NYPD Shooting Incident Data Report

2024-09-18

Importing Libraries

```r
library(nnet)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Loading Data

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

Introduction and Question of Interest

The data used in the report below describes a collection of reports of shooting incidents taking place from 2006 to 2024 in New York City. The data is reviewed by the Office of Management Analysis and Planning and uploaded to New York Police Department's website. This report will focus on analyzing the spacial data presented to the end of identifying where incidents are most prevalent, and how these incidents are distributed across the region.

An additional analysis will be done on a model using logistic regression to predict the Borough in which a shooting incident took place based on certain characteristics of the perpetrator, namely sex, race and age group.

Data Cleaning and Data Description For Visualizations

The visualization portion of this report will work with spacial data exclusively, to that end the columns used in this analysis will be the following:

1. **Longitude:** Longitude coordinates in decimal degrees
2. **Latitude:** Latitude coordinates in decimal degrees
3. **BORO:** Borough where the shooting incident occurred
4. **PRECINCT:** Precinct where the shooting incident occurred

Luckily, the data in these columns is clean. This means that we can use it as is. The code below is simply selecting the columns that will be used for this analysis.

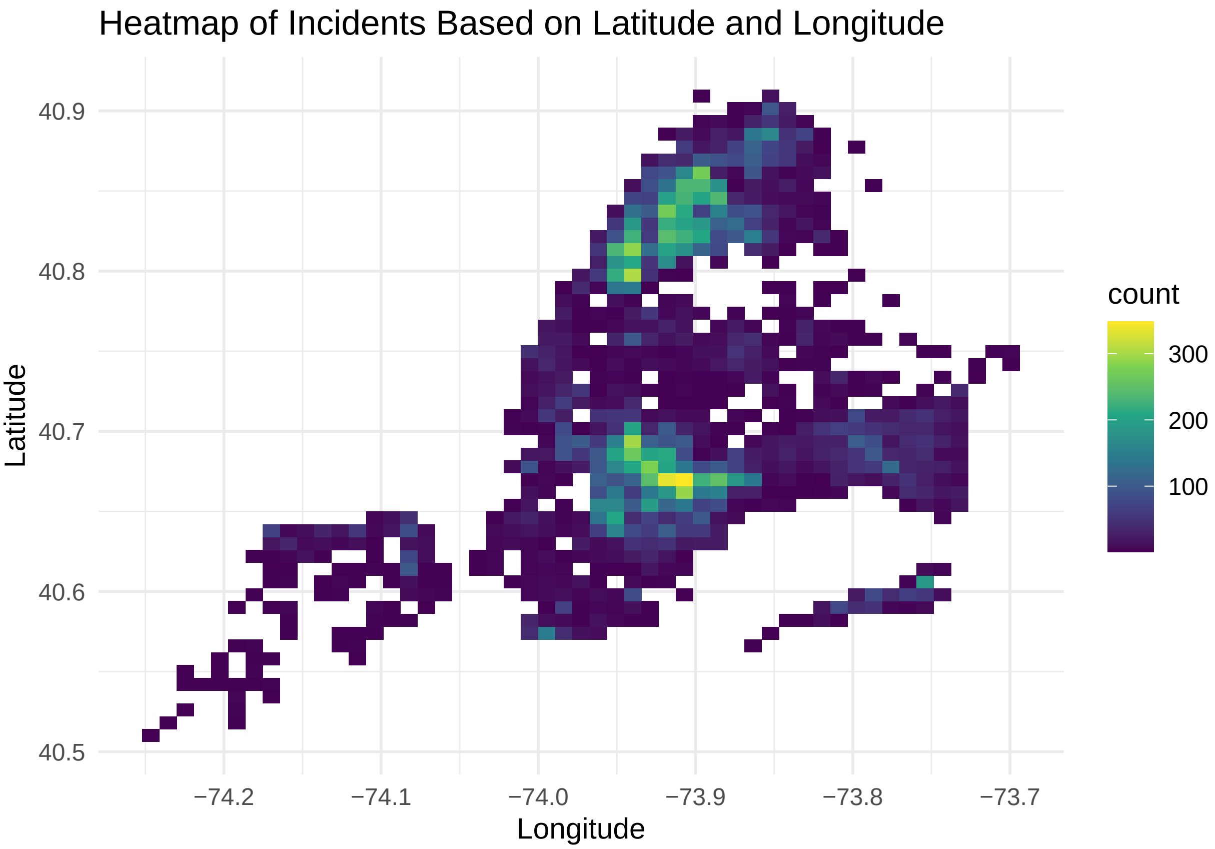
```
nypd_data <- nypd_data %>% select(c(
  Longitude,Latitude,PRECINCT,BORO))
```

Visualizations and Analysis

The heat map below displays the concentration of shooting incidents based on their reported latitude and longitude. This allows us to generate an approximation of the geographical map of NYC with which we can determine the areas with highest activity.

It seems like the frequency of incidents is highest in a couple areas. One of these areas is located further north while the other is more central. Let's look at the precincts per borough to identify precisely where these locations are.

```
  ggplot(nypd_data, aes(x = Longitude, y = Latitude)) +
  geom_bin2d(bins = 50) +
  scale_fill_continuous(type = "viridis") +
  labs(title = "Heatmap of Incidents Based on Latitude and Longitude",
       x = "Longitude",
       y = "Latitude") +
  theme_minimal()
```

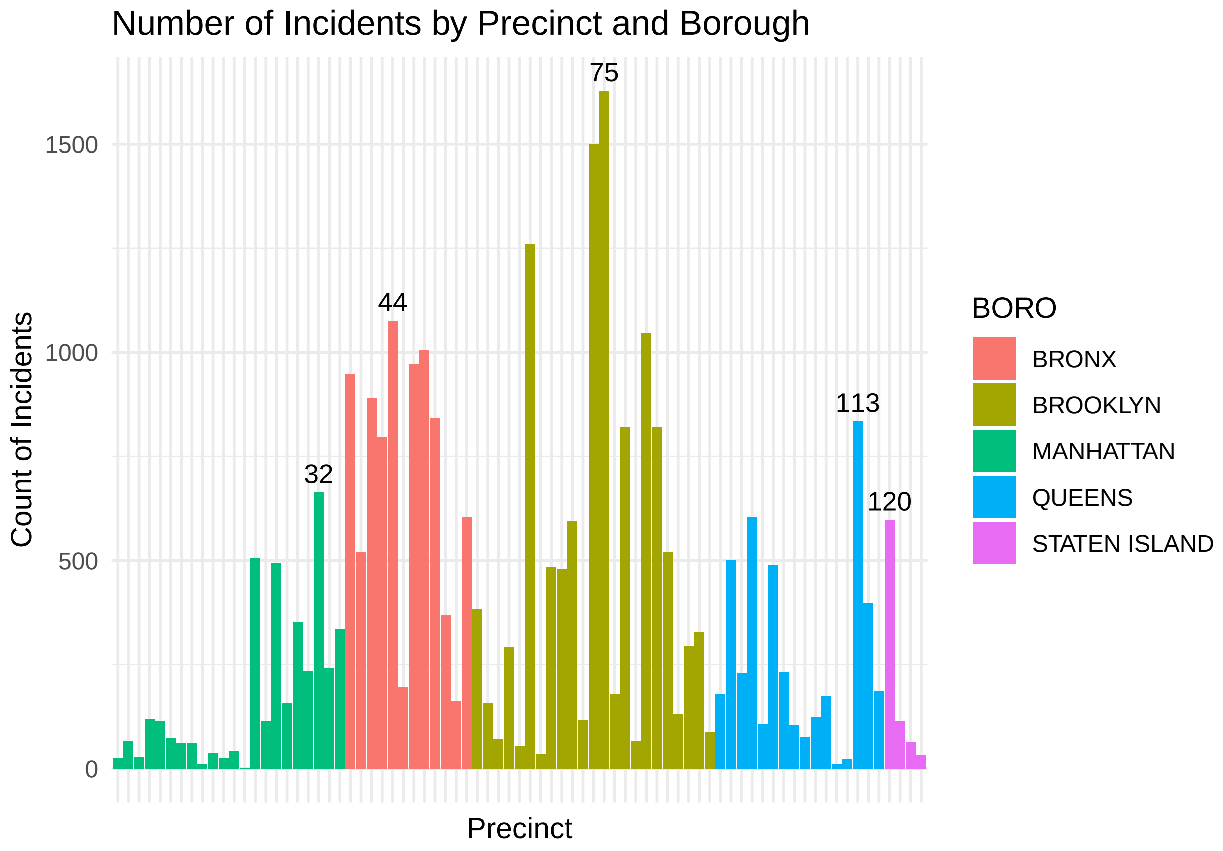


Based on the bar graph below it seems like the two hotspots seen on the heat map correspond to the boroughs of Bronx and Brooklyn. The constant height of the bars across Bronx's precincts implies that it is the borough with highest average level of incidents; however, the largest concentrations of incidents per individual precinct can be found in Brooklyn in precinct 75.

Queens and Staten island had much lower average counts of incidents per precinct, but they each had one precinct with significantly increased counts of incidents in comparison to the rest of their precincts. This was precinct 113 for Queens and precinct 120 for Staten Island.

```
  suppressMessages(most_incidents_per_borough <- nypd_data %>%
  group_by(BORO, PRECINCT) %>%
  summarise(incident_count = n()) %>%
  top_n(1, wt = incident_count) %>%
  arrange(BORO, desc(incident_count)))

ggplot(nypd_data, aes(x = factor(PRECINCT), fill = BORO)) +
  geom_bar(stat = 'count') +
  geom_text(data = most_incidents_per_borough,
  aes(x = factor(PRECINCT), y = incident_count, label = PRECINCT),
            vjust = -0.5, color = "black", size = 3.5) +
  labs(title = "Number of Incidents by Precinct and Borough",
       x = "Precinct",
       y = "Count of Incidents") +
  theme_minimal() +
  theme(axis.text.x =element_blank(),
        axis.ticks.x=element_blank())
```



Manhattan has the lowest average frequency of shooting incidents per precinct, with around 200, which is approximately 3.5 times less than Bronx's average frequency and approximately 2.5 times less than Brooklyn's.

```
suppressMessages(average_incidents_by_borough_precinct <- nypd_data %>%
  group_by(BORO, PRECINCT) %>%
  summarise(incident_count = n()) %>%
  group_by(BORO) %>%
  summarise(avg_incidents_per_precinct = mean(incident_count)))
```

```r
knitr::kable(average_incidents_by_borough_precinct,
             col.names = c("Borough", "Average Incidents per Precinct"),
             caption = "Average Incidents per Precinct by Borough")
```

Table 1: Average Incidents per Precinct by Borough

Borough	Average Incidents per Precinct
BRONX	698.0000
BROOKLYN	493.3043
MANHATTAN	171.0000
QUEENS	266.9375
STATEN ISLAND	201.7500

Data Cleaning and Data Description For Model

This section will use the characteristics of the perpetrator to attempt to predict the borough in which the shooting incident took place. Since this model will use a largely different group of columns compared to the analysis above, a new data cleansing process must be done for those columns.

The data cleansing below will class values that have an ambiguous classification as "Unknown", and then remove rows containing those values.

These will be the columns used for this model:

1. **PERP_SEX:** Sex of the perpetrator
2. **PERP_RACE:** Race of the perpetrator
3. **PERP_AGE_GROUP:** Age group of the perpetrator
4. **BORO:** Borough where the shooting incident occurred

```r
nypd_data <- read.csv(data_url)

# Cleaning PERP_SEX by putting missing values under "Unknown"
nypd_data$PERP_SEX[nypd_data$PERP_SEX == ""] <- "Unknown"
nypd_data$PERP_SEX[nypd_data$PERP_SEX == "U"] <- "Unknown"
nypd_data$PERP_SEX[nypd_data$PERP_SEX == "(null)"] <- "Unknown"

# Cleaning PERP_RACE by putting missing values under "Unknown"
nypd_data$PERP_RACE[nypd_data$PERP_RACE == "(null)"] <- "Unknown"
nypd_data$PERP_RACE[nypd_data$PERP_RACE == "UNKNOWN"] <- "Unknown"

# Cleaning PERP_AGE_GROUP by putting missing values under "Unknown"
nypd_data$PERP_AGE_GROUP[nypd_data$PERP_AGE_GROUP == "(null)"] <- "Unknown"
nypd_data$PERP_AGE_GROUP[nypd_data$PERP_AGE_GROUP == ""] <- "Unknown"
nypd_data$PERP_AGE_GROUP[nypd_data$PERP_AGE_GROUP == "UNKNOWN"] <- "Unknown"

# Some age ranges had 1 entry and weird categories,
# there were likely errors so they will be placed under "Unknown" as well
age_group_counts <- nypd_data %>%
  group_by(PERP_AGE_GROUP) %>%
  tally() %>%
  filter(n == 1)
nypd_data <- nypd_data %>%
  mutate(PERP_AGE_GROUP = ifelse(PERP_AGE_GROUP %in%
  age_group_counts$PERP_AGE_GROUP, "Unknown", PERP_AGE_GROUP))

# Selecting only the columns that will be used
nypd_data <- nypd_data %>% select(c(
  BORO,PERP_AGE_GROUP,PERP_SEX,PERP_RACE))

# Removing rows where any column contains "Unknown"
nypd_data_clean <- nypd_data %>%
  filter_all(all_vars(. != "Unknown"))

# Converting to factor
nypd_data_clean$BORO <- as.factor(nypd_data_clean$BORO)
nypd_data_clean$PERP_SEX <- as.factor(nypd_data_clean$PERP_SEX)
nypd_data_clean$PERP_RACE <- as.factor(nypd_data_clean$PERP_RACE)
nypd_data_clean$PERP_AGE_GROUP <- as.factor(nypd_data_clean$PERP_AGE_GROUP)
```

Model Set-up and Evaluation

The type of model used to predict the borough is a multinomial logistic regression model. This is a type of logistic regression that is generalized for problems with more than two possible discrete outcomes. A confusion matrix will be used to evaluate the model.

```
# Setting up model
multinom_model <- multinom(BORO ~ PERP_AGE_GROUP + PERP_SEX + PERP_RACE, data = nypd_data_clean)
```

```
## # weights:  60 (44 variable)
## initial  value 23764.960215
## iter  10 value 21080.049339
## iter  20 value 20645.130549
## iter  30 value 20433.626785
## iter  40 value 20423.401252
## iter  50 value 20422.413516
## iter  60 value 20421.985348
## final  value 20421.970079
## converged
```

```
# Making predictions
predicted_classes <- predict(multinom_model)

# Creating confusion matrix
confusion_matrix <- table(nypd_data_clean$BORO, predicted_classes)

# Displaying confusion matrix
confusion_matrix
```

```
##                 predicted_classes
##                  BRONX BROOKLYN MANHATTAN QUEENS STATEN ISLAND
##   BRONX           1793     2803         0     31             0
##   BROOKLYN         689     4452         0     42             0
##   MANHATTAN        653     1510         0     13             0
##   QUEENS           464     1651         1     74             0
##   STATEN ISLAND     79      499         0     12             0
```

The confusion matrix shows the predicted values as columns and the real values as rows. It looks like the model had a significant amount of false positives in Brooklyn and Bronx. This is understandable since these are the two regions with most incidents; however this model cannot be determined to be sufficiently reliable. It also worth noting that it almost entirely failed to assign any incidents to Manhattan or Staten Island.

I believe that the performance of this model could be improved by selecting a different type of model, such as random forest or decision tree. Adding the characteristics of the victim may also provide the model with useful data to help it more accurately assign outcomes.

Conclusion and Bias Analysis

The data analyzed in this report describes the areas with the least shooting incidents to be Manhattan, Staten Island and Queens. Brooklyn and Bronx have larger concentrations of shooting incidents than the other regions, with Bronx having the highest average count of incidents per precinct and Brooklyn having the most total incidents.

The multinominal logistic regression model used seems to imply that the characteristics of the perpetrator are not enough to identify under which borough the incident took place, but this could be a limitation in the execution of the model rather than an inherent lack of predictive between the selected variables.

Additionally, it is important to note that the methodology of the data collection, validation, and cleansing processes could have affected the reported values significantly.

We are unaware of what the data looked like before it was reviewed by the Office of Management Analysis and Planning, and more importantly we do not know the conditions that the data was collected under.

A bias could exist causing certain areas to be reported more often, or inversely, a lack of police presence may cause a lack of reporting in certain areas. Without context and further details of the conditions in which this data was recorded it is difficult to ascertain what biases could have had a significant impact.

Some additional data that could supplement this report to provider a more complete picture of the situation is data describing population levels, and data describing the presence of police officers in the region.

I do not believe that I have a bias that could significantly affect the analysis provided since I am not local to the region, or even particularly familiar with New York City's boroughs. As a person raised outside of the United States, I may be qualified to provide a somewhat unbiased perspective of the data analyzed in this report.

Resources:

- https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8/about_data