NYPD shooting cases analysis

Alex Zhukov

2022-06-27

Overview

This project is to show the simple analysis of the shooting cases based on NYPD data collected from 2006 to 2021.

We'll use `R` and the following libraries for data cleaning, analysis and visualization are needed:

```
library(ggplot2)
library(dplyr)
library(lubridate)
library(tidyr)
library(vcd)
```

Importing Data

First, let's import data from https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

```
original_data <- read.csv('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOA
```

Let's have a glance on the data

```
str(original_data)
```

```
## 'data.frame':    25596 obs. of  19 variables:
##  $ INCIDENT_KEY           : int  24050482 77673979 226950018 237710987 224701998 225295736 231190175
##  $ OCCUR_DATE             : chr  "08/27/2006" "03/11/2011" "04/14/2021" "12/10/2021" ...
##  $ OCCUR_TIME             : chr  "05:35:00" "12:03:00" "21:08:00" "19:30:00" ...
##  $ BORO                   : chr  "BRONX" "QUEENS" "BRONX" "BRONX" ...
##  $ PRECINCT               : int  52 106 42 52 34 75 32 26 41 67 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 2 2 0 ...
##  $ LOCATION_DESC          : chr  "" "" "COMMERCIAL BLDG" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "true" "false" "true" "false" ...
##  $ PERP_AGE_GROUP         : chr  "" "" "" "" ...
##  $ PERP_SEX               : chr  "" "" "" "" ...
##  $ PERP_RACE              : chr  "" "" "" "" ...
##  $ VIC_AGE_GROUP          : chr  "25-44" "65+" "18-24" "25-44" ...
##  $ VIC_SEX                : chr  "F" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK HISPANIC" "WHITE" "BLACK" "BLACK" ...
##  $ X_COORD_CD             : num  1017542 1027543 1009489 1017440 1005426 ...
```

```
## $ Y_COORD_CD             : num  255919 186095 243050 256046 254690 ...
## $ Latitude               : num  40.9 40.7 40.8 40.9 40.9 ...
## $ Longitude              : num  -73.9 -73.8 -73.9 -73.9 -73.9 ...
## $ Lon_Lat                : chr  "POINT (-73.87963173099996 40.86905819000003)" "POINT (-73.8439201
```

There are in total 19 variables. Some of them can be interesting and some of them we won't use.

Tidying and Transforming Data

Let's remove some of the variables, rename those that we can be interested in, change empty and incorrect values to 'Unknown' and put it all in a new data frame.

```
cleaned_df <- original_data %>%
  select(-c(INCIDENT_KEY, OCCUR_TIME, PRECINCT, JURISDICTION_CODE, LOCATION_DESC, X_COORD_CD, Y_COORD_C
  rename(Date = OCCUR_DATE, Area = BORO, Murdered = STATISTICAL_MURDER_FLAG) %>%
  mutate(Murdered = if_else(Murdered == 'true', 1, 0),
         PERP_SEX = if_else(PERP_SEX == '', 'UNKNOWN', PERP_SEX),
         PERP_RACE = if_else(PERP_RACE == '', 'UNKNOWN', PERP_RACE),
         PERP_AGE_GROUP = if_else(PERP_AGE_GROUP %in% c('', '224', '1020', '940'), 'UNKNOWN', PERP_AGE_

cleaned_df$PERP_RACE[cleaned_df$PERP_RACE == 'AMERICAN INDIAN/ALASKAN NATIVE'] <- 'AM. INDIAN/ALASKAN'
cleaned_df$VIC_RACE[cleaned_df$VIC_RACE == 'AMERICAN INDIAN/ALASKAN NATIVE'] <- 'AM. INDIAN/ALASKAN'
```

Let's see what we have now:

```
str(cleaned_df)
```

```
## 'data.frame':    25596 obs. of  9 variables:
## $ Date           : chr  "08/27/2006" "03/11/2011" "04/14/2021" "12/10/2021" ...
## $ Area           : chr  "BRONX" "QUEENS" "BRONX" "BRONX" ...
## $ Murdered       : num  1 0 1 0 0 1 0 0 1 0 ...
## $ PERP_AGE_GROUP: chr  "UNKNOWN" "UNKNOWN" "UNKNOWN" "UNKNOWN" ...
## $ PERP_SEX       : chr  "UNKNOWN" "UNKNOWN" "UNKNOWN" "UNKNOWN" ...
## $ PERP_RACE      : chr  "UNKNOWN" "UNKNOWN" "UNKNOWN" "UNKNOWN" ...
## $ VIC_AGE_GROUP : chr  "25-44" "65+" "18-24" "25-44" ...
## $ VIC_SEX        : chr  "F" "M" "M" "M" ...
## $ VIC_RACE       : chr  "BLACK HISPANIC" "WHITE" "BLACK" "BLACK" ...
```

Now, let's format dates to date type and factorize variables of age groups, sex and race for both perpetrators and victims.

```
cleaned_df$Date <- mdy(cleaned_df$Date)
cols <- c('Area', 'PERP_AGE_GROUP','PERP_SEX', 'PERP_RACE', 'VIC_AGE_GROUP', 'VIC_SEX', 'VIC_RACE')
cleaned_df[cols] <- lapply(cleaned_df[cols], factor)

str(cleaned_df)
```

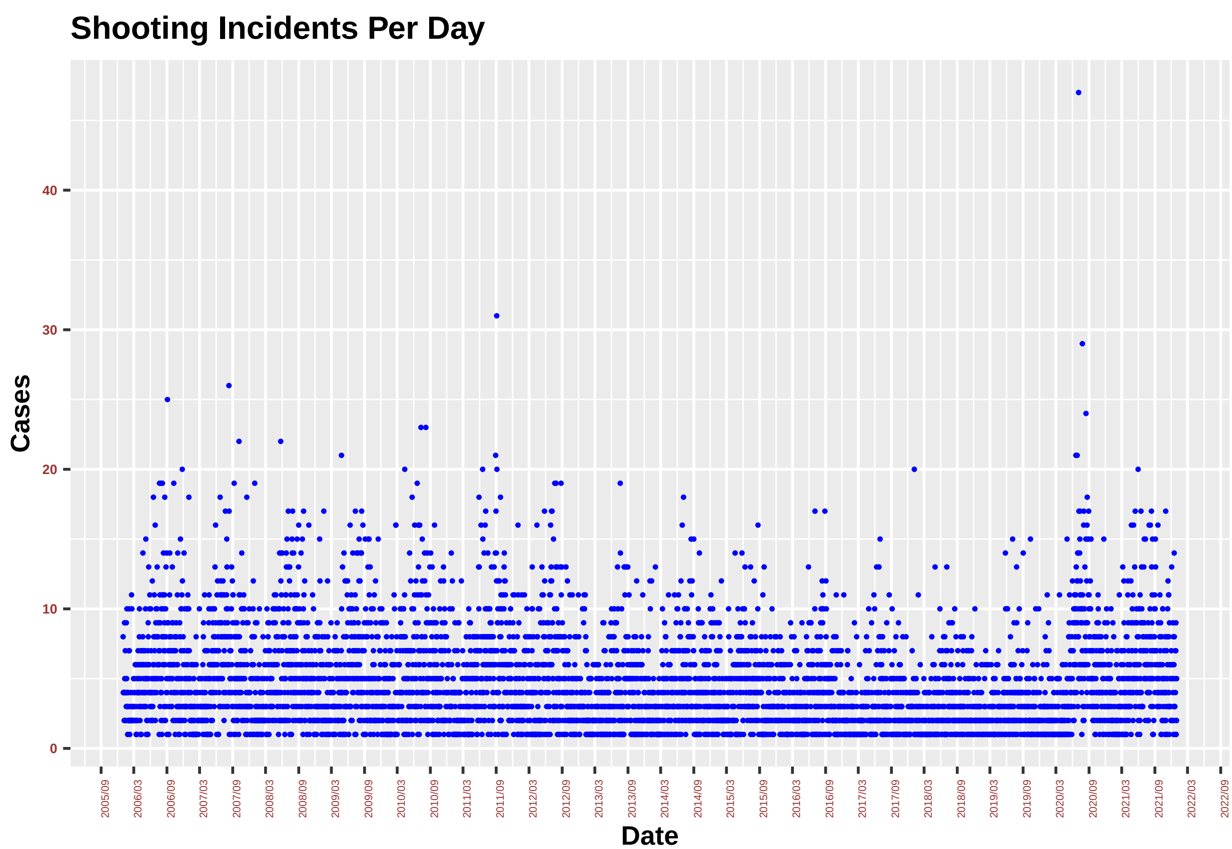
```
## 'data.frame':    25596 obs. of  9 variables:
## $ Date           : Date, format: "2006-08-27" "2011-03-11" ...
## $ Area           : Factor w/ 5 levels "BRONX","BROOKLYN",..: 1 4 1 1 3 2 3 3 1 2 ...
## $ Murdered       : num  1 0 1 0 0 1 0 0 1 0 ...
```

```
##  $ PERP_AGE_GROUP: Factor w/ 6 levels "<18","18-24",..: 6 6 6 6 6 3 3 6 3 6 ...
##  $ PERP_SEX      : Factor w/ 4 levels "F","M","U","UNKNOWN": 4 4 4 4 4 2 2 4 2 4 ...
##  $ PERP_RACE     : Factor w/ 7 levels "AM. INDIAN/ALASKAN",..: 5 5 5 5 5 4 3 5 3 5 ...
##  $ VIC_AGE_GROUP : Factor w/ 6 levels "<18","18-24",..: 3 5 2 3 3 3 3 2 3 2 ...
##  $ VIC_SEX       : Factor w/ 3 levels "F","M","U": 1 2 2 2 2 2 2 2 2 2 ...
##  $ VIC_RACE      : Factor w/ 7 levels "AM. INDIAN/ALASKAN",..: 4 6 3 3 4 7 3 3 4 3 ...
```

Visualising Data

Let's first take a look at the number of cases per day during the entire period.

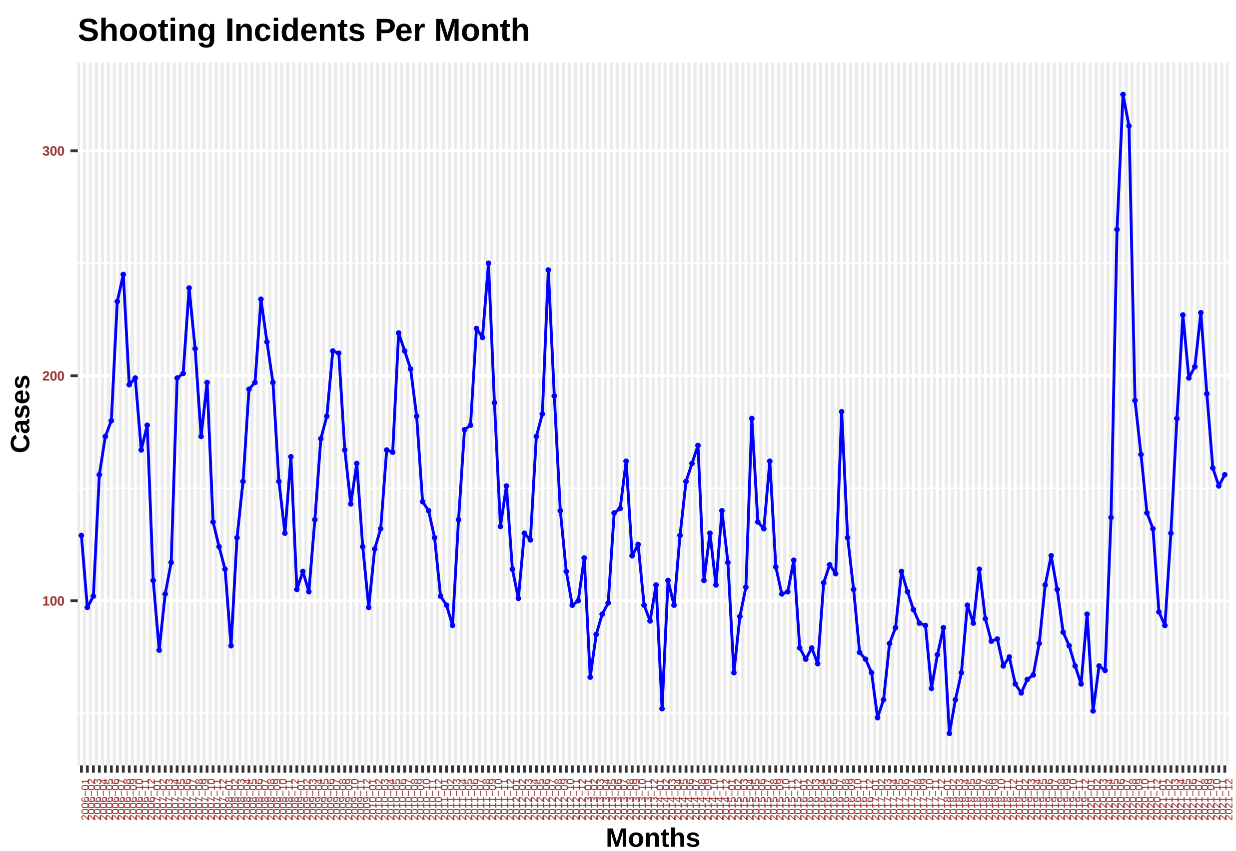
```
cleaned_df %>%
  group_by(Date) %>%
  summarize(Cases = n()) %>%
  ggplot(aes(x = Date, y = Cases))+
  geom_point(size=.3, color='blue')+
  ggtitle('Shooting Incidents Per Day') +
  scale_x_date(date_breaks = "6 month", date_labels = "%Y/%m") + p
```



It's rather difficult to look at all of the daily cases, so let's group them by months and see the trends if any.

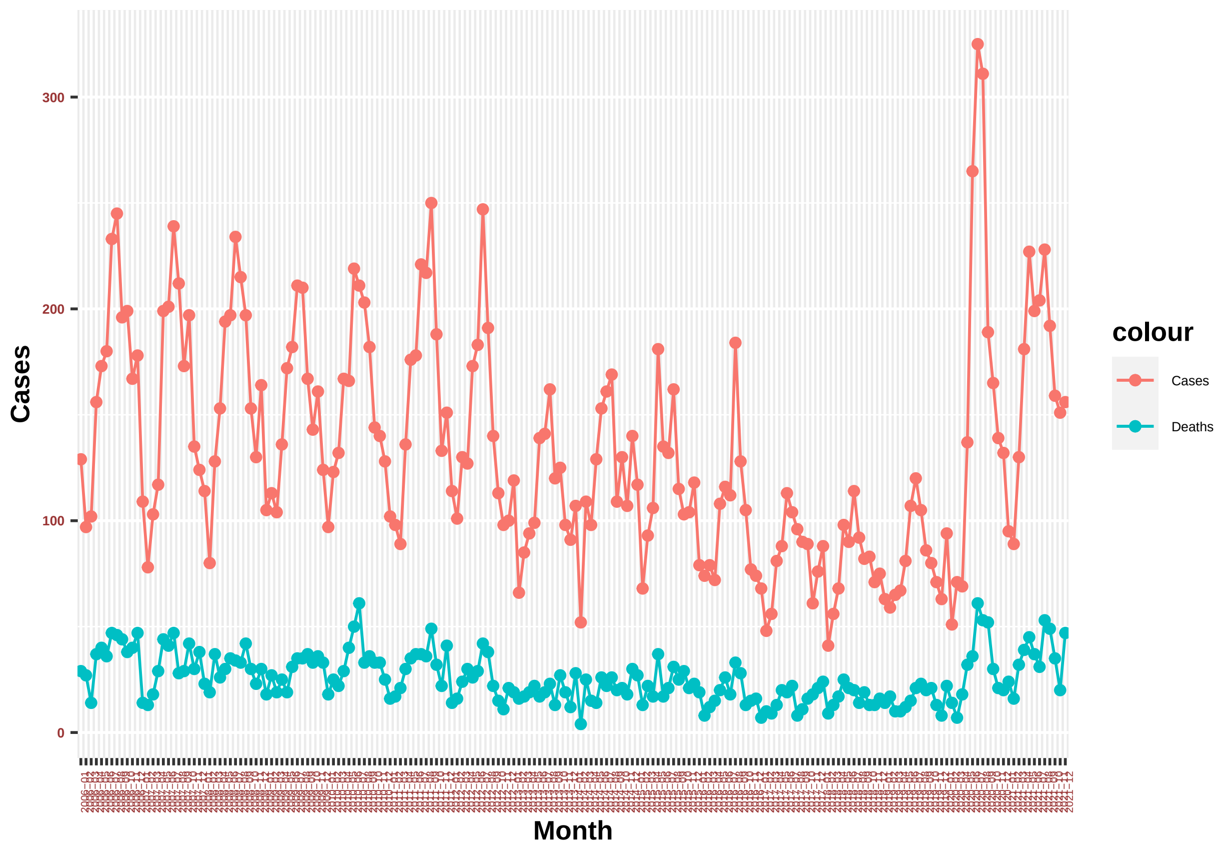
```
cleaned_df %>%
  mutate(Month = format(cleaned_df$Date, "%Y-%m")) %>%
  group_by(Month) %>%
  summarize(Cases = n()) %>%
```

```
  ggplot(aes(x = Month, y = Cases))+
  geom_point(size=.3, color='blue')+
  geom_line(group=1, color='blue')+
  ggtitle('Shooting Incidents Per Month')+
  xlab('Months')+ p
```



Now let's add number of murders in that monthly plot.

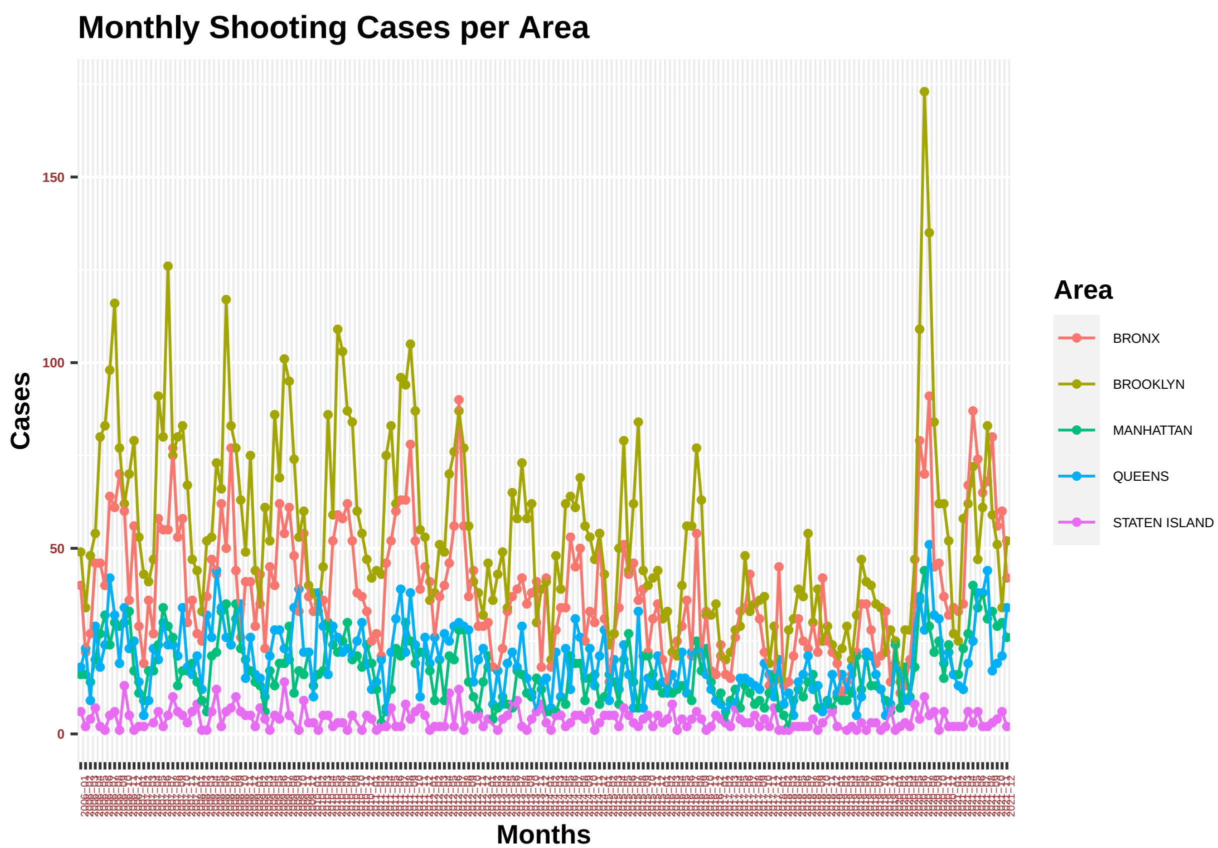
```
cleaned_df %>%
  mutate(Month = format(cleaned_df$Date, "%Y-%m")) %>%
  group_by(Month) %>%
  summarise(Cases = n(), Deaths = sum(Murdered)) %>%
  filter(Cases > 0 & Deaths > 0) %>%
  ggplot(aes(x = Month, y= Cases))+
  geom_line(aes(group = 1, color='Cases')) +
  geom_point(aes(color='Cases')) +
  geom_line(aes(y=Deaths, group = 2, color='Deaths')) +
  geom_point(aes(y=Deaths, color='Deaths'))+ p
```



Seems that there are some fluctuations in the number of incidents every six months, plus we see that there were minimum number of incidents between 2017 and 2019, and then it drastically increased.

Now, let's split the observations by the areas and analyze it.

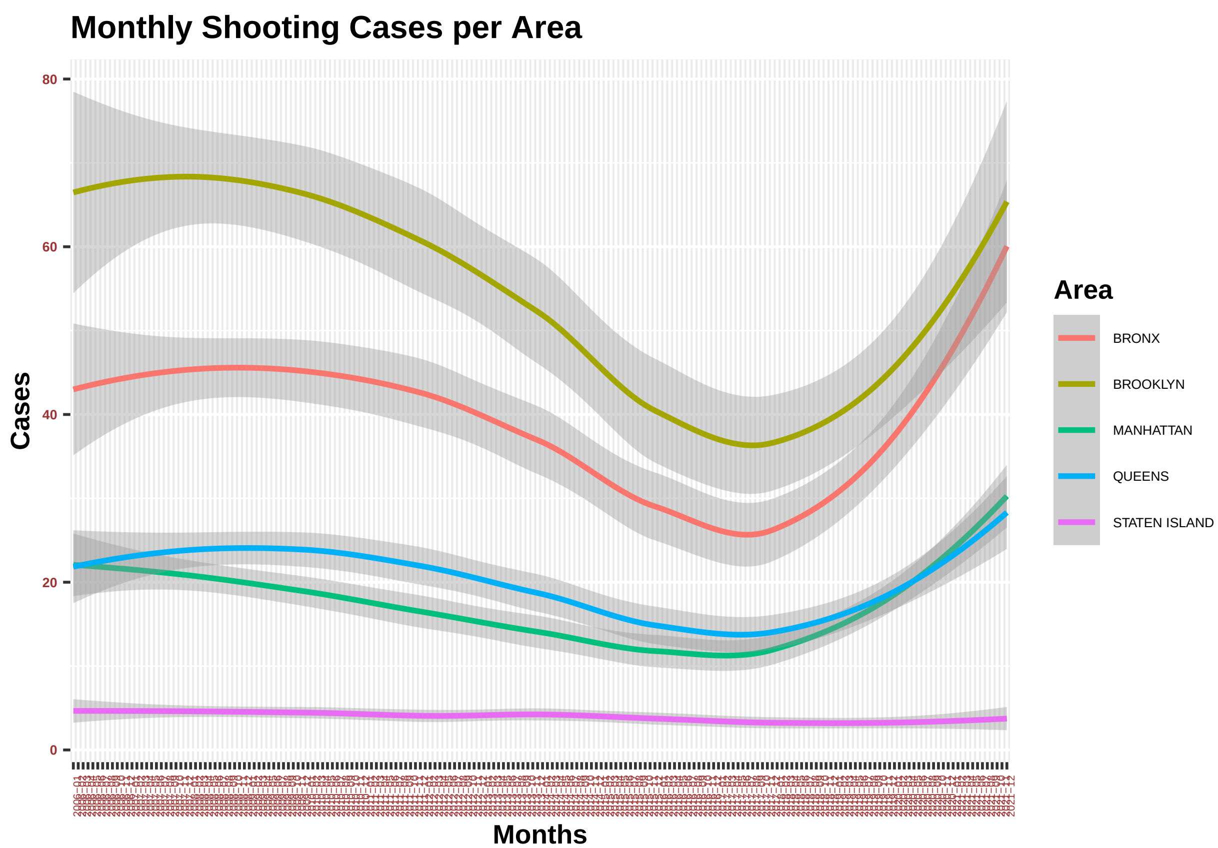
```
cleaned_df %>%
  mutate(Month = format(cleaned_df$Date, "%Y-%m")) %>%
  group_by(Month, Area) %>%
  summarise(Cases = n()) %>%
  filter(Cases > 0) %>%
  ggplot(aes(x = Month, y = Cases, col = Area, group = Area))+
  geom_point(size=1)+
  geom_line()+
  ggtitle('Monthly Shooting Cases per Area')+
  xlab('Months') +p
```



According to the plot, the area with a maximum number of monthly incidents is Brooklyn, followed by Bronx. Staten Island seems to be an area with a least number of cases.

Now let's see the trends.

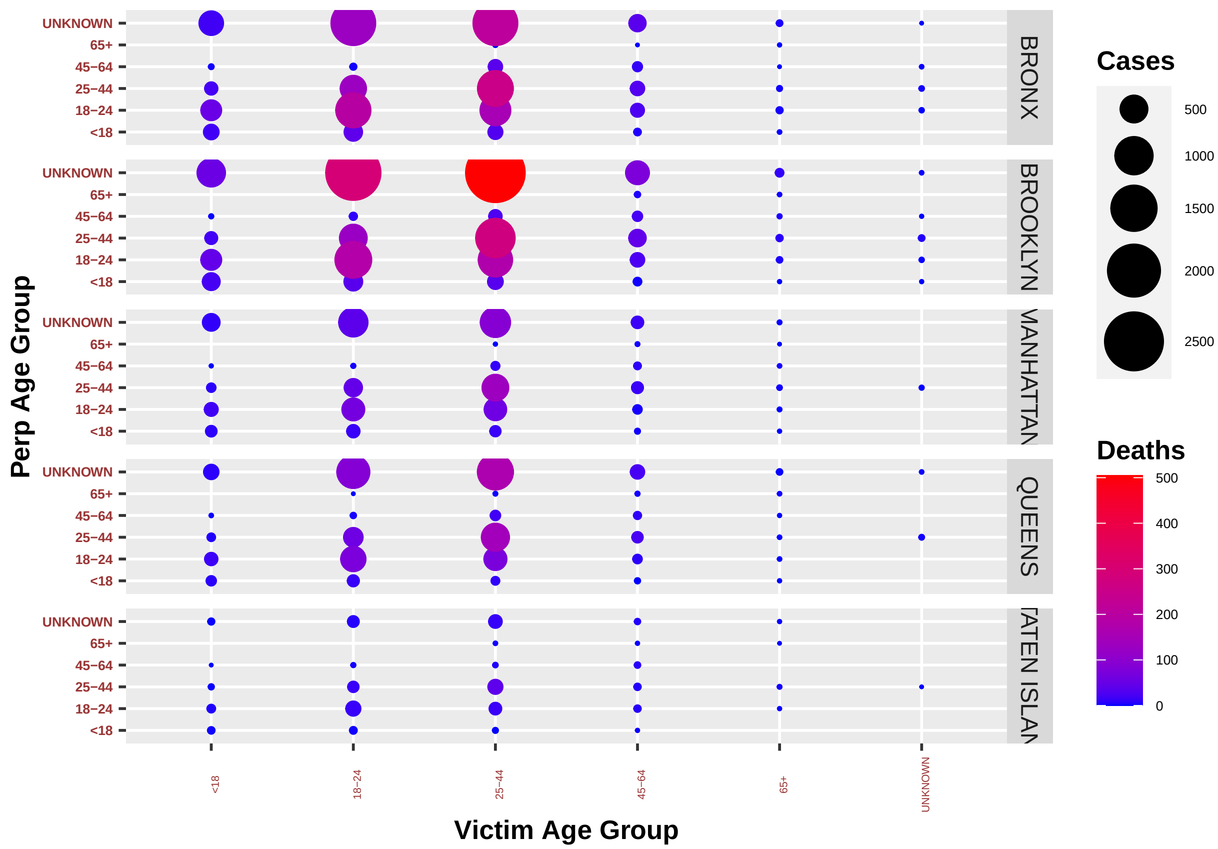
```
cleaned_df %>%
  mutate(Month = format(cleaned_df$Date, "%Y-%m")) %>%
  group_by(Month, Area) %>%
  summarise(Cases = n()) %>%
  ggplot(aes(x = Month, y = Cases, col = Area, group = Area))+
  geom_smooth()+
  ggtitle('Monthly Shooting Cases per Area')+
  xlab('Months') + p
```



As it was noted earlier, there's an increase of cases in the last 3 years in all areas except Staten Island, where the number of incidents is constantly low and even throughout the entire timeframe.

Now let's look at the age groups of the participants in the incident in the different areas.

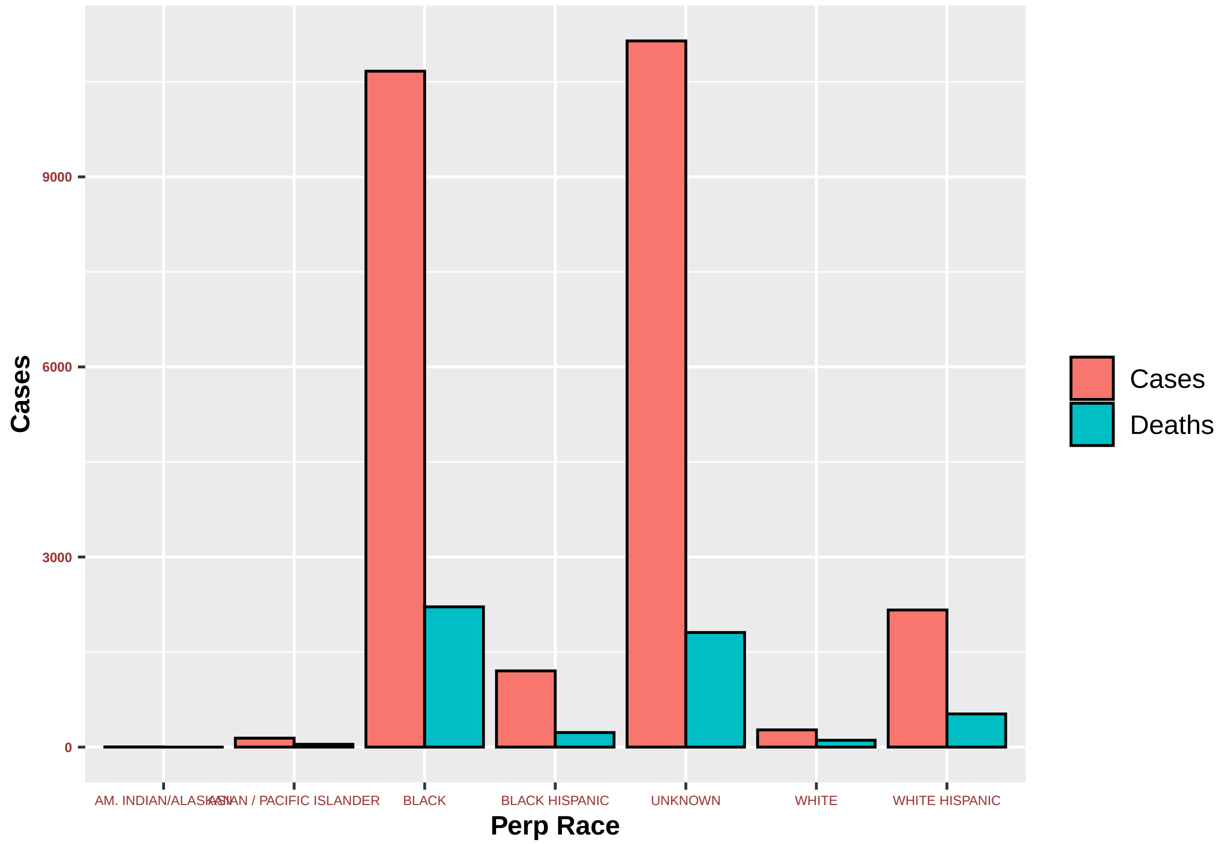
```
cleaned_df %>%
  group_by(Area, PERP_AGE_GROUP, VIC_AGE_GROUP) %>%
  summarise(Cases=n(), Deaths=sum(Murdered)) %>%
  ggplot(aes(x = VIC_AGE_GROUP, y = PERP_AGE_GROUP, col=Deaths))+
  geom_count(aes(col=Deaths, size = Cases)) +
  scale_size_area(max_size = 10) +
  scale_colour_gradient(low="blue", high="red")+
  facet_grid(Area~.) +
  xlab('Victim Age Group')+
  ylab('Perp Age Group') + p
```



Unfortunately, there are many incidents in which we don't know the age of the criminal, and these incidents seem to be the most deadly.

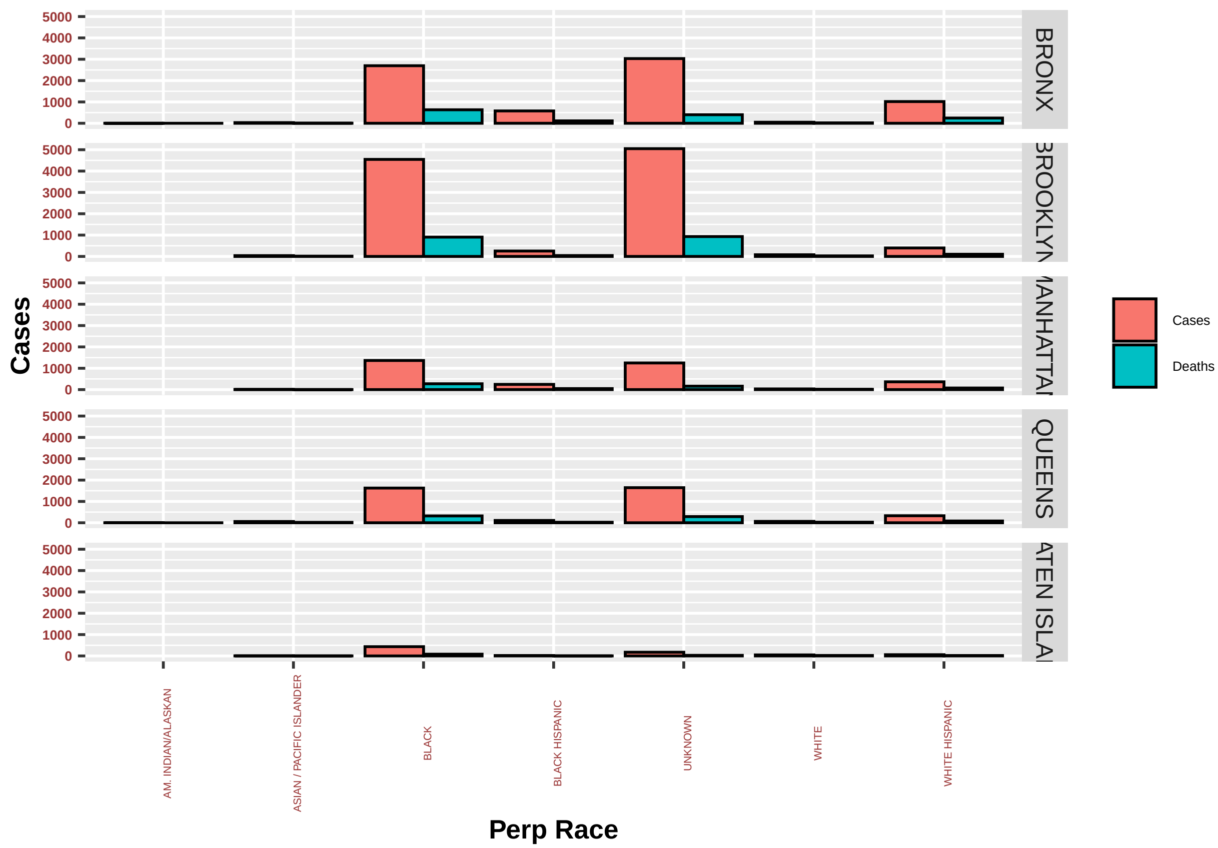
Next we'll look at the race of the perpetrator in all incidents.

```
cleaned_df %>%
  group_by(PERP_RACE) %>%
  summarise(Cases=n(), Deaths=sum(Murdered)) %>%
  gather(key = var, value = value, Cases, Deaths) %>%
  ggplot(color='black', aes(x = PERP_RACE, y = value, fill = var)) +
  geom_col(group=1, position='dodge', color='black') +
  xlab('Perp Race')+
  ylab('Cases') +
  theme(axis.text.x = element_text(color="#993333", size=5),
        axis.text.y = element_text(face="bold", color="#993333", size=5),
        legend.text = element_text(size=10),
        legend.title = element_blank(),
        title = element_text(face='bold', size=10))
```



And here is a split by area:

```
cleaned_df %>%
  group_by(Area, PERP_RACE) %>%
  summarise(Cases=n(), Deaths=sum(Murdered)) %>%
  gather(key = var, value = value, Cases, Deaths) %>%
  ggplot(aes(x = PERP_RACE, y = value, fill = var)) +
  geom_bar(stat = 'identity', position='dodge', color='black')+
  facet_grid(Area~.) +
  xlab('Perp Race')+
  ylab('Cases') + p + theme(legend.title = element_blank())
```



Modeling data

The main idea is to calculate the odds of being murdered during the shooting incidents depending on the predictors in the original data. Here we will not count the probability to get into an incident in a certain area, as the number of cases depending on the area was clearly seen earlier. We will focus on the model showing what factors significantly change the chances to be killed.

So, let's first put some nominative predictors from original data into a separate data frame to make a logistic regression and remove observations where age values are unknown.

And here we look at the cases indicated with murdered flag which include sex and age group of both perpetrator and victim, and also an area where the incident happened.

```
df_for_models <- cleaned_df %>%
  filter(VIC_AGE_GROUP != 'UNKNOWN', PERP_AGE_GROUP != 'UNKNOWN', VIC_SEX != 'U') %>%
  mutate(Murdered = factor(if_else(Murdered == 1, 'Y', 'N'))) %>%
  select(Murdered, Area, VIC_SEX, VIC_AGE_GROUP, PERP_AGE_GROUP)

df_for_models$VIC_SEX <- droplevels(df_for_models$VIC_SEX)
df_for_models$VIC_AGE_GROUP <- droplevels(df_for_models$VIC_AGE_GROUP)
df_for_models$PERP_AGE_GROUP <- droplevels(df_for_models$PERP_AGE_GROUP)

summary(df_for_models)
```

```
##   Murdered             Area       VIC_SEX    VIC_AGE_GROUP PERP_AGE_GROUP
```

```
##  N:9946   BRONX        :4023   F: 1478   <18  :1453    <18  :1461
##  Y:3100   BROOKLYN     :4691   M:11568   18-24:4669    18-24:5830
##           MANHATTAN    :1832             25-44:5840    25-44:5168
##           QUEENS       :1955             45-64: 977    45-64: 530
##           STATEN ISLAND: 545             65+  : 107    65+  :  57
```

Let's take an intercept only model:

```r
simple_fit <- glm(Murdered ~ 1, df_for_models, family = "binomial")
coef(simple_fit)
```

```
## (Intercept)
##   -1.165768
```

This negative number is a logarithm of odds of being murdered in a shooting incident regardless of any other influencing factors. In other words this the logarithm of a fraction of total murders and incidents.

```r
table(df_for_models$Murdered)
```

```
##
##    N    Y
## 9946 3100
```

```r
odds <- 3100 / 9946
odds
```

```
## [1] 0.3116831
```

```r
log(odds)
```

```
## [1] -1.165768
```

Let's take a look on a logistic regression where a dependent variable is a factor 'Murdered' and it depends on all other predictors.

```r
fits <- glm(Murdered ~ . , df_for_models, family = "binomial")
summary(fits)
```

```
##
## Call:
## glm(formula = Murdered ~ ., family = "binomial", data = df_for_models)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2040  -0.7837  -0.6930  -0.5634   1.9692
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.53227    0.10622 -14.425  < 2e-16 ***
## AreaBROOKLYN        -0.13815    0.05063  -2.728 0.006364 **
```

```
## AreaMANHATTAN       -0.16128    0.06710  -2.403 0.016240 *
## AreaQUEENS          -0.09719    0.06467  -1.503 0.132841
## AreaSTATEN ISLAND   -0.11934    0.10822  -1.103 0.270112
## VIC_SEXM            -0.08992    0.06505  -1.382 0.166897
## VIC_AGE_GROUP18-24   0.22679    0.07887   2.876 0.004034 **
## VIC_AGE_GROUP25-44   0.33895    0.07839   4.324 1.53e-05 ***
## VIC_AGE_GROUP45-64   0.34847    0.10325   3.375 0.000739 ***
## VIC_AGE_GROUP65+     0.62309    0.22100   2.819 0.004811 **
## PERP_AGE_GROUP18-24  0.12070    0.07617   1.585 0.113058
## PERP_AGE_GROUP25-44  0.42145    0.07757   5.433 5.53e-08 ***
## PERP_AGE_GROUP45-64  0.76230    0.11721   6.504 7.84e-11 ***
## PERP_AGE_GROUP65+    0.97153    0.28343   3.428 0.000609 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14307  on 13045  degrees of freedom
## Residual deviance: 14142  on 13032  degrees of freedom
## AIC: 14170
##
## Number of Fisher Scoring iterations: 4
```

```
anova(fits, test="Chisq")
```

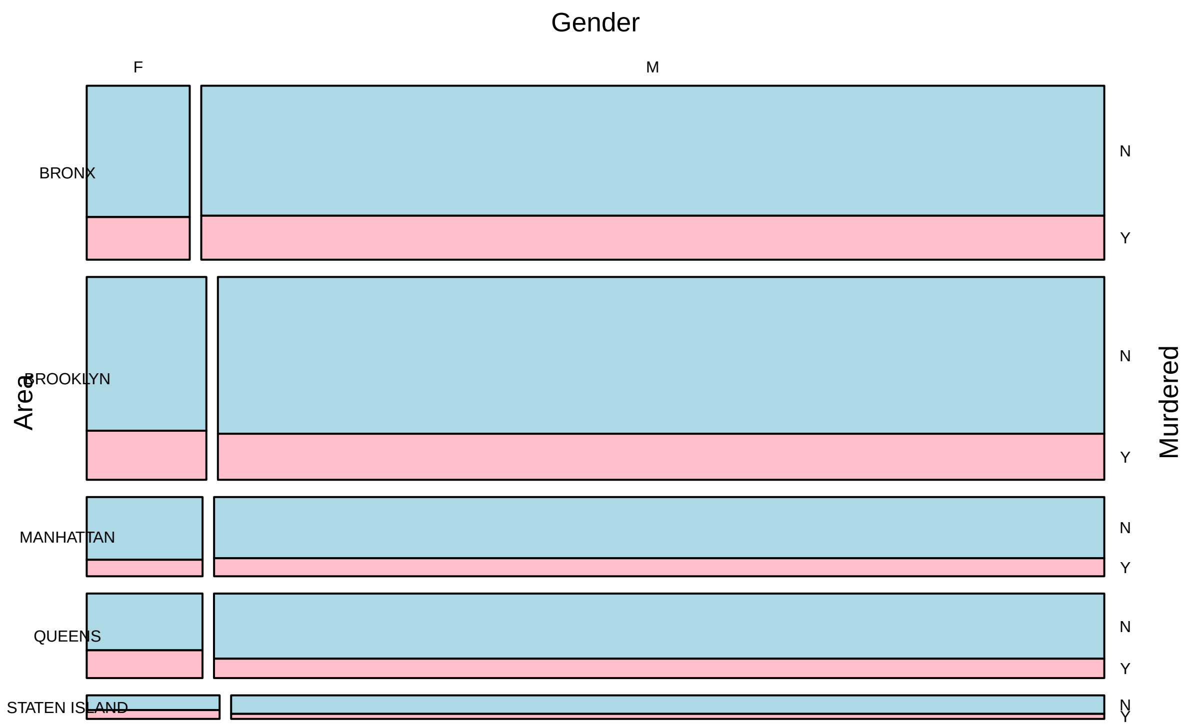
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Murdered
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           13045      14307
## Area            4    8.150     13041      14299   0.08625 .
## VIC_SEX         1    3.920     13040      14295   0.04772 *
## VIC_AGE_GROUP   4   67.444     13036      14227  7.86e-14 ***
## PERP_AGE_GROUP  4   85.205     13032      14142 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Firstly, we see that this model significantly improves our intercept only model. Also, we can see that victim sex and area don't have an impact on the model, whereas age groups of both perpetrators and victims have a significant influence.

Indeed, if we look at the mosaic plot, we can see that although the amount of cases and deaths of males and females vary a lot, the incident-murders ratio remains almost the same in each area for both genders

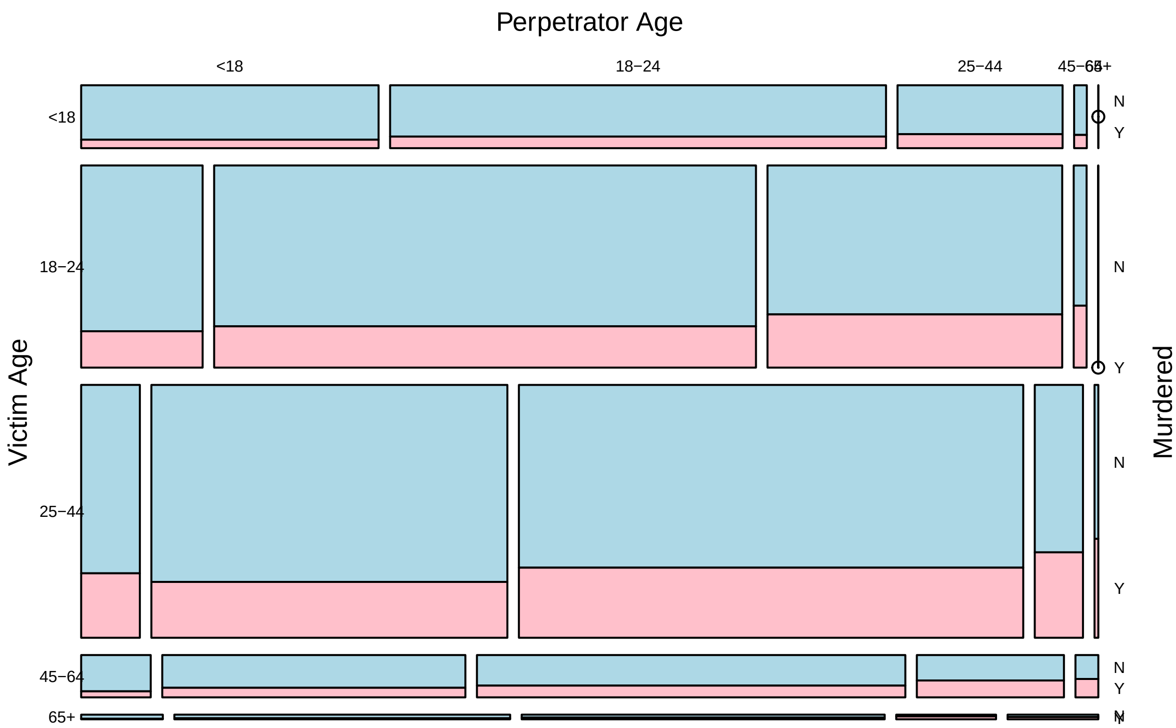
```
mosaic(Murdered ~ VIC_SEX | Area, data=df_for_models,
       highlighting_fill = c("lightblue", "pink"),
       labeling = labeling_border(rot_labels = c(0),
```

```
                                  gp_labels = gpar(fontsize = 6),
                                  set_varnames = c(VIC_SEX="Gender"),
                                  gp_varnames = gpar(fontsize = 10)
                                  )
       )
```



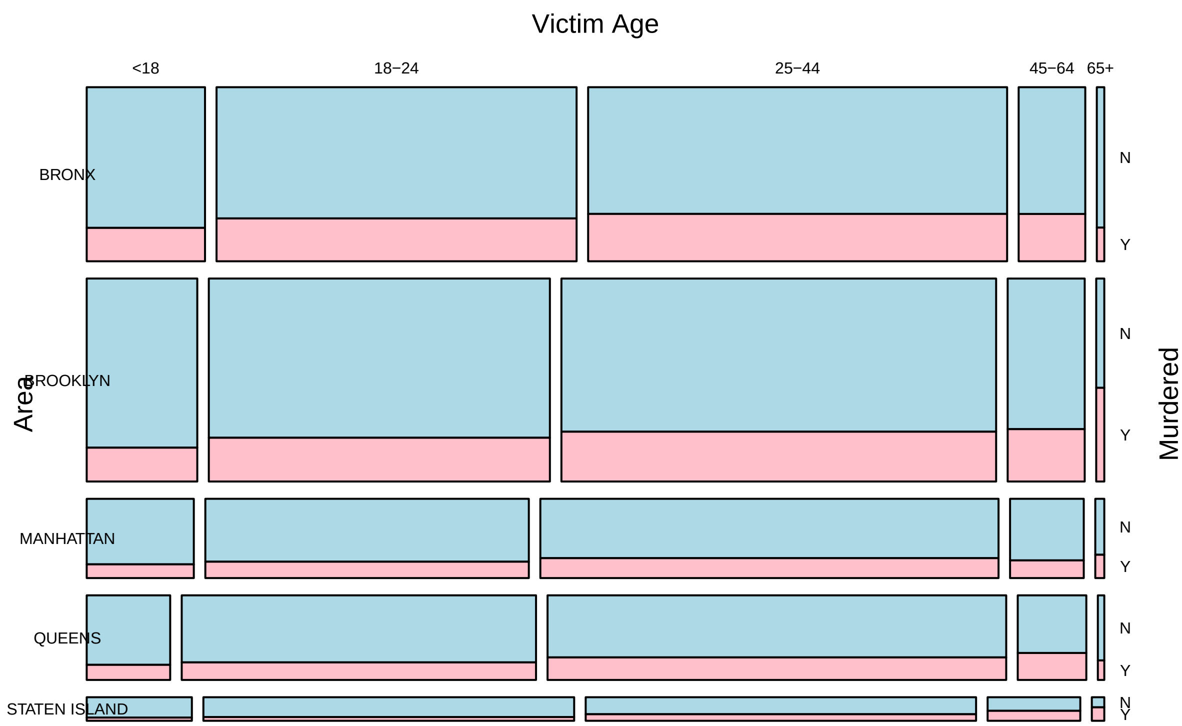
The most interesting deviations are in age groups:

```
mosaic(Murdered ~ PERP_AGE_GROUP | VIC_AGE_GROUP, data=df_for_models,
       highlighting_fill = c("lightblue", "pink"),
       labeling = labeling_border(rot_labels = c(0),
                                  gp_labels = gpar(fontsize = 6),
                                  set_varnames = c(PERP_AGE_GROUP='Perpetrator Age', VIC_AGE_GROUP='Vic
                                  gp_varnames = gpar(fontsize = 10)
       ))
```



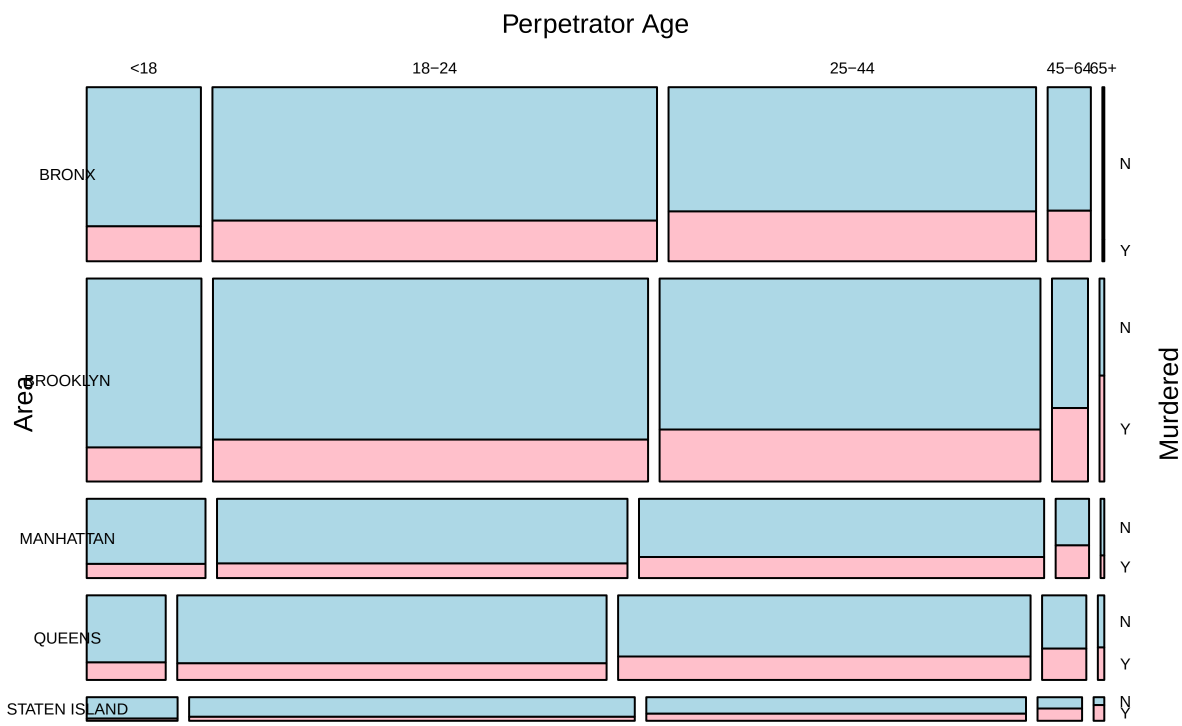
Moreover, although the area of an incident doesn't significantly impacts on the odds of being killed, we can see below that victims of different age group have different case-death ratio depending on the region.

```
mosaic(Murdered ~ VIC_AGE_GROUP | Area, data=df_for_models,
       highlighting_fill = c("lightblue", "pink"),
       labeling = labeling_border(rot_labels = c(0),
                                  gp_labels = gpar(fontsize = 6),
                                  set_varnames = c(VIC_AGE_GROUP='Victim Age'),
                                  gp_varnames = gpar(fontsize = 10)
       ))
```



And the same is for perpetrators of different age groups:

```
mosaic(Murdered ~ PERP_AGE_GROUP | Area, data=df_for_models,
       highlighting_fill = c("lightblue", "pink"),
       labeling = labeling_border(rot_labels = c(0),
                                  gp_labels = gpar(fontsize = 6),
                                  set_varnames = c(PERP_AGE_GROUP='Perpetrator Age'),
                                  gp_varnames = gpar(fontsize = 10)
       ))
```



All of these observations have to be analyzed and checked.

First, let's prove that victim sex is not statistically significant and it doesn't improve our model:

```
fit_sex <- glm(Murdered ~ VIC_SEX, df_for_models, family = "binomial")
coef(fit_sex)
```

```
## (Intercept)     VIC_SEXM
##  -1.0575351  -0.1225052
```

```
summary(fit_sex)
```

```
##
## Call:
## glm(formula = Murdered ~ VIC_SEX, family = "binomial", data = df_for_models)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7722  -0.7320  -0.7320  -0.7320   1.7018
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.05754    0.05947 -17.784   <2e-16 ***
## VIC_SEXM    -0.12251    0.06338  -1.933   0.0533 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14307  on 13045  degrees of freedom
## Residual deviance: 14303  on 13044  degrees of freedom
## AIC: 14307
##
## Number of Fisher Scoring iterations: 4
```

```
table(df_for_models$Murdered, df_for_models$VIC_SEX)
```

```
##
##         F    M
##   N 1097 8849
##   Y  381 2719
```

```
anova(simple_fit, fit_sex, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Murdered ~ 1
## Model 2: Murdered ~ VIC_SEX
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     13045      14307
## 2     13044      14303  1   3.6791   0.0551 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here our Intercept is a female and changing the gender to male doesn't have a significant impact.

Secondly, let's prove that age groups of both perpetrators and victims significantly improve our model:

```
fit_vic_age <- glm(Murdered ~ VIC_AGE_GROUP, df_for_models, family = "binomial")
coef(fit_vic_age)
```

```
##       (Intercept) VIC_AGE_GROUP18-24 VIC_AGE_GROUP25-44 VIC_AGE_GROUP45-64
##        -1.5423963          0.2790000          0.4898314          0.5695277
##   VIC_AGE_GROUP65+
##         0.8632353
```

```
summary(fit_vic_age)
```

```
##
## Call:
## glm(formula = Murdered ~ VIC_AGE_GROUP, family = "binomial",
##     data = df_for_models)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9057  -0.7738  -0.7056  -0.6226   1.8634
##
```

```
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.54240    0.06886 -22.399  < 2e-16 ***
## VIC_AGE_GROUP18-24    0.27900    0.07738   3.605 0.000312 ***
## VIC_AGE_GROUP25-44    0.48983    0.07506   6.526 6.77e-11 ***
## VIC_AGE_GROUP45-64    0.56953    0.09942   5.729 1.01e-08 ***
## VIC_AGE_GROUP65+      0.86324    0.21588   3.999 6.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14307  on 13045  degrees of freedom
## Residual deviance: 14240  on 13041  degrees of freedom
## AIC: 14250
##
## Number of Fisher Scoring iterations: 4
```

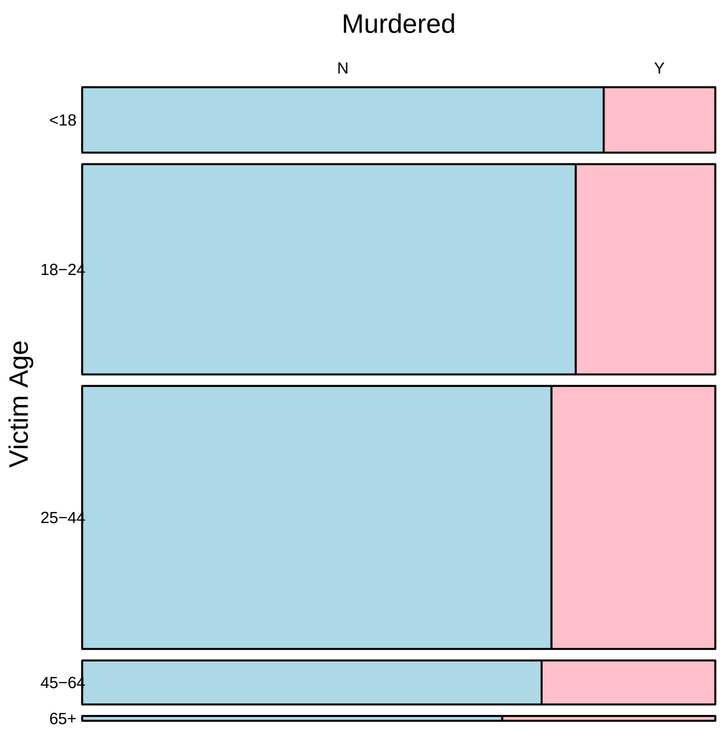
```
table(df_for_models$Murdered, df_for_models$VIC_AGE_GROUP)
```

```
##
##      <18 18-24 25-44 45-64  65+
##   N 1197  3640  4329   709   71
##   Y  256  1029  1511   268   36
```

```
anova(simple_fit, fit_vic_age, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Murdered ~ 1
## Model 2: Murdered ~ VIC_AGE_GROUP
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     13045      14307
## 2     13041      14240  4   66.602 1.183e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mosaic(Murdered ~ VIC_AGE_GROUP, data=df_for_models,
      highlighting_fill = c("lightblue", "pink"),
       labeling = labeling_border(rot_labels = c(0),
                                  gp_labels = gpar(fontsize = 6),
                                  set_varnames = c(VIC_AGE_GROUP="Victim Age"),
                                  gp_varnames = gpar(fontsize = 10)
                                  )
       )
```




```
fit_perp_age <- glm(Murdered ~ PERP_AGE_GROUP, df_for_models, family = "binomial")
coef(fit_perp_age)
```

```
##         (Intercept) PERP_AGE_GROUP18-24 PERP_AGE_GROUP25-44 PERP_AGE_GROUP45-64
##          -1.5024052           0.1719913           0.5162482           0.8957833
##   PERP_AGE_GROUP65+
##           1.1839514
```

```
summary(fit_perp_age)
```

```
##
## Call:
## glm(formula = Murdered ~ PERP_AGE_GROUP, family = "binomial",
##     data = df_for_models)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0455  -0.7963  -0.6849  -0.6340   1.8457
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.50241    0.06780 -22.161  < 2e-16 ***
## PERP_AGE_GROUP18-24  0.17199    0.07506   2.291   0.0219 *
## PERP_AGE_GROUP25-44  0.51625    0.07466   6.915 4.69e-12 ***
```

```
## PERP_AGE_GROUP45-64  0.89578    0.11340   7.899 2.80e-15 ***
## PERP_AGE_GROUP65+    1.18395    0.27671   4.279 1.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14307  on 13045  degrees of freedom
## Residual deviance: 14177  on 13041  degrees of freedom
## AIC: 14187
##
## Number of Fisher Scoring iterations: 4
```

```r
table(df_for_models$Murdered, df_for_models$PERP_AGE_GROUP)
```

```
##
##      <18 18-24 25-44 45-64  65+
##   N 1195  4611  3764   343   33
##   Y  266  1219  1404   187   24
```

```r
anova(simple_fit, fit_perp_age, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Murdered ~ 1
## Model 2: Murdered ~ PERP_AGE_GROUP
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     13045      14307
## 2     13041      14177  4   130.22 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, let's make the model with two categorical predictors of age groups, which should be significantly important.

```r
fit_ages <- glm(Murdered ~ VIC_AGE_GROUP * PERP_AGE_GROUP, df_for_models, family = "binomial")
coef(fit_ages)
```

```
##                            (Intercept)                    VIC_AGE_GROUP18-24
##                             -1.8588988                             0.3411585
##                     VIC_AGE_GROUP25-44                    VIC_AGE_GROUP45-64
##                              0.7865544                             0.0671393
##                      VIC_AGE_GROUP65+                   PERP_AGE_GROUP18-24
##                             -0.2205428                             0.3736512
##                    PERP_AGE_GROUP25-44                   PERP_AGE_GROUP45-64
##                              0.6087366                             0.5371429
##                      PERP_AGE_GROUP65+ VIC_AGE_GROUP18-24:PERP_AGE_GROUP18-24
##                              2.0794415                            -0.2160196
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP18-24 VIC_AGE_GROUP45-64:PERP_AGE_GROUP18-24
##                             -0.5615921                             0.1883829
##   VIC_AGE_GROUP65+:PERP_AGE_GROUP18-24 VIC_AGE_GROUP18-24:PERP_AGE_GROUP25-44
##                              0.4179360                            -0.1202327
```

```
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP25-44 VIC_AGE_GROUP45-64:PERP_AGE_GROUP25-44
##                             -0.4943702                              0.2307217
##   VIC_AGE_GROUP65+:PERP_AGE_GROUP25-44 VIC_AGE_GROUP18-24:PERP_AGE_GROUP45-64
##                              0.9598793                              0.1638362
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP45-64 VIC_AGE_GROUP45-64:PERP_AGE_GROUP45-64
##                             -0.1373265                              0.8435176
##   VIC_AGE_GROUP65+:PERP_AGE_GROUP45-64   VIC_AGE_GROUP18-24:PERP_AGE_GROUP65+
##                              2.1019144                            -11.1277290
##   VIC_AGE_GROUP25-44:PERP_AGE_GROUP65+   VIC_AGE_GROUP45-64:PERP_AGE_GROUP65+
##                             -1.4489299                             -0.5500463
##     VIC_AGE_GROUP65+:PERP_AGE_GROUP65+
##                                     NA
```

```
summary(fit_ages)
```

```
##
## Call:
## glm(formula = Murdered ~ VIC_AGE_GROUP * PERP_AGE_GROUP, family = "binomial",
##     data = df_for_models)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4224  -0.7816  -0.6759  -0.5382   2.0963
##
## Coefficients: (1 not defined because of singularities)
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -1.85890    0.13880 -13.393  < 2e-16
## VIC_AGE_GROUP18-24                       0.34116    0.17571   1.942  0.05219
## VIC_AGE_GROUP25-44                       0.78655    0.18487   4.255 2.09e-05
## VIC_AGE_GROUP45-64                       0.06714    0.36869   0.182  0.85550
## VIC_AGE_GROUP65+                        -0.22054    1.06970  -0.206  0.83666
## PERP_AGE_GROUP18-24                      0.37365    0.16798   2.224  0.02612
## PERP_AGE_GROUP25-44                      0.60874    0.20653   2.947  0.00320
## PERP_AGE_GROUP45-64                      0.53714    0.57960   0.927  0.35405
## PERP_AGE_GROUP65+                        2.07944    1.23491   1.684  0.09220
## VIC_AGE_GROUP18-24:PERP_AGE_GROUP18-24  -0.21602    0.20540  -1.052  0.29294
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP18-24  -0.56159    0.21411  -2.623  0.00872
## VIC_AGE_GROUP45-64:PERP_AGE_GROUP18-24   0.18838    0.40449   0.466  0.64141
## VIC_AGE_GROUP65+:PERP_AGE_GROUP18-24     0.41794    1.14573   0.365  0.71528
## VIC_AGE_GROUP18-24:PERP_AGE_GROUP25-44  -0.12023    0.24063  -0.500  0.61732
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP25-44  -0.49437    0.24334  -2.032  0.04219
## VIC_AGE_GROUP45-64:PERP_AGE_GROUP25-44   0.23072    0.41336   0.558  0.57674
## VIC_AGE_GROUP65+:PERP_AGE_GROUP25-44     0.95988    1.12886   0.850  0.39515
## VIC_AGE_GROUP18-24:PERP_AGE_GROUP45-64   0.16384    0.65071   0.252  0.80121
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP45-64  -0.13733    0.60519  -0.227  0.82049
## VIC_AGE_GROUP45-64:PERP_AGE_GROUP45-64   0.84352    0.69338   1.217  0.22379
## VIC_AGE_GROUP65+:PERP_AGE_GROUP45-64     2.10191    1.36154   1.544  0.12264
## VIC_AGE_GROUP18-24:PERP_AGE_GROUP65+   -11.12773  119.47447  -0.093  0.92579
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP65+    -1.44893    1.31242  -1.104  0.26959
## VIC_AGE_GROUP45-64:PERP_AGE_GROUP65+    -0.55005    1.34855  -0.408  0.68336
## VIC_AGE_GROUP65+:PERP_AGE_GROUP65+            NA         NA      NA       NA
##
## (Intercept)                            ***
## VIC_AGE_GROUP18-24                     .
```

```
## VIC_AGE_GROUP25-44                     ***
## VIC_AGE_GROUP45-64
## VIC_AGE_GROUP65+
## PERP_AGE_GROUP18-24                    *
## PERP_AGE_GROUP25-44                    **
## PERP_AGE_GROUP45-64
## PERP_AGE_GROUP65+                      .
## VIC_AGE_GROUP18-24:PERP_AGE_GROUP18-24
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP18-24 **
## VIC_AGE_GROUP45-64:PERP_AGE_GROUP18-24
## VIC_AGE_GROUP65+:PERP_AGE_GROUP18-24
## VIC_AGE_GROUP18-24:PERP_AGE_GROUP25-44
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP25-44 *
## VIC_AGE_GROUP45-64:PERP_AGE_GROUP25-44
## VIC_AGE_GROUP65+:PERP_AGE_GROUP25-44
## VIC_AGE_GROUP18-24:PERP_AGE_GROUP45-64
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP45-64
## VIC_AGE_GROUP45-64:PERP_AGE_GROUP45-64
## VIC_AGE_GROUP65+:PERP_AGE_GROUP45-64
## VIC_AGE_GROUP18-24:PERP_AGE_GROUP65+
## VIC_AGE_GROUP25-44:PERP_AGE_GROUP65+
## VIC_AGE_GROUP45-64:PERP_AGE_GROUP65+
## VIC_AGE_GROUP65+:PERP_AGE_GROUP65+
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14307  on 13045  degrees of freedom
## Residual deviance: 14136  on 13022  degrees of freedom
## AIC: 14184
##
## Number of Fisher Scoring iterations: 9
```

```
anova(simple_fit, fit_ages, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Murdered ~ 1
## Model 2: Murdered ~ VIC_AGE_GROUP * PERP_AGE_GROUP
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     13045      14307
## 2     13022      14136 23   170.91 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the next step, we can have a logistic regression depending on an area predictor. Overall, this model doesn't have a significant improvement compared to intercept only model, however we see that changing to some areas may influence on estimated odds.

```
fit_area <- glm(Murdered ~ Area, df_for_models, family = "binomial")
coef(fit_area)
```

```
##       (Intercept)      AreaBROOKLYN     AreaMANHATTAN        AreaQUEENS
##       -1.08770489       -0.12775803       -0.14343184       -0.05993344
## AreaSTATEN ISLAND
##       -0.09335107
```

```r
summary(fit_area)
```

```
##
## Call:
## glm(formula = Murdered ~ Area, family = "binomial", data = df_for_models)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7621  -0.7425  -0.7207  -0.7158   1.7247
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.08770    0.03631 -29.955   <2e-16 ***
## AreaBROOKLYN      -0.12776    0.05027  -2.542   0.0110 *
## AreaMANHATTAN     -0.14343    0.06663  -2.153   0.0313 *
## AreaQUEENS        -0.05993    0.06415  -0.934   0.3502
## AreaSTATEN ISLAND -0.09335    0.10737  -0.869   0.3846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14307  on 13045  degrees of freedom
## Residual deviance: 14299  on 13041  degrees of freedom
## AIC: 14309
##
## Number of Fisher Scoring iterations: 4
```

```r
table(df_for_models$Murdered, df_for_models$Area)
```

```
##
##     BRONX BROOKLYN MANHATTAN QUEENS STATEN ISLAND
##   N  3009     3618      1418   1484           417
##   Y  1014     1073       414    471           128
```

```r
anova(simple_fit, fit_area, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Murdered ~ 1
## Model 2: Murdered ~ Area
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     13045      14307
## 2     13041      14299  4   8.1496  0.08625 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, let's create a model with two predictors - age group of perpetrators and area with their dependencies. And that works fine.

```
fit_perp_age_area <- glm(Murdered ~ PERP_AGE_GROUP * Area, df_for_models, family = "binomial")
coef(fit_perp_age_area)
```

```
##                          (Intercept)                  PERP_AGE_GROUP18-24
##                          -1.38101730                           0.19274856
##                  PERP_AGE_GROUP25-44                  PERP_AGE_GROUP45-64
##                           0.46840305                           0.48807394
##                    PERP_AGE_GROUP65+                         AreaBROOKLYN
##                          -0.56489284                          -0.21978125
##                        AreaMANHATTAN                           AreaQUEENS
##                          -0.14503900                           0.04124296
##                    AreaSTATEN ISLAND     PERP_AGE_GROUP18-24:AreaBROOKLYN
##                          -0.83818615                           0.06231364
##     PERP_AGE_GROUP25-44:AreaBROOKLYN     PERP_AGE_GROUP45-64:AreaBROOKLYN
##                           0.06614850                           0.54632914
##       PERP_AGE_GROUP65+:AreaBROOKLYN    PERP_AGE_GROUP18-24:AreaMANHATTAN
##                           2.25270277                          -0.15115349
##    PERP_AGE_GROUP25-44:AreaMANHATTAN    PERP_AGE_GROUP45-64:AreaMANHATTAN
##                           0.04113833                           0.68516099
##      PERP_AGE_GROUP65+:AreaMANHATTAN       PERP_AGE_GROUP18-24:AreaQUEENS
##                           1.17465842                          -0.25964774
##       PERP_AGE_GROUP25-44:AreaQUEENS       PERP_AGE_GROUP45-64:AreaQUEENS
##                          -0.08747905                           0.32285628
##         PERP_AGE_GROUP65+:AreaQUEENS PERP_AGE_GROUP18-24:AreaSTATEN ISLAND
##                           1.43466356                           0.42658644
## PERP_AGE_GROUP25-44:AreaSTATEN ISLAND PERP_AGE_GROUP45-64:AreaSTATEN ISLAND
##                           0.90573719                           1.81117223
##   PERP_AGE_GROUP65+:AreaSTATEN ISLAND
##                           3.47724348
```

```
summary(fit_perp_age_area)
```

```
##
## Call:
## glm(formula = Murdered ~ PERP_AGE_GROUP * Area, family = "binomial",
##     data = df_for_models)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4823  -0.7692  -0.6803  -0.6062   2.1552
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -1.38102    0.11477 -12.033  < 2e-16 ***
## PERP_AGE_GROUP18-24                    0.19275    0.12730   1.514 0.130003
## PERP_AGE_GROUP25-44                    0.46840    0.12801   3.659 0.000253 ***
## PERP_AGE_GROUP45-64                    0.48807    0.20069   2.432 0.015017 *
## PERP_AGE_GROUP65+                     -0.56489    1.07518  -0.525 0.599311
## AreaBROOKLYN                          -0.21978    0.16154  -1.361 0.173648
## AreaMANHATTAN                         -0.14504    0.20882  -0.695 0.487332
```

```
## AreaQUEENS                             0.04124    0.22674   0.182 0.855665
## AreaSTATEN ISLAND                     -0.83819    0.48463  -1.730 0.083713 .
## PERP_AGE_GROUP18-24:AreaBROOKLYN       0.06231    0.17897   0.348 0.727713
## PERP_AGE_GROUP25-44:AreaBROOKLYN       0.06615    0.17934   0.369 0.712241
## PERP_AGE_GROUP45-64:AreaBROOKLYN       0.54633    0.27943   1.955 0.050564 .
## PERP_AGE_GROUP65+:AreaBROOKLYN         2.25270    1.15896   1.944 0.051928 .
## PERP_AGE_GROUP18-24:AreaMANHATTAN     -0.15115    0.23499  -0.643 0.520066
## PERP_AGE_GROUP25-44:AreaMANHATTAN      0.04114    0.23136   0.178 0.858874
## PERP_AGE_GROUP45-64:AreaMANHATTAN      0.68516    0.36905   1.857 0.063377 .
## PERP_AGE_GROUP65+:AreaMANHATTAN        1.17466    1.37348   0.855 0.392418
## PERP_AGE_GROUP18-24:AreaQUEENS        -0.25965    0.24853  -1.045 0.296153
## PERP_AGE_GROUP25-44:AreaQUEENS        -0.08748    0.24625  -0.355 0.722409
## PERP_AGE_GROUP45-64:AreaQUEENS         0.32286    0.35592   0.907 0.364347
## PERP_AGE_GROUP65+:AreaQUEENS           1.43466    1.23258   1.164 0.244445
## PERP_AGE_GROUP18-24:AreaSTATEN ISLAND  0.42659    0.51625   0.826 0.408627
## PERP_AGE_GROUP25-44:AreaSTATEN ISLAND  0.90574    0.51031   1.775 0.075919 .
## PERP_AGE_GROUP45-64:AreaSTATEN ISLAND  1.81117    0.64979   2.787 0.005315 **
## PERP_AGE_GROUP65+:AreaSTATEN ISLAND    3.47724    1.45867   2.384 0.017133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14307  on 13045  degrees of freedom
## Residual deviance: 14140  on 13021  degrees of freedom
## AIC: 14190
##
## Number of Fisher Scoring iterations: 4
```

```
anova(simple_fit, fit_perp_age_area, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Murdered ~ 1
## Model 2: Murdered ~ PERP_AGE_GROUP * Area
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     13045      14307
## 2     13021      14140 24   167.17 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, our model will include victim age group, perpetrator age group, area and their dependencies:

```
fit_long <- glm(Murdered ~ VIC_AGE_GROUP + PERP_AGE_GROUP * Area, df_for_models, family = "binomial")
summary(fit_long)
```

```
##
## Call:
## glm(formula = Murdered ~ VIC_AGE_GROUP + PERP_AGE_GROUP * Area,
##     family = "binomial", data = df_for_models)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.4623  -0.7832  -0.6764  -0.5560   2.2252
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -1.57981    0.12788 -12.354  < 2e-16 ***
## VIC_AGE_GROUP18-24                    0.21720    0.07869   2.760  0.00578 **
## VIC_AGE_GROUP25-44                    0.33376    0.07827   4.264 2.01e-05 ***
## VIC_AGE_GROUP45-64                    0.33847    0.10359   3.267  0.00109 **
## VIC_AGE_GROUP65+                      0.63216    0.22203   2.847  0.00441 **
## PERP_AGE_GROUP18-24                   0.15147    0.12782   1.185  0.23604
## PERP_AGE_GROUP25-44                   0.38288    0.12947   2.957  0.00310 **
## PERP_AGE_GROUP45-64                   0.37428    0.20274   1.846  0.06488 .
## PERP_AGE_GROUP65+                    -0.78142    1.07828  -0.725  0.46865
## AreaBROOKLYN                         -0.20899    0.16179  -1.292  0.19645
## AreaMANHATTAN                        -0.14757    0.20911  -0.706  0.48037
## AreaQUEENS                            0.04673    0.22711   0.206  0.83698
## AreaSTATEN ISLAND                    -0.80799    0.48505  -1.666  0.09576 .
## PERP_AGE_GROUP18-24:AreaBROOKLYN      0.04515    0.17926   0.252  0.80115
## PERP_AGE_GROUP25-44:AreaBROOKLYN      0.04740    0.17961   0.264  0.79186
## PERP_AGE_GROUP45-64:AreaBROOKLYN      0.52983    0.27981   1.894  0.05829 .
## PERP_AGE_GROUP65+:AreaBROOKLYN        2.28268    1.16096   1.966  0.04928 *
## PERP_AGE_GROUP18-24:AreaMANHATTAN    -0.15651    0.23530  -0.665  0.50594
## PERP_AGE_GROUP25-44:AreaMANHATTAN     0.03482    0.23166   0.150  0.88051
## PERP_AGE_GROUP45-64:AreaMANHATTAN     0.66982    0.36961   1.812  0.06995 .
## PERP_AGE_GROUP65+:AreaMANHATTAN       1.21122    1.37570   0.880  0.37862
## PERP_AGE_GROUP18-24:AreaQUEENS       -0.27370    0.24896  -1.099  0.27161
## PERP_AGE_GROUP25-44:AreaQUEENS       -0.10235    0.24665  -0.415  0.67818
## PERP_AGE_GROUP45-64:AreaQUEENS        0.31293    0.35630   0.878  0.37979
## PERP_AGE_GROUP65+:AreaQUEENS          1.44715    1.23436   1.172  0.24104
## PERP_AGE_GROUP18-24:AreaSTATEN ISLAND  0.39173    0.51672   0.758  0.44838
## PERP_AGE_GROUP25-44:AreaSTATEN ISLAND  0.86680    0.51078   1.697  0.08969 .
## PERP_AGE_GROUP45-64:AreaSTATEN ISLAND  1.79422    0.65059   2.758  0.00582 **
## PERP_AGE_GROUP65+:AreaSTATEN ISLAND    3.47945    1.46060   2.382  0.01721 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14307  on 13045  degrees of freedom
## Residual deviance: 14116  on 13017  degrees of freedom
## AIC: 14174
##
## Number of Fisher Scoring iterations: 4
```

```r
anova(fit_long, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Murdered
##
## Terms added sequentially (first to last)
```

```
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                13045      14307
## VIC_AGE_GROUP        4   66.602     13041      14240 1.183e-13 ***
## PERP_AGE_GROUP       4   86.815     13037      14153 < 2.2e-16 ***
## Area                 4    9.412     13033      14144   0.05159 .
## PERP_AGE_GROUP:Area 16   28.222     13017      14116   0.02974 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running The Model

Based on our final model let's make some predictions. First, let's imagine that a shooting incident happened in Manhattan area, both victim and perpetrator are less than 18 y.o. Seems that this would be a scenario, when a victim has maximum chances to stay alive in the shooting incident

```
new_df_min <- data.frame(VIC_SEX = 'M',
                     VIC_AGE_GROUP = "<18",
                     PERP_AGE_GROUP = "<18", Area = 'STATEN ISLAND')
predict(fit_long, newdata = new_df_min, type = "response")
```

```
##          1
## 0.08410847
```

And otherwise, if an accident happen in Staten Island and both victim and perpetrator are older than 65, then there are least chances for a victim to stay alive.

```
new_df_max <- data.frame(VIC_SEX = 'M',
                         VIC_AGE_GROUP = "65+",
                         PERP_AGE_GROUP = "65+", Area = 'STATEN ISLAND')
predict(fit_long, newdata = new_df_max, type = "response")
```

```
##         1
## 0.7195847
```

Conclusion

Based on the analysis, there are significant difference in the number of shooting incidents in different areas. There's also an increasing trend seen in the last 3 years.

In most of the cases males involved, however the chances of being murdered are the same for both genders.

The number of incidents resulting death significantly depends on the age of a perpetrator and a victim.

Incidents in which both criminal and a victim are younger than 18 are the least deadly.

And the most deadly cases happened in the most 'calm' area of Staten Island where participants are older than 65.

In this data analysis project all the categorical variables may seem to be a source of biases. Everything here, including sex, age, race and area is a very sensitive subject.

As a data scientist one have to mitigate his own personal biases and look on the data as on the list of numbers and variables regardless of his own experience, opinion etc.

Appendix

Session info:

```
## R version 4.2.0 (2022-04-22 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] vcd_1.4-10      tidyr_1.2.0     lubridate_1.8.0 dplyr_1.0.9
## [5] ggplot2_3.3.6
##
## loaded via a namespace (and not attached):
##  [1] highr_0.9        pillar_1.7.0     compiler_4.2.0   tools_4.2.0
##  [5] digest_0.6.29    nlme_3.1-157     lattice_0.20-45  evaluate_0.15
##  [9] lifecycle_1.0.1  tibble_3.1.7     gtable_0.3.0     mgcv_1.8-40
## [13] pkgconfig_2.0.3  rlang_1.0.2      Matrix_1.4-1     cli_3.3.0
## [17] DBI_1.1.3        rstudioapi_0.13  yaml_2.3.5       xfun_0.31
## [21] fastmap_1.1.0    withr_2.5.0      stringr_1.4.0    knitr_1.39
## [25] generics_0.1.2   vctrs_0.4.1      lmtest_0.9-40    tidyselect_1.1.2
## [29] glue_1.6.2       R6_2.5.1         fansi_1.0.3      rmarkdown_2.14
## [33] farver_2.1.0     purrr_0.3.4      magrittr_2.0.3   splines_4.2.0
## [37] MASS_7.3-56      scales_1.2.0     ellipsis_0.3.2   htmltools_0.5.2
## [41] assertthat_0.2.1 colorspace_2.0-3 labeling_0.4.2   utf8_1.2.2
## [45] stringi_1.7.6    munsell_0.5.0    crayon_1.5.1     zoo_1.8-10
```