NYPD_DataAnalysis

pzuloaga

2024-03-01

NYPD Shooting Incident Data Analysis

Objetives

The objective of this study is to analyze two specific issues on the statistics of the data shooting New York City, data provided by the New York Police Department. The first issue is to know at what time and in which period of the year do the most shooting incidents occurs, so this can further provide hints on the main drivers and can help to propose measures and policies to reduce the gun violence. The second issue will be to analyze the demographics of the incident, specially the ethnicity of the suspects and victims involved.

Importing and reading

First, we will begin by importing and reading the NPYD data from the provided URL

```
##Get NYPD data in the url
url<-"https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
data<-read.csv(url)
```

Then, we will select the columns we are interested in (Date and time of incident occurrence and demographics of perpetrator and victim). We will later calculate two additional variables (hour and month of the occurrence).

Tidying and transforming

```
# Select and tidy data
data <- data %>%
  select('OCCUR_DATE','OCCUR_TIME','PERP_RACE','VIC_RACE') %>%
  mutate (
    OCCUR_DATE = mdy(OCCUR_DATE),
    # Convert month numbers to factor to maintain order and use as categorical data
    OCCUR_MONTH = factor(month(OCCUR_DATE), levels = 1:12, labels = month.abb),
    OCCUR_TIME = parse_hms(OCCUR_TIME),
    OCCUR_HOUR = hour(OCCUR_TIME)
)
head (data)
```

```
##   OCCUR_DATE OCCUR_TIME PERP_RACE       VIC_RACE OCCUR_MONTH OCCUR_HOUR
## 1 2021-05-27   21:30:00                   BLACK         May         21
```

```
## 2 2014-06-27   17:40:00                    BLACK         Jun          17
## 3 2015-11-21   03:56:00                    WHITE         Nov           3
## 4 2015-10-09   18:30:00           WHITE HISPANIC         Oct          18
## 5 2009-02-19   22:58:00     BLACK          BLACK         Feb          22
## 6 2020-10-21   21:36:00                    BLACK         Oct          21
```

```
#summary (data)
```

As it can be seen, there is some missing data regarding the ethnicity of the perpetrator. So, we will combine the blank entries with nulls, other and unknown.

We will also change the variable to factor and date types.

```
#replacing values
data$PERP_RACE[data$PERP_RACE==""] = "UNKNOWN"
data$PERP_RACE[data$PERP_RACE=="(null)"] = "UNKNOWN"
#setting factor for further plots
data$PERP_RACE<- factor(data$PERP_RACE, levels = names(sort(table(data$PERP_RACE), decreasing = TRUE)))
data$VIC_RACE <- factor(data$VIC_RACE, levels = names(sort(table(data$VIC_RACE), decreasing = TRUE)))
```

Here there is a summary of the final data.

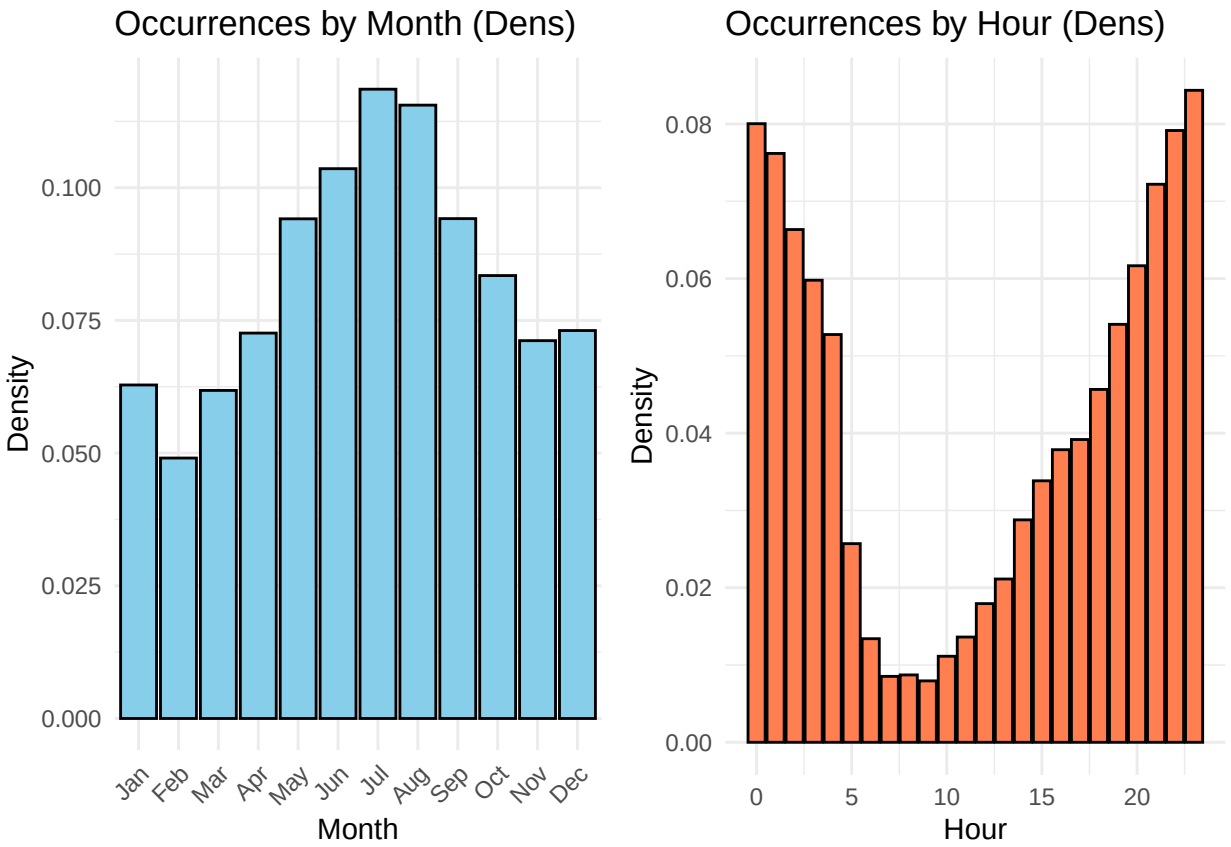
Visualization and Analysis

```
p1<-data %>%
  # Create the bar plot for incidents by month
  ggplot(aes(x = OCCUR_MONTH)) +
    geom_bar(aes(y = (..count..)/sum(..count..)),fill = "skyblue", color = "black")+
    xlab("Month") +
    ylab("Density") +
    ggtitle("Occurrences by Month (Dens)") +  # Title of the plot
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

p2<-data %>%
  # Create the bar plot for incidents by hour
  ggplot(aes(x = OCCUR_HOUR)) +
    geom_bar(aes(y = (after_stat(count))/sum(after_stat(count))),fill = "coral", color = "black")+
    xlab("Hour") +
    ylab("Density") +
    ggtitle("Occurrences by Hour (Dens)") +  # Title of the plot
    theme_minimal()

grid.arrange(p1, p2, ncol = 2)
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

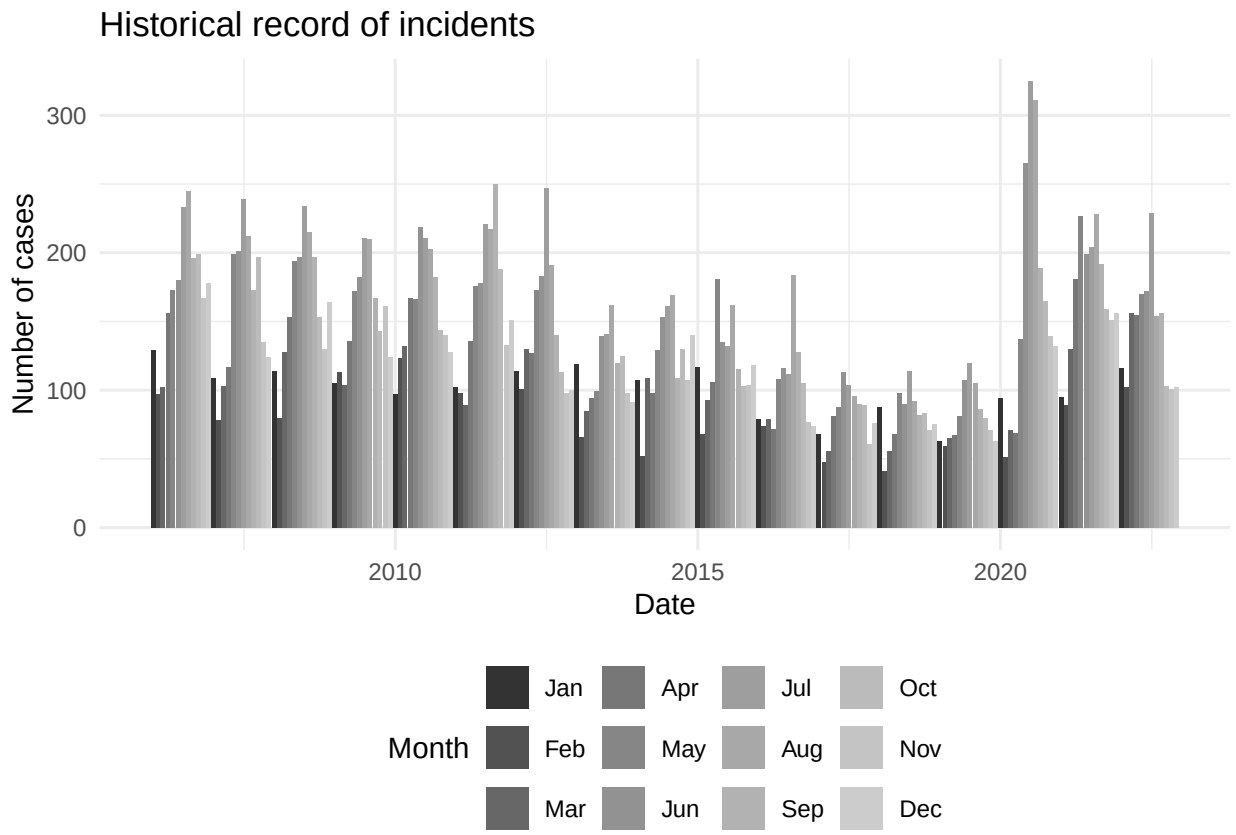


From the two previous plots, we can see that the month with the higher percentage of occurrences is July. This may be related to the fact that the summer months have more activity and person-to-person interaction in comparison to the winter months.

We may analyze this same information year by year to see if this same pattern is repeatable through time. It can also be noted that most of the incidents occur during the night with a peak around midnight.

```
data %>%
  group_by(month=lubridate::floor_date(OCCUR_DATE, "month")) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
    NUM_CASES = 1) %>%
  summarize(num_cases=sum(NUM_CASES),
  .groups="drop_last") %>%
  mutate(Month=factor(month.abb[month(month)],
    levels=c("Jan", "Feb", "Mar",
      "Apr", "May", "Jun",
      "Jul", "Aug", "Sep",
      "Oct", "Nov", "Dec"))) %>%
  ggplot(aes(fill=Month, y=num_cases, x=month)) +
    geom_bar(position="dodge", stat="identity") +
    ggtitle("Historical record of incidents") +
    theme_minimal() +
    theme(legend.position="bottom", legend.box = "horizontal") +
    scale_fill_grey() +
    xlab("Date") +
    ylab("Number of cases")
```

```
## Warning: There were 204 warnings in `mutate()`.
## The first warning was:
## i In argument: `OCCUR_DATE = mdy(OCCUR_DATE)`.
## i In group 1: `month = 2006-01-01`.
## Caused by warning:
## ! All formats failed to parse. No formats found.
## i Run `dplyr::last_dplyr_warnings()` to see the 203 remaining warnings.
```



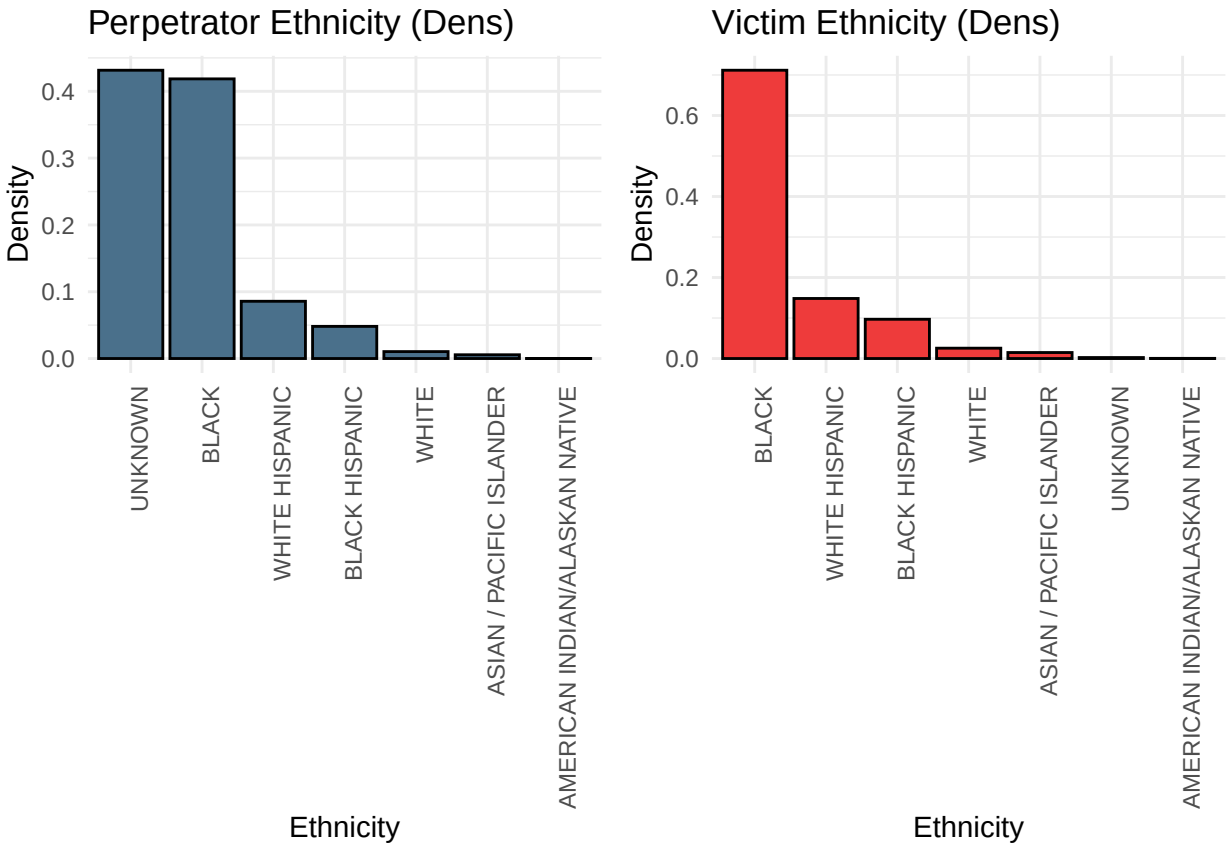
We can see in this plot there is actually a seasonal trend in the number of incidents, and the observation of summer as the period with more incidents is a pattern that repeats every year.We can additionally recognize there was a decrease in shooting incidents in 2013/2014 an 2020. The first change my be related to some change in crime policy, since there was a change in the city government in 2014, and the second one is most likely related to the COVID pandemic.

```
p3<-data %>%
  # Create the bar plot for perpetrators
  ggplot(aes(x = PERP_RACE)) +
    geom_bar(aes(y = (..count..)/sum(..count..)),fill = "skyblue4", color = "black")+
    xlab("Ethnicity") +
    ylab("Density") +
    ggtitle("Perpetrator Ethnicity (Dens)") +  # Title of the plot
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

p4<-data %>%
  # Create the bar plot for victims
```

```
  ggplot(aes(x = VIC_RACE)) +
    geom_bar(aes(y = (..count..)/sum(..count..)),fill = "brown2", color = "black")+
    xlab("Ethnicity") +
    ylab("Density") +
    ggtitle("Victim Ethnicity (Dens)") +  # Title of the plot
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

grid.arrange(p3, p4, ncol = 2)
```



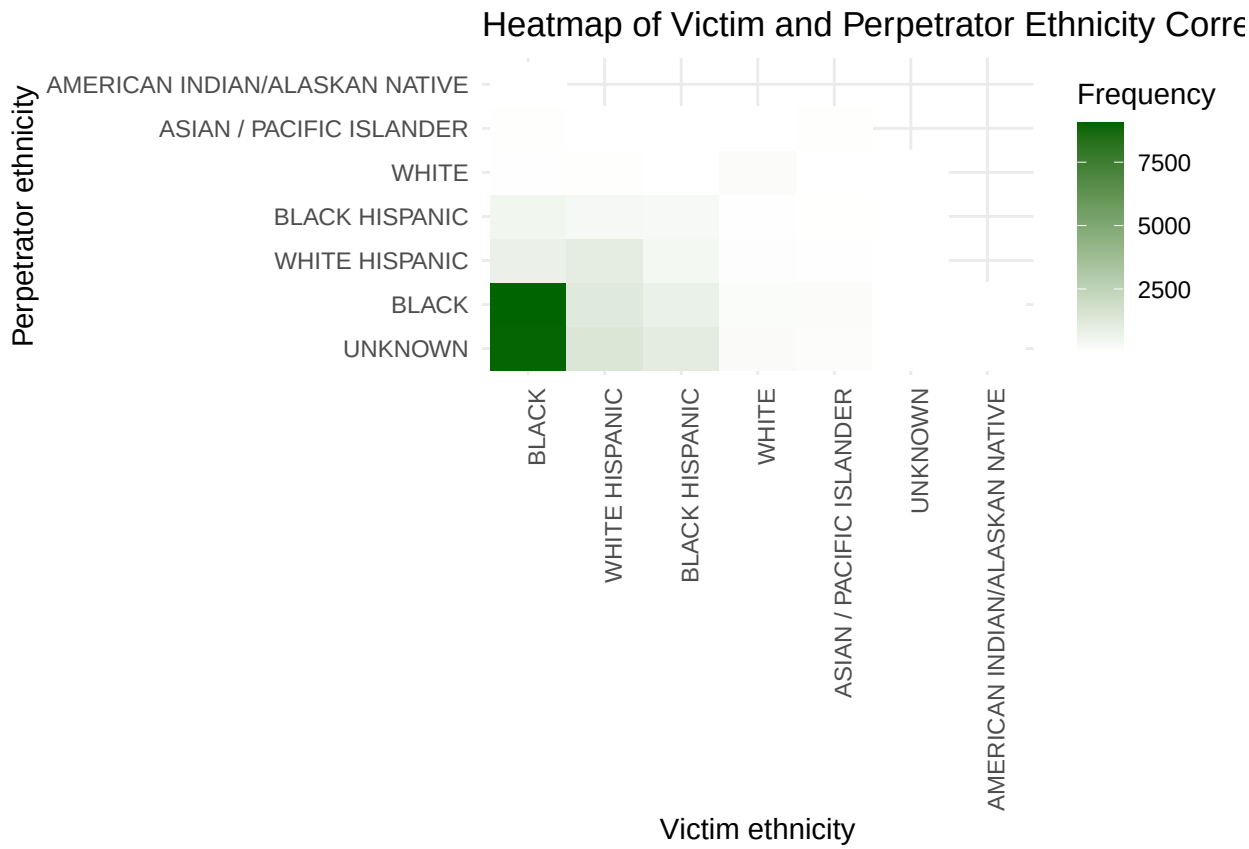
According to the data, African Americans are the group most involved with the incidents, both as perpetrator and as victim. This data may not be complete if don't know the total amount of population and their distribution trough the city according to ethnicity. One possible source of bias is the high number of Unknown entries. We could also be in a case of over reporting or under reporting when some particular groups are involved. This may explain the high amount of missing data.

Data modeling

It we useful for the analysis to find a correlation between the ethnicity of the perpetrator and the victim. Specially to know in the missing data can be interfering in the analysis.

```
# Preprocess the data to calculate frequencies
eth_correlation <- data %>%
  group_by(VIC_RACE, PERP_RACE) %>%
  summarise(Frequency = n(), .groups = 'drop')
```

```r
# Create the heatmap
ggplot(eth_correlation, aes(x = VIC_RACE, y = PERP_RACE, fill = Frequency)) +
  geom_tile() +
  scale_fill_gradient(low = "white", high = "darkgreen") +
  labs(title = "Heatmap of Victim and Perpetrator Ethnicity Correlation",
       x = "Victim ethnicity",
       y = "Perpetrator ethnicity",
       fill = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



From the analysis of heat map it does not seem the unknown variables is affecting in particular to the African american group, since there is a correlation of the involvement in the incidents between all the groups, i.e. if we put in order the ethnic groups according to the frequency of the incident we basically get the same order in the two axes. So, further analysis will be required to give an explanation to these findings.

In general, to have a better understanding of this data we should also consider the location of the incidents and the social economic conditions that may influence the data. For example, if the poorest zones are inherently more violent from the recent history and have a record of higher rate crimes, but happens to be also more affordable for marginalized groups such as African Americans and Hispanics, we may expect to see higher shooting incidents in these groups. We will need data beyond what is available at the NYPD database.

Conclusions

The analysis of the data reveals a notable seasonal trend in the incident occurrences, with July showing the highest frequency. This suggest a potential link between warmer months and higher amount of social interactions, and de incident rates. The data also shows the incidents predominantly occurs at night. This could also suggest the city may require to enhance safety measures and interventions at this time.

Furthermore, the demographic analysis point to a disproportionate involvement of African Americans in these incidents. However, the observation requires a caution interpretation, considering for potential bias, specially due to the significant number of entries classified as Unknown. It is also important a comprehensive approach to data analysis to incorporates socioeconomic perspective to provide a more nuanced understanding of the underlying dynamic and to mitigate the risk of oversimplification and bias.