Analysis on NYPD Shoooting Incident Historical Data

2024-01-24

Load Libraries

```
library(readr)
library(tidyverse,)
library(janitor)
library(ggrepel)
library(float)
```

Import Data

The NYPD Shooting Incident Data lists every shooting incident that occurred in New York City (NYC) during the calendar year. Metadata for this dataset is available on NYC Open Data

The exploratory analysis will focus on showing the relationship between time and number of arrests related to shooting incidents. I will use logistic regression to model the relationship between sex and counts of shooting incidents.

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
NYPD_historical_data <- read.csv(url)

# transform date columns for visualization
NYPD_data <-
NYPD_historical_data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME),
         YEAR = year(OCCUR_DATE),
         MONTH = month(OCCUR_DATE),
         MONTH_label = month(OCCUR_DATE, label = TRUE),
         WEEK = week(OCCUR_DATE),
         WEEKDAY = weekdays(OCCUR_DATE),
         HOUR = hour(OCCUR_TIME))
```

Exploratory Analysis

I am interested in visualizing police enforcement activity at different time levels - year, month, week, day.

Enforcement Activity per Year (2006 - 2022)

```
NYPD_data %>%
  group_by(YEAR) %>%
  summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY)) %>%
  ggplot(aes(YEAR, INCIDENT_COUNTS)) +
  geom_line(color = "grey") +
  geom_point(size = 4, fill = "#69b3a2", shape = 21) +
  geom_hline(aes(yintercept = mean(INCIDENT_COUNTS)), color = "grey", lty = "dashed")+
  scale_x_continuous(breaks = seq(2006, 2022, 2)) +
  theme(
    axis.text.x = element_text(size = 16, color = 'black'),
    axis.text.y = element_text(size = 16, color = 'black')) +
  theme_bw()+
  labs(title = "New York City Shooting Incidents per Year",
       x = "Year",
       y = "Count of shooting incidents")
```

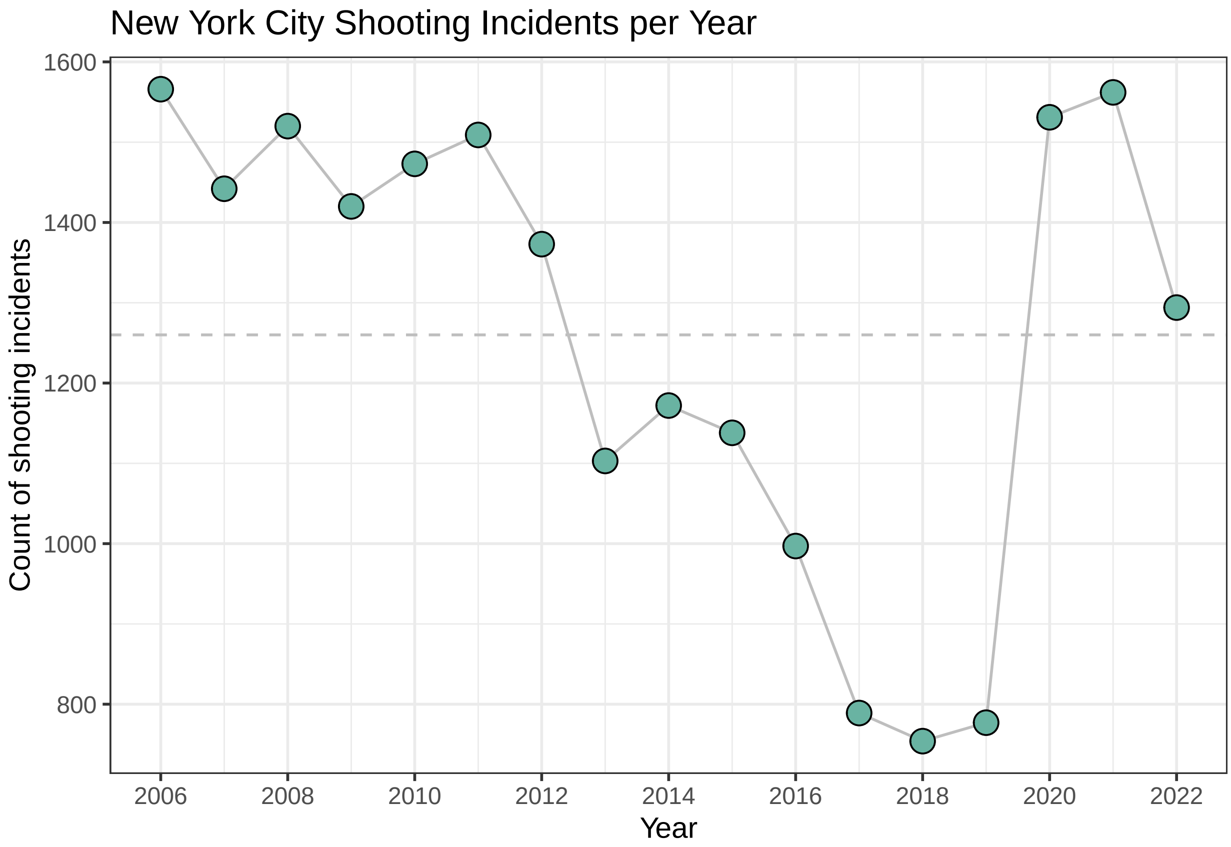


Figure 1: New York City Shooting Incidents (2006 - 2022)

Figure 1. Number of arrests related to shooting incidents that occurred in NYC between 2006 - 2022. NYC shooting incidents were at a steady decline between 2006 and 2019. The year 2018 had the lowest number of reported incidents. There is a sudden increase in shooting incidents in 2020, they rise up to the 2006 levels. This is likely related to the shutdown of American business in the wake of the COVID-19 pandemic.

```r
Labels <- NYPD_data %>%
          group_by(YEAR, BORO) %>%
          summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY)) %>%
          filter(YEAR ==2022)

NYPD_data %>%
  group_by(YEAR, BORO) %>%
  summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY)) %>%
  ggplot(aes(YEAR, INCIDENT_COUNTS, group = BORO, color = BORO)) +
  geom_line(size = 1.5) +
  geom_point(size = 1) +
  geom_hline(aes(yintercept = mean(INCIDENT_COUNTS)), color = "grey", lty = "dashed")+
  geom_text_repel(data = Labels,  aes(x = 2022, y = INCIDENT_COUNTS,label = BORO),
                  show.legend = FALSE, hjust = -0.5, vjust = -0.5, inherit.aes = FALSE) +
  scale_x_continuous(breaks = scales::pretty_breaks(10)) +
  scale_y_continuous(sec.axis = sec_axis(~ ., breaks = Labels$INCIDENT_COUNTS)) +
  expand_limits(x = 2023)+
  scale_color_viridis_d() +
  theme_bw()+
  theme(
    axis.text.x = element_text(size = 12, color = 'black'),
    axis.text.y = element_text(size = 12, color = 'black'),
    title = element_text(size = 18),
    legend.position = "bottom",
    legend.title = element_blank(),
    legend.text = element_text(size = 12)) +
  labs(title = "New York City Shooting Incidents per Year per Borough",
      x = "Year",
      y = "Count of shooting incidents")
```

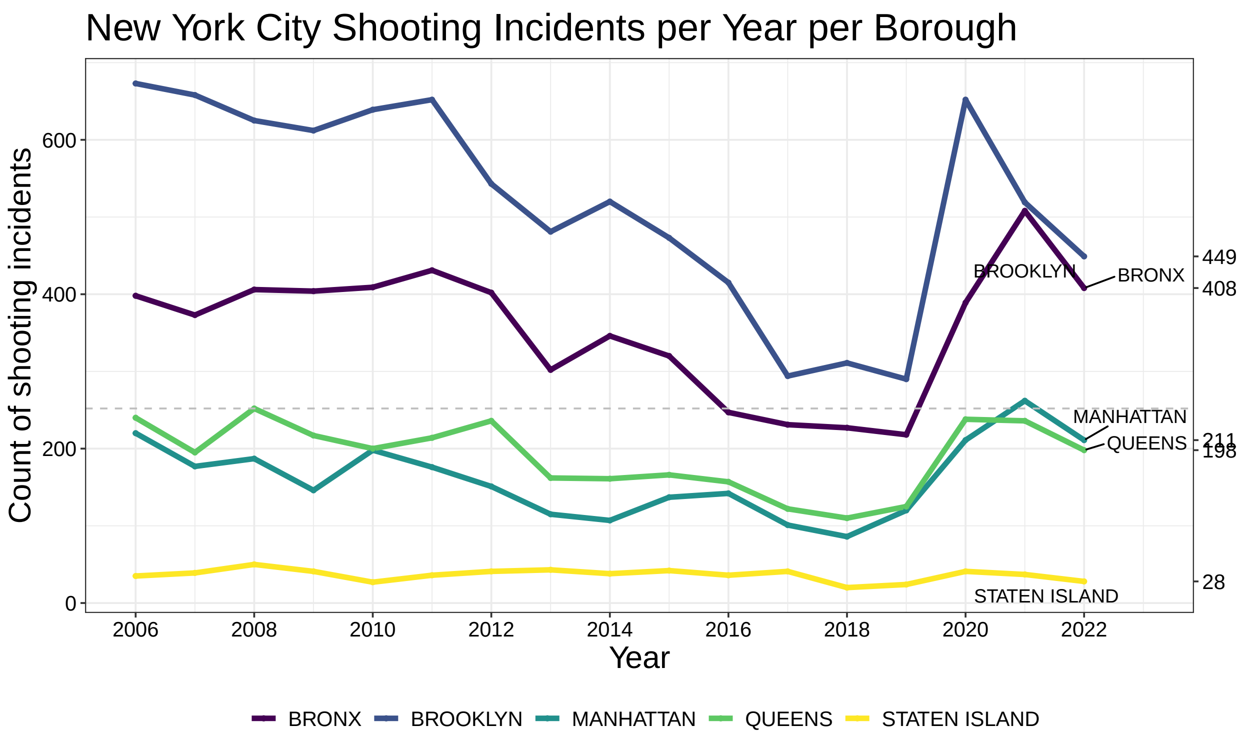


Figure 2: New York City Shooting Incidents per Borough

Figure 2. Number of arrests for shooting incidents that occurred within NYC's five boroughs between 2006 - 2022. Shooting incidents have historically been highest in Brooklyn. However, after 2020, the number of arrests in the Bronx are near Brooklyn levels. Staten Island has lowest level of shooting incidents, it is also un-usual in that it the only borough missing a noticeable spike in number of arrests coinciding the COVID-19 pandemic.

Enforcement Activity by month for all reported years

```
max_year <-
  NYPD_data %>%
  group_by(YEAR, MONTH_label) %>%
  summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY)) %>%
  filter(INCIDENT_COUNTS == max(INCIDENT_COUNTS))


NYPD_data %>%
  group_by(YEAR, MONTH, MONTH_label) %>%
  summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY)) %>%
  ggplot(aes(MONTH_label, INCIDENT_COUNTS, group = YEAR)) +
  geom_line(color = "grey") +
  geom_point(data = max_year, aes(x = MONTH_label, y = INCIDENT_COUNTS)) +
  geom_text(data = max_year,
            aes(x = MONTH_label, y = INCIDENT_COUNTS,
                label = MONTH_label), vjust = 1.5, size = 10) +
  facet_wrap(~YEAR) +
  theme_bw()+
  theme(
    axis.text.x = element_text(size = 10, color = 'black', angle = 90),
    axis.text.y = element_text(size = 10, color = 'black'),
    axis.title.y = element_text(size = 18, color = 'black'),
    title  = element_text(size = 24)
      ) +
  labs(title = "Monthly police enforcement activity (2006 - 2022)",
      x = NULL,
      y = "Count of shooting incidents")
```

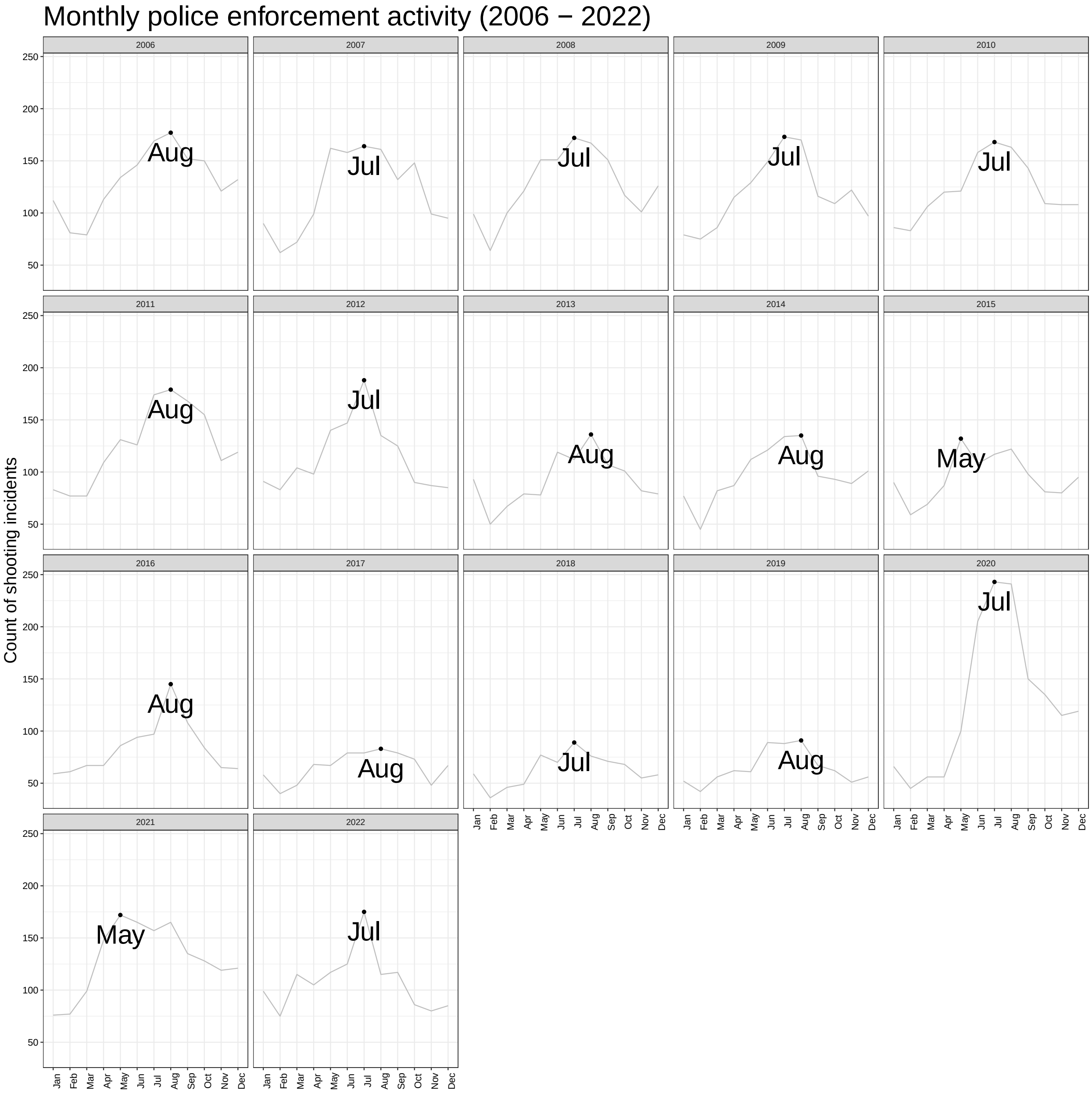


Figure 3: New York City Shooting Incidents per Month (2006 - 2022)

Figure 3. The figure below displays number of arrests related to shooting incidents for each year between 2006 and 2022. It shows that within a given calendar year, there is likely to be a peak number of police enforcement activity in the summer months of July or August. Is there a correlation between warm weather and criminal activity?

Enforcement Activity by the Week

```
# Define the custom order for weekdays
custom_order <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")


NYPD_data %>%
  group_by(WEEKDAY) %>%
  summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY)) %>%
  ggplot(aes( x = factor(WEEKDAY, levels = custom_order),
              y = INCIDENT_COUNTS,
              group = 1)) +
  geom_line() +
  geom_point(size = 4, fill = "#69b3a2", shape = 21) +
  theme(
    axis.text.x = element_text(size = 16, color = 'black'),
    axis.text.y = element_text(size = 16, color = 'black'),
    title  = element_text(size = 18)) +
  theme_bw()+
  labs(title = "New York City Shooting Incidents per week",
      x = NULL,
      y = "Count of shooting incidents")
```

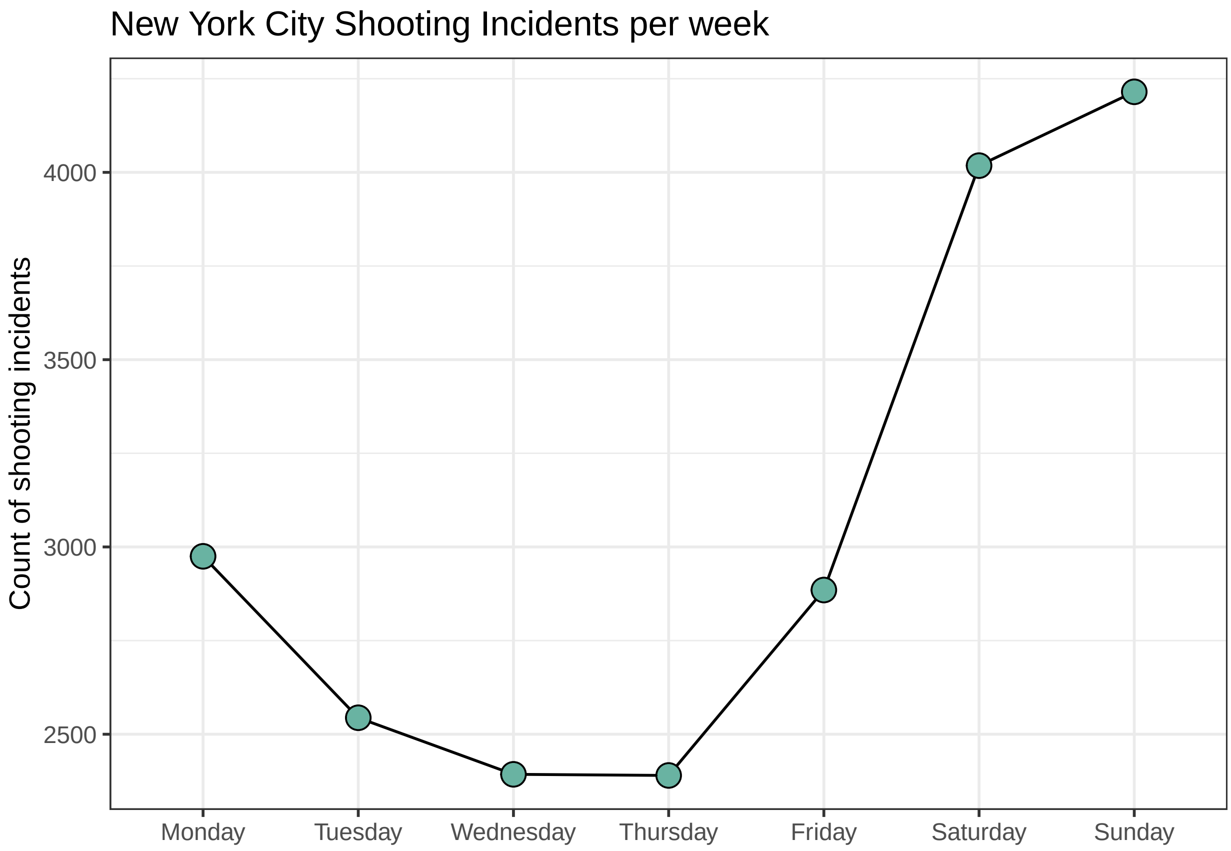


Figure 4: New York City Shooting Incidents per Week

Figure 4. The figure below displays the total number of shooting incidents that were reported for each day

of the week. In the city of New York, enforcement activity of NYC police officers is highest on Saturdays and Sundays.

Enforcement Activity by the Hour

```r
By_hour <-
  NYPD_data %>%
  group_by(HOUR) %>%
  summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY)) %>%
  mutate(time = sprintf("%02d:00", HOUR))

By_hour %>%
  ggplot(aes(x = time,
             y = INCIDENT_COUNTS,
           group = 1)) +
  geom_line() +
  geom_point(size = 4, fill = "#69b3a2", shape = 21) +
  theme_bw()+
  theme(
    axis.text.x = element_text(size = 10, color = 'black'),
    axis.text.y = element_text(size = 11, color = 'black'),
    axis.title.x = element_text(size = 14, color = "black"),
    axis.title.y = element_text(size = 14, color = "black"),
    title  = element_text(size = 18)) +
  labs(title = "New York City Shooting Incidents by the Hour",
      x = "Time",
      y = "Number of Shooting Incidents")
```

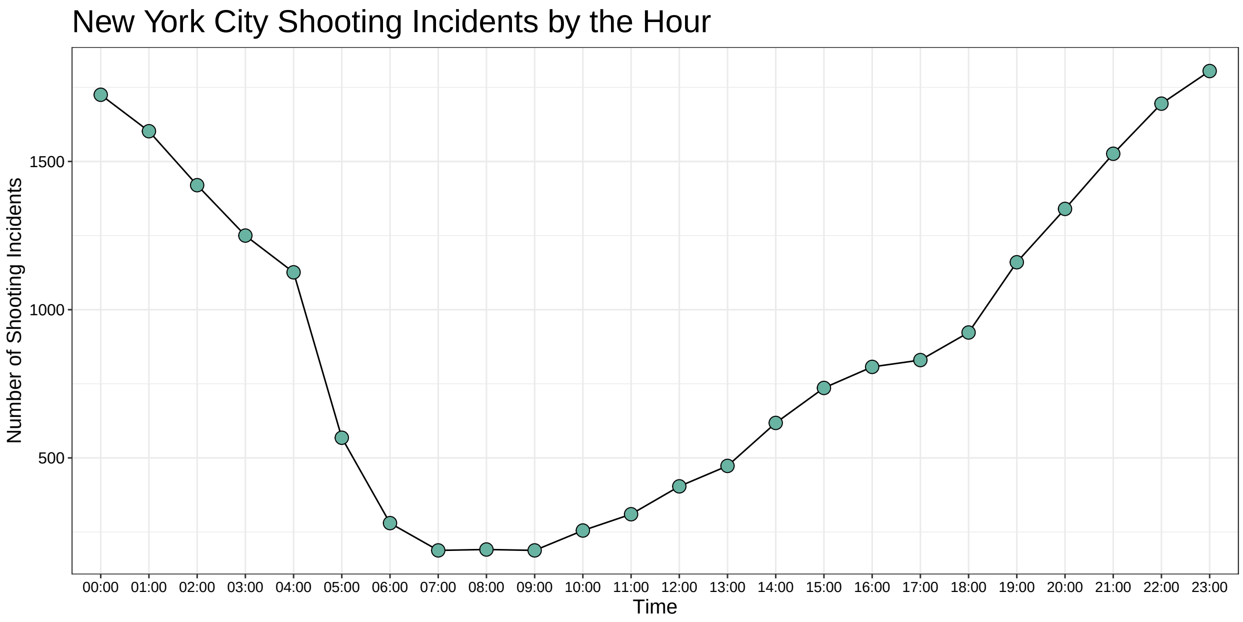


Figure 5: New York City Shooting Incidents by the hour

```
NYPD_data %>%
  group_by(WEEKDAY, HOUR) %>%
  summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY)) %>%
  ggplot(aes(x =  sprintf("%02d:00", HOUR),
             y = INCIDENT_COUNTS,
             group = 1)) +
  geom_line() +
  facet_wrap(~factor(WEEKDAY, levels = custom_order)) +
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90),
        title = element_text(size = 18)) +
  labs(title = "New York City Shooting Incidents by Hour per Day of the Week",
      x = "Year",
      y = "Count of shooting incidents")
```

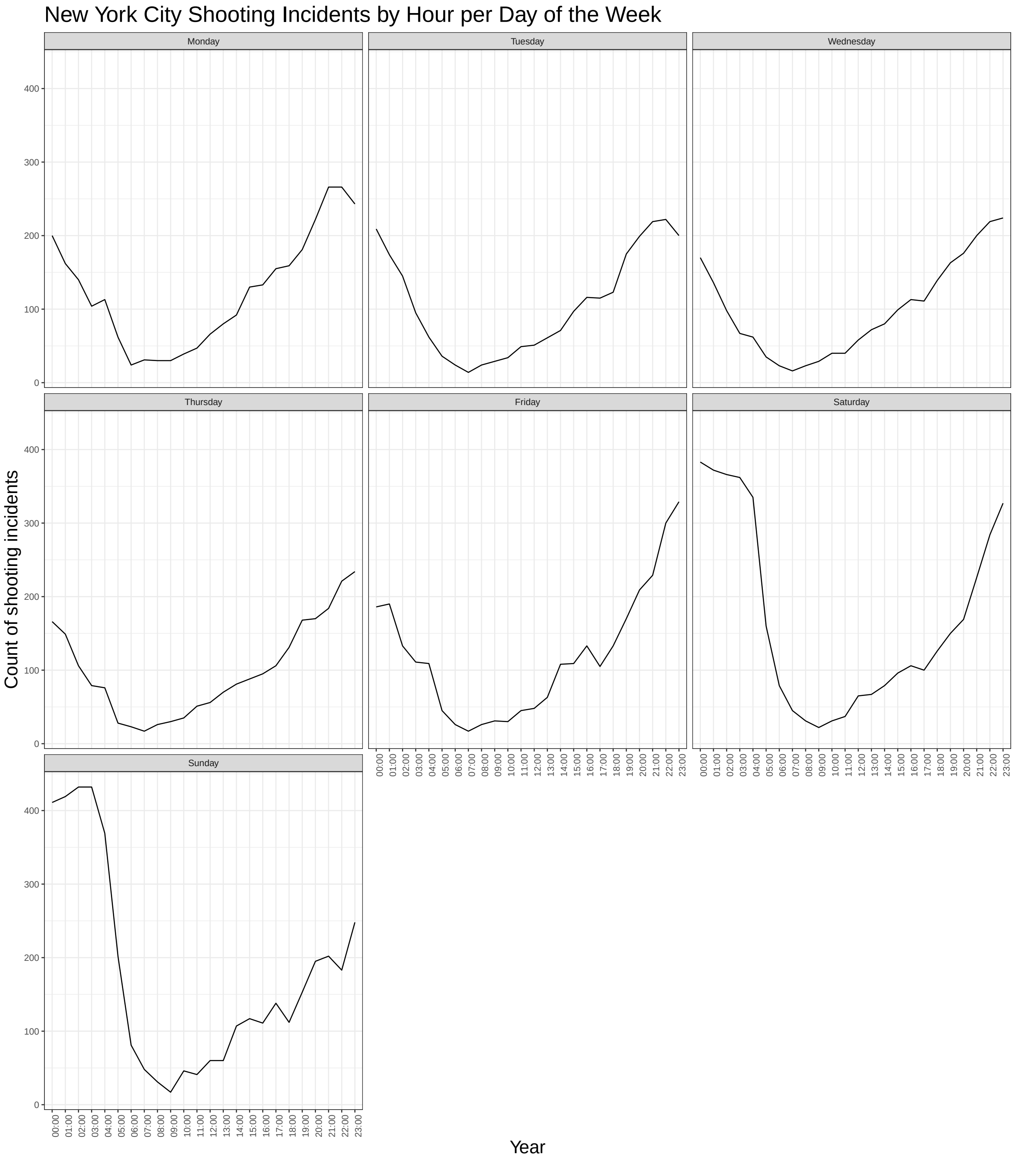


Figure 6: New York City Shooting Incidents by the hour for each weekday

Figure 5 and 6. In the 16 years that NYPD has recorded shooting incidents, irrespective of the day of the week, police enforcement activity related shooting incidents is high between 6PM and 4AM, and lowest between 6AM and 1PM.

LOGISTIC REGRESSION

Logistic regression allows us to estimate the probability of a categorical response based on one or more predictor variables (X). It allows one to say that the presence of a predictor increases (or decreases) the probability of a given outcome by a specific percentage. Poisson regression is used to model count variables, this is useful to model incident counts against any of the other categorical variables in the dataset.

```
mod_data <-
  NYPD_data %>%
   filter(PERP_SEX %in% c("M", "F"),
         PERP_RACE != "UNKNOWN",
         PERP_AGE_GROUP %in% c("<18", "18-24", "25-44", "45-64", "65+"),
         VIC_SEX %in% c("M", "F"),
         VIC_RACE != "UNKNOWN",
         VIC_AGE_GROUP %in% c("<18", "18-24", "25-44", "45-64", "65+")) %>%
  mutate(PERP_SEX = case_when(PERP_SEX == "F" ~ 1,
                              PERP_SEX == "M"~ 0)) %>%
  mutate(VIC_SEX = case_when(VIC_SEX == "F" ~ 1,
                              VIC_SEX == "M"~ 0)) %>%
  mutate(STATISTICAL_MURDER_FLAG  = case_when(STATISTICAL_MURDER_FLAG == "false" ~ 0,
                              STATISTICAL_MURDER_FLAG == "true" ~ 1 ))

perp_mod_data <-
  mod_data  %>%
  group_by(PERP_AGE_GROUP, PERP_SEX, PERP_RACE) %>%
  summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY))
```

Fitting the Model

The model will examine the effect of on unit increase in sex on the odds of a shooting incident. The expected change in log odds for one unit increase in perp sex is -3.2.

```
perp_model <-
  glm(INCIDENT_COUNTS ~ PERP_SEX,
      data = perp_mod_data ,
      family = poisson)


summary(perp_model)
```

```
##
## Call:
## glm(formula = INCIDENT_COUNTS ~ PERP_SEX, family = poisson, data = perp_mod_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.087411   0.009531  638.66   <2e-16 ***
## PERP_SEX    -3.205407   0.053777  -59.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 38476  on 44  degrees of freedom
## Residual deviance: 28131  on 43  degrees of freedom
## AIC: 28361
##
## Number of Fisher Scoring iterations: 6
```

Given that the PERP_SEX variable has values 0 (Male) and 1 (Female), and the coefficient for PERP_SEX is -3.2, we can interpret the odds ratio as follows: Calculate the Odds ratio by obtaining the exponential of the PERP_SEX coefficient (-3.205). The result is 0.0405. Since the odds ratio is less than one, this indicates a decrease in the odds of shooting incidence for the female group compared to the reference category (males)

```
exp_coef_perp_sex <- exp(coef(perp_model)["PERP_SEX"])
exp_coef_perp_sex
```

```
##   PERP_SEX
## 0.04054238
```

Sources of Potential Bias

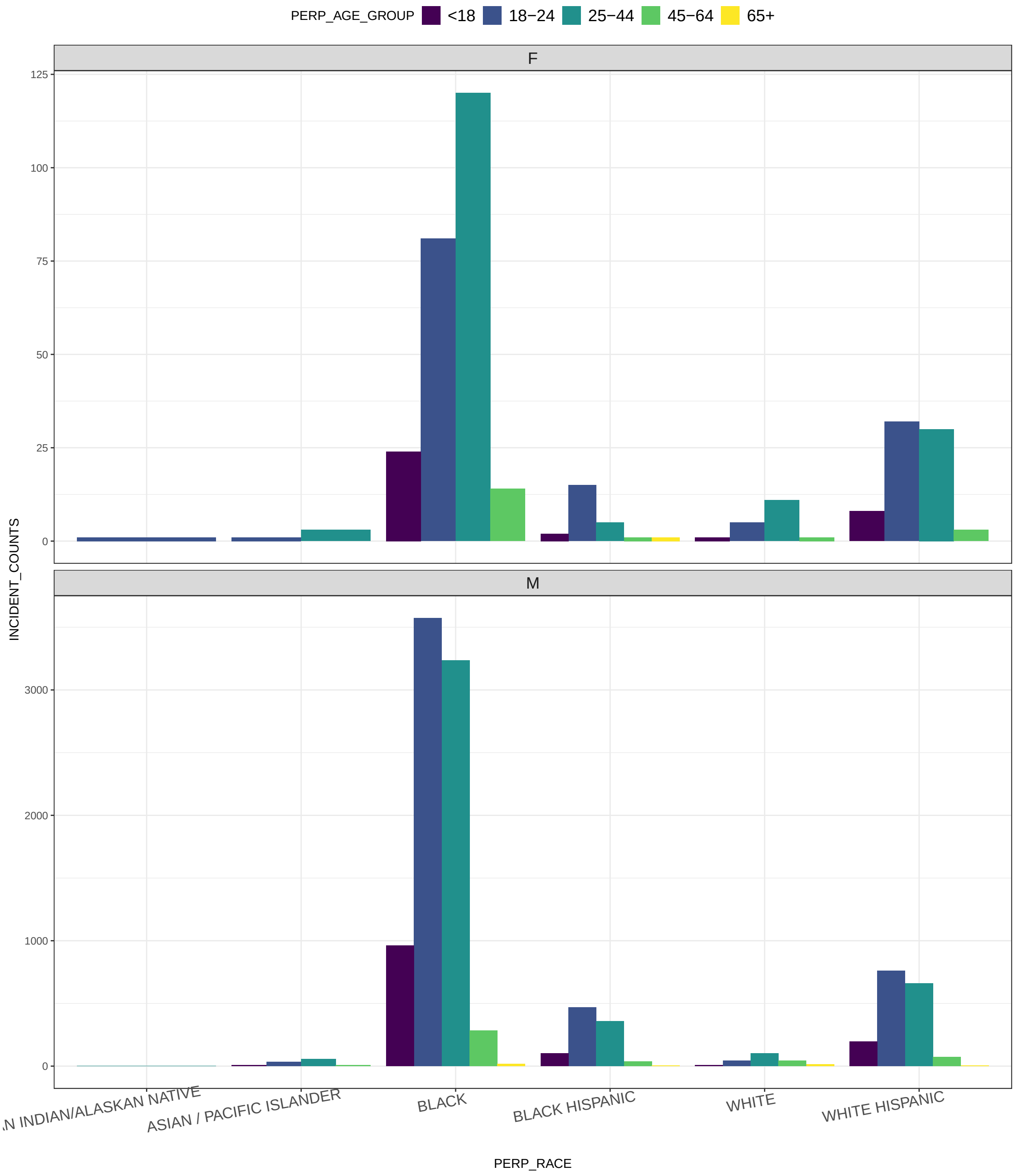
1. One observation I discovered that stands out the most as a source of bias is the demographics data. There is a disproportionate representation of one demographic group - individuals identified in the race columns as "BLACK", and sex columns as "Male". The table below illustrates this point by showing demographic information of the "perp". There may be bias in the data collection process that overrepresents these groups.

2. It is also unknown to us whether there is a sampling bias in the data. Since this data is reported to the public, it is likely that shooting incidents are under reported to demonstrate a trend in decrease in crimes within the city. It is likely that some crimes have gone undocumented - that is the police did not arrest a perp for a shooting incident.

3. The dataset it missing data for variables that provide information on the location where the incident took place. This make's it very difficult to understand trends in the location since the dataset is lacking representation for most arrests.

```
perp <-
  NYPD_data %>%
  group_by(PERP_AGE_GROUP , PERP_SEX, PERP_RACE) %>%
  summarise(INCIDENT_COUNTS = n_distinct(INCIDENT_KEY))


perp %>%
  filter(PERP_SEX %in% c("M", "F"),
        PERP_AGE_GROUP %in% c("<18", "18-24", "25-44", "45-64", "65+"),
         PERP_RACE != "UNKNOWN")%>%
  ggplot(aes(x = PERP_RACE, y = INCIDENT_COUNTS, fill = PERP_AGE_GROUP))+
  geom_bar(position = "dodge", stat = "identity")+
  scale_fill_viridis_d() +
  facet_wrap(~PERP_SEX, ncol = 1, scales = "free_y")+
   theme_bw() +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 10, hjust = 0.8, vjust = 1, size = 13),
        legend.text = element_text(size = 14),
        strip.text = element_text(size = 14))
```


PERP_AGE_GROUP
<18
F
M
INCIDENT_COUNTS
N INDIAN/ALASKAN NATIVE
ASIAN / PACIFIC ISLANDER
BLACK
BLACK HISPANIC
WHITE
WHITE HISPANIC
PERP_RACE