NYPD Shooting

Sean Baier

2022-11-23

Assignment

Import, tidy and analyze the NYPD Shooting Incident data set obtained.

- Be sure your project is reproducible and contains some visualization and analysis.
- You may use the data to do any analysis that is of interest to you.
- You should include at least two visualizations and one model.
- Be sure to identify any bias possible in the data and in your analysis.

Load dependencies

```
library(tidyverse)
library(lubridate)
library(modelr)
```

Raw Data for NYPD shootings

Import raw data from url.

https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

```
shootings_data <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNL
```

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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.
```

Tidy Raw Shootings Data

- Some of the `perp_race` values are implicitly `UNKNOWN` due to them having the value `NA`. We are going to make these values explicitly `UNKNOWN` by mutating them.

- Normalize column names to lowercase and snake_case format.
- Remove unnecessary coordinate related columns since they are not needed for the following analysis.

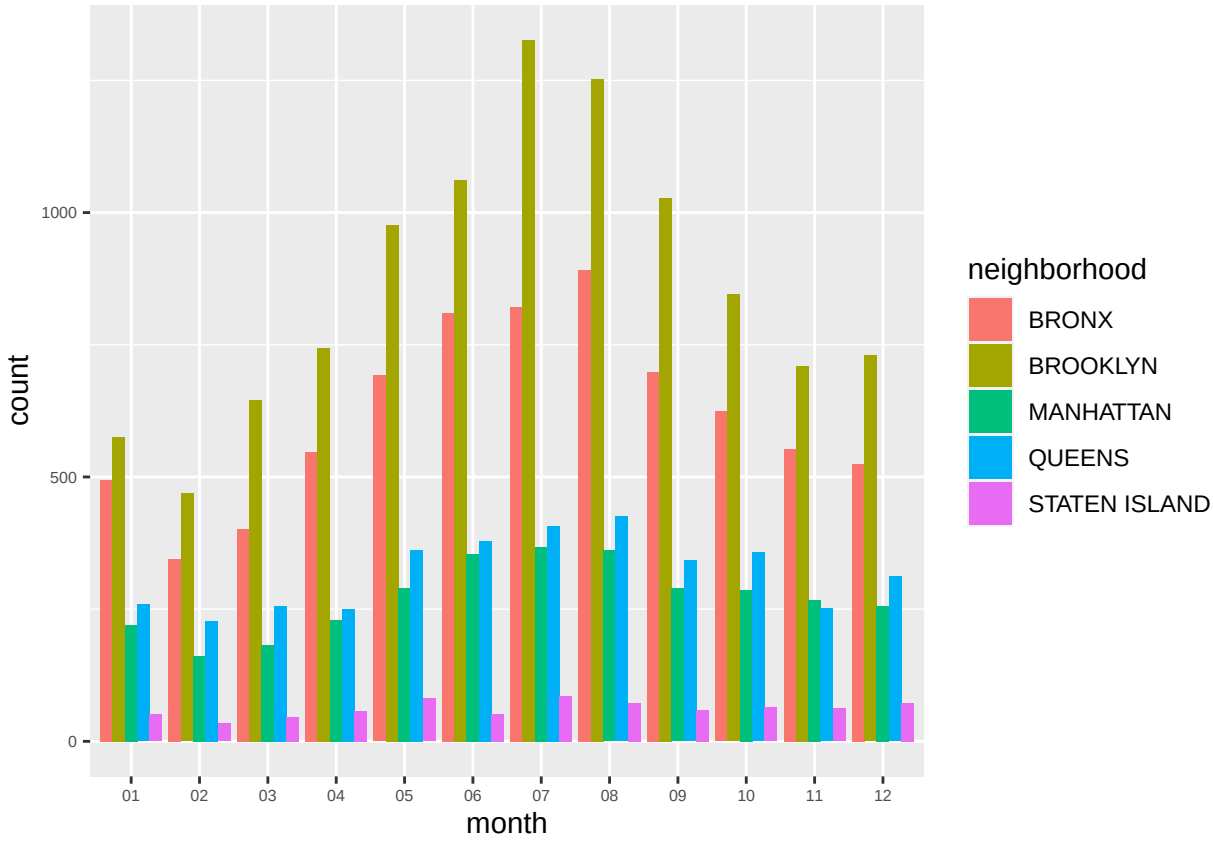
```
tidy_data <- shootings_data %>%
  rename(
    full_date = OCCUR_DATE,
    time = OCCUR_TIME,
    neighborhood = BORO,
    precinct = PRECINCT,
    jurisdiction_code = JURISDICTION_CODE,
    statistical_murder = STATISTICAL_MURDER_FLAG,
    perp_age = PERP_AGE_GROUP,
    perp_sex = PERP_SEX,
    perp_race = PERP_RACE,
    vic_age = VIC_AGE_GROUP,
    vic_sex = VIC_SEX,
    vic_race = VIC_RACE) %>%
  mutate(date = mdy(full_date)) %>%
  separate(date, into =c("year", "month", "day")) %>%
  replace_na(list(perp_race = "UNKNOWN")) %>%
  select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat, INCIDENT_KEY, LOCATION_DESC))
tidy_data
```

```
## # A tibble: 25,596 x 15
##    full_~1 time  neigh~2 preci~3 juris~4 stati~5 perp_~6 perp_~7 perp_~8 vic_age
##    <chr>   <tim> <chr>     <dbl>   <dbl> <lgl>   <chr>   <chr>   <chr>   <chr>
##  1 11/11/~ 15:04 BROOKL~      79       0 FALSE   <NA>    <NA>    UNKNOWN 18-24
##  2 07/16/~ 22:05 BROOKL~      72       0 FALSE   45-64   M       ASIAN ~ 25-44
##  3 07/11/~ 01:09 BROOKL~      79       0 FALSE   <18     M       BLACK   25-44
##  4 12/11/~ 13:42 BROOKL~      81       0 FALSE   <NA>    <NA>    UNKNOWN 25-44
##  5 02/16/~ 20:00 QUEENS      113       0 FALSE   <NA>    <NA>    UNKNOWN 25-44
##  6 05/15/~ 04:13 QUEENS      113       0 TRUE    <NA>    <NA>    UNKNOWN 25-44
##  7 04/14/~ 21:08 BRONX        42       0 TRUE    <NA>    <NA>    UNKNOWN 18-24
##  8 12/10/~ 19:30 BRONX        52       0 FALSE   <NA>    <NA>    UNKNOWN 25-44
##  9 02/22/~ 00:18 MANHAT~      34       0 FALSE   <NA>    <NA>    UNKNOWN 25-44
## 10 03/07/~ 06:15 BROOKL~      75       0 TRUE    25-44   M       BLACK ~ 25-44
## # ... with 25,586 more rows, 5 more variables: vic_sex <chr>, vic_race <chr>,
## #   year <chr>, month <chr>, day <chr>, and abbreviated variable names
## #   1: full_date, 2: neighborhood, 3: precinct, 4: jurisdiction_code,
## #   5: statistical_murder, 6: perp_age, 7: perp_sex, 8: perp_race
```

Total number of shootings per month by neighborhood

We are also going to separate the `date` into `month`, `day`, and `year` in order to analyze the per month shootings per neighborhood.

```
plot <- ggplot(data = tidy_data) +
  geom_bar(mapping = aes(x = month, fill = neighborhood), position = "dodge")

plot + theme(
  axis.text = element_text(size = rel(0.5))
)
```

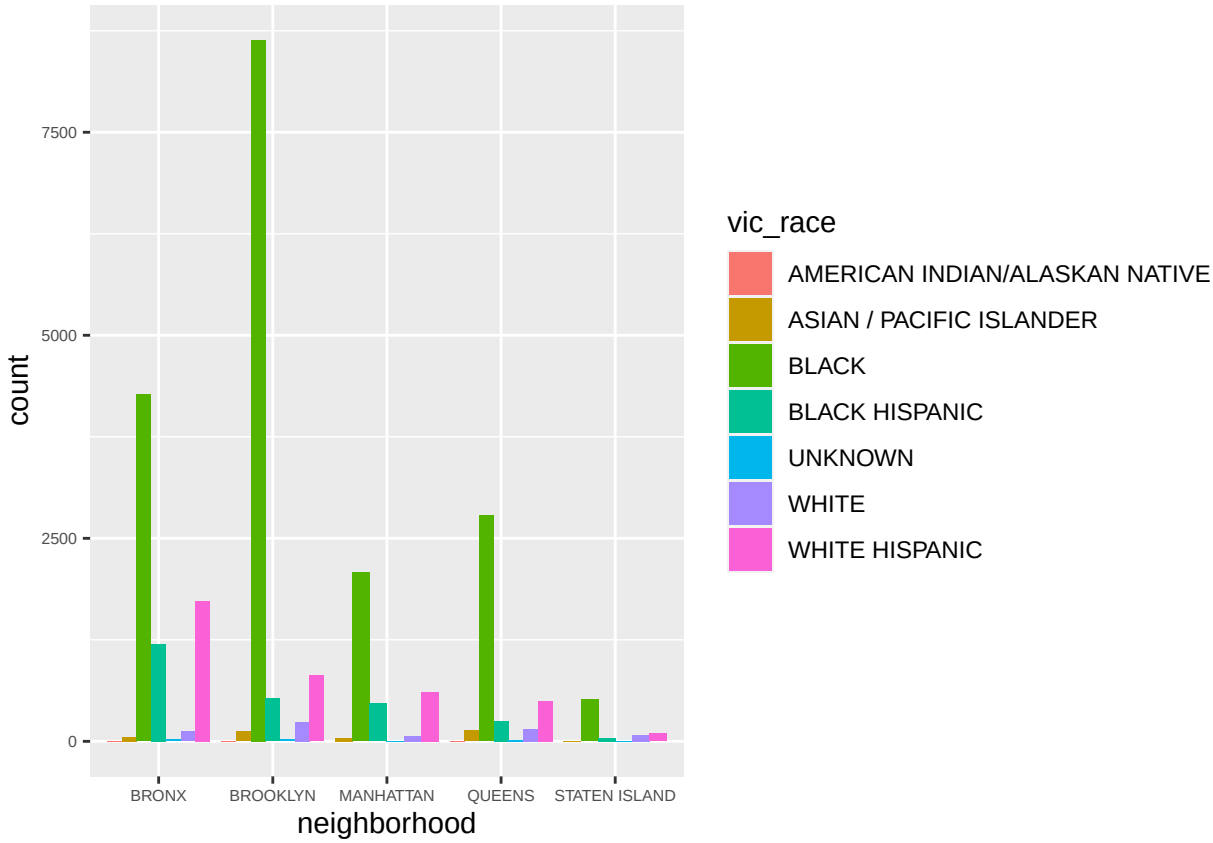


Shootings by race in neighborhoods

- Some potential bias might be to assume certain races might have higher numbers due to the fact that there is a higher population of those races in the neighborhood.
- It might also be assumed that the reason for less complete data on the `perp_race` is because they may have escaped arrest.

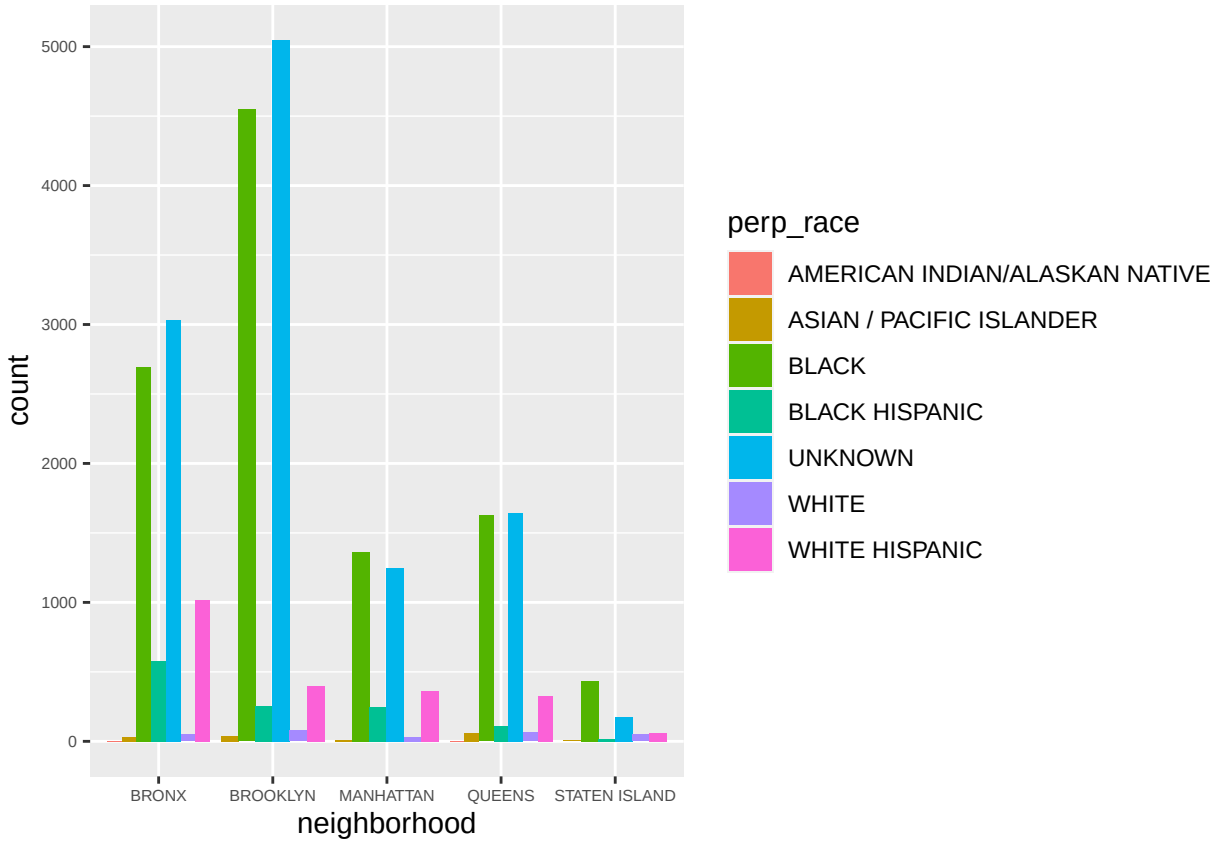
Shootings by victim's race in neighborhoods

```
plot <- ggplot(data = tidy_data) +
  geom_bar(mapping = aes(x = neighborhood, fill = vic_race), position = "dodge")

plot + theme(
  axis.text = element_text(size = rel(0.5))
)
```



Shootings by perpetrator race in neighborhoods

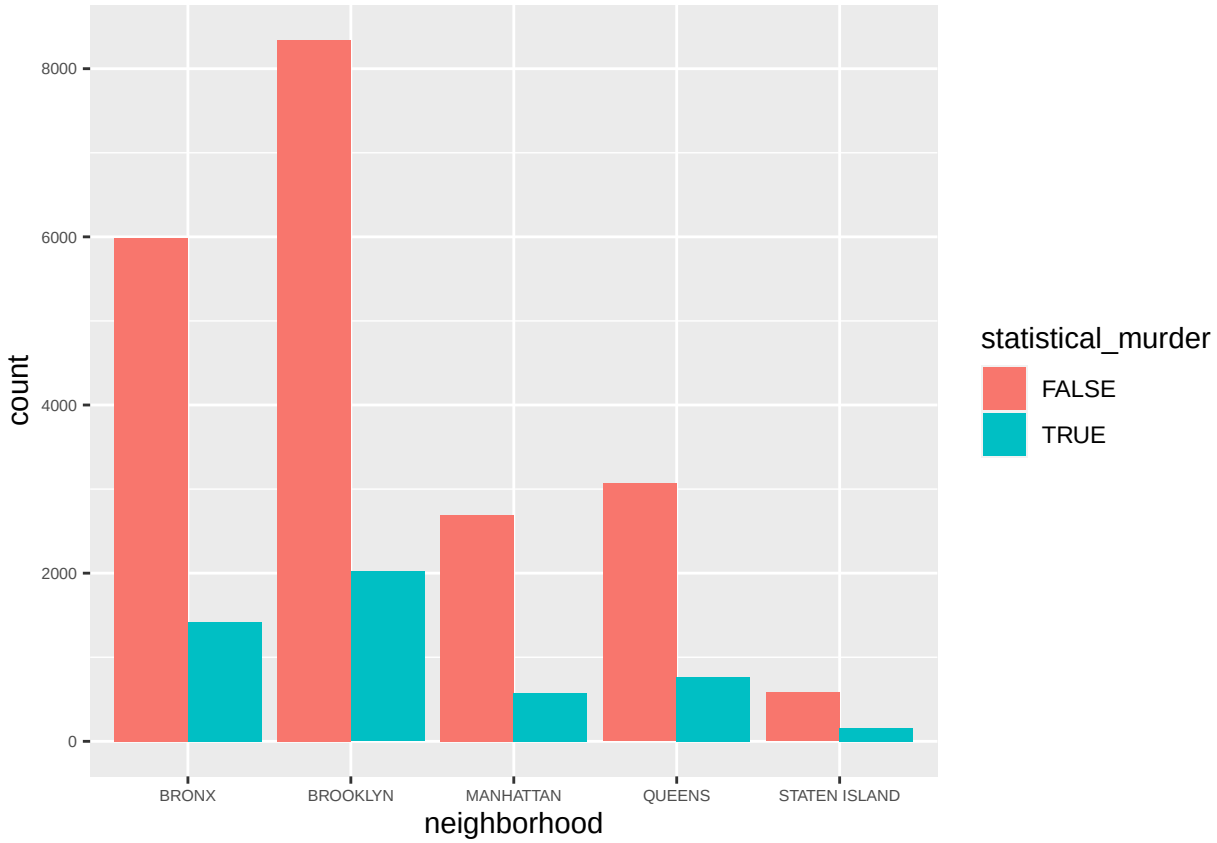
```
plot <- ggplot(data = tidy_data) +
  geom_bar(mapping = aes(x = neighborhood, fill = perp_race), position = "dodge")

plot + theme(
  axis.text = element_text(size = rel(0.5))
)
```



Compare shootings to deaths per neighborhood

```
shootings_to_deaths <- tidy_data %>%
  group_by(neighborhood, month) %>%
  mutate(shootings = n(), deaths = sum(statistical_murder)) %>%
  select(neighborhood, shootings, statistical_murder, deaths, month, full_date) %>%
  ungroup() %>%
  summarize(neighborhood, month, shootings, statistical_murder, deaths, full_date)
```

```
plot <- ggplot(data = shootings_to_deaths) +
  geom_bar(mapping = aes(x = neighborhood, fill = statistical_murder), position = "dodge")

plot + theme(
  axis.text = element_text(size = rel(0.5))
)
```



Model shootings to deaths

```
data <- tidy_data %>%
  group_by(neighborhood, month) %>%
  mutate(neighborhood, month, shootings = n(), deaths = sum(statistical_murder)) %>%
  ungroup()

mod = lm(deaths ~ shootings, data = data)
summary(mod)
```

```
##
## Call:
## lm(formula = deaths ~ shootings, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.730 -10.336   1.184   9.064  22.031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.759510   0.178113   21.11   <2e-16 ***
## shootings   0.185513   0.000241  769.65   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.85 on 25594 degrees of freedom
## Multiple R-squared:  0.9586, Adjusted R-squared:  0.9586
## F-statistic: 5.924e+05 on 1 and 25594 DF,  p-value: < 2.2e-16
```

```
data %>%
  mutate(pred = predict(mod)) %>%
  ggplot() +
  geom_point(aes(x = shootings, y = deaths), color = "blue") +
  geom_point(aes(x = shootings, y = pred), color = "red")
```