NYPD Shooting Incident Data

Almohanned Harfoush

2023-05-03

NYPD Shooting Incident

This document is describing the work on NYPD shooting incident project. Data used in this project is imported from https://catalog.data.gov/dataset and we used the dataset titled NYPD Shooting Incident Data (Historic).

Data is list of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. you can find more details from https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

Setup (Project Step 1)

- Importing libraries.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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
```

```
library(lubridate)
```

- Loading Data.

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

```
## Rows: 27312 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.
```

Tidy and Transform Data (Project Step 2)

- Verify imported data variables and data types.

```
spec(nypd_shooting)
```

```
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   LOC_OF_OCCUR_DESC = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOC_CLASSFCTN_DESC = col_character(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_double(),
##   Y_COORD_CD = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

- Change OCCUR_DATE variable to date type.

```
nypd_shooting <- nypd_shooting %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

- Select only columns that we will use.

```
nypd_shooting_final <- nypd_shooting[c("OCCUR_DATE", "BORO", "VIC_RACE", "STATISTICAL_MURDER_FLAG")]
```

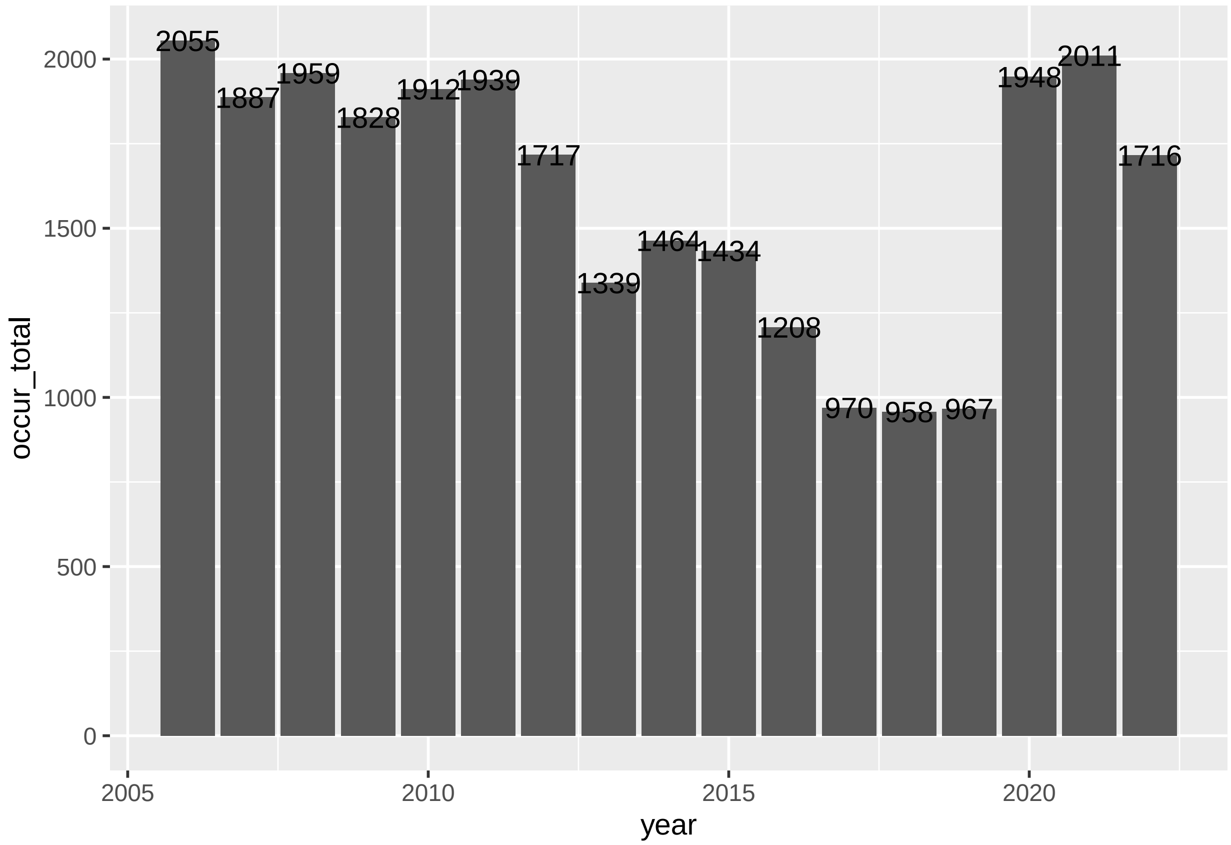
Add Visualizations and Analysis (Project Step 3)

- Compare number of incidents per year.

```
nypd_shooting_year <- nypd_shooting_final %>%
  group_by(VIC_RACE, BORO, OCCUR_DATE, STATISTICAL_MURDER_FLAG) %>%
  summarise(Frequency=n()) %>%
  group_by(year = lubridate::floor_date(OCCUR_DATE, 'year')) %>%
  summarise(occur_total = sum(Frequency)) %>%
  select(year, occur_total) %>% ungroup()
```

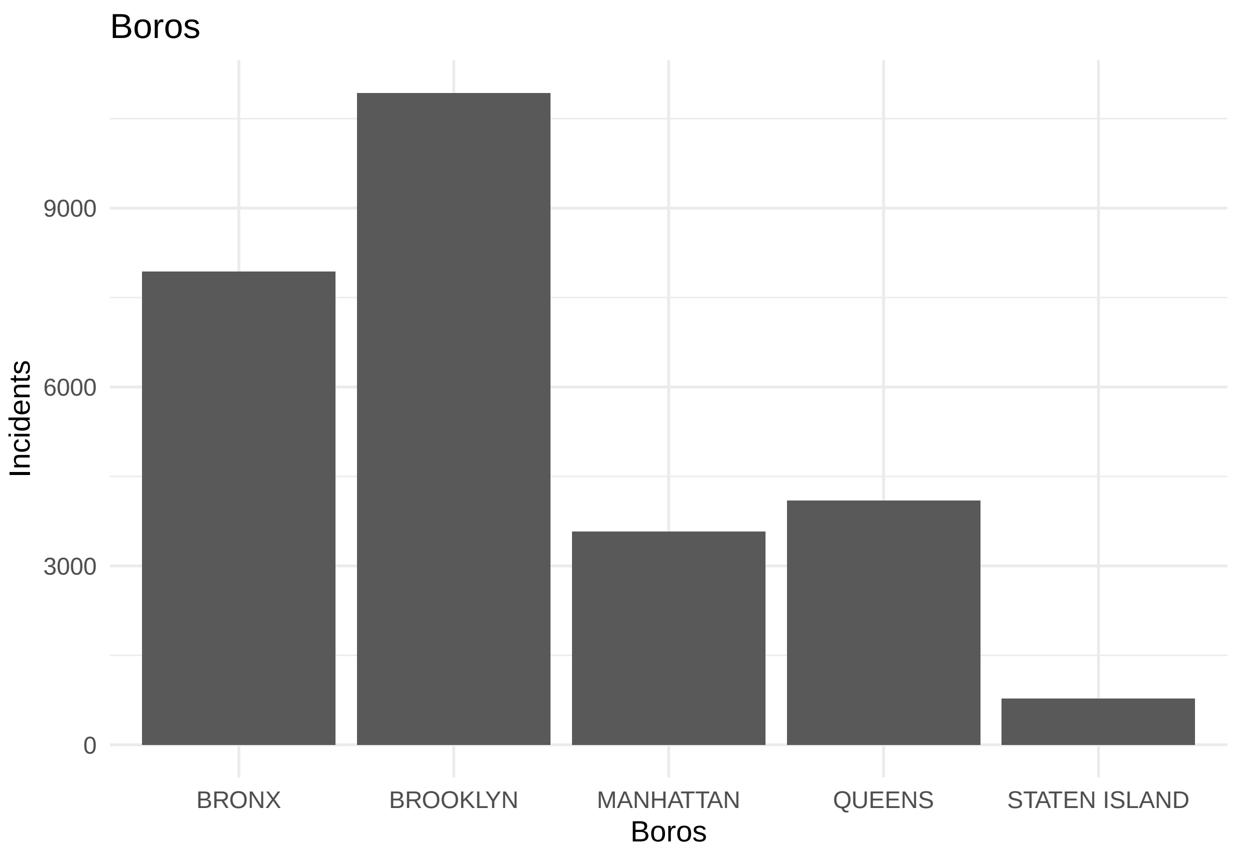
```
## ‘summarise()‘ has grouped output by ’VIC_RACE’, ’BORO’, ’OCCUR_DATE’. You can
## override using the ‘.groups‘ argument.
```

```
ggplot(data = nypd_shooting_year, mapping = aes(x=year, y=occur_total)) +
  geom_bar(stat='identity') +  geom_text(aes(label=occur_total))
```



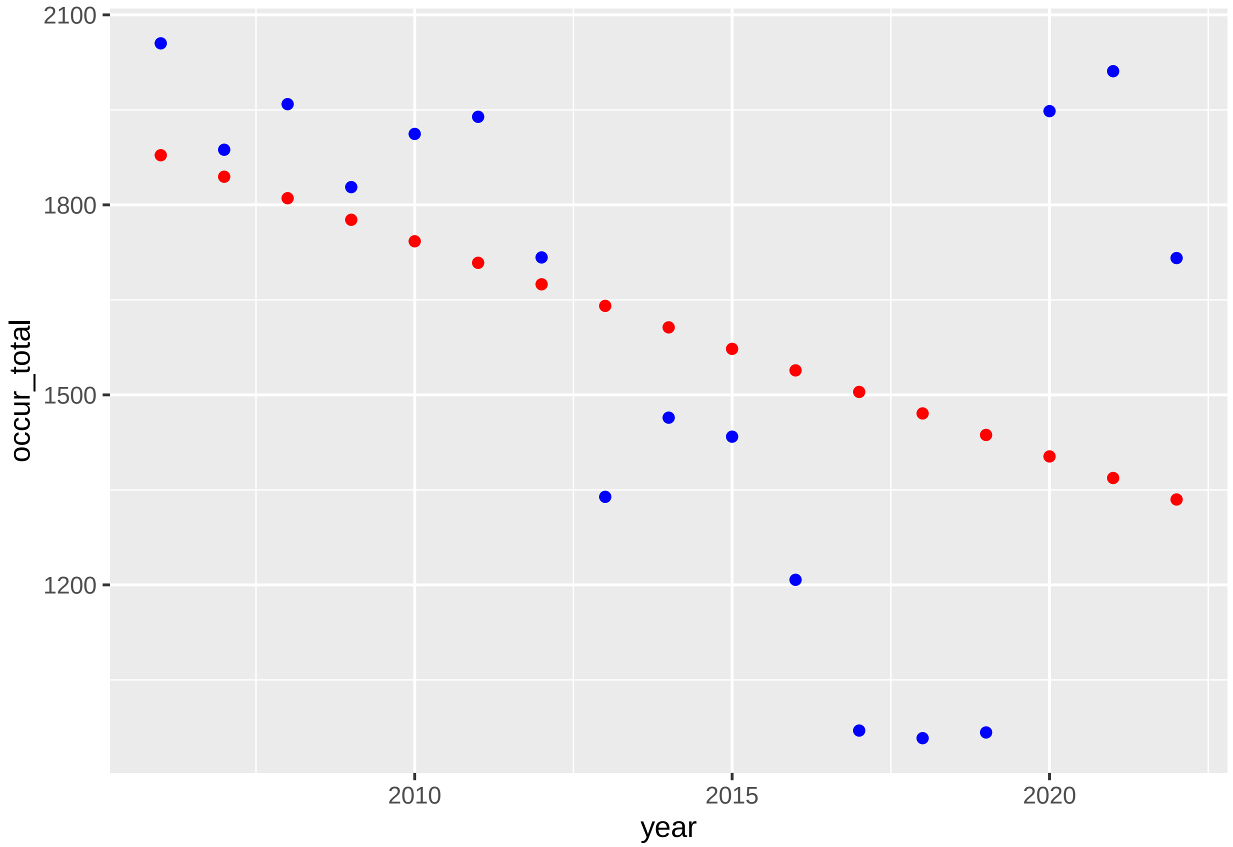
- Compare number of incidents per Boro

```
ggplot(nypd_shooting_final, aes(x = BORO)) +
  geom_bar() +
  labs(title = "Boros",
       x = "Boros",
       y = "Incidents") +
  theme_minimal()
```



- Try to predict the increase in shooting incidents over time. (creating a model)

```
mod <- lm(occur_total ~ year, data=nypd_shooting_year)
nypd_shooting_year <- nypd_shooting_year %>% mutate(pred = predict(mod))
nypd_shooting_year %>% ggplot() + geom_point(aes(x= year, y= occur_total), color = "blue") + geom_point
```



You can simply notice from this plot. that in recent years 2020, 2021, 2022. has clearly exceeded the expected numbers of incidents. which might be interesting to investigate.

Add Bias Identification (Project Step 4)

the source of bias in this analysis might come clearly from ignoring boro population and race and gender distribution. we have used only the number of incidents, but we did not make the required connection between boro population and the number of incidents. that will make the assumption from the second graph 'number of incidents per Boro' that Brooklyn is the most dangerous or violent, might not be fair.