NYPD shooting incident report

2024-02-24

Problem

Utilizing data from NYPD shooting incidents, I aim to address the following inquiries: 'Which locations and times pose the greatest risk to young women in New York?' Additionally, I seek to explore the correlation between the number of incidents and the number of murders in specific boroughs.

Data Description

The data being analyzed was collected and provided by the NYPD and spans from 2006 to 2022. Data to be used in the analysis:

Column Name	Data Type	Description
OCCUR_DATE	chr	Exact date of the shooting incident
OCCUR_TIME	S3: hms	Exact time of the shooting incident
BORO	chr	Borough where the shooting incident occurred
VIC_AGE_GROUP	chr	Victim's age within a category
VIC_SEX	chr	Victim's sex description
VIC_RACE	chr	Victim's race description
STATISTICAL_MURDER_FLAG	lgl	Shooting resulted in the victim's death which would be counted as a murder

Import Data

The data is initially imported allowing it to be analyzed.

During data tiding and cleaning I choose to remove such variables INCIDENT_KEY, STATISTICAL_MURDER_FLAG, Latitude, Longitude, Lon_Lat, X_COORD_CD, Y_COORD_CD, PRECINCT, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOC_OF_OCCUR_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, LOCATION_DESC, VIC_RACE.

Additional changes:

- add OCCUR_YEAR (year of the shooting incident)
- add OCCUR_HOUR (hour of the shooting incident)

```
ny_inc <- ny_inc_raw %>%
  select(-c(INCIDENT_KEY, Latitude, Longitude, Lon_Lat, X_COORD_CD, Y_COORD_CD, PRECINCT, JURISDICTION_C
  # change to date format and add hour, month, year of indecent
  mutate(OCCUR_YEAR = as.factor(year(mdy(OCCUR_DATE))), OCCUR_HOUR=hour(OCCUR_TIME))
head(ny_inc)
```

```
## # A tibble: 6 x 8
##   OCCUR_DATE OCCUR_TIME BORO     STATISTICAL_MURDER_FLAG VIC_AGE_GROUP VIC_SEX
##   <chr>      <time>     <chr>    <lgl>                   <chr>         <chr>
## 1 05/27/2021 21:30      QUEENS   FALSE                   18-24         M
## 2 06/27/2014 17:40      BRONX    FALSE                   18-24         M
## 3 11/21/2015 03:56      QUEENS   TRUE                    25-44         M
## 4 10/09/2015 18:30      BRONX    FALSE                   <18           M
## 5 02/19/2009 22:58      BRONX    TRUE                    45-64         M
## 6 10/21/2020 21:36      BROOKLYN TRUE                    25-44         M
## # i 2 more variables: OCCUR_YEAR <fct>, OCCUR_HOUR <int>
```

Summary to check missing data

```
ny_inc %>%
  summarize(OCCUR_DATE_NA = sum(is.na(ny_inc$OCCUR_DATE)),
            OCCUR_TIME_NA = sum(is.na(ny_inc$OCCUR_TIME)),
            OCCUR_YEAR_NA = sum(is.na(ny_inc$OCCUR_YEAR)),
            OCCUR_HOUR_NA = sum(is.na(ny_inc$OCCUR_HOUR)),
            BORO_NA = sum(is.na(ny_inc$BORO)),
            VIC_AGE_GROUP_NA = sum(is.na(ny_inc$VIC_AGE_GROUP)),
            VIC_SEX_NA = sum(is.na(ny_inc$VIC_SEX))
  )
```

```
## # A tibble: 1 x 7
##   OCCUR_DATE_NA OCCUR_TIME_NA OCCUR_YEAR_NA OCCUR_HOUR_NA BORO_NA
##           <int>         <int>         <int>         <int>   <int>
## 1             0             0             0             0       0
## # i 2 more variables: VIC_AGE_GROUP_NA <int>, VIC_SEX_NA <int>
```

As result we don't have missing data in the cleaned dataset.

Analysis

Let's filter data for young female (<45 years).

```
inc_vic_female <- ny_inc %>%
  filter(VIC_SEX=="F", VIC_AGE_GROUP=="18-24"|VIC_AGE_GROUP=="25-44"|VIC_AGE_GROUP=="<18")
head(inc_vic_female)
```

```
## # A tibble: 6 x 8
##   OCCUR_DATE OCCUR_TIME BORO      STATISTICAL_MURDER_FLAG VIC_AGE_GROUP VIC_SEX
##   <chr>      <time>     <chr>     <lgl>                   <chr>         <chr>
## 1 02/01/2015 23:16      MANHATTAN TRUE                    18-24         F
## 2 11/21/2017 22:25      BROOKLYN  TRUE                    25-44         F
## 3 09/01/2009 16:00      BROOKLYN  FALSE                   18-24         F
## 4 09/06/2011 02:20      QUEENS    FALSE                   18-24         F
## 5 02/09/2006 14:55      QUEENS    TRUE                    25-44         F
## 6 09/28/2021 20:40      MANHATTAN FALSE                   25-44         F
## # i 2 more variables: OCCUR_YEAR <fct>, OCCUR_HOUR <int>
```

Group by borough and hour of shooting incidence.

```r
inc_vic_female_by_boro <- inc_vic_female %>%
  group_by(BORO, OCCUR_HOUR) %>%
  summarise(N_INC=n(), N_MURDER=sum(STATISTICAL_MURDER_FLAG))
```

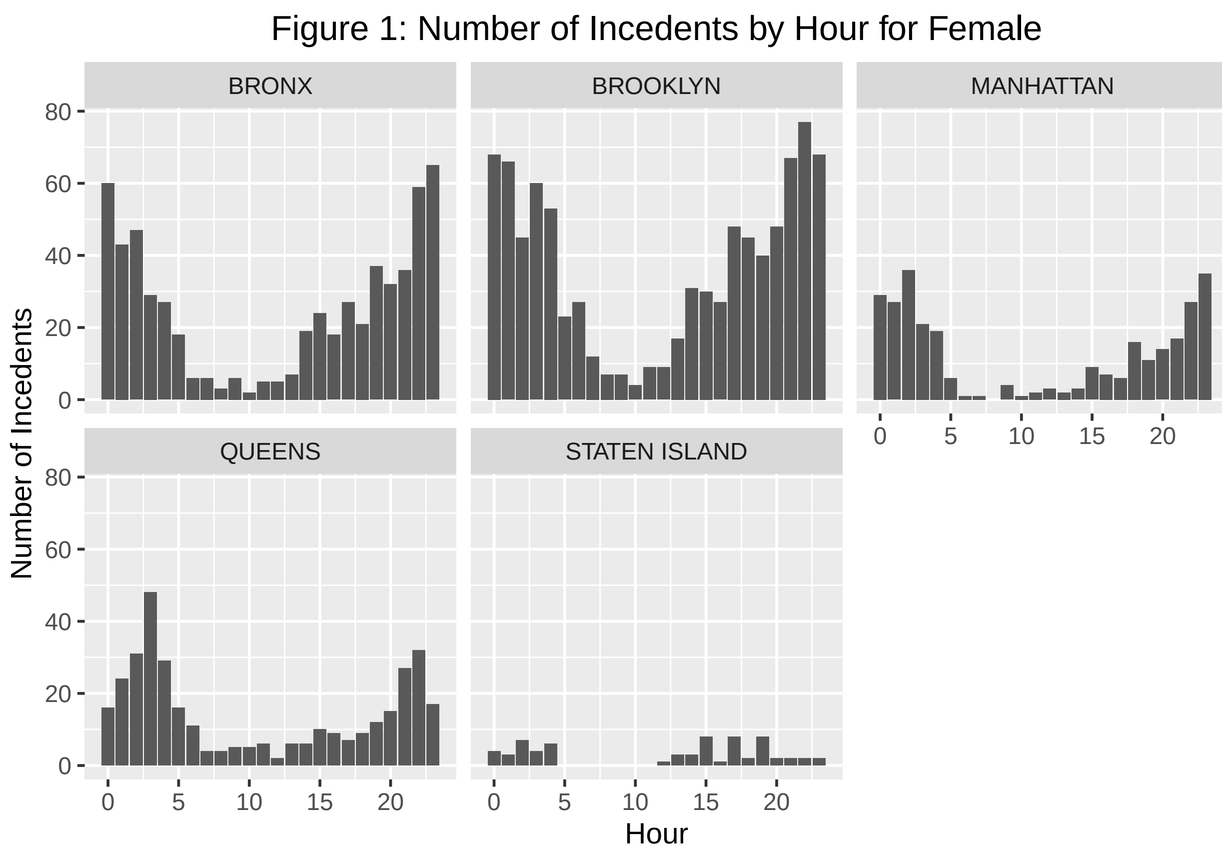
```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```

```r
head(inc_vic_female_by_boro)
```

```
## # A tibble: 6 x 4
## # Groups:   BORO [1]
##   BORO  OCCUR_HOUR N_INC N_MURDER
##   <chr>      <int> <int>    <int>
## 1 BRONX          0    60        8
## 2 BRONX          1    43        9
## 3 BRONX          2    47        9
## 4 BRONX          3    29        4
## 5 BRONX          4    27        5
## 6 BRONX          5    18        4
```

Visualize how many female victims where in years from 2006 to 2021 for each borough.

```r
ggplot(inc_vic_female_by_boro, aes(x=OCCUR_HOUR, y=N_INC)) +
  labs(x="Hour", y="Number of Incedents") +
  ggtitle("Figure 1: Number of Incedents by Hour for Female", ) +
  # center title
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_col() +
  facet_wrap(~BORO)
```

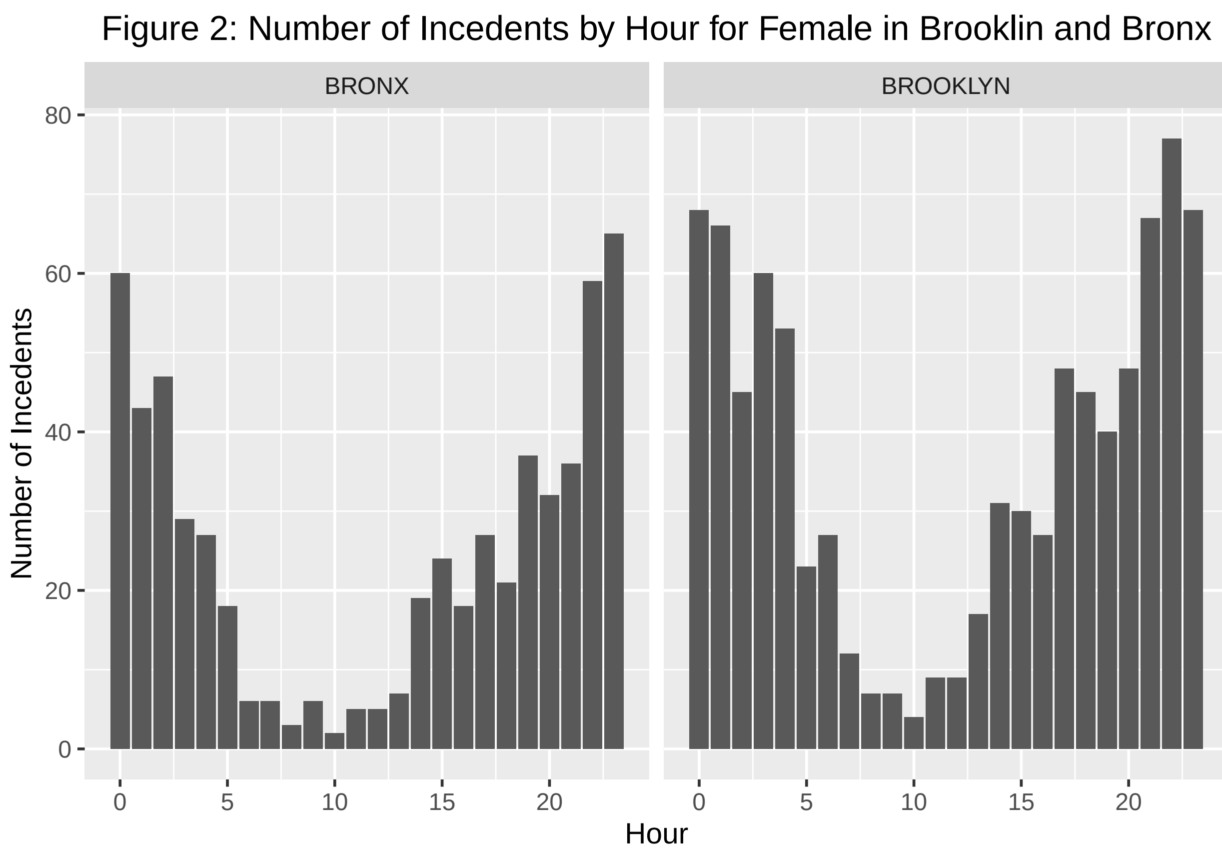


As we see, the most dangerous boroughs where Bronx and Brooklyn, the safest borough is Staten Island. Focus on Bronx and Brooklyn areas for analysis.

```
inc_brooklyn_bronx <- inc_vic_female_by_boro %>%
  filter(BORO=="BROOKLYN"|BORO=="BRONX")
head(inc_brooklyn_bronx)
```

```
## # A tibble: 6 x 4
## # Groups:   BORO [1]
##   BORO  OCCUR_HOUR N_INC N_MURDER
##   <chr>      <int> <int>    <int>
## 1 BRONX          0    60        8
## 2 BRONX          1    43        9
## 3 BRONX          2    47        9
## 4 BRONX          3    29        4
## 5 BRONX          4    27        5
## 6 BRONX          5    18        4
```

```
ggplot(inc_brooklyn_bronx, aes(x=OCCUR_HOUR, y=N_INC)) +
  labs(x="Hour", y="Number of Incedents") +
  ggtitle("Figure 2: Number of Incedents by Hour for Female in Brooklin and Bronx", ) +
  # center title
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_col() +
  facet_wrap(~BORO)
```



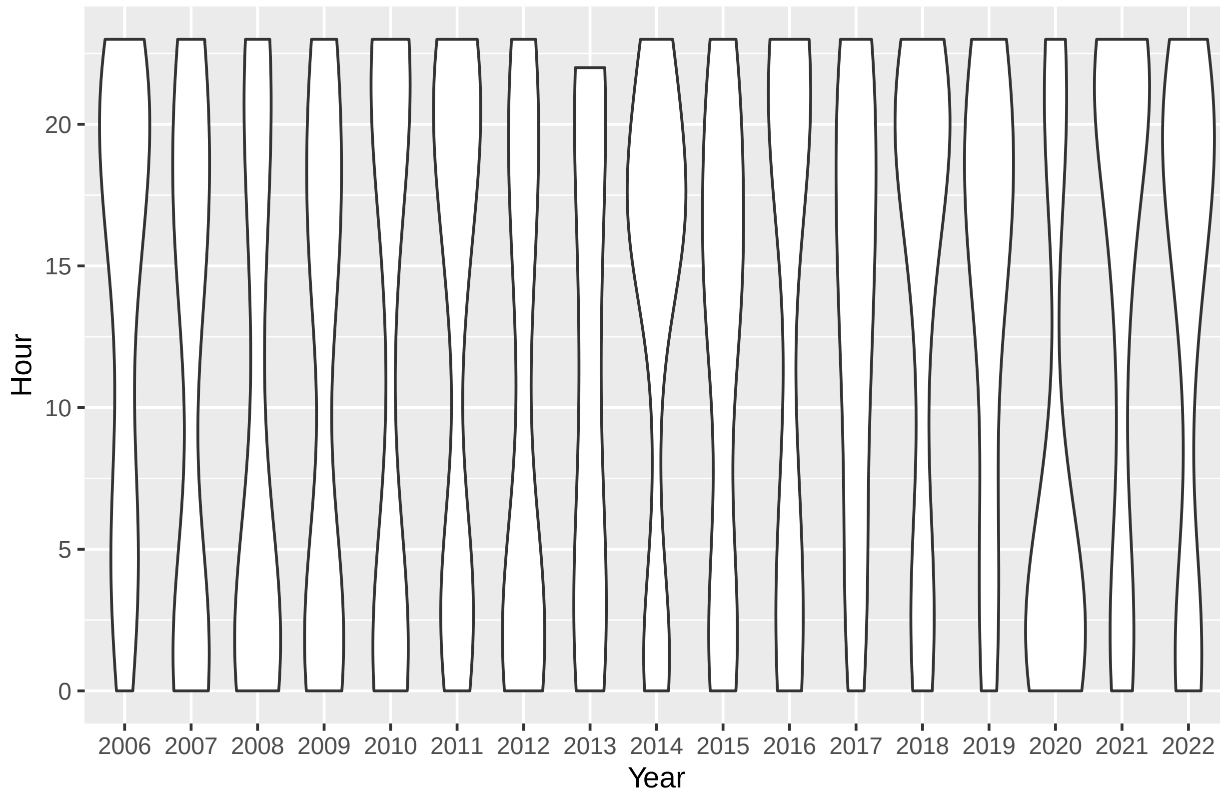
After reviewing Figure 2, it's evident that the peak of violence occurs between 8pm and 4am. Further investigation is needed to determine if this pattern remains consistent across the years.

```
inc_vic_female_bronx <- inc_vic_female %>%
  filter(BORO=="BRONX")
```

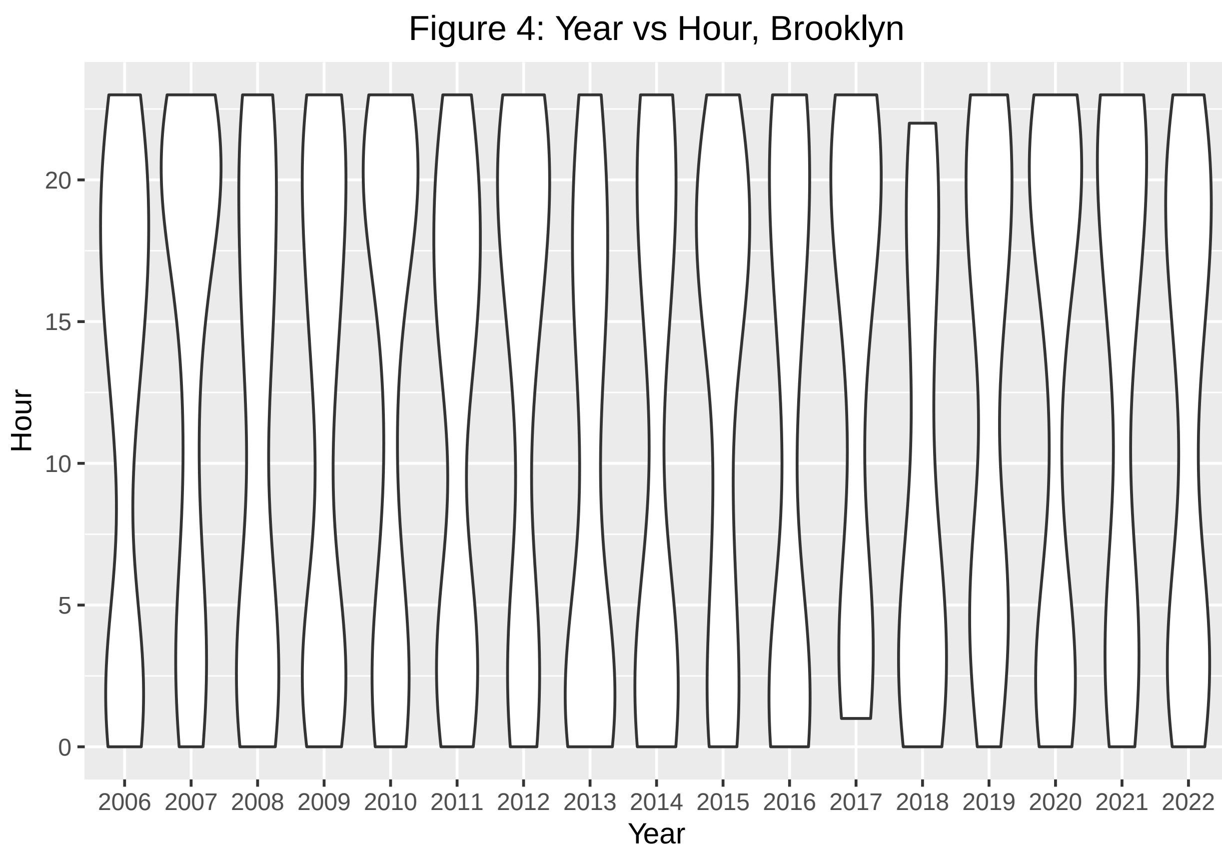
```
inc_vic_female_brook <- inc_vic_female %>%
  filter(BORO=="BROOKLYN")
```

```
ggplot(inc_vic_female_bronx, aes(x=OCCUR_YEAR, y=OCCUR_HOUR)) +
  labs(x="Year", y="Hour") +
  ggtitle("Figure 3: Year vs Hour, Bronx", ) +
  # center title
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_violin()
```

Figure 3: Year vs Hour, Bronx




```r
ggplot(inc_vic_female_brook, aes(x=OCCUR_YEAR, y=OCCUR_HOUR)) +
  labs(x="Year", y="Hour") +
  ggtitle("Figure 4: Year vs Hour, Brooklyn", ) +
  # center title
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_violin()
```



We observe that the majority of incidents in the Bronx were reported during the late evening hours. However, the data for the year 2020 deviates from this trend, suggesting the need for additional data from other sources to facilitate further analysis. In Brooklyn, the distribution of incidents is consistent, with the most dangerous hours for young women being late evening and early morning.

Model

The final step in the analysis involved creating a model to assess the relationship between the number of shootings and the corresponding number of murders per year for young women in both the safest borough (Staten Island) and the most dangerous borough (Brooklyn).

First, filter data and group by year.

```
inc_murd_f_st <- inc_vic_female %>%
  filter(BORO=="STATEN ISLAND") %>%
  group_by(OCCUR_YEAR) %>%
  summarise(N_INC=n(), N_MURDER=sum(STATISTICAL_MURDER_FLAG))
head(inc_murd_f_st)
```

```
## # A tibble: 6 x 3
##   OCCUR_YEAR N_INC N_MURDER
##   <fct>      <int>    <int>
## 1 2006           7        3
## 2 2007           5        1
## 3 2008           5        2
## 4 2009           1        0
```

```
## 5 2010           1        1
## 6 2011           6        1
```

```r
inc_murd_f_brook <- inc_vic_female %>%
  filter(BORO=="BROOKLYN") %>%
  group_by(OCCUR_YEAR) %>%
  summarise(N_INC=n(), N_MURDER=sum(STATISTICAL_MURDER_FLAG))
head(inc_murd_f_brook)
```

```
## # A tibble: 6 x 3
##   OCCUR_YEAR N_INC N_MURDER
##   <fct>      <int>    <int>
## 1 2006          54       10
## 2 2007          57       11
## 3 2008          59       11
## 4 2009          71       18
## 5 2010          70       13
## 6 2011          65       12
```

Create models and get summary.

```r
mod_brook <- lm(N_MURDER~N_INC, inc_murd_f_brook)
summary(mod_brook)
```

```
##
## Call:
## lm(formula = N_MURDER ~ N_INC, data = inc_murd_f_brook)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3880 -0.7777 -0.1674  0.6378  4.5203
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.35409    2.14449  -0.165 0.871057
## N_INC        0.19484    0.03965   4.915 0.000187 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.296 on 15 degrees of freedom
## Multiple R-squared:  0.6169, Adjusted R-squared:  0.5913
## F-statistic: 24.15 on 1 and 15 DF,  p-value: 0.000187
```

```r
mod_st <- lm(N_MURDER~N_INC, inc_murd_f_st)
summary(mod_st)
```

```
##
## Call:
## lm(formula = N_MURDER ~ N_INC, data = inc_murd_f_st)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.27119 -0.38136 -0.09322  0.30720  1.55085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20339    0.37184   0.547    0.593
## N_INC        0.17797    0.07996   2.226    0.043 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6867 on 14 degrees of freedom
## Multiple R-squared:  0.2613, Adjusted R-squared:  0.2086
## F-statistic: 4.953 on 1 and 14 DF,  p-value: 0.04298
```

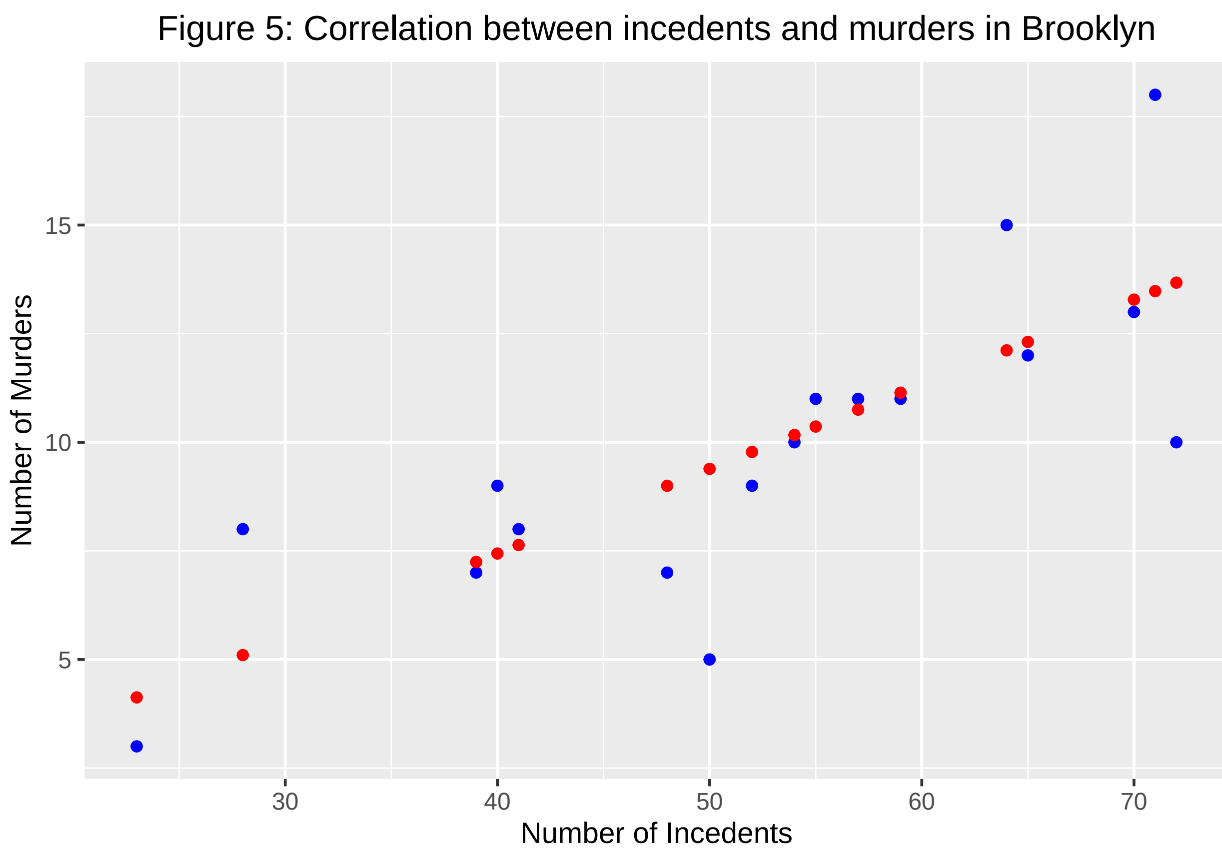
Lastly, I will be using my results to make predictions.

```
inc_murd_f_brook_pred <- inc_murd_f_brook %>%
  mutate(pred = predict(mod_brook))
```

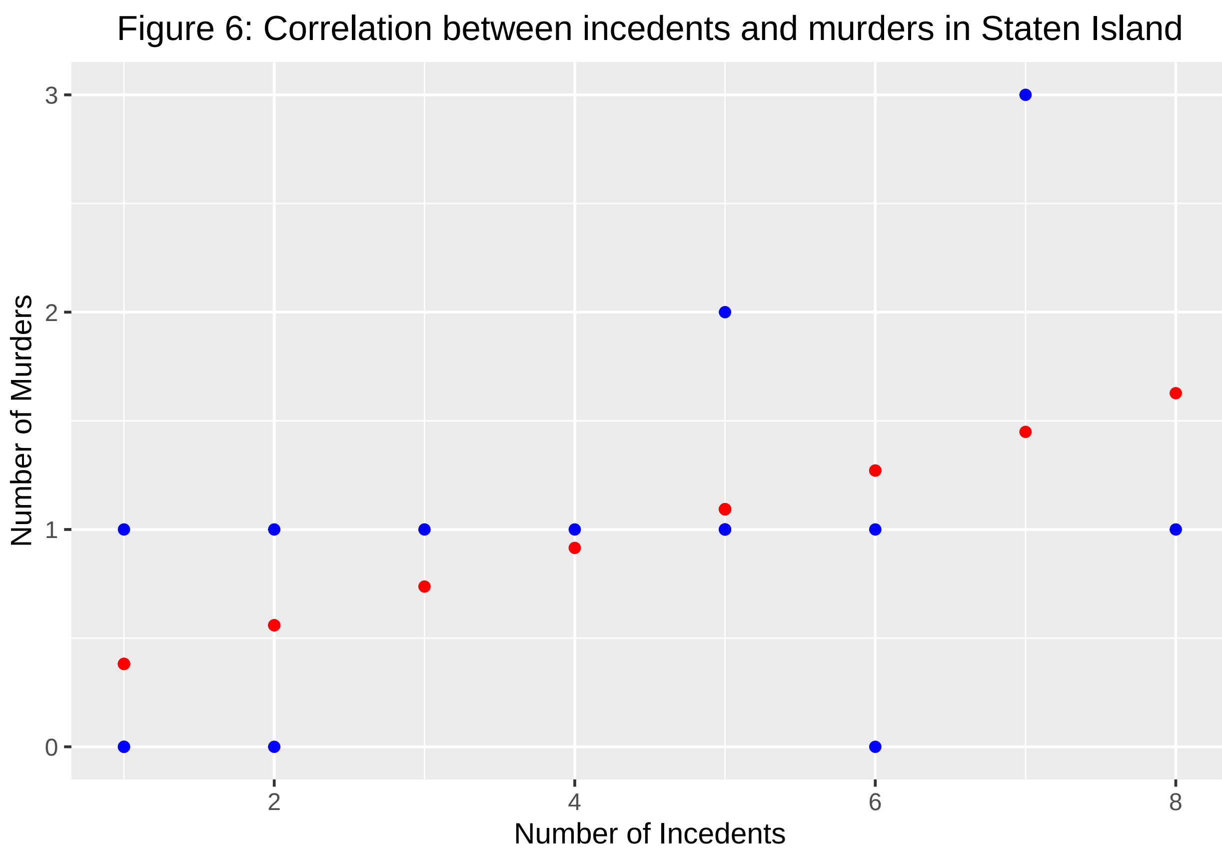
```
inc_murd_f_st_pred <- inc_murd_f_st %>%
  mutate(pred = predict(mod_st))
```

Visualize real data and perdiction.

```
ggplot(inc_murd_f_brook_pred) +
  labs(x="Number of Incedents", y="Number of Murders") +
  ggtitle("Figure 5: Correlation between incedents and murders in Brooklyn", ) +
  # center title
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_point(aes(x = N_INC, y = N_MURDER), color = "blue") +
  geom_point(aes(x = N_INC, y = pred), color = "red")
```




```r
ggplot(inc_murd_f_st_pred) +
  labs(x="Number of Incedents", y="Number of Murders") +
  ggtitle("Figure 6: Correlation between incedents and murders in Staten Island", ) +
  # center title
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_point(aes(x = N_INC, y = N_MURDER), color = "blue") +
  geom_point(aes(x = N_INC, y = pred), color = "red")
```



For Brooklyn model:

The model explains approximately 61.69% of the variance in the number of murders (N_MURDER) based on the number of incidents (N_INC) in Brooklyn.

The intercept term is not statistically significant ($p = 0.871057$), indicating that when the number of incidents is zero, the expected number of murders is not significantly different from zero.

The coefficient for number of incidents is zero is statistically significant ($p = 0.000187$), suggesting that for each additional incident in Brooklyn, the expected number of murders increases by approximately 0.19484.

For the Staten Island model:

The model explains approximately 26.13% of the variance in the number of murders based on the number of incidents in Staten Island.

The intercept term is not statistically significant ($p = 0.593$), indicating that when the number of incidents is zero, the expected number of murders is not significantly different from zero.

The coefficient for N_INC is statistically significant ($p = 0.043$), suggesting that for each additional incident in Staten Island, the expected number of murders increases by approximately 0.17797.

Overall, both models indicate a positive relationship between the number of incidents and the number of murders, but the model for Brooklyn explains a larger proportion of the variance and has a higher coefficient for number of incidents, indicating a stronger relationship compared to the model for Staten Island.

To enhance safety for young women in New York, we recommend allocating additional resources towards bolstering measures such as heightened police presence, community outreach programs, and crime prevention initiatives. These efforts aim to effectively address safety concerns and foster a more secure environment for young women across the city.

Bias

Variations in geographical factors such as population density, urban infrastructure, and neighborhood characteristics could introduce biases into the analysis. Differences in policing strategies or community resources between boroughs may also affect the observed correlations.

Conclusion

The analysis reveals significant spatial disparities in safety for young women across New York City boroughs. Brooklyn emerges as the most dangerous location, while Staten Island is deemed the safest.

The analysis indicates that the peak of violence typically occurs between 8 pm and 4 am. This suggests a need for heightened vigilance and increased police presence during these hours to ensure the safety of young women.

There is a notable correlation between the number of shooting incidents and the number of murders in specific boroughs, particularly in Brooklyn, where the correlation is strong. This underscores the importance of targeted interventions to address underlying factors contributing to violence in these areas.

To mitigate safety risks and create a safer environment for young women in New York City, it is recommended to allocate additional resources towards initiatives such as increased police presence, community outreach programs, and crime prevention strategies. These efforts should be tailored to address the unique spatial and temporal patterns of violence identified in the analysis.