NYPD Shooting Incidents Project Analysis

2024-10-09

Introduction

In this report, statistics from shooting incidents within New York City will be imported, tidied, and analyzed. Data was provided by the NYPD/ City of New York website and includes a wide array of data points, all of which can be seen in a summary below. The two main questions that will be answered are whether the borough in which the incident occurred or the race of the victims have a large impact on overall incidence and mortality rates.

Data Manipulation

We first will add the necessary libraries.

```
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Next, we will input the data from the City of New York's website

```
nypd_shootings <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv")
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

After that, we will remove all data that we don't want and change changes dates and times into objects.

```
nypd_shootings <- nypd_shootings %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME)) %>%
  select(-c(INCIDENT_KEY, LOC_OF_OCCUR_DESC, PRECINCT, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOCATION_
  rename( MURDER_OCCURED = `STATISTICAL_MURDER_FLAG` )

summary(nypd_shootings)
```

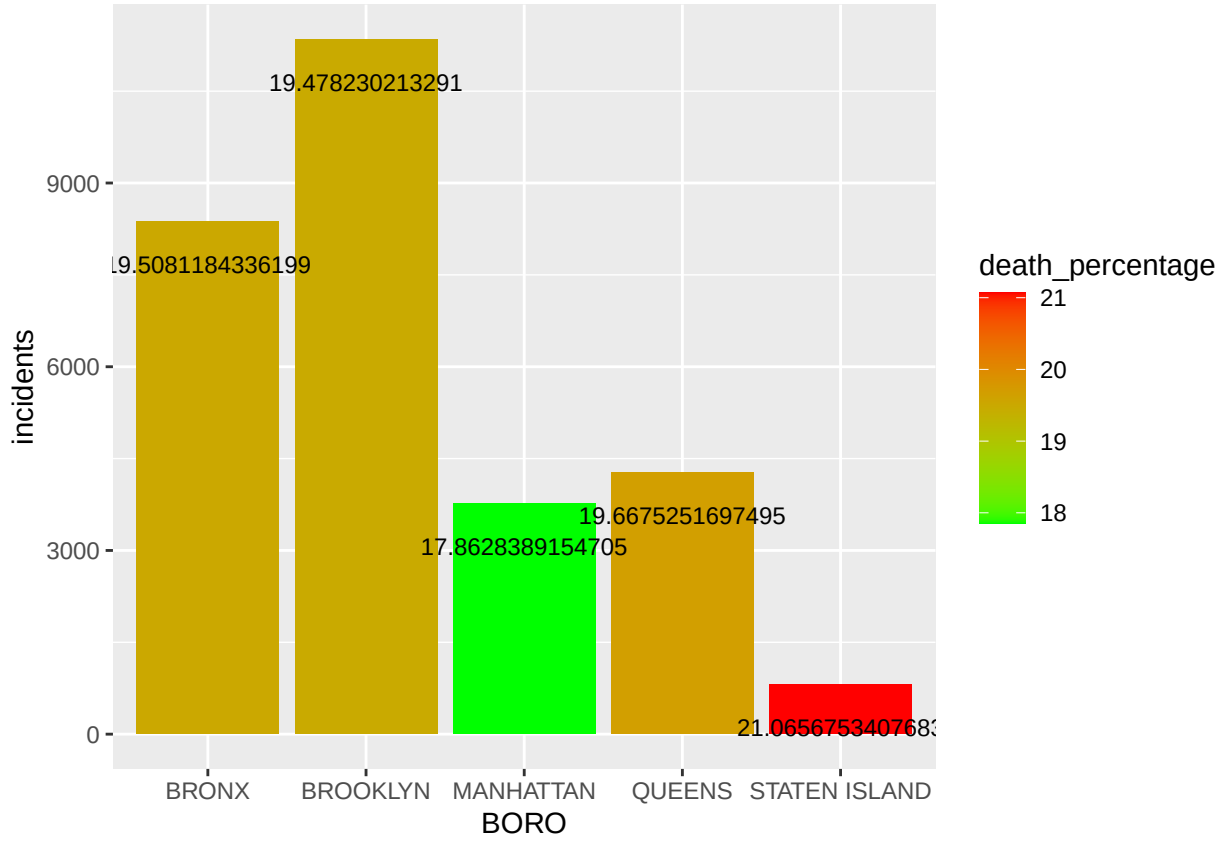
```
##    OCCUR_DATE           OCCUR_TIME                          BORO
##  Min.   :2006-01-01   Min.   :0S                        Length:28562
##  1st Qu.:2009-09-04   1st Qu.:3H 30M 0S                 Class :character
##  Median :2013-09-20   Median :15H 15M 0S                Mode  :character
##  Mean   :2014-06-07   Mean   :12H 44M 16.713115328057S
##  3rd Qu.:2019-09-29   3rd Qu.:20H 45M 0S
##  Max.   :2023-12-29   Max.   :23H 59M 0S
##  MURDER_OCCURED  PERP_AGE_GROUP       PERP_SEX          PERP_RACE
##  Mode :logical   Length:28562       Length:28562       Length:28562
##  FALSE:23036     Class :character   Class :character   Class :character
##  TRUE :5526      Mode  :character   Mode  :character   Mode  :character
##
##
##
##  VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Data Visualizations

After cleaning our data, we can manipulate it further to group statistics by borough that we are interested in. By calculating the deaths per incident in each borough, we can come to the conclusion that Staten Island has the most deadly shooting incidents (while also having the fewest incidents in total). We can also see that Manhattan has the least deadly shooting incidents, while having considerably more overall. This leads to the question- which borough is actually safest? With the data here, an argument could be made for either, but other statistics could sway that perception greatly- maybe to even one of the others boroughs completely.

```
boro_stats <- nypd_shootings %>%
  group_by(BORO) %>%
  summarize(incidents = n(), deaths = sum(MURDER_OCCURED==TRUE)) %>%
  mutate(death_percentage = (deaths / incidents * 100))

ggplot(data=boro_stats, aes(x=BORO, y=incidents, fill = death_percentage)) +
geom_bar(stat="identity") + scale_fill_gradient(low="green", high="red") +
geom_text(aes(label = death_percentage), size = 3, hjust = 0.5, vjust = 3, position = "stack")
```

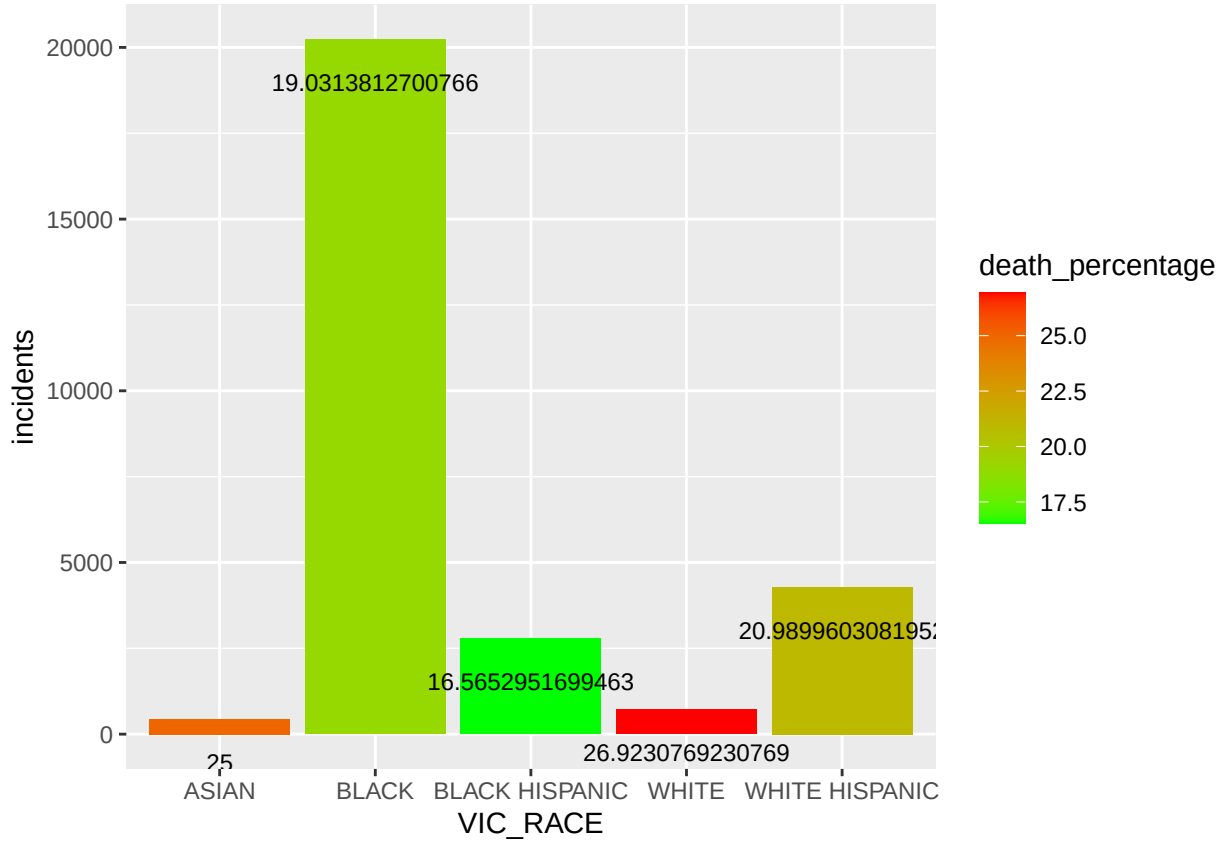


In another visualization, we look at deaths and incident statistics based on the victim's race. By looking at the data broken into each victim race group, we can see that black individuals are the most common victims. However, white and Asian individuals are more likely to be fatally wounded when shot. Why does this occur? What effects do population and cultural/societal impacts have on this data? Could it be misleading at all? We could certainly add more to this data set and find out.

```
victim_race_stats <- nypd_shootings %>%
  group_by(VIC_RACE) %>%
  filter(VIC_RACE != 'UNKNOWN' & VIC_RACE != 'AMERICAN INDIAN/ALASKAN NATIVE' ) %>%
  summarize(incidents = n(), deaths = sum(MURDER_OCCURED==TRUE)) %>%
  mutate(death_percentage = (deaths / incidents * 100))

victim_race_stats[victim_race_stats == "ASIAN / PACIFIC ISLANDER"] <- "ASIAN"

ggplot(data=victim_race_stats, aes(x=VIC_RACE, y=incidents, fill = death_percentage)) +
geom_bar(stat="identity") + scale_fill_gradient(low="green", high="red") +
geom_text(aes(label = death_percentage), size = 3, hjust = 0.5, vjust = 3, position = "stack")
```



Data Models

Here, we can make a model that predicts the the likely amount of deaths per incidents for each borough. The model proves to be quite accurate, with the greatest variances coming from the two boroughs we questioned most in our first analysis. Even though Manhattan and Staten Island have the greatest variance from our predicted model, they still are quite close (within two percent) of their expected value. Given this, I would say that boroughs do not have a large impact on deadliness of incidents in New York City.

```
mod <- lm(death_percentage ~ incidents, data = boro_stats)
boro_stats_w_pred <- boro_stats %>% mutate(pred = predict(mod))

boro_stats_w_pred
```

```
## # A tibble: 5 x 5
##   BORO          incidents deaths death_percentage  pred
##   <chr>             <int>  <int>            <dbl> <dbl>
## 1 BRONX              8376   1634             19.5  19.3
## 2 BROOKLYN          11346   2210             19.5  19.1
## 3 MANHATTAN          3762    672             17.9  19.7
## 4 QUEENS             4271    840             19.7  19.6
## 5 STATEN ISLAND       807    170             21.1  19.9
```

Bias and Conlusion

In this data, there are quite a few possible sources of bias. To start, the data is sourced from the NYPD who know these statistics will be looked upon by many. Given this, it is certainly possible that they could misrepresent certain data points. Two that come to mind are the shootings that involve NYPD members and shootings that involve friends and family of the recorders. In each case, the officers recording the data could have personal bias that could skew data. Another source of bias here would be the data interpreter (me). I am a white male from a small town, so I could have a narrow view when it comes to interpreting opinions about a large, diverse city. Regardless, I tried my best to broaden my head space when looking through the data and not let my personal bias effect my analysis.

To conclude, It would appear that while boroughs and victim's races greatly impacted the number of incidents, neither strongly affected the mortality rates in these instances.