NYPD Incidents Project Report

Javel Williamson

2/20/2022

Analysis topic

The analysis of murder cases in New York seeks to find the correlation between the nature of deaths to the race, age, sex,or location of the individuals. Are murders on the rise or the decline? What possible sources of bias remain in the analysis?

Library Setup

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
```

Import NYPD Incident Data

The data is supplied by data.gov. The dataset used in this analysis is the NYPD Shooting Incident Data (Historic), which provides a breakdown of shooting incidents in NYC from 2006–2020 and can be found here data.url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD" data.raw <- read.csv(data.url)

Clean Data: Remove Unnecessary Columns

```
data.raw.cols <- colnames(data.raw) data.raw.cols
## [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"
```

```
# Determine the columns we don't need to keep data.raw.cols.bad <- c(1, 5, 6,
7, 17, 18, 19)

# We'll change our naming style a bit data.clean.cols <- c(
      "Date", "Time", "Borough", "Murder", "PerpAge", "PerpSex", "PerpRace", "VictimAge", "VictimSex",
      "VictimRace", "Xcoord", "Ycoord"
)

# Create cleaned data
data.full <- data.raw[-data.raw.cols.bad] %>%
      setNames(data.clean.cols) %>% arrange(Date,
      Time) %>% mutate(Date=mdy(Date)) %>%
      mutate(Time=hms(Time)) %>% arrange(Date,
      Time)

# Replace all empty cells with NA data.full.dims <-
dim(data.full) data.full.nrows <- data.full.dims[1]
data.full.ncols <- data.full.dims[2]

# For some reason 2 values in age are 940 and 224?.. Set to NA as well
data.full.bad.values <- c("", "UNKNOWN", "U", "940", "224") for (i in
1:data.full.nrows) { for (j in 1:data.full.ncols) { val <- data.full[i, j]
            if (val %in% data.full.bad.values) { data.full[i, j] = NA
            }
      }
}
```

Data Summary

```
##
## Unique Values: Murder:
## false true

##
## Unique Values: PerpAge:
## 18-24 25-44 NA 45-64 <18 65+ 1020

##
## Unique Values: PerpSex:
## M NA F
##
## Unique Values: PerpRace:
## BLACK NA WHITE HISPANIC BLACK HISPANIC WHITE ASIAN / PACIFIC ISLANDER AMERICAN INDIAN/ALASKAN NATIVE

##
## Unique Values: VictimAge:
## <18 25-44 18-24 45-64 65+ NA

##
## Unique Values: VictimSex:
```

```
## M F NA
##
## Unique Values: VictimRace:
## BLACK WHITE HISPANIC BLACK HISPANIC WHITE ASIAN / PACIFIC ISLANDER NA AMERICAN INDIAN/ALASKAN NATIVE
##       Date                    Time                               Borough
## Min.      :2006-01-01Min.     :0S                                Length:23585
## 1st Qu.:2008-12-31      1st Qu.:3H 20M 0S                        Class :character
## Median :2012-02-27      Median :15H 0M 0S                        Mode :character
## Mean      :2012-10-05Mean     :12H 33M 7.48187407250225S

## 3rd Qu.:2016-03-02      3rd Qu.:20H 45M 0S

## Max.      :2020-12-31Max.     :23H 59M 0S

##       Murder            PerpAge           PerpSex           PerpRace
## Length:23585          Length:23585      Length:23585      Length:23585
## Class :character      Class :character  Class :character  Class :character
## Mode :character       Mode :character   Mode :character   Mode :character
##
##
##
##     VictimAge          VictimSex         VictimRace           Xcoord
## Length:23585          Length:23585      Length:23585      Min.      : 914928
## Class :character      Class :character  Class :character  1st Qu.: 999925
## Mode :character       Mode :character   Mode :character   Median :1007654
##                                                           Mean      :1009379

##                                                           3rd Qu.:1016782

##                                                           Max.      :1066815
##        Ycoord
## Min. :125757 ## 1st
Qu.:182539 ## Median
:193470    ##    Mean
:207300     ##     3rd
Qu.:239163   ##   Max.
:271128
```

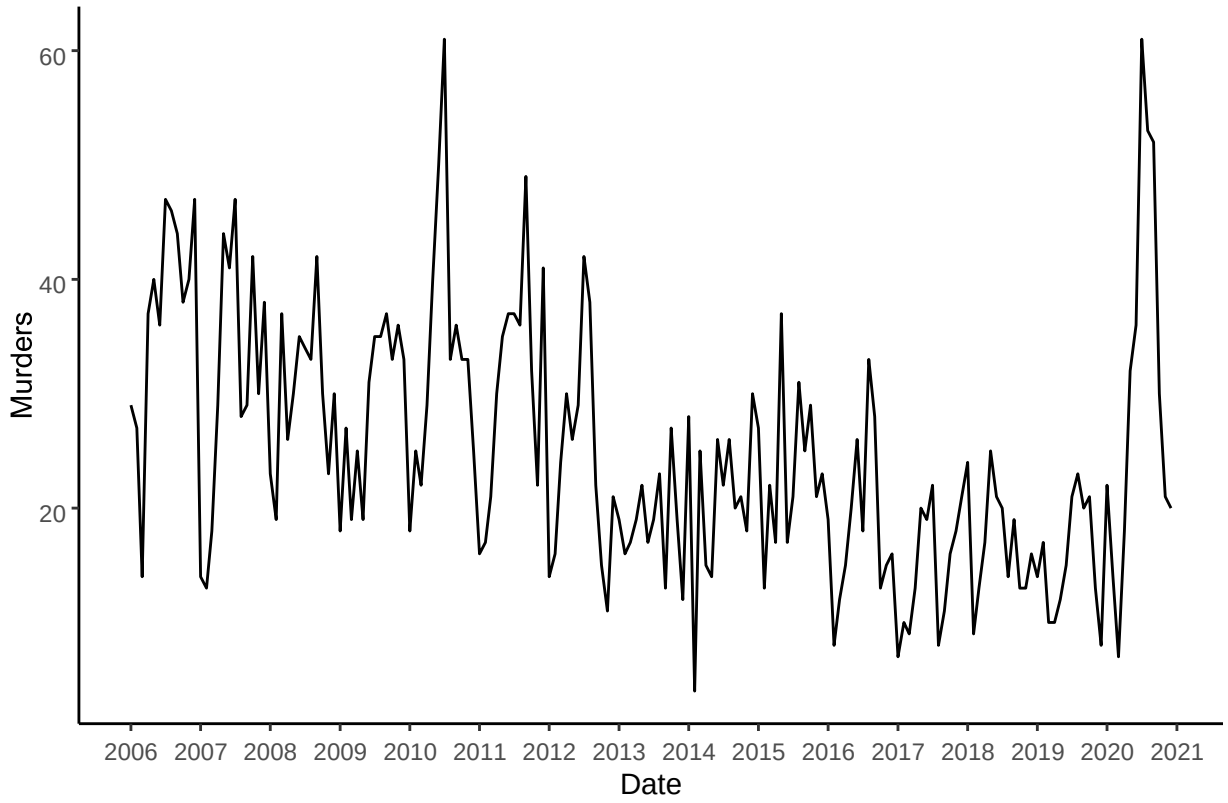
Create Additional DataFrames for Analysis and Modeling

```
# Select all murders and sort by date and time

data.murders <- data.full %>% filter(Murder == 'true')

# Count how many murders occurred each month
data.murders.by.month <- as.data.frame( table(
data.murders %>%
```

```
            mutate(Year=year(Date), Month=month(Date)) %>%
            unite(YearMonth, Year, Month, sep="-") %>%
            mutate(YearMonth=ym(YearMonth)) %>% select(YearMonth)
      )
) %>% setNames(c("Date", "Murders")) %>%
      mutate(Date=ymd(Date))

# Count how many murders occured in each borough
data.murders.by.borough <- as.data.frame( table(data.murders %>%
select(Borough))) %>% setNames(c("Borough", "Murders")
) %>%
      arrange(-Murders)

# Is there a correlation between perpetrator and victim age?
data.murders.by.age <- as.data.frame( table(data.murders %>%
select(c(PerpAge, VictimAge)))
) %>% setNames(c("PerpAge", "VictimAge", "Murders"))


# Is there a correlation between perpetrator and victim race?
data.murders.by.race <- as.data.frame( table(data.murders %>%
select(c(PerpRace, VictimRace)))
) %>% setNames(c("PerpRace", "VictimRace", "Murders"))


# Is there a correlation between perpetrator and victim sex?
data.murders.by.sex <- as.data.frame( table(data.murders %>%
select(c(PerpSex, VictimSex)))
) %>% setNames(c("PerpSex", "VictimSex", "Murders"))
```

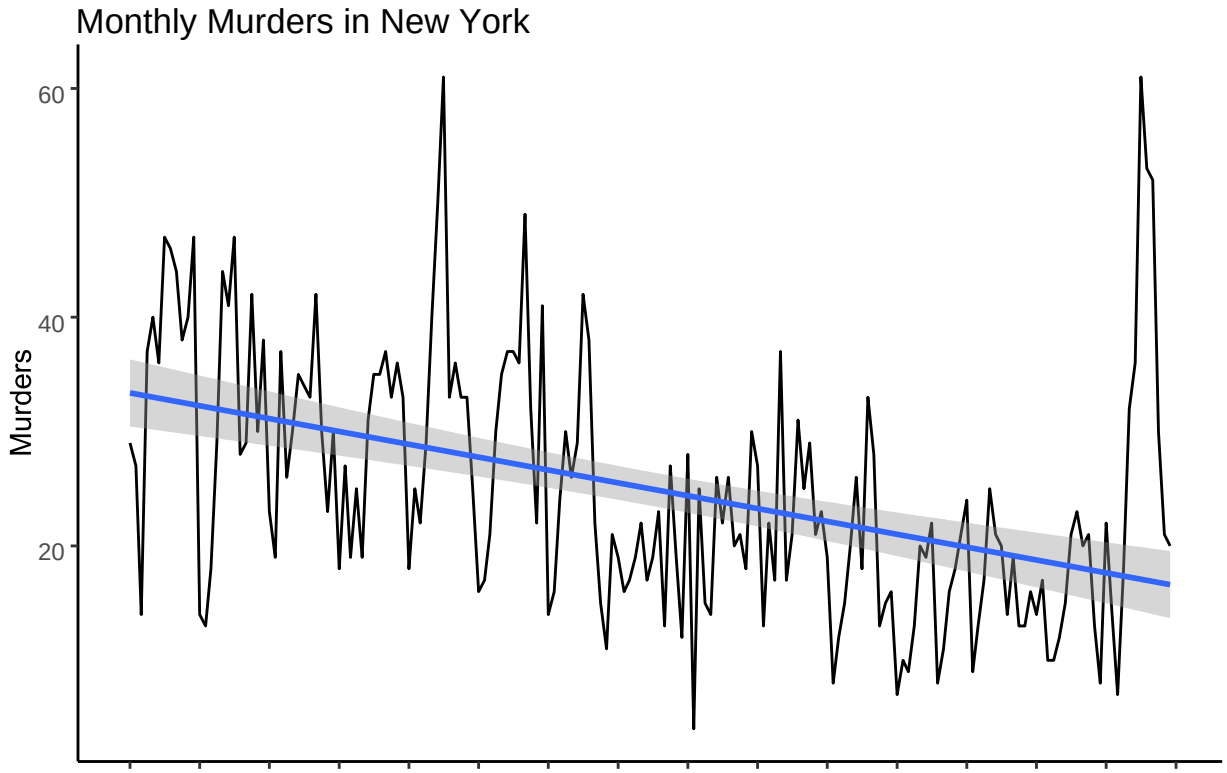
Data Analysis and Modeling

```
# [1] Create a line plot to show the number of murders each month data.murders.by.month.plot <-
ggplot(data=data.murders.by.month) + aes(x=Date, y=Murders) +
      scale_x_date(date_labels="%Y", date_breaks="1 year") +
      theme(axis.text.x=element_text(angle=45, hjust=1)) + geom_line() +
      theme_classic() +
      ggtitle("Monthly Murders in New York")

# It seems like we can fit a linear regression model relatively well data.murders.by.month.mod <-
data.murders.by.month.plot + geom_smooth(method="lm")

data.murders.by.month.plot
```



data.murders.by.month.mod

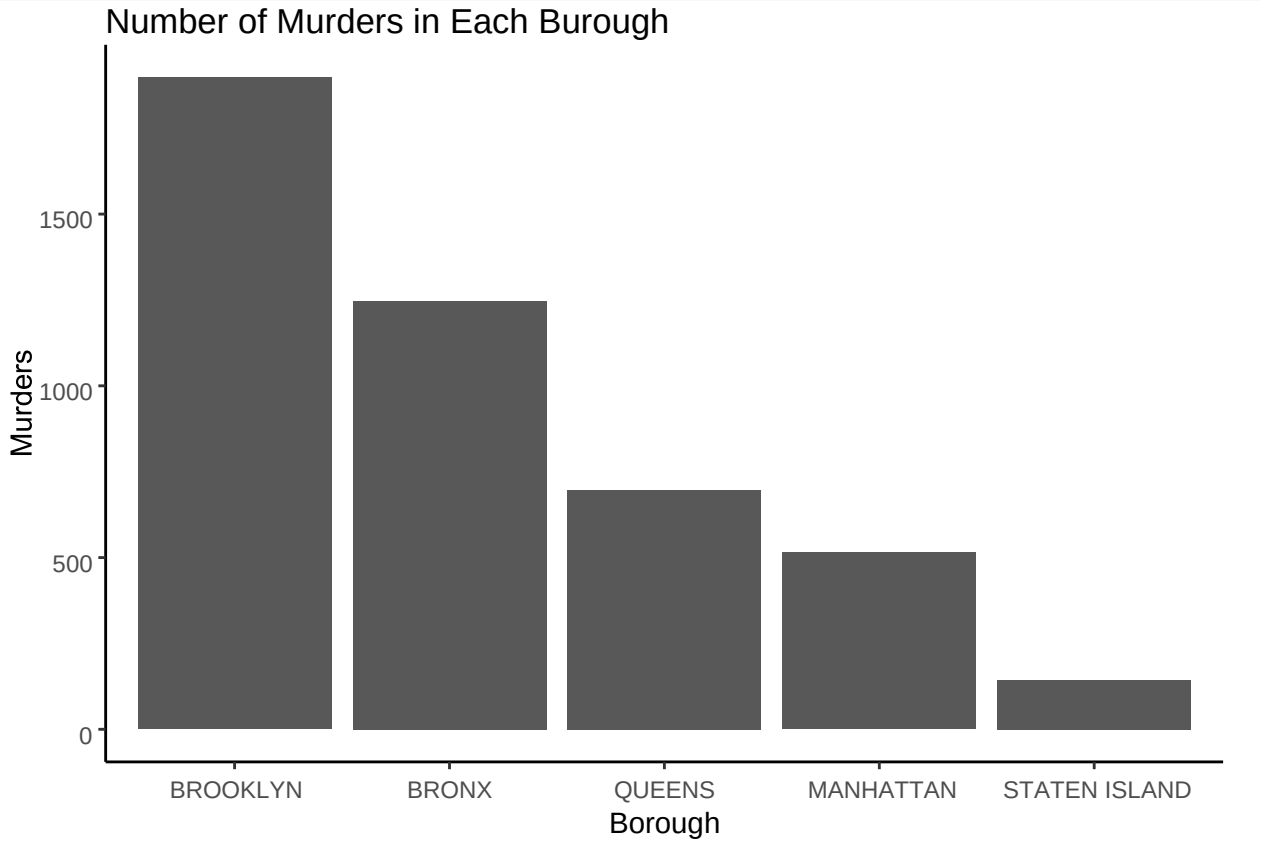
`geom_smooth()` using formula 'y ~ x'



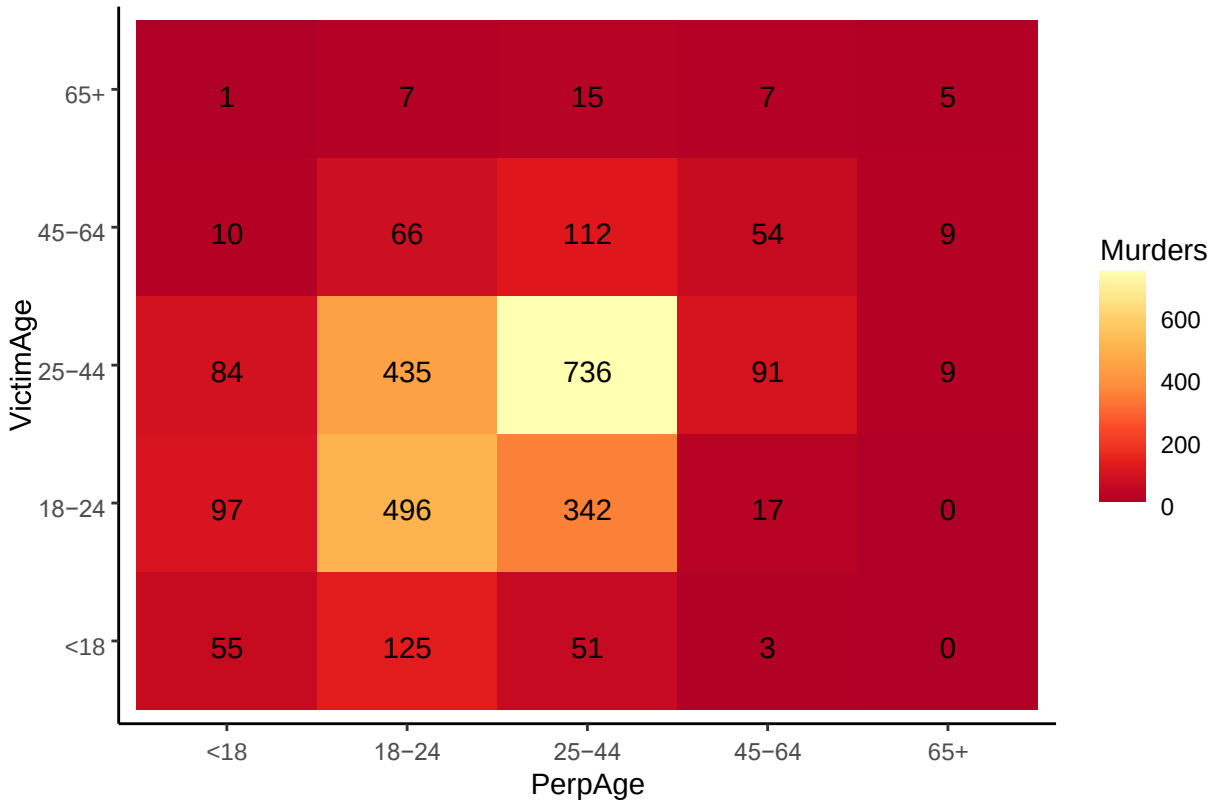
2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021

Date

```
# [2] Create a simple bar chart to visualize murders by borough
data.murders.by.borough.plot <- ggplot(data=data.murders.by.borough) +
aes(x=reorder(Borough, -Murders), y=Murders) + geom_col() + theme_classic() +
      ggtitle("Number of Murders in Each Burough") + xlab("Borough")

data.murders.by.borough.plot
```

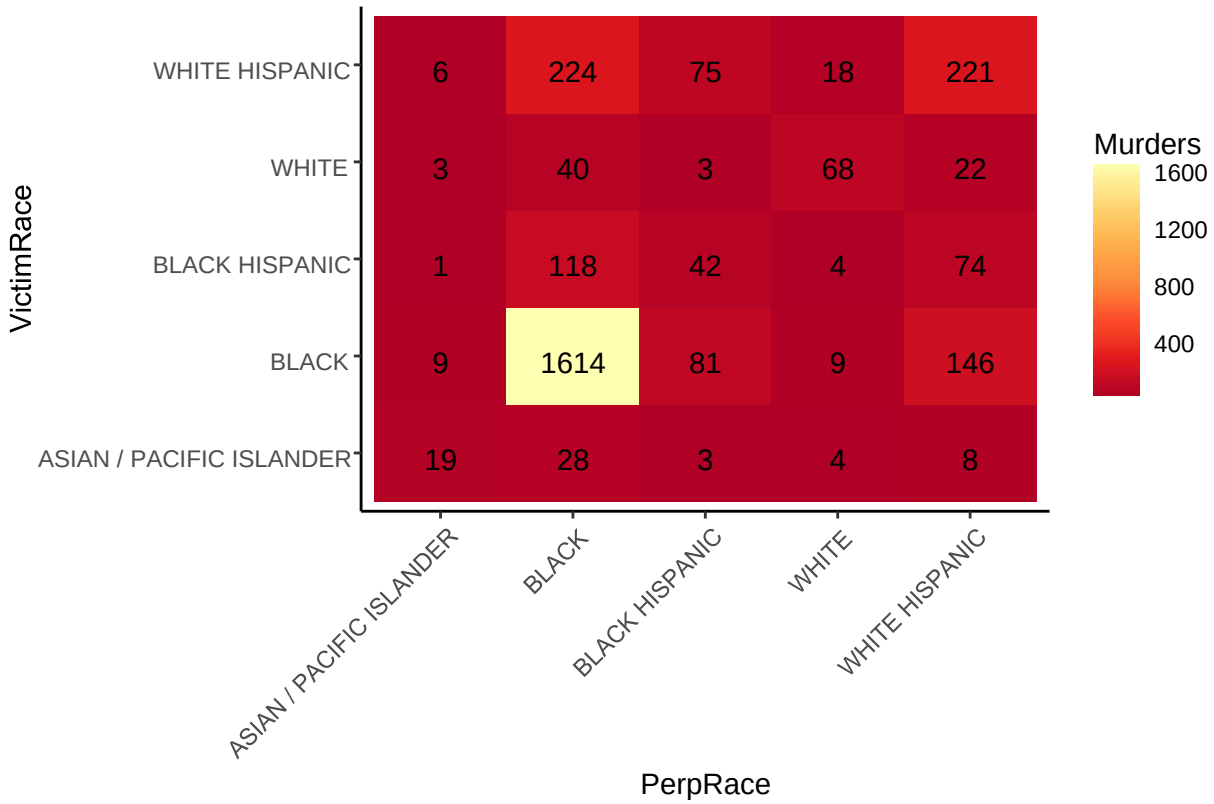



```
# [3] Create a heatmap to visualize perpetrator Vs. victim age
data.murders.by.age.plot <- ggplot(data=data.murders.by.age) +
aes(x=PerpAge, y=VictimAge, fill=Murders) + geom_tile() +
      scale_fill_distiller(palette="YlOrRd") + theme_classic() +
      geom_text(aes(label=Murders)) + ggtitle("Ages Involved
      in Murders")

data.murders.by.age.plot
```

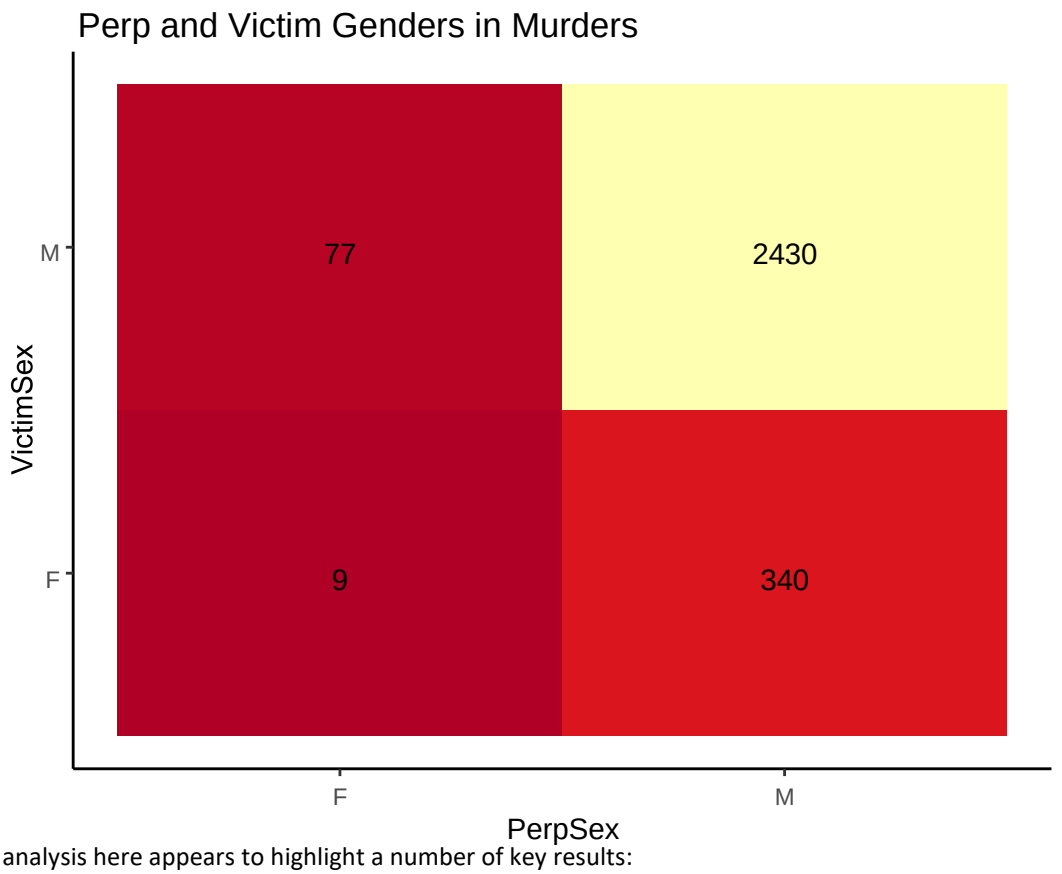



```
# [4] Create a heatmap to visualize perpetrator Vs. victim race
data.murders.by.race.plot <- ggplot(data=data.murders.by.race) +
aes(x=PerpRace, y=VictimRace, fill=Murders) + geom_tile() +
      scale_fill_distiller(palette="YlOrRd") + theme_classic() +
      theme(axis.text.x=element_text(angle=45, hjust=1)) +
      geom_text(aes(label=Murders)) + ggtitle("Races Involved in
      Murders")

data.murders.by.race.plot
```

Races Involved in Murders




```
# [5] Create a heatmap to visualize perpetrator Vs. victim sex
data.murders.by.sex.plot <- ggplot(data=data.murders.by.sex) + aes(x=PerpSex,
y=VictimSex, fill=Murders) + geom_tile() +
      scale_fill_distiller(palette="YlOrRd") + theme_classic() +
      geom_text(aes(label=Murders)) +
      ggtitle("Perp and Victim Genders in Murders")

data.murders.by.sex.plot
```



Conclusion and Sources of Bias

The analysis here appears to highlight a number of key results:

1. The number of murders occurring in New York has been steadily decreasing each month since 2006, albeit some distinct outliers.
2. It appears that young to middle-aged (18–44) are more likely to both be a murder victim and a perpetrator. Younger and older individuals are less likely to be involved in a murder
3. Black-on-black, black-on-white hispanic and white hispanic-on-white hispanic are the most common races involved in murders in New York.
4. Males are much more likely to be the perpetrators than females with male-on-male murder being significantly more frequent than any other.
5. Murders occur most-frequently in Brooklyn and least-frequently in Staten Island.

It is important to acknowledge likely sources of bias from these results to avoid forming potentially erroneous conclusions. A likely mistake on all fronts is that the data is not conditioned on total population statistics. That is, simply because black-on-black murders are much more prominent than other combinations does not mean that other races are less likely to murder than black individuals. To assign causation without any further evidence would be ridiculous It may be, for instance, that a much higher percentage of the population in New York is black and, as such, we expect to see more black-on-black murders even if the probability any given race would murder another is uniform across all races.

Additionally, while young to middle-aged (18–44) are more likely to be involved in a murder, it does not mean that they are inherently more violenced than those who are younger or older. It may be the case that individuals aged 18–44 are simply more exposed to a harsh environment and are required to hard life decisions regarding how they live their lives. In other words, many poor individuals in New York are stuck between “a rock and a hard place”: they must make hard decisions that promote a risky lifestyle. This risky lifestyle increases the probability that they will

eventually become involved in a murder, either as the victim or the perpetrator. To that end, there is nothing to say here that individuals aged 18–44 are inherently more violent that any others.

All in all, this analysis is a first-step in examining the muddy world of murder incidents. There are many variables to consider and condition upon, and any suggestion of causation must be met with suspicion. Much more work would be needed with input from a number of domain experts to extract causation from a study such as this.