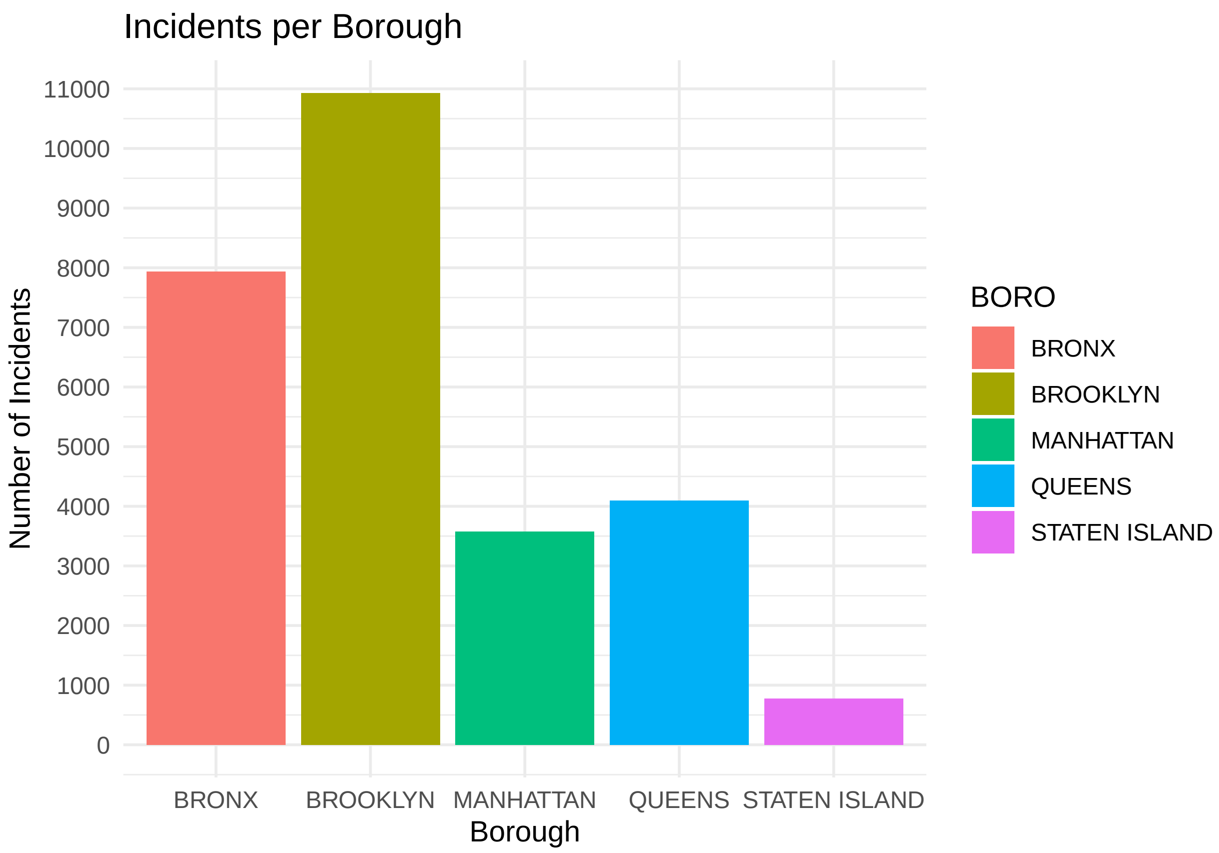
Visualization of Data

The first thing we're going to see is the total number of incidents by bourough. We will also look at the shooting incidents over time using a bar chart. I also want to see the number of missing data for the incidents for each borogh. We will also look at the age group and gender of victims per borough.

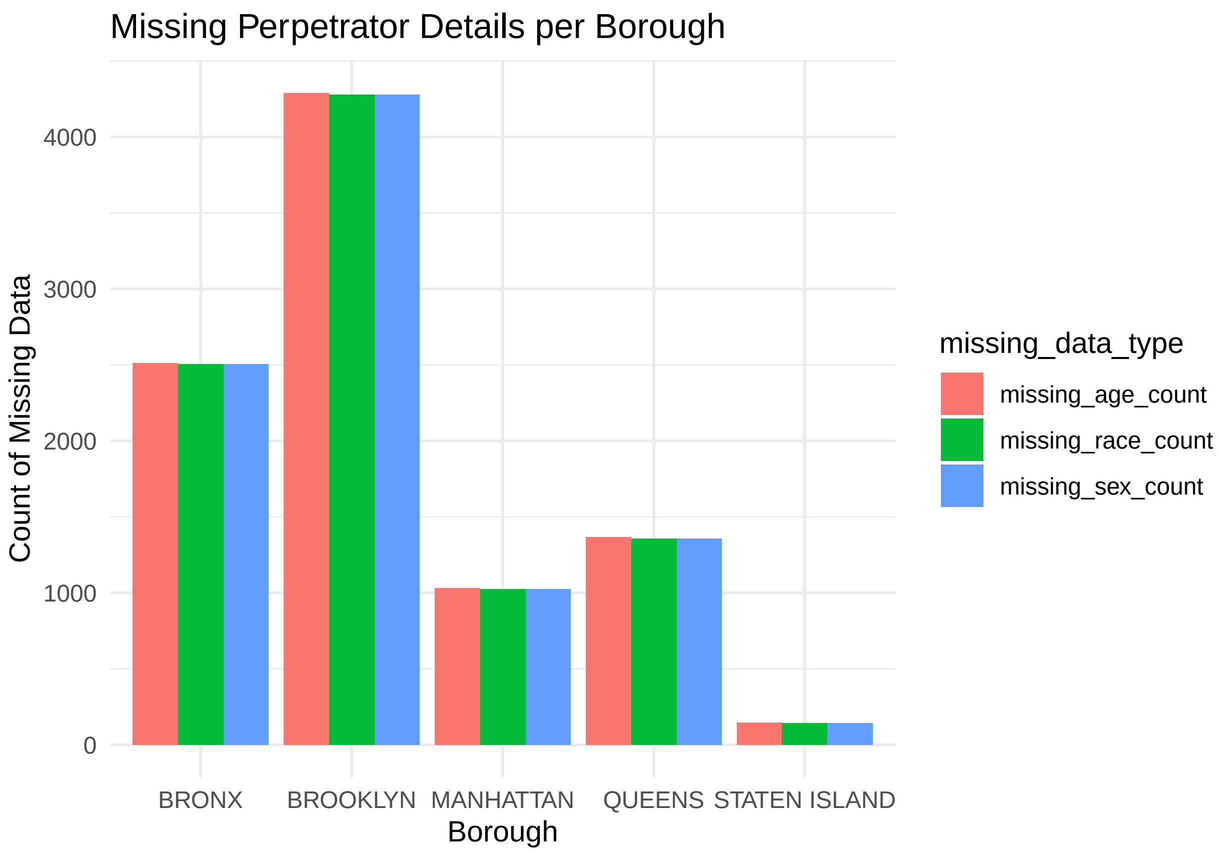
```
#Creating a count of shooting incidents grouped by month
monthly_counts <- shooting_data %>%
  mutate(month = floor_date(OCCUR_DATE, "month")) %>%
  group_by(month) %>%
  summarise(count = n())

# Creating a bar chart to look at the incident distribution over time
monthly_counts %>%
  ggplot(aes(x = month, y = count))+
  geom_col() +
  scale_x_date(date_breaks = "1 year", date_labels = "%Y")+
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Daily Shooting Incidents Over Time",
       x = "Date",
       y = "Count of Incidents")
```



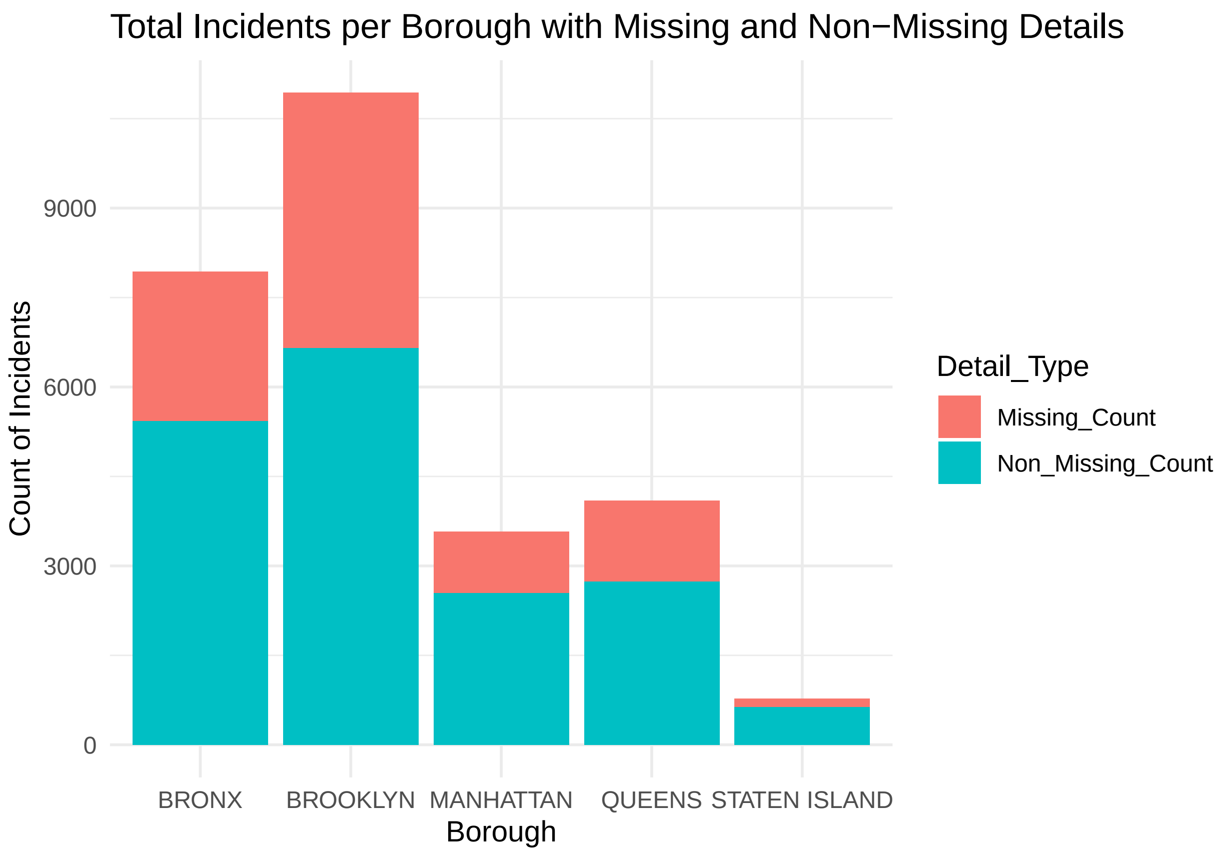
```
#Creating a bar chart to see the number of incidents per borough
borough_count %>%
  ggplot(aes(x = BORO, y = count, fill = BORO)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = seq(0, 11000, by = 1000))+
  theme_minimal() +
  labs(title = "Incidents per Borough",
       x = "Borough",
       y = "Number of Incidents")
```



```
#Creating a bar chart to see the number of missing data per borough
missing_by_location %>%
  ggplot(aes(x = BORO, y = count, fill = missing_data_type)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Missing Perpetrator Details per Borough",
       x = "Borough",
       y = "Count of Missing Data")
```



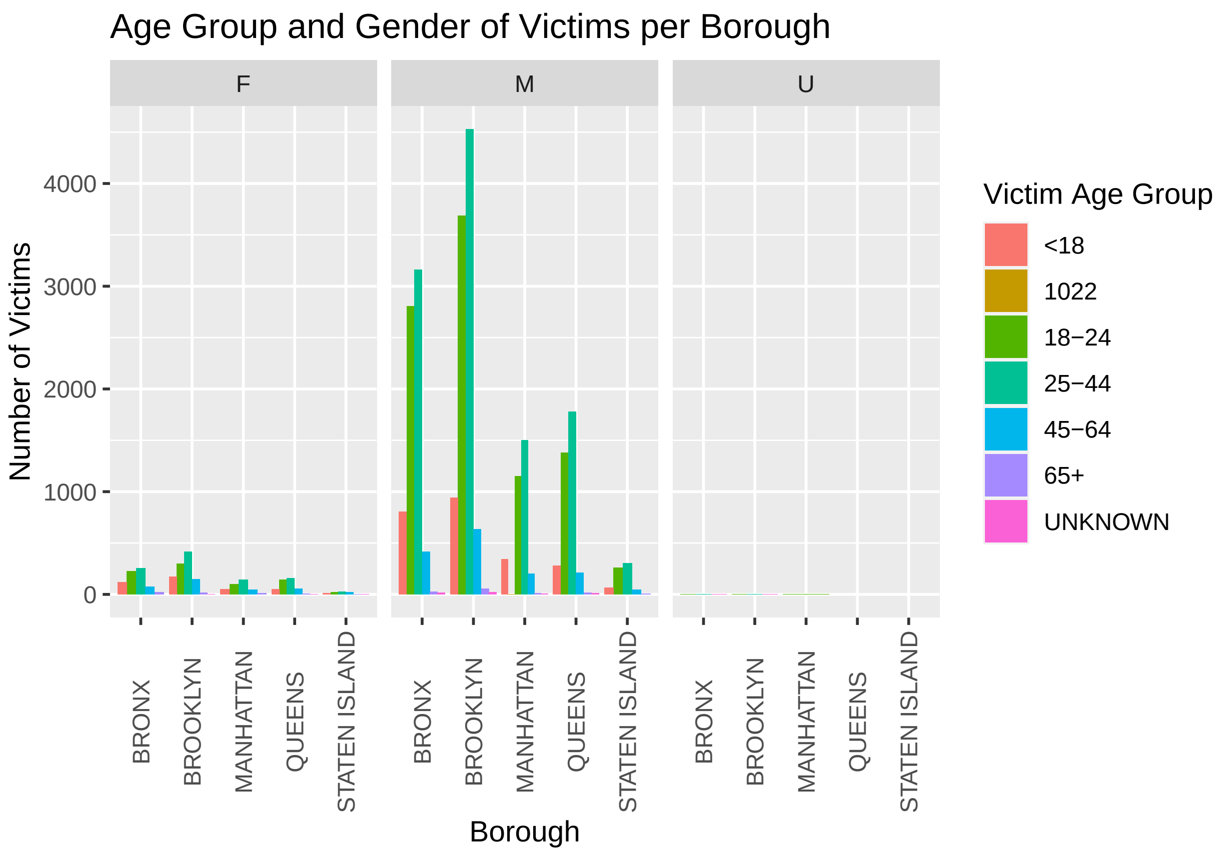
```
#Bar chart to see the missing data compared to the non-missing data per borough
borough_incident %>%
  ggplot(aes(x = BORO, y = Count, fill = Detail_Type)) +
  geom_bar(stat = "identity", position = "stack") +
  theme_minimal() +
  labs(title = "Total Incidents per Borough with Missing and Non-Missing Details",
       x = "Borough",
       y = "Count of Incidents")
```



```r
#Creating a visualization to examine the age group and gender of victims per borough
# Aggregate the data
agg_data <- shooting_data %>%
  group_by(BORO, VIC_AGE_GROUP, VIC_SEX) %>%
  summarise(count = n()) %>%
  ungroup()
```

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

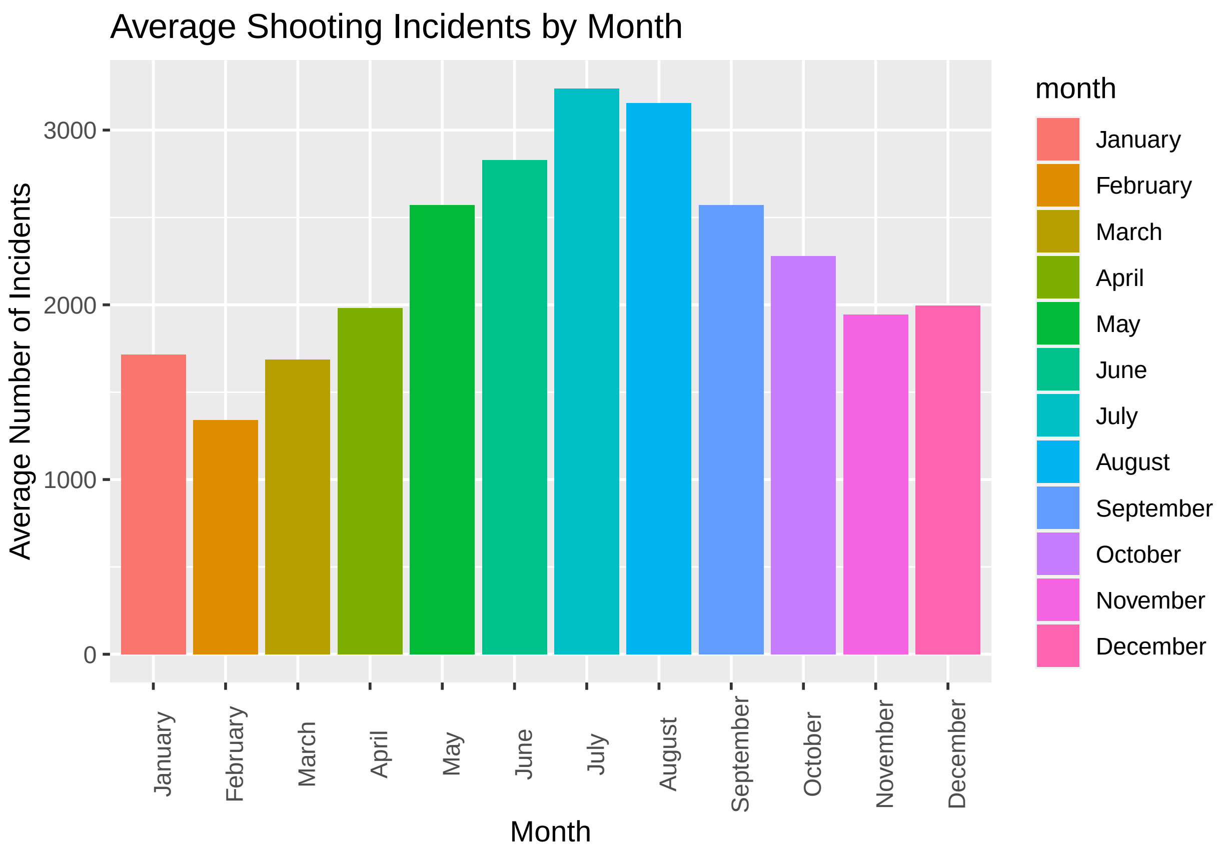
```r
# Create a stacked bar chart
ggplot(agg_data, aes(x = BORO, y = count, fill = VIC_AGE_GROUP)) +
  geom_bar(stat = "identity", position = position_dodge()) +  # Use position_dodge to separate bars for
  facet_wrap(~VIC_SEX) +  # Add this line to create a separate plot for each gender
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +  # Rotate x-axis labels to vertical
  labs(title = "Age Group and Gender of Victims per Borough",
       x = "Borough",
       y = "Number of Victims",
       fill = "Victim Age Group")
```



Analysis of Data After Visualization

1. There seems to be a pattern to the shooting incidents. It seems like there are certain months that the shooting incidents are the highest and it always seems to decrease towards the end of the year.
2. To do a fair analysis of number of incidents per borough, I might need to account for the population size of each borough.

```
## Getting the average incident per month
shooting_data$month <- format(shooting_data$OCCUR_DATE, "%m")
monthly_average <- shooting_data %>%
  group_by(month) %>%
  summarise(avg_incident = mean(n()))

#Converting the numbered month to the name of the month
monthly_average <- monthly_average %>%
  mutate(month = factor(month, levels = sprintf("%02d", 1:12), labels = base::month.name))

#Bar chart showing the average number of incidents per month
monthly_average %>%
  ggplot(aes(x = month, y = avg_incident, fill = month)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Average Shooting Incidents by Month",
       x = "Month",
       y = "Average Number of Incidents")
```



Modeling Data

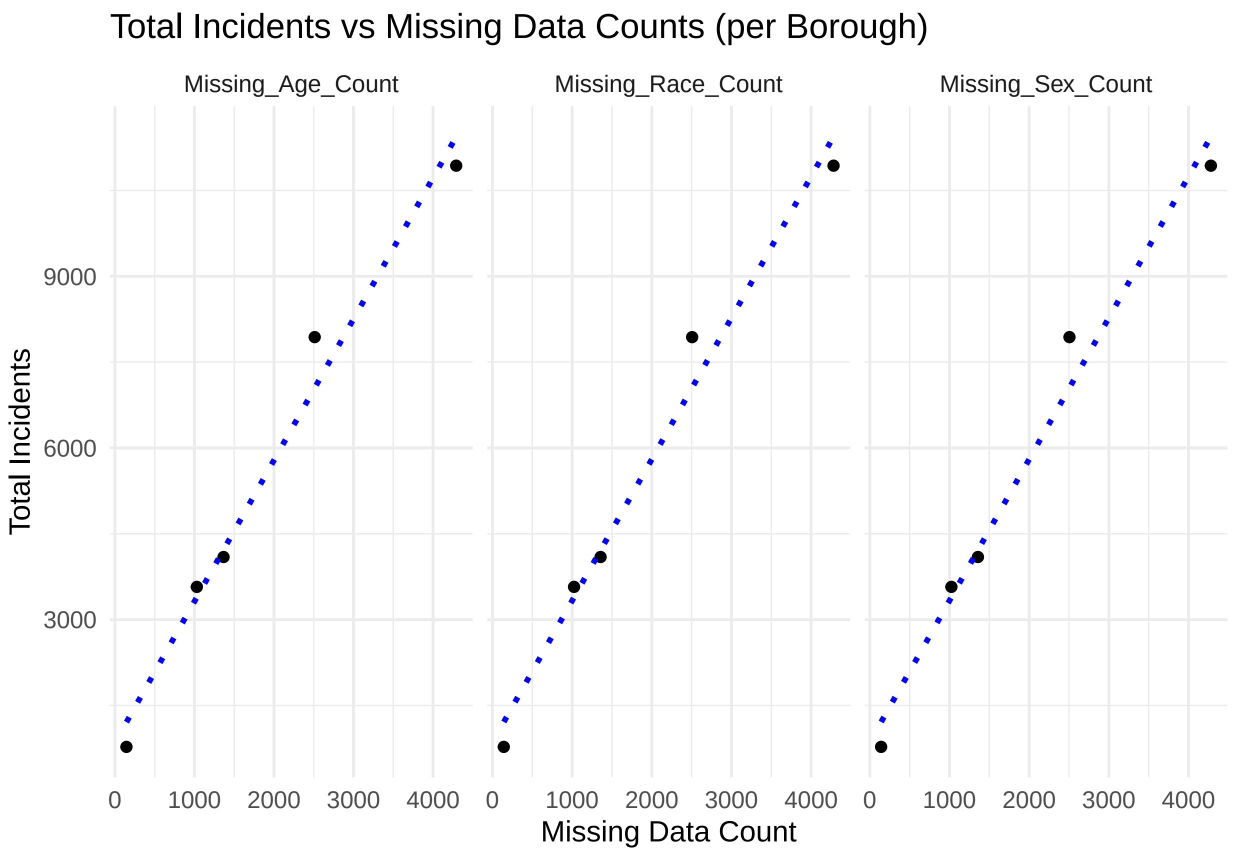
```
# Dividing up the data according to missing and non-missing variables for each category
borough_data <- shooting_data %>%
  group_by(BORO) %>%
  summarise(
    Total_Incidents = n(),
    Missing_Age_Count = sum(missing_age_group),
    Non_Missing_Age_Count = Total_Incidents - Missing_Age_Count,
    Missing_Sex_Count = sum(missing_sex),
    Non_Missing_Sex_Count = Total_Incidents - Missing_Sex_Count,
    Missing_Race_Count = sum(missing_race),
    Non_Missing_Race_Count = Total_Incidents - Missing_Race_Count
  )

model <- lm(Total_Incidents ~ Missing_Age_Count + Missing_Sex_Count + Missing_Race_Count, data = boroug
summary(model)
```

```
##
## Call:
## lm(formula = Total_Incidents ~ Missing_Age_Count + Missing_Sex_Count +
##     Missing_Race_Count, data = borough_data)
##
## Residuals:
```

```
##      1      2      3      4      5
##  749.9 -504.7  204.6  111.5 -561.4
##
## Coefficients: (1 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1214.56    1079.59   1.125    0.377
## Missing_Age_Count    -81.06     210.63  -0.385    0.737
## Missing_Sex_Count     83.64     210.90   0.397    0.730
## Missing_Race_Count       NA         NA      NA       NA
##
## Residual standard error: 770.2 on 2 degrees of freedom
## Multiple R-squared:  0.9813, Adjusted R-squared:  0.9626
## F-statistic: 52.49 on 2 and 2 DF,  p-value: 0.0187
```

```r
long_borough_data <- borough_data %>%
  pivot_longer(cols = c("Missing_Age_Count", "Missing_Sex_Count", "Missing_Race_Count"),
               names_to = "Missing_Data_Type",
               values_to = "Missing_Count")

# Combined plot
ggplot(long_borough_data, aes(x = Missing_Count, y = Total_Incidents)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue", linetype = "dotted") +
  facet_wrap(~ Missing_Data_Type, scales = "free_x") +
  labs(title = "Total Incidents vs Missing Data Counts (per Borough)",
       x = "Missing Data Count", y = "Total Incidents") +
  theme_minimal()
```

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



Conclusion

Some of the key findings from our analysis of NYPD Shooting Incident (Historic) dataset from DATA.GOV include:

1. There is a noticeable fluctuation in shooting incidents over time, with certain months showing higher incident rates. This suggests a possible seasonal or temporal pattern that could be influenced by various external factors such as weather, holidays, or police activity.
2. Each borough exhibits a distinct pattern in terms of shooting incidents. However, a comprehensive analysis requires considering the population size of each borough to ensure fair comparisons.
3. A notable correlation exists between the number of incidents in a borough and the missing details on perpetrators. This could indicate areas with higher crime rates also face challenges in crime reporting and perpetrator identification.

The potential sources of biases include:

- The significant amount of missing data, especially regarding perpetrator details, could skew the analysis. This missing data might be non-random and could be related to the efficiency of law enforcement in different areas.
- The data is dependent on the accuracy and completeness of the NYPD's reporting. Any systemic biases in police reporting practices could affect the findings.
- Not accounting for population size and density in each borough may lead to misleading conclusions about the relative safety or risk in these areas.

It is very possible for personal biases to influence the analysis. Bias mitigation is crucial for ensuring the integrity and objectivity of the findings especially in areas such as crime statistics.