NYPD Shooting Incident Analysis

Christopher Washington

2024-06-01

Introduction

Violent crime is something that has the potential to affect each one of us every single day. It is important to analyze violent crime data in order to allocate resources to mitigate crime in every neighborhood in the world. Luckily for us, there is data publicly available for shootings in New York City provided by the New York Police Department (NYPD). The report will take a small look into this data. ## Required R Libraries The following R libraries are required to run the R code contained in this document:

```
library(tidyverse)
library(knitr)
library(lubridate)
library(reshape2)
set.seed(10)
```

Dataset

The dataset that will be used for this report is the NYPD Shooting Incident Data (Historic) located at https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD. This dataset contains a comprehensive list of shootings in New York City from 2006 to the end of the previous calendar year. It includes data on the perpetrator and victim of the shooting, such as age, gender, and race. The dataset also contains data on location such as the borough that the incident occurred in, as well data about the date and time of the incident.

There is also population data used in this analysis that is obtained from the census.gov data located at https://www.census.gov/quickfacts/fact/table/newyorkcountynewyork,richmondcountynewyork,kingscountynewyork,queenscountynewyork,bronxcountynewyork,newyorkcitynewyork/PST045222

Data Retrieval

We will begin by loading the dataset

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

We are only interested in the date and borough for this analysis, so we will remove the unnecessary columns. We will also rename the columns to simpler names for later use.

```
shootings = shootings %>%
  rename(date = 'OCCUR_DATE') %>%
  rename(boro = 'BORO') %>%
  select(date, boro)
```

We will continue to preprocess our data by converting columns to the appropriate data types. Categorical data columns will become factors, and date/time columns will be converted to the appropriate date/time data type.

```
shootings = shootings %>% mutate_at(
  vars(boro),
  factor
)
shootings = shootings %>% mutate(date = mdy(date))
```

Finally, we will filter out any null and UNKNOWN records from the dataset:

```
shootings = shootings %>%
  filter(boro != "(null)") %>%
  filter(boro != "UNKNOWN")
```

The final dataset looks like the following snapshot where each row contains the date and borough for a single shooting:

date	boro
2022-05-05	MANHATTAN
2022-07-04	BRONX
2012-05-27	QUEENS
2019-09-24	BRONX
2007-02-25	BROOKLYN
2021-07-01	MANHATTAN

Visualization 1: Shootings Per 1000 Residents in Each Borough

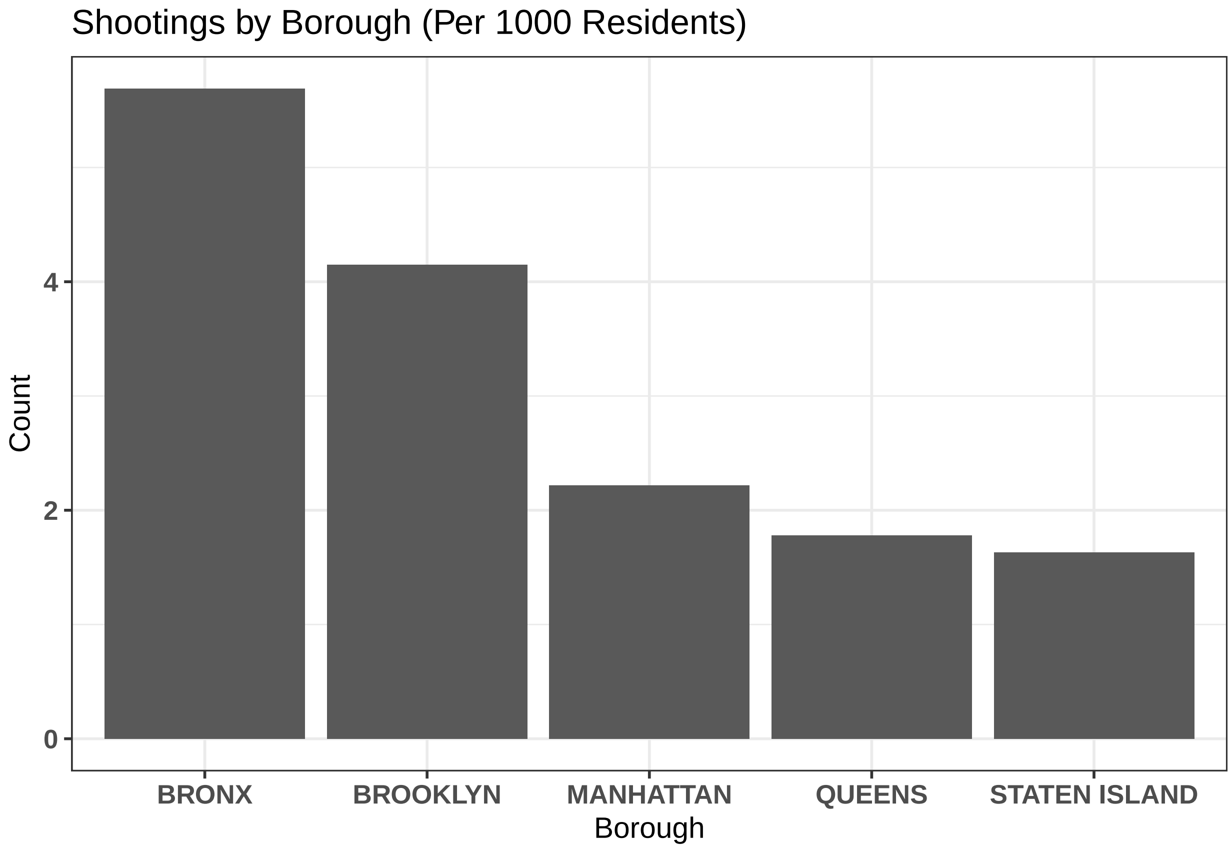
The first visualization that we will consider is the number of shootings per 1000 residents in each New York City borough. We compute the per 1000 value as follows

```
shootings_by_borough_per_1000 = shootings %>%
  group_by(boro) %>%
  summarise(n = n()) %>%
  mutate(pop_n =
    ifelse(boro == 'BRONX', 1472654,
        ifelse(boro == 'BROOKLYN', 2736074,
           ifelse(boro == 'MANHATTAN', 1694251,
              ifelse(boro == 'QUEENS', 2405464,
                 ifelse(boro == 'STATEN ISLAND', 495747, n)
              )
           )
        )
    )
  ) %>%
  mutate(per_1000 = round(1000*(n/pop_n), 2)) %>%
  select(boro, per_1000)
```

and we obtain the following data

boro	per_1000
BRONX	5.69
BROOKLYN	4.15
MANHATTAN	2.22
QUEENS	1.78
STATEN ISLAND	1.63

We then plot this data on a bar chart



Model 1: Shootings by Borough vs Population by Borough

We will now conduct some analysis of the distribution of shootings in each borough. If there are no factors contributing to increases or decreases in shootings in each borough, then we can assume that the proportions of shootings in each borough will reflect the populations in each borough. We compute the proportions of shootings in each borough, and the proportion of population in each borough as follows

```
shootings_by_borough = shootings %>%
  group_by(boro) %>%
  summarise(n = n()) %>%
  mutate(p = round(n/sum(n), 4)) %>%
  mutate(pop_n =
    ifelse(boro == 'BRONX', 1472654,
        ifelse(boro == 'BROOKLYN', 2736074,
           ifelse(boro == 'MANHATTAN', 1694251,
              ifelse(boro == 'QUEENS', 2405464,
                 ifelse(boro == 'STATEN ISLAND', 495747, n)
             )
          )
       )
    )
  ) %>%
  mutate(pop_p =
    ifelse(boro == 'BRONX', round(1472654/sum(pop_n), 4),
        ifelse(boro == 'BROOKLYN', round(2736074/sum(pop_n), 4),
           ifelse(boro == 'MANHATTAN', round(1694251/sum(pop_n), 4),
              ifelse(boro == 'QUEENS', round(2405464/sum(pop_n), 4),
                 ifelse(boro == 'STATEN ISLAND', round(495747/sum(pop_n), 4), n)
             )
          )
       )
    )
  )
```

From this computation we obtain the following data

boro	n	p	pop_n	pop_p
BRONX	8376	0.2933	1472654	0.1673
BROOKLYN	11346	0.3972	2736074	0.3108
MANHATTAN	3762	0.1317	1694251	0.1924
QUEENS	4271	0.1495	2405464	0.2732
STATEN ISLAND	807	0.0283	495747	0.0563

To test if it is indeed the case that the proportion of shootings in each borough matches the proportion of population in each borough, we will use the Chi-Squared Goodness of Fit test. The test statistic we will consider is

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

. We compute the test statistic as follows

```
shootings_by_borough_analysis = shootings_by_borough %>%
  mutate(exp_n = round(sum(n)*pop_p, 0)) %>%
  mutate(diff = n-exp_n) %>%
  mutate(diff_sq = diff^2) %>%
  mutate(diff_sq_scl = round(diff_sq/exp_n, 0)) %>%
  select(boro, n, exp_n, diff, diff_sq, diff_sq_scl)
```

We obtain the following data

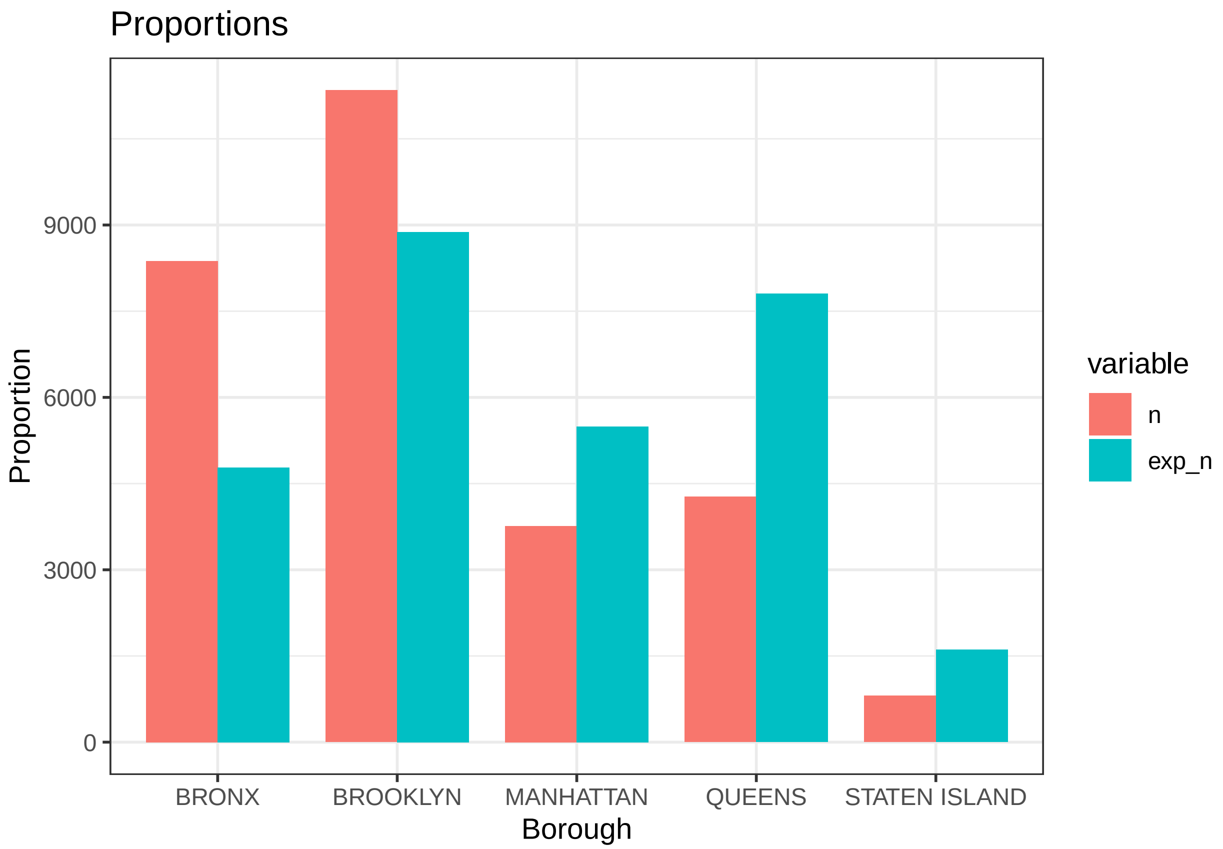
boro	n	exp_n	diff	diff_sq	diff_sq_scl
BRONX	8376	4778	3598	12945604	2709
BROOKLYN	11346	8877	2469	6095961	687
MANHATTAN	3762	5495	-1733	3003289	547
QUEENS	4271	7803	-3532	12475024	1599
STATEN ISLAND	807	1608	-801	641601	399

Then we sum up the last column of the previous table to obtain the test statistic value of 5,941. The p-value of this test statistic is given by

```
1-pchisq(5941, 4)
```

```
## [1] 0
```

Since the p-value is extremely small, we may reject the assumption that the proportion of shootings in each borough matches the population in each borough. We visualize this different in the following bar chart displaying the number of shootings in each borough (n) next to the expected number of shootings in each borough under the previous assumption (exp_n).



Analysis

Both visualtions displayed here raise questions about the distributions of shootings in each borough. It is clear that some boroughs like Bronx and Brooklyn have more shooting incidents than others. Reasons for these discrepancies are numerous, and may include socioeconomic differences or geographic differences. In future work we can explore these other topics to attempt to zero in on the causes of the differences.

Bias Identification

We do not see any bias in the dataset that we have constructed, or the analyses we have made here. The dataset simply contains a date and location for each shooting. However, it is possible that bias occurs in the underlying dataset. It may be the case that the dataset does not contain every shooting instance in New York City, but only contains the incidents that the police were called and investigated for. This could certainly upset the visualizations and analysis that we have done here.

Conclusion

There are many other visualization and analyses that can be brought from this dataset. Two of those have been provided here. These visualizations and analyses provide a starting point for further research and anlysis regarding shootings in New York City.