NYPD Shooting Incident Data (Historic)

Kath Nguyen

12/5/2021

## Project Step 1: Start an Rmd Document

For this portion of the project, the first step was to ensure that the `tidyverse` package would be accessible to run the codes necessary to analyze this data using the `library()` function. From there, the data set, “NYPD Shooting Incident Data (Historic)”, was imported using the url (https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD). The url was then assigned the variable “NYPD_data_url” to simplify the csv file to tibble conversion.

```
#Import NYPD Shooting Incident Data (Historic) URL

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     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
```

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

Thereafter, the **NYPD_data_url** variable was placed into a `read_csv()` function to properly convert the data csv file into a tibble.

```
#Import NYPD Shooting Incident Data (Historic) as csv
#Read as tibble
NYPD_data <- read_csv(NYPD_data_url)
```

```
## Rows: 23585 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.
```

```
NYPD_data
```

```
## # A tibble: 23,585 x 19
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     PRECINCT JURISDICTION_CODE
##           <dbl> <chr>      <time>     <chr>       <dbl>             <dbl>
##  1     24050482 08/27/2006 05:35      BRONX          52                 0
##  2     77673979 03/11/2011 12:03      QUEENS        106                 0
##  3    203350417 10/06/2019 01:09      BROOKLYN       77                 0
##  4     80584527 09/04/2011 03:35      BRONX          40                 0
##  5     90843766 05/27/2013 21:16      QUEENS        100                 0
##  6     92393427 09/01/2013 04:17      BROOKLYN       67                 0
##  7     73057167 06/05/2010 21:16      BROOKLYN       77                 0
##  8    211362213 03/20/2020 21:27      BROOKLYN       81                 0
##  9    137564752 07/04/2014 00:25      QUEENS        101                 0
## 10    147024011 10/18/2015 01:33      QUEENS        106                 0
## # ... with 23,575 more rows, and 13 more variables: 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>
```

This data set lists all shooting incident that occurred in New York City from last year dating back to 2006. Furthermore, the data set provides numeric and non-numeric data for each variable related to the shooting incidents. Thus far in this data set, there should be the following variables/factors:

- INCIDENT KEY
- OCCUR_DATE
- OCCUR_TIME
- BORO

- PRECINCT
- JURISDICTION_CODE
- LOCATION_DESC
- STATISTICAL_MURDER_FLAG

- PERP_AGE_GROUP
- PERP_SEX
- PERP_RACE
- VIC_AGE_GROUP
- VIC_SEX
- VIC_RACE
- X_COORD_CD
- Y_COORD_CD
- Latitude
- Longitude
- Lon_Lat

Given this data, it will be used to investigate whether race, sex, and borough location are factors that may have led to these shooting incidents over the years.

## Project Step 2: Tidy and Transform Data

For this second portion, data needs to be wrangled in a way that is cleaner and more concise. Before doing so, a preview of the data is necessary to check if there is missing data.

```
View(NYPD_data) #preview whole data
summary(NYPD_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23585       Length:23585      Length:23585
##  1st Qu.: 55322804   Class :character   Class1:hms        Class :character
##  Median : 83435362   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :102280741                      Mode  :numeric
##  3rd Qu.:150911774
##  Max.   :230611229
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.000     Length:23585       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.000     Class :character   FALSE:19085
##  Median : 69.00   Median :0.000     Mode  :character   TRUE :4500
##  Mean   : 66.21   Mean   :0.333
##  3rd Qu.: 81.00   3rd Qu.:0.000
##  Max.   :123.00   Max.   :2.000
##                   NA’s   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23585       Length:23585       Length:23585       Length:23585
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:23585       Length:23585       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999925   1st Qu.:182539
##  Mode  :character   Mode  :character   Median :1007654   Median :193470
##                                        Mean   :1009379   Mean   :207300
##                                         3rd Qu.:1016782   3rd Qu.:239163
##                                         Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23585
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Based on the data/summary and the purpose of this investigation, evidently coordinates, latitude/longitude, time, age, locations, codes, and incident key information will not be needed; therefore they will be removed. Furthermore, considering that majority of the data listed for perpetrators are N/A, it will not be used for analysis and will be removed as well. This first code helps tidy up all the excess data using the functions in the `tidyverse` library.

```
#Remove: INCIDENT_KEY, OCCUR_TIME, BORO, PRECINCT, JURISDICTION_CODE, LOCATION_DESC, STATISTICAL_MURDER_

tidy_NYPD_data <- NYPD_data %>%
  select(-c(INCIDENT_KEY,
            OCCUR_TIME,
            PRECINCT,
            JURISDICTION_CODE,
            LOCATION_DESC,
            STATISTICAL_MURDER_FLAG,
            PERP_AGE_GROUP,
            PERP_SEX,
            PERP_RACE,
            VIC_AGE_GROUP,
            X_COORD_CD,
            Y_COORD_CD,
            Latitude,
            Longitude,
            Lon_Lat))

tidy_NYPD_data
```

```
## # A tibble: 23,585 x 4
##    OCCUR_DATE BORO     VIC_SEX VIC_RACE
##    <chr>      <chr>    <chr>   <chr>
##  1 08/27/2006 BRONX    F       BLACK HISPANIC
##  2 03/11/2011 QUEENS   M       WHITE
##  3 10/06/2019 BROOKLYN F       BLACK
##  4 09/04/2011 BRONX    M       BLACK
##  5 05/27/2013 QUEENS   M       BLACK
##  6 09/01/2013 BROOKLYN M       BLACK
##  7 06/05/2010 BROOKLYN M       BLACK
##  8 03/20/2020 BROOKLYN M       BLACK
##  9 07/04/2014 QUEENS   M       BLACK
## 10 10/18/2015 QUEENS   M       BLACK
## # ... with 23,575 more rows
```

```
summary(tidy_NYPD_data) #summarize the data
```

```
##   OCCUR_DATE            BORO              VIC_SEX            VIC_RACE
##  Length:23585       Length:23585       Length:23585       Length:23585
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

After condensing as much as possible in the previous code, it was necessary to find summarize the data, using the `summary()` function, to ensure all the data is correct. Unfortunately, since the **OCCUR_DATE** column was not classified as a *date* but as a *character*, this column needed to be manipulated into a *date* classified column. This process can be seen in the following code chunk.

```
#convert the OCCUR_DATE column from character to date class
tidy_NYPD_data_date <- tidy_NYPD_data %>%
  rename(Borough = BORO,
         Victim_Sex = VIC_SEX,
         Victim_Race = VIC_RACE) %>%
  mutate(Date = as.Date(OCCUR_DATE, "%m/%d/%Y")) %>%
  select(-c(OCCUR_DATE))
tidy_NYPD_data_date
```

```
## # A tibble: 23,585 x 4
##    Borough  Victim_Sex Victim_Race    Date
##    <chr>    <chr>      <chr>          <date>
##  1 BRONX    F          BLACK HISPANIC 2006-08-27
##  2 QUEENS   M          WHITE          2011-03-11
##  3 BROOKLYN F          BLACK          2019-10-06
##  4 BRONX    M          BLACK          2011-09-04
##  5 QUEENS   M          BLACK          2013-05-27
##  6 BROOKLYN M          BLACK          2013-09-01
##  7 BROOKLYN M          BLACK          2010-06-05
##  8 BROOKLYN M          BLACK          2020-03-20
##  9 QUEENS   M          BLACK          2014-07-04
## 10 QUEENS   M          BLACK          2015-10-18
## # ... with 23,575 more rows
```

As shown above, column **Date** was created in place of **OCCUR_DATE**.

For the next portion, to simplify the data for analysis, rather than using the actual dates in the the **Date** column, it was simpler to utilize only the years. So the **Date** column needed to be simplified into a **Year** column.

```
tidy_NYPD_data_final <- tidy_NYPD_data_date%>%
  mutate(Year = format(Date, "%Y")) %>%
  group_by(Borough, Victim_Sex, Victim_Race, Year) %>%
  summarise(name_count = n())%>%
  select(Borough,
         Victim_Sex,
         Victim_Race,
         Year)%>%
  ungroup()
tidy_NYPD_data_final
```

```
## # A tibble: 649 x 4
##    Borough Victim_Sex Victim_Race Year 
##    <chr>   <chr>      <chr>       <chr>
##  1 BRONX   F          BLACK       2006 
##  2 BRONX   F          BLACK       2007 
##  3 BRONX   F          BLACK       2008 
##  4 BRONX   F          BLACK       2009 
##  5 BRONX   F          BLACK       2010 
##  6 BRONX   F          BLACK       2011 
##  7 BRONX   F          BLACK       2012 
##  8 BRONX   F          BLACK       2013 
##  9 BRONX   F          BLACK       2014 
## 10 BRONX   F          BLACK       2015 
## # ... with 639 more rows
```

Alas, the data was finally being tidy and ready for analysis, which leads to the next part of the data science process: visualize and model the data.

## Project Step 3: Add Visualizations and Analysis

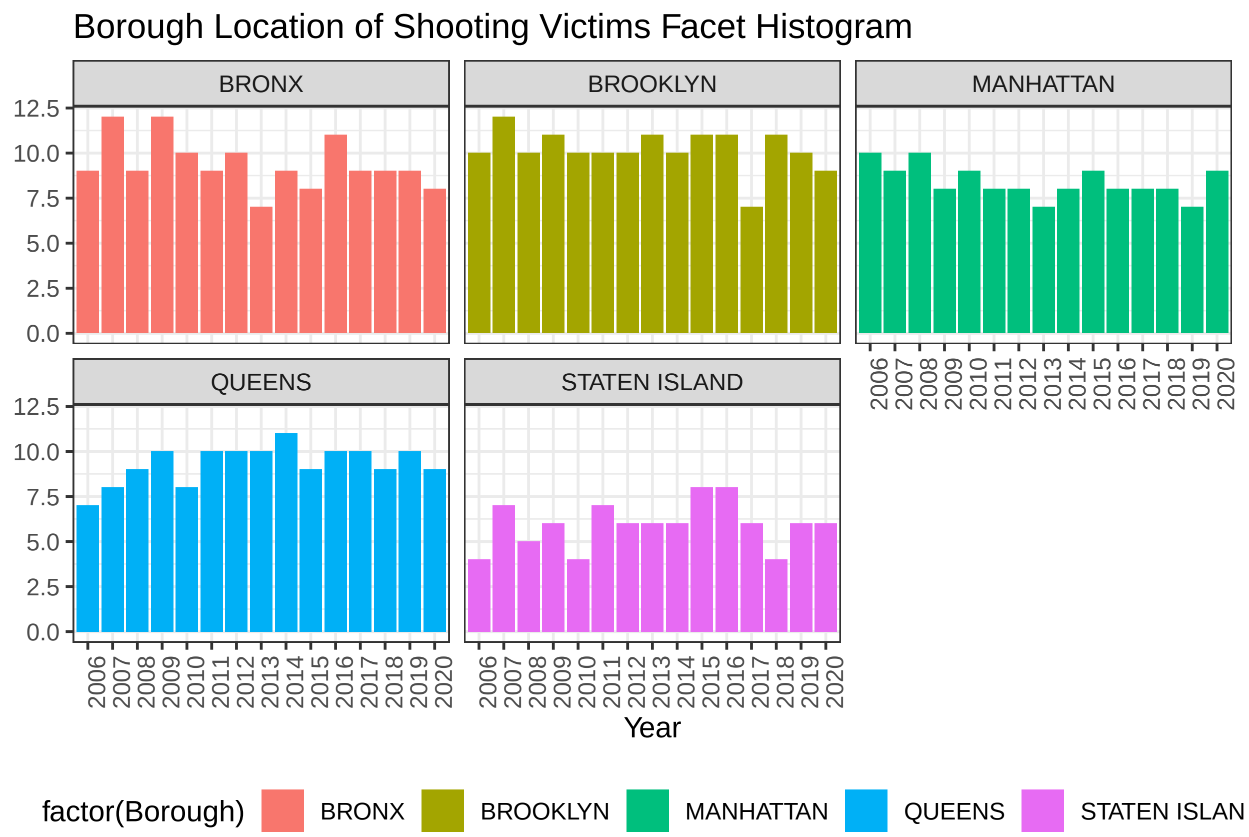
For this portion of the project, visualizations in relation to the investigation are necessary for understanding what the data representing, as well as potentially acknowledging potential external factors. To investigate whether race, sex, and borough location are contributing factors of shooting incidents, histograms were made for each variable, in respect to **Year**.

Visualization 1: Borough Location of Shooting Victims Histogram

Histograms

```
ggplot(tidy_NYPD_data_final,aes(x=Year, group=Borough, fill=factor(Borough)))+
   geom_histogram(stat = "count",
   position="dodge", bins = 10)+
   theme_bw()+
   facet_wrap(Borough~.) +
   theme(legend.position = "bottom",
        axis.text.x = element_text (angle = 90)) +
   labs(title = str_c("Borough Location of Shooting Victims Facet Histogram"),
       y = NULL)
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

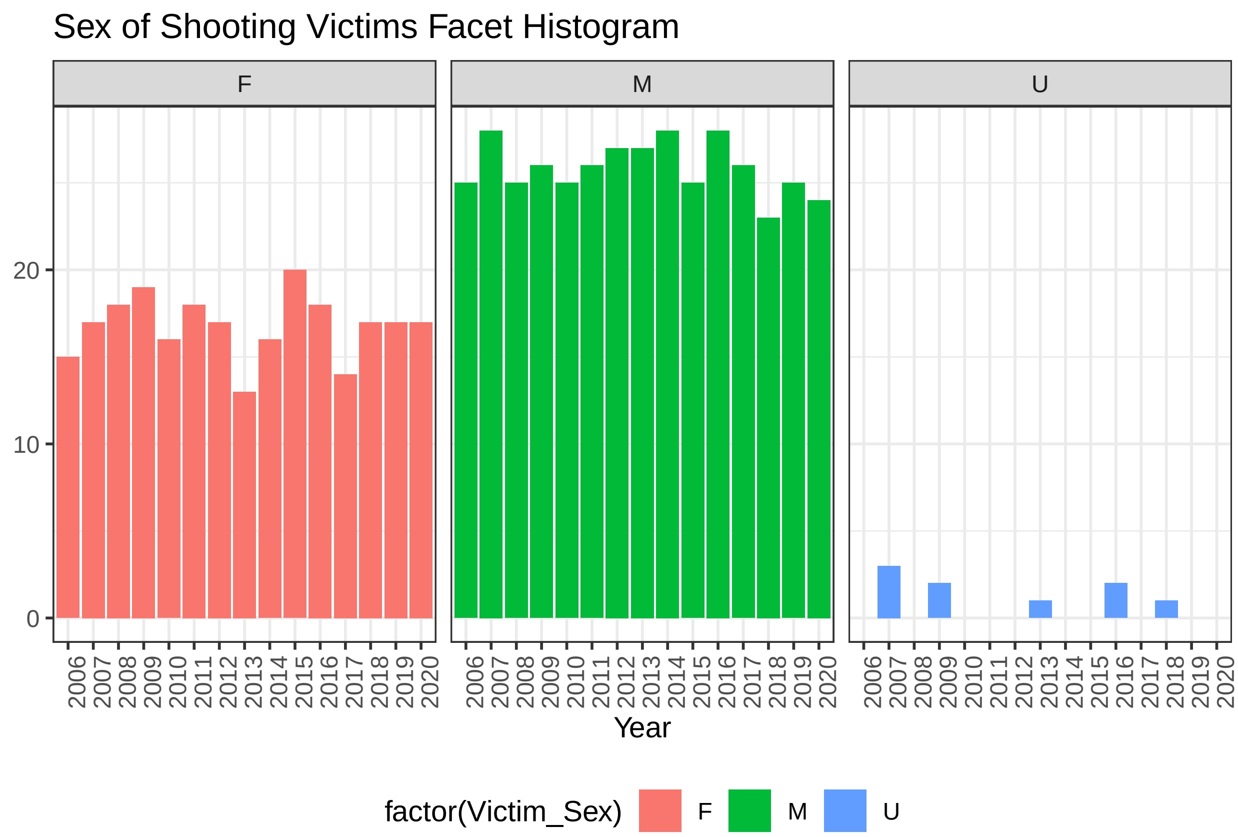


Analysis 1: Borough Location of Shooting Victims Histogram Using the given histogram visualization above, it is evident by the heights of these bars and high count overall that Bronx and Brooklyn are hotspots for shooting incidents throughout the last 15 years. This could indicate that location can determine higher risk for shooting incidents. However, given that the consistent and uniform shape of the bars over time, it indicates that the amount of shootings for Brooklyn and Bronx are on average about the same. This trend is also seen in the histograms for Manhattan, Queens, and Staten Island. In result, this also means that if there are increased or decreased amounts of crimes in NYC, evidently location does not strongly impact the causes of shooting incidents; however, location does play a role in increased risk.

Visualization 2: Sex of Shooting Victims Histogram

```
ggplot(tidy_NYPD_data_final,aes(x=Year,group=Victim_Sex, fill=factor(Victim_Sex)))+
   geom_histogram(stat = "count",
   position="dodge", bins = 10)+
   theme_bw()+
   facet_wrap(Victim_Sex~.) +
   theme(legend.position = "bottom",
        axis.text.x = element_text (angle = 90)) +
   labs(title = str_c("Sex of Shooting Victims Facet Histogram"),
       y = NULL)
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

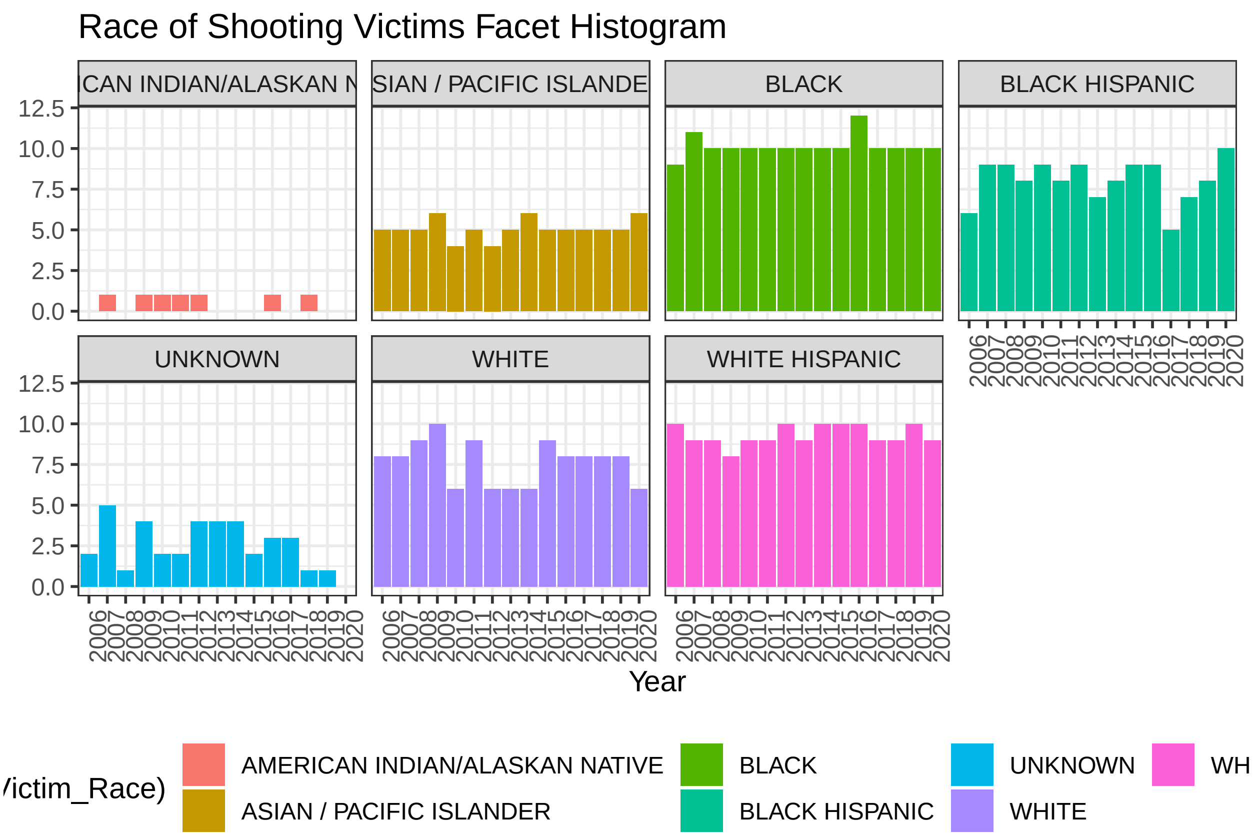


Analysis 2: Sex of Shooting Victims Histogram For this graph, the male histogram of NYC shootings are clearly higher than females, no matter what year it was for the past 15 years. Since all the male histogram bars in each year are taller than female, it is evident that between the two sexes, male have a higher chance of encountering a shooting incident compared to a women. In result, sex is likely a factor that plays into how likely one is going to become a victim in shootings.

Visualization 3: Race of Shooting Victims Facet Histogram

```
ggplot(tidy_NYPD_data_final,aes(x=Year,group=Victim_Race, fill=factor(Victim_Race)))+
   geom_histogram(stat = "count",
   position="dodge", bins = 10)+
   theme_bw()+
   facet_wrap(Victim_Race~., nrow=2) +
   theme(legend.position = "bottom",
        axis.text.x = element_text (angle = 90)) +
   labs(title = str_c("Race of Shooting Victims Facet Histogram"),
       y = NULL)
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



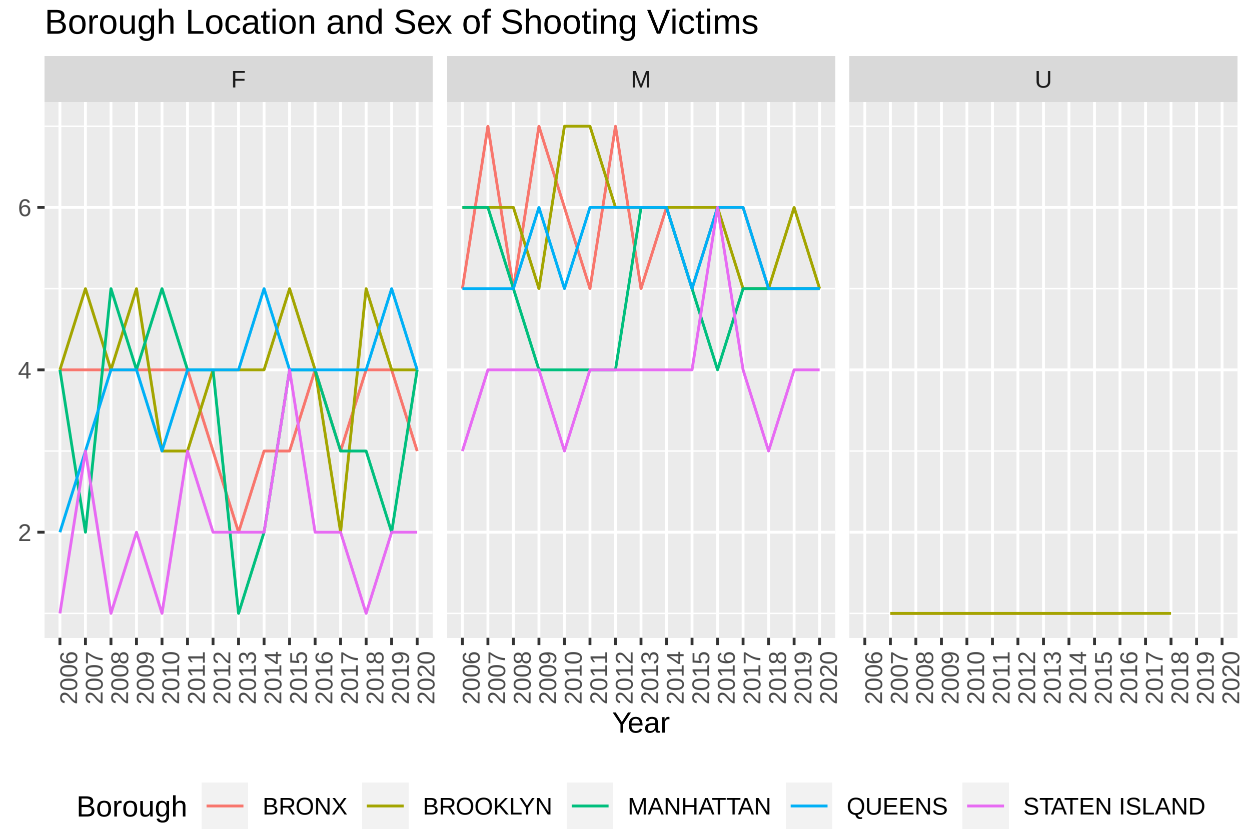
Analysis 3: Race of Shooting Victims Facet Histogram Similarly to the histograms for Borough Locations, the average count for each respective racial group is about the same, meaning all the racial groups have some sort of uniformity when it comes to representation in NYC. Based on solely this information, this shows consistency throughout the years of racial groups involved in shooting incidents, which could mean that racism does not necessarily play a role in shooting incidents because the proportion of those victims remain the same over the years. However, on the contrary, considering that Blacks, White Hispanics, and Black Hispanics are the ones with the highest count shooting victims, it could indicate racism plays a role in these incidents because they have the highest counts consistently. Without any knowledge of how prominent these races are relative to their population, it could easily be misunderstood. Therefore, based on this, race can play a role towards who the victims are in these shooting accidents; however, there is not enough evidence in this data set to prove this as a strong factor. In result, this variable is likely a potential factor but not a definite factor for instigating these shooting incidents. Therefore, among these three histograms, it is likely there are other potential factors that contribute to the shooting incidents overall.

Plot Comparisons ### Visualization 4: Borough Location and Sex of Shooting Victims In this data, the Borough locations were graphed by year relative to sex.

```
ggplot(tidy_NYPD_data_final,aes(x=Year,
                                group=Borough,
                                colour = Borough))+
   geom_freqpoly(stat = "count",
   position="dodge", bins = 10)+
  facet_wrap(Victim_Sex~.)+
  theme(legend.position = "bottom",
        axis.text.x = element_text (angle = 90)) +
  labs(title = str_c("Borough Location and Sex of Shooting Victims"),
       y = NULL)
```

```
## Warning: Ignoring unknown parameters: bins

## Warning: Width not defined. Set with ‘position_dodge(width = ?)‘
```



Analysis 4: Borough Location and Sex of Shooting Victims With two variables displayed in this plot, it is evident that the count of victims based on borough location is about the same hierarchy as the one demonstrated in visualization 1. However, despite some differences in plot for each location, the overall count of male victims stays true across all places. This means that regardless of where the shooting happens, males are at high risk of becoming victims to them, which also matches the analysis of those factors individually on the histogram. Therefore, sex can determine the chances of a person becoming a victim to shooting, regardless of that individuals where the crime may take place.

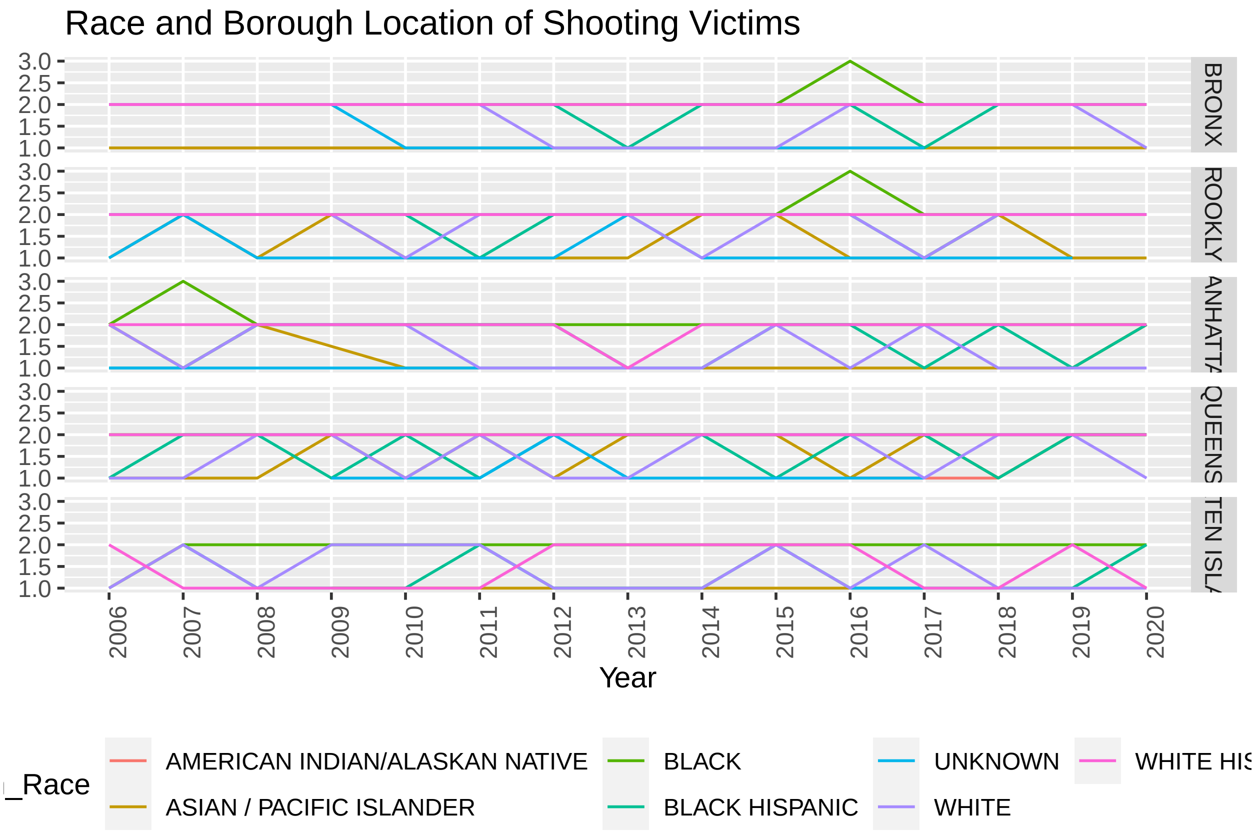
Visualization 5: Race and Borough Location of Shooting Victims

In this data, each race was graphed by year, relative to Borough Location..

```
ggplot(tidy_NYPD_data_final,aes(x=Year,
                                group=Victim_Race,
                                colour = Victim_Race))+
   geom_freqpoly(stat = "count",
   position="dodge", bins = 10)+
  facet_grid(Borough~.)+
  theme(legend.position = "bottom",
       axis.text.x = element_text (angle = 90)) +
  labs(title = str_c("Race and Borough Location of Shooting Victims"),
      y = NULL)
```

```
## Warning: Ignoring unknown parameters: bins

## Warning: Width not defined. Set with ‘position_dodge(width = ?)‘
```



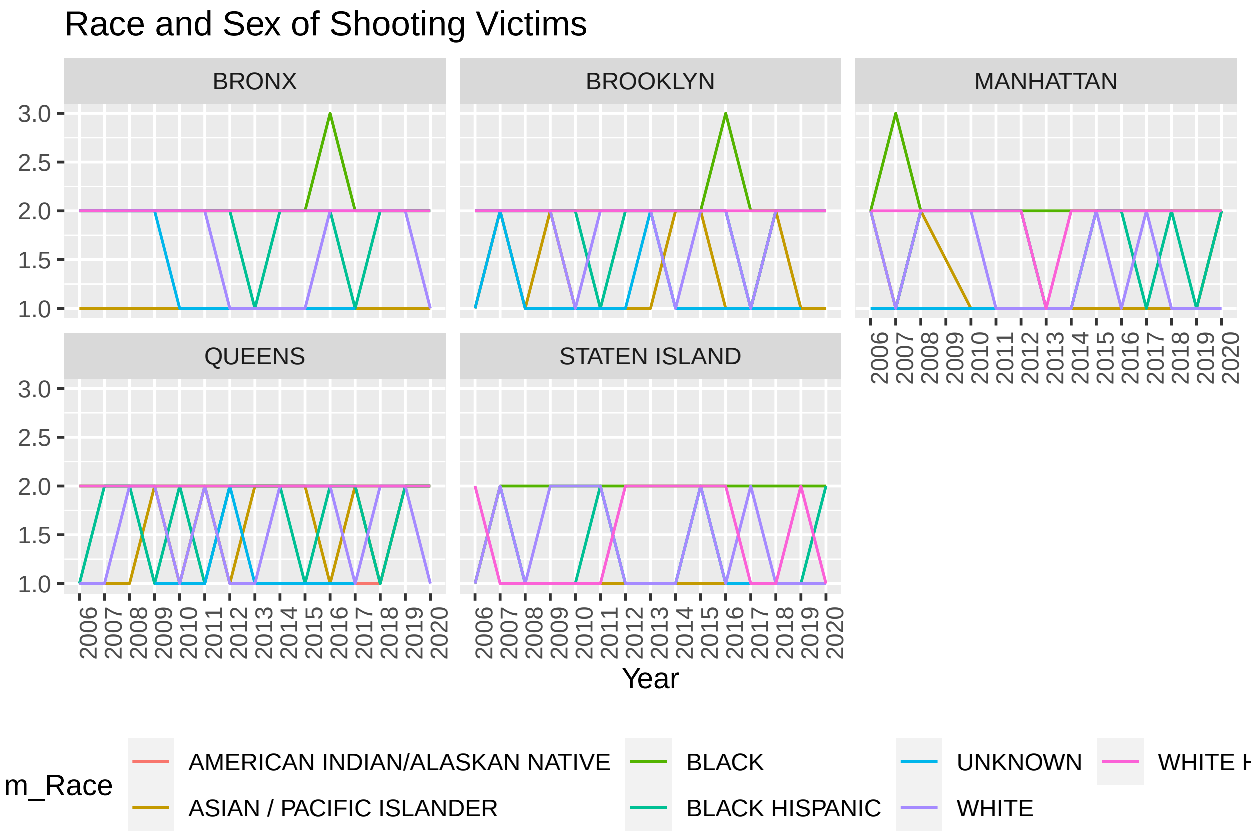
Analysis 5: Race and Borough Location of Shooting Victims The outcome of this graph does not indicate any sort of difference between victims of differing races in regards to the differences in location. This indicates that both Borough location and race demonstrate little to no effect on whether a person becomes a victim to shooting because there is not significant difference for either data sets. This also indicates there are likely other factors that likely appeal to how these shooting incidents happen.

Visualization 6: Race and Sex of Shooting Victims

In this data, each race was graphed by year, relative to Borough Location..

```
ggplot(tidy_NYPD_data_final,aes(x=Year,
                                group=Victim_Race,
                                colour = Victim_Race))+
   geom_freqpoly(stat = "count",
   position="dodge", bins = 10)+
  facet_wrap(Borough~.)+
  theme(legend.position = "bottom",
       axis.text.x = element_text (angle = 90)) +
  labs(title = str_c("Race and Sex of Shooting Victims"),
      y = NULL)
```

```
## Warning: Ignoring unknown parameters: bins
## Warning: Width not defined. Set with `position_dodge(width = ?)`
```



Analysis 6: Race and Sex of Shooting Victims Similarly to the histograms for Borough Locations, the hierarchy of race is somewhat the same as as visualization 1 for both males and females. Based on this information, the relative amount of people by race that become victims are about the same between both sexes. However, as previous data has shown, males are at higher risk than females to be in the range of fire, so based on that, it is clear that despite the difference in count for both sexes, the racial proportions that fall victim are approximately the same between both genders. This highly indicates that race does not help contribute to the causes of the these shooting incidents when sex is already a factor.

Reflection Based on the analysis conducted on this trial, it is clear that sex is predominantly on of the factors that potentially lead to these shooting incidents. However, the relationship between both borough locations and race to shooting incidents overall were not as impactful as sex was. In result, it led to many questions, such as whether differences between racial groups were just relative to the proportion of the whole population in New York; whether the Borough location size contributed towards slight differences in count; and whether there are other factors like: criminal records, living situation, and the perpetrator that makes a huge difference for how shooting incidents occur.

Conclusion Overall, to answer investigation question for thsi project, evidently only sex prove to be a factor that largely contributed to some of these shooting incidents, while borough location and race did not highly contribute. However, due to some of the data, there is a chance they both serve as a minor factor towards these incidents instead. Throughout this research, there were some issues with bias, considering race is one of the larger topics in the social world today. However, unlike the common knowledge or racism and prejudice towards those of minority communities, especially Hispanic and Black communities, this data did not represent that outlook, which actually made analyzing this project a little harder. Other sources of bias lie in the data itself, for it only represents the victims and not the prepertrators, so it could have lead to biases supporting th vitims protrayal of the shooting incidents over those who were deemed the prepratrator. Lastly, there is bias towards identifying minority groups without acknowledging the relative proportion of minorities within a population. Without this acknowledgement, it essentially favours the minority group by name rather than the data, which was a mistake almost made during this project. However, with thorough re-evaluation and reflection, all these biases were acknowledged and emphasized by means of writing the analysis and reflection.

Session Info

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
## 
## Matrix products: default
## 
## locale:
## [1] LC_COLLATE=English_United States.1252 
## [2] LC_CTYPE=English_United States.1252   
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C                          
## [5] LC_TIME=English_United States.1252    
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] lubridate_1.8.0 forcats_0.5.1   stringr_1.4.0   dplyr_1.0.7    
##  [5] purrr_0.3.4     readr_2.1.1     tidyr_1.1.4     tibble_3.1.6   
##  [9] ggplot2_3.3.5   tidyverse_1.3.1
## 
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.7       assertthat_0.2.1 digest_0.6.29    utf8_1.2.2      
##  [5] R6_2.5.1         cellranger_1.1.0 backports_1.3.0  reprex_2.0.1    
##  [9] evaluate_0.14    highr_0.9        httr_1.4.2       pillar_1.6.4    
## [13] rlang_0.4.12     curl_4.3.2       readxl_1.3.1     rstudioapi_0.13 
## [17] rmarkdown_2.11   labeling_0.4.2   bit_4.0.4        munsell_0.5.0   
## [21] broom_0.7.10     compiler_4.1.2   modelr_0.1.8     xfun_0.28       
## [25] pkgconfig_2.0.3  htmltools_0.5.2  tidyselect_1.1.1 fansi_0.5.0     
## [29] crayon_1.4.2     tzdb_0.2.0       dbplyr_2.1.1     withr_2.4.3     
## [33] grid_4.1.2       jsonlite_1.7.2   gtable_0.3.0     lifecycle_1.0.1 
## [37] DBI_1.1.1        magrittr_2.0.1   scales_1.1.1     cli_3.1.0       
## [41] stringi_1.7.6    vroom_1.5.7      farver_2.1.0     fs_1.5.1        
## [45] xml2_1.3.3       ellipsis_0.3.2   generics_0.1.1   vctrs_0.3.8     
## [49] tools_4.1.2      bit64_4.0.5      glue_1.5.1       hms_1.1.1       
## [53] parallel_4.1.2   fastmap_1.1.0    yaml_2.2.1       colorspace_2.0-2
## [57] rvest_1.0.2      knitr_1.36       haven_2.4.3
```