NYPD Shooting Incident Data Analysis

Introduction

This data report aims to find out the trend and relationship of the number of victims and number of criminals for the past 21 years using the data set of NYPD Shooting Incident Data. The hypothesis is that the number of victims of the shooting incident in New York gradually decreased for the past 21 years, and the number of arrested criminals increased for the past 21 years.

Import Data

First Step: importing NYPD Shooting Incident Data from an online resource, “https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD”.

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

```
shooting_data <-readr:::read_csv(raw_data)
```

Clearing Data

Second step: I cleared data for analyzing data of shooting incidents in New York in the past 21 years. The data columns I need for this analysis are “OCCUR_DATE”,“OCCUR_TIME”,“BORO”,“PERP_AGE_GROUP”,“PERP_S “PERP_RACE”, “VIC_AGE_GROUP”, “VIC_SEX”, “VIC_RACE”.

- The data type of “OCCUR_DATE” is “character.” To better sort this data set, I changed its data type to “DATE.”
- I only want to study the shooting incident for the past 21 years, so I filtered data, which its “OCCUR_DATE” is between “2000-01-01” and “2020-12-30”.
- Also, to clear out the invalid data, I filtered the rows that contain the NULL VIC_AGE_GROUP, VIC_SEX, and VIC_RACE.

```
shooting_data<-shooting_data %>%
  select("OCCUR_DATE","OCCUR_TIME","BORO","PERP_AGE_GROUP","PERP_SEX", "PERP_RACE", "VIC_AGE_GROUP", "VI
```

```
shooting<-shooting_data %>%
  mutate(OCCUR_DATE=mdy(OCCUR_DATE))%>%
  filter(OCCUR_DATE >= ymd("2000-01-01")&OCCUR_DATE <= ymd("2020-12-30"))%>%
  filter(VIC_AGE_GROUP!="NA"& VIC_SEX!="NA"& VIC_RACE!="NA" )
```

Transforming Data

Third step: I transformed data in order to better study the shooting data. * To better count the number of victim, I first added the number of victim (1) as “countVit” for each record.

- To have a general view of shooting data, I created “shooting_per_month”, which summarizes the total shooting incident of each month for the past 21 years.
- To compare shooting data among different years, I created “shooting_per_year”, which summarizes the total shooting incident, total escaped criminal, and total criminal caught.
- “shooting_per_Year_per_AgeGroup” is created by grouping the month, year, and age group of victim.
- “shooting_compare” is created by combining the “shooting_NA_prep” table, which summarizes the total escaped criminal, and the “shooting_per_year”, which summarizes total criminal caught.

```
shooting_modify<-shooting%>%
  mutate(countVit=1)

shooting_per_month<-shooting_modify%>%
  mutate( year = format(OCCUR_DATE, "%Y"), month = format(OCCUR_DATE, "%m")) %>%
  group_by(year, month) %>%
  summarise(total_shooting=sum(countVit))
```

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

```
shooting_per_month<-shooting_per_month%>%
  unite("month_year", c(year,month), sep=",", na.rm=TRUE, remove=FALSE)

shooting_per_year<-shooting_modify%>%
  mutate( year = format(OCCUR_DATE, "%Y")) %>%
  group_by(year) %>%
  summarise(total_shooting=sum(countVit))

shooting_NA_prep<-shooting_modify%>%
  filter(PERP_AGE_GROUP!="NA")%>%
  mutate(year = format(OCCUR_DATE, "%Y")) %>%
  group_by(year) %>%
  summarise(total_escaped_criminal=sum(countVit))

shooting_per_Year_per_AgeGroup<-shooting_modify%>%
  mutate( year = format(OCCUR_DATE, "%Y"), month = format(OCCUR_DATE, "%m")) %>%
  group_by(year, month, VIC_AGE_GROUP) %>%
  summarise(total_shooting=sum(countVit))
```

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

```
shooting_compare<-shooting_NA_prep%>%
  left_join(shooting_per_year, by=c("year"))%>%
  mutate(shooting_criminal_caught=total_shooting-total_escaped_criminal)
```

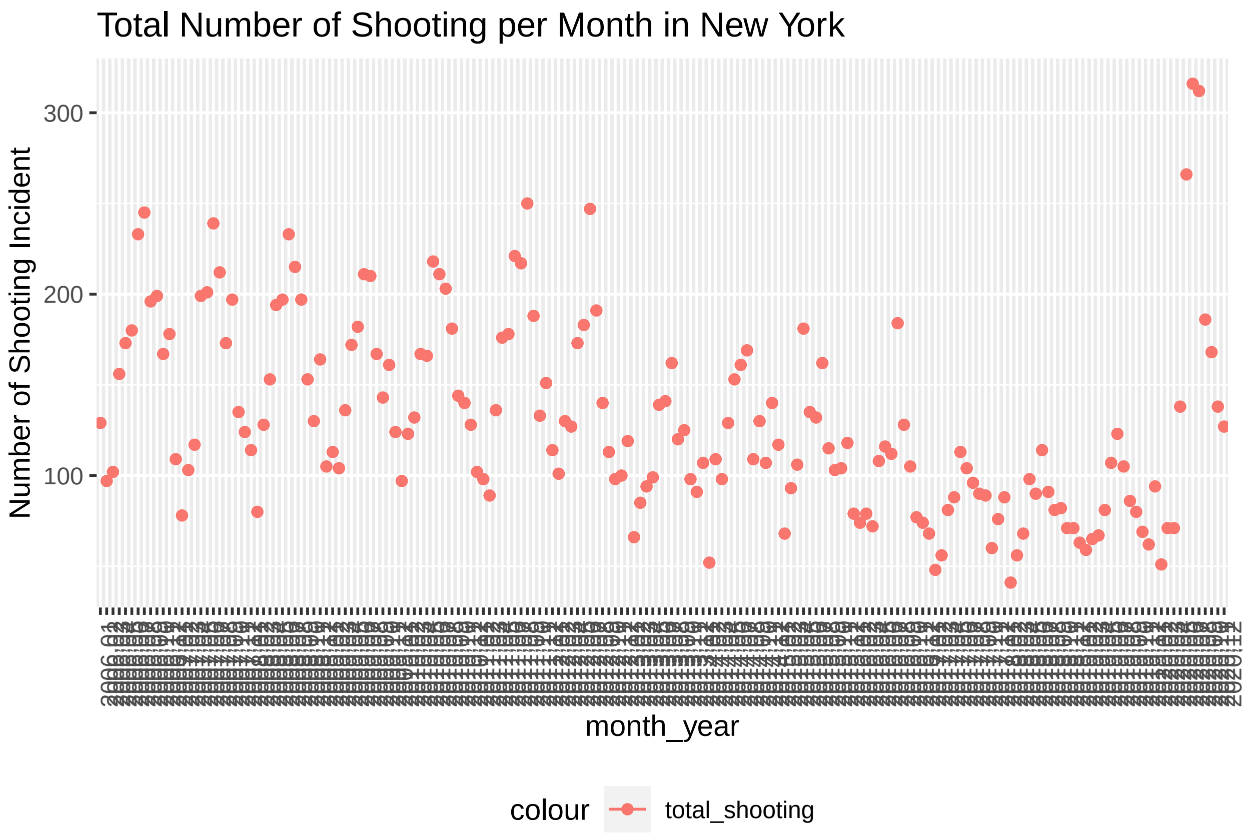
Plots

The first plot shows the total shooting incident per month from the year 2000 to 2021. * The first plot shows that the number of shooting incidents in New York decreases from January to December every year. The second plot shows the total escaped criminal, total arrested criminal, and total_shooting from the year 2000 to 2021. * The second plot shows that the general trends of the total number of shooting incidents and the total number of escaped criminals are decreasing from the year 2000 to 2019.

- The second plot also shows that the total number of arrested criminals increases from the year 2000 to 2011 but gradually decreases from the year 2012 to 2019. The decreasing trend of the number of arrested criminals may be caused by the decreased number of shooting incidents.
- The total number of arrested criminals, the total number of shooting incidents and the total number of escaped criminals of the year 2020 are outliers of the second plot.

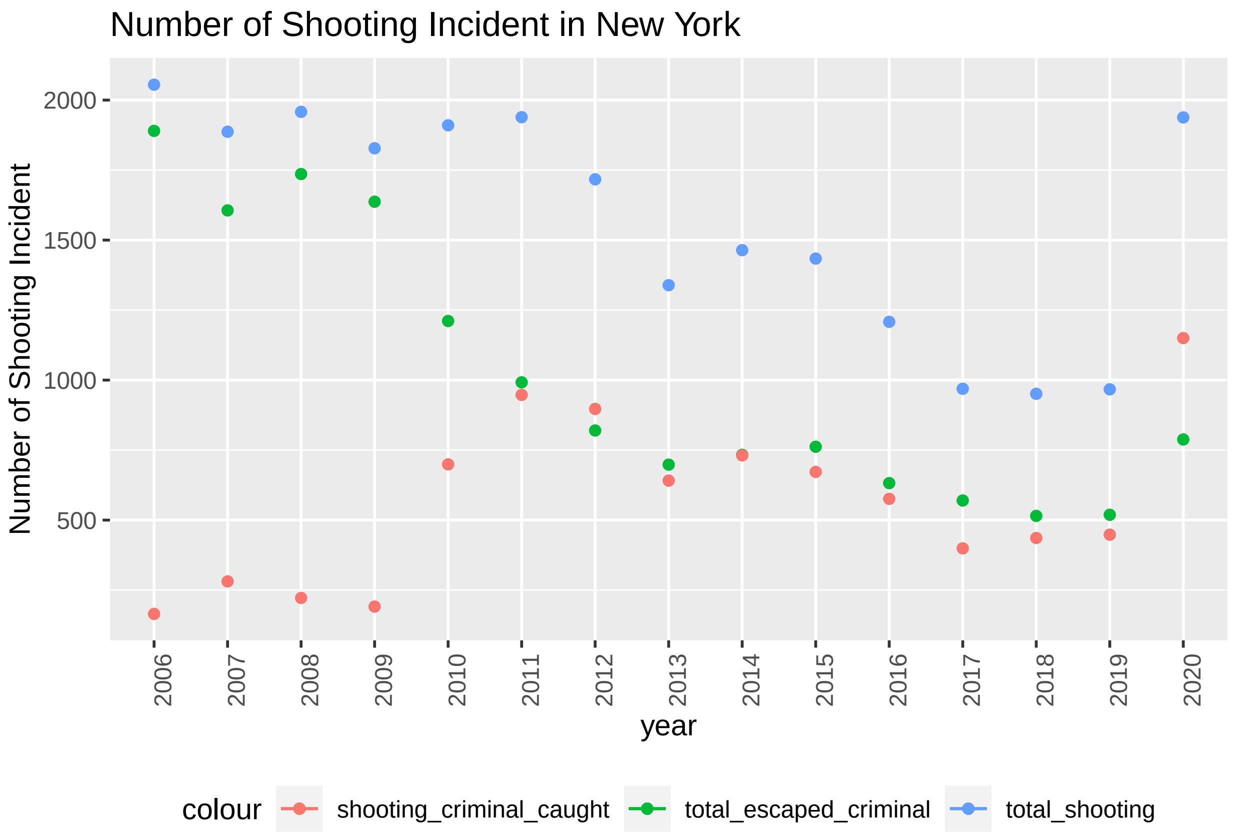
```
shooting_per_month%>%
  filter(total_shooting>0) %>%
  ggplot(aes(x=month_year, y=total_shooting)) +
  geom_line(aes(color="total_shooting"))+
  geom_point(aes(color="total_shooting"))+
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90))+
  labs(title = "Total Number of Shooting per Month in New York", y="Number of Shooting Incident")
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```



```
shooting_compare%>%
  filter(total_shooting>0) %>%
  ggplot(aes(x=year, y=total_shooting)) +
  geom_line(aes(color="total_shooting"))+
  geom_point(aes(color="total_shooting"))+
  geom_line(aes(y=total_escaped_criminal,color="total_escaped_criminal"))+
  geom_point(aes(y=total_escaped_criminal,color="total_escaped_criminal"))+
  geom_line(aes(y=shooting_criminal_caught,color="shooting_criminal_caught"))+
  geom_point(aes(y=shooting_criminal_caught,color="shooting_criminal_caught"))+
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90))+
  labs(title = "Number of Shooting Incident in New York", y="Number of Shooting Incident")
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```



Analyzing Data

The groups of people who are most often being shot are male African Americans, age 18-44.

```
number_vic<-shooting%>%
  count(VIC_AGE_GROUP,VIC_SEX, VIC_RACE, sort = TRUE)
number_vic
```

```
## # A tibble: 78 x 4
##    VIC_AGE_GROUP VIC_SEX VIC_RACE           n
##    <chr>         <chr>   <chr>          <int>
##  1 25-44         M       BLACK           6813
##  2 18-24         M       BLACK           6075
##  3 <18           M       BLACK           1586
##  4 25-44         M       WHITE HISPANIC  1348
##  5 18-24         M       WHITE HISPANIC  1221
##  6 25-44         M       BLACK HISPANIC   876
##  7 18-24         M       BLACK HISPANIC   800
##  8 45-64         M       BLACK            783
##  9 25-44         F       BLACK            547
## 10 18-24         F       BLACK            485
## # ... with 68 more rows
```

The maximum number of criminal's age_group is 8421. The criminal group of males, African Americans, age 18-44 created about 50% of the shooting incidents for the past 21 years.

```
number_pert<-shooting%>%
  count(PERP_AGE_GROUP,PERP_SEX,PERP_RACE, sort = TRUE)

number_pert
```

```
## # A tibble: 74 x 4
##    PERP_AGE_GROUP PERP_SEX PERP_RACE          n
##    <chr>          <chr>    <chr>          <int>
##  1 <NA>           <NA>     <NA>            8421
##  2 18-24          M        BLACK           3844
##  3 25-44          M        BLACK           3249
##  4 UNKNOWN        U        UNKNOWN         1436
##  5 UNKNOWN        M        BLACK           1233
##  6 <18            M        BLACK            975
##  7 18-24          M        WHITE HISPANIC   831
##  8 25-44          M        WHITE HISPANIC   654
##  9 18-24          M        BLACK HISPANIC   490
## 10 25-44          M        BLACK HISPANIC   323
## # ... with 64 more rows
```

The area, where shooting incidents most often happened in New York, is Brooklyn. The number of shooting incident happened in Brooklyn occupies 41% of the overall shooting incident for the past 21 years.

```
number_shooting_per_area<-shooting%>%
  count(BORO, sort = TRUE)
number_shooting_per_area
```

```
## # A tibble: 5 x 2
##   BORO              n
##   <chr>         <int>
## 1 BROOKLYN       9722
## 2 BRONX          6699
## 3 QUEENS         3525
## 4 MANHATTAN      2920
## 5 STATEN ISLAND   698
```

In Brooklyn district, New York, there are 3968 escape criminals, and the top characteristics of the main criminal group are male, age 15-44, African American.

```
number_shooting_in_brooklyn<-shooting%>%
  filter(BORO=="BROOKLYN")%>%
  count(BORO,PERP_AGE_GROUP,PERP_SEX,PERP_RACE, sort = TRUE)
number_shooting_in_brooklyn
```

```
## # A tibble: 58 x 5
##    BORO     PERP_AGE_GROUP PERP_SEX PERP_RACE          n
##    <chr>    <chr>          <chr>    <chr>          <int>
##  1 BROOKLYN <NA>           <NA>     <NA>            3968
##  2 BROOKLYN 18-24          M        BLACK           1668
##  3 BROOKLYN 25-44          M        BLACK           1359
##  4 BROOKLYN UNKNOWN        U        UNKNOWN          628
##  5 BROOKLYN UNKNOWN        M        BLACK            599
##  6 BROOKLYN <18            M        BLACK            437
##  7 BROOKLYN 18-24          M        WHITE HISPANIC   135
##  8 BROOKLYN 25-44          M        WHITE HISPANIC   124
##  9 BROOKLYN 45-64          M        BLACK            108
## 10 BROOKLYN 18-24          M        BLACK HISPANIC   100
## # ... with 48 more rows
```

Modeling Data

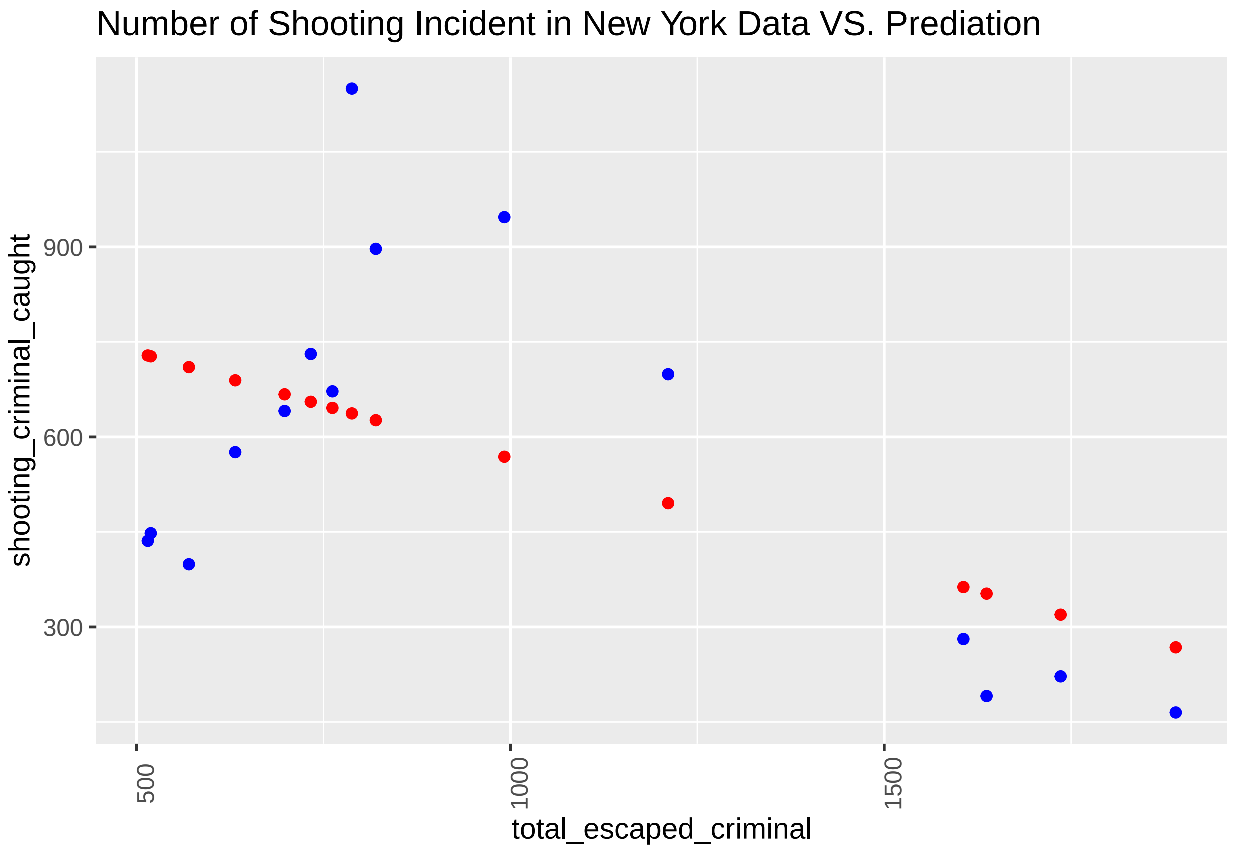
I used a linear model to fit the total shooting criminal and the total escaped criminal. In summary of this linear model, we know that shooting_criminal_caught=-0.3351*total_escaped_criminal+901.22.

```
mod<-lm(shooting_criminal_caught~total_escaped_criminal, data=shooting_compare)
summary(mod)
```

```
## 
## Call:
## lm(formula = shooting_criminal_caught ~ total_escaped_criminal, 
##     data = shooting_compare)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -311.20 -137.53  -82.02  139.51  512.85 
## 
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            901.2218   158.0859   5.701 7.28e-05 ***
## total_escaped_criminal  -0.3351     0.1425  -2.351   0.0351 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 256.3 on 13 degrees of freedom
## Multiple R-squared:  0.2984, Adjusted R-squared:  0.2444
## F-statistic: 5.528 on 1 and 13 DF,  p-value: 0.03515
```

```
shooting_compare_pred<-shooting_compare%>% mutate(pred=predict(mod))
```

```
shooting_compare_pred%>%
  filter(total_shooting>0) %>%
  ggplot() +
  geom_point(aes(x=total_escaped_criminal, y=shooting_criminal_caught), color="blue")+
  geom_point(aes(x=total_escaped_criminal, y=pred), color="red")+
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90))+
  labs(title = "Number of Shooting Incident in New York Data VS. Prediation")
```



Conclusion

According to the plot, this model is not good to predict shooting_criminal_caught when the total_escaped_criminal is smaller than 1500. This model shows that there are other factors that influence the shooting_criminal_caught and total_escaped_criminal. Even though previous plots show that the number of victims of the shooting incident in New York gradually decreased for the past 21 years, and the number of arrested criminals increased for the past 21 years, there are not enough evidences to prove the relationship between the relationship of the number of victims and number of criminals for the past 21 years.

Bias Interpration

The possible bias sources in this report are:

- This set of data may include data produced by humans which may contain bias against groups of people.
- Because this shooting incident data does not include variables that properly capture the phenomenon I want to predict, it may results in selection bias while doing data clearing and transforming.
- Last bias may be omitted variable bias because while clearing data, I only select few columns, it may cause the critical attributes that influence the outcome to be missing during analysis.