NYPD Report

Vincent Zagala

2024-08-18

Introduction

Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity.

Including the libraries as well as importing the data from the website

```r
library(dplyr)
library(hms)
library(lubridate)

# Taking the Data from the Site and saving each column saved into a variable
nypd_cases <- read.csv(url("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNL
```

Summarized data from the data set before any changes is made

```r
summary(nypd_cases)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class :character   Class :character
##  Median : 92711254   Mode  :character   Mode  :character   Mode  :character
##  Mean   :127405824
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA’s   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Length:28562            Length:28562
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

Renaming Columns for Easier Reference

```
nypd_cases_1 <- data.frame(
  INCIDENT_KEY = Incident_Key,
  OCCUR_DATE = Occur_Date,
  OCCUR_TIME = Occur_Time,
  BORO = Location_Of_Incident,
  PRECINCT = Precinct,
  JURISDICTION_CODE = Jurisdiction_Code,
  PERP_AGE_GROUP = Suspect_Age_Group,
  PERP_SEX = Suspect_Sex,
  PERP_RACE = Suspect_Race,
  VIC_AGE_GROUP = Victim_Age_Group,
  VIC_SEX = Victim_Sex,
  VIC_RACE = Victim_Race,
  X_COORD_CD = X_Coord,
  Y_COORD_CD = Y_Coord,
  Latitude = Lat,
  Longitude = Long
)
```

```r
nypd_cases_1 <- nypd_cases %>%
    select(
        INCIDENT_KEY,
        OCCUR_DATE,
        OCCUR_TIME,
        BORO,
        PRECINCT,
        JURISDICTION_CODE,
        PERP_AGE_GROUP,
        PERP_SEX,
        PERP_RACE,
        VIC_AGE_GROUP,
        VIC_SEX,
        VIC_RACE,
        X_COORD_CD,
        Y_COORD_CD,
        Latitude,
        Longitude
    )
```

I edited columns that needed to be converted into the appropriate data type such as Occur_Date converted into Date dataype and changed the murder flag from just characters into a boolean value based on true or false

```r
nypd_cases_1 <- nypd_cases %>%
    rename(
        Incident_Key = INCIDENT_KEY,
        Occur_Date = OCCUR_DATE,
        Occur_Time = OCCUR_TIME,
        Location_Of_Incident = BORO,
        Precinct = PRECINCT,
        Jurisdiction_Code = JURISDICTION_CODE,
        Suspect_Age_Group = PERP_AGE_GROUP,
        Suspect_Sex = PERP_SEX,
        Suspect_Race = PERP_RACE,
        Victim_Age_Group = VIC_AGE_GROUP,
        Victim_Sex = VIC_SEX,
        Victim_Race = VIC_RACE,
        X_Coord = X_COORD_CD,
        Y_Coord = Y_COORD_CD,
        Lat = Latitude,
        Long = Longitude
    ) %>%
    mutate(
        Occur_Date = as.Date(Occur_Date, format = "%m/%d/%Y"),
        Occur_Time = hms::as_hms(Occur_Time),
        STATISTICAL_MURDER_FLAG = STATISTICAL_MURDER_FLAG == "Y"
    )
```

Bias

Possible biases in the data can arise from various sources. For example, if the data analyst is familiar with the area or has a background in criminal justice, their knowledge of criminal tendencies could introduce bias.

Reporting Bias: If crimes are not reported, they will not be included in the dataset. Conversely, if crimes are reported excessively, this can lead to an inaccurate representation of the number of cases for a given year or location.

Jurisdictional Bias: In the NYPD dataset, if multiple precincts report the same crime, it can result in duplicate entries. Alternatively, if two precincts are involved, each might assume the other will make the report, leading to no report being filed. This can result in underreporting of cases.

Linear Model Analysis

```r
nypd_cases_1 <- nypd_cases_1 %>%
mutate(Occur_Date = as.Date(Occur_Date, format = "%m/%d/%Y")) %>%
mutate(YEAR = format(Occur_Date, "%Y"))

cases_per_year <- nypd_cases_1 %>%
  group_by(YEAR) %>%
  summarise(CASE_COUNT = n())


lm_model <- lm(CASE_COUNT ~ as.numeric(YEAR), data = cases_per_year)


summary(lm_model)
```

The model I'm looking at is the Linear model in which I'm trying to find the number of cases per year

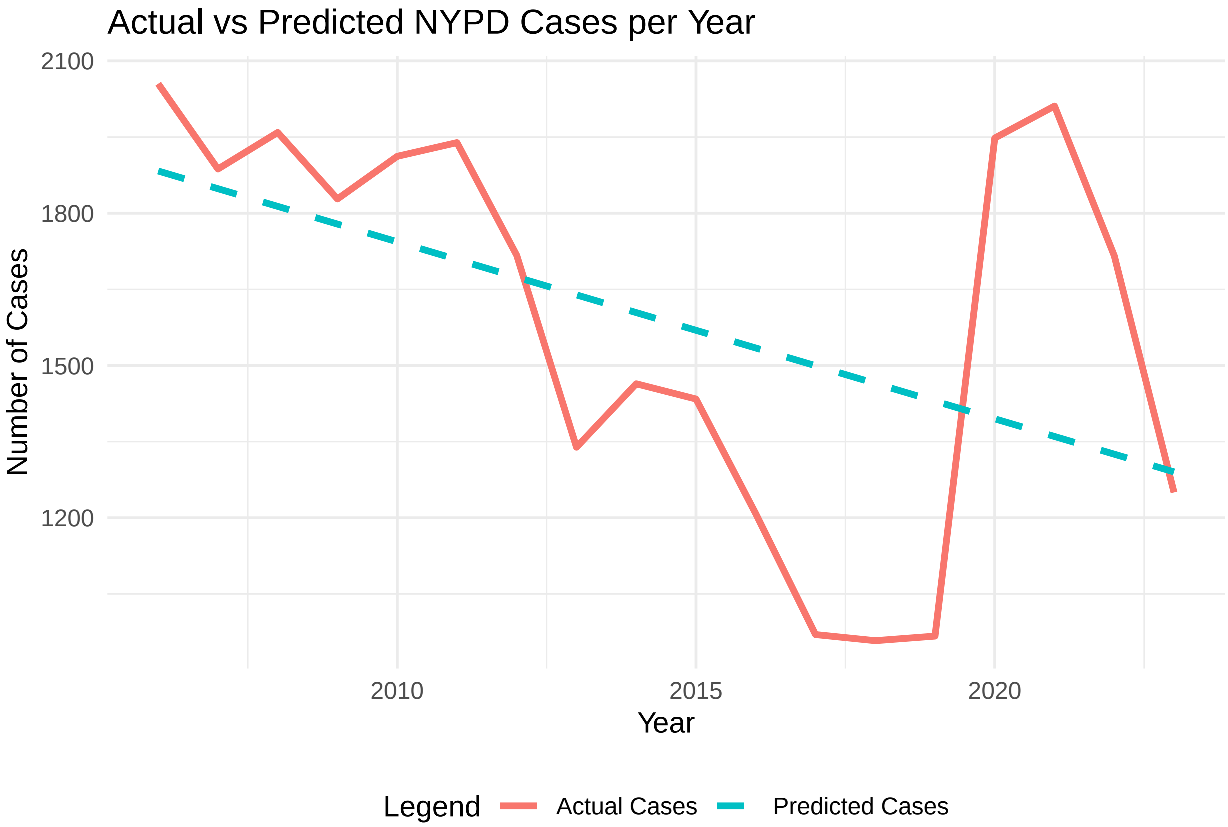
```
##
## Call:
## lm(formula = CASE_COUNT ~ as.numeric(YEAR), data = cases_per_year)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -529.6 -260.1   40.9  171.0  650.8
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      71821.94   32419.87   2.215   0.0416 *
## as.numeric(YEAR)   -34.86      16.09  -2.166   0.0457 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 354.2 on 16 degrees of freedom
## Multiple R-squared:  0.2268, Adjusted R-squared:  0.1785
## F-statistic: 4.693 on 1 and 16 DF,  p-value: 0.04572
```

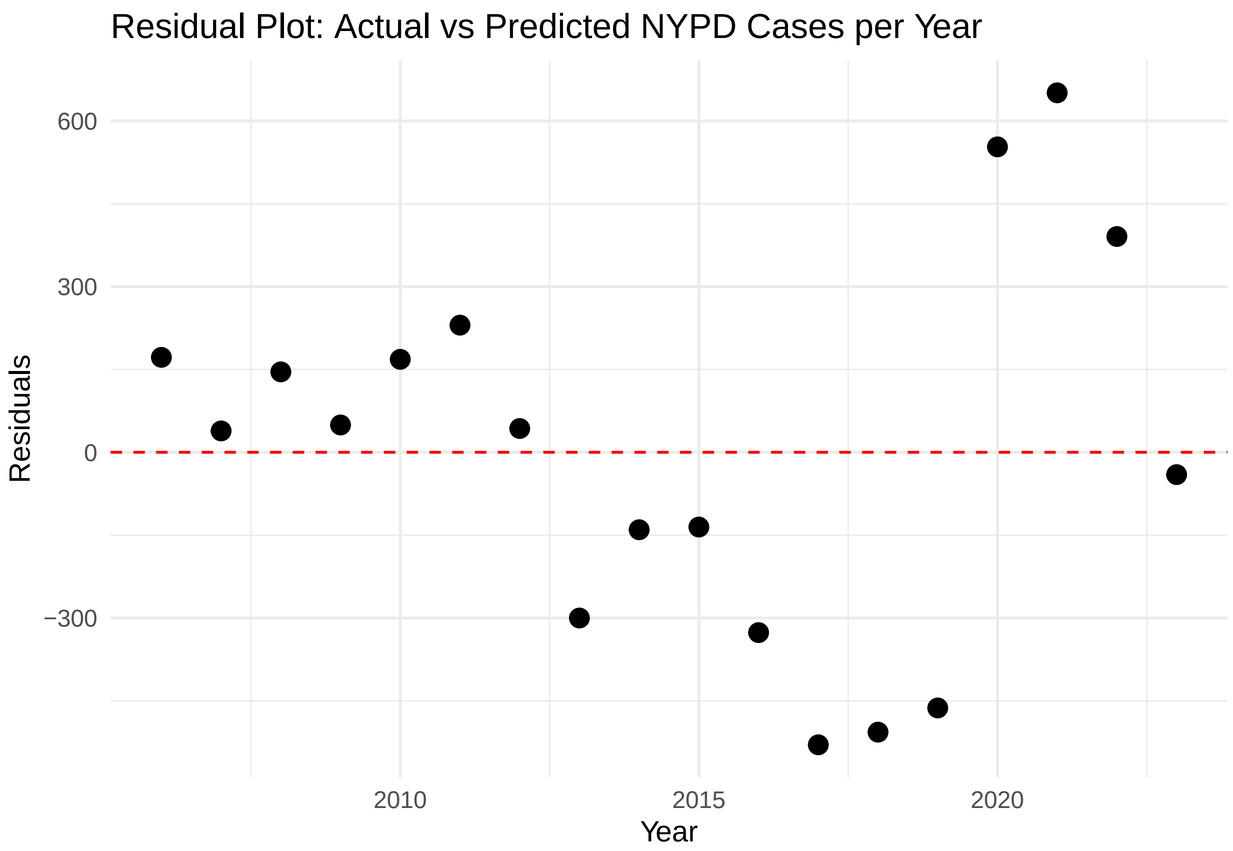
The linear model shows that the cases per year has been decreasing over time and that since theres a big decrease throughout the recent years. other factors may affect it that was not captured in the years.

```r
cases_per_year <- cases_per_year %>%
  mutate(PREDICTED_CASE_COUNT = predict(lm_model, newdata = cases_per_year))

ggplot(cases_per_year, aes(x = as.numeric(YEAR))) +
  geom_line(aes(y = CASE_COUNT, color = "Actual Cases"), linewidth = 1.2) +
  geom_line(aes(y = PREDICTED_CASE_COUNT, color = "Predicted Cases"),
            linetype = "dashed", linewidth = 1.2) +
  labs(
    title = "Actual vs Predicted NYPD Cases per Year",
    x = "Year",
    y = "Number of Cases",
    color = "Legend"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

I then made a linear graph regarding the actual case numbers per year with the predicted number of cases.



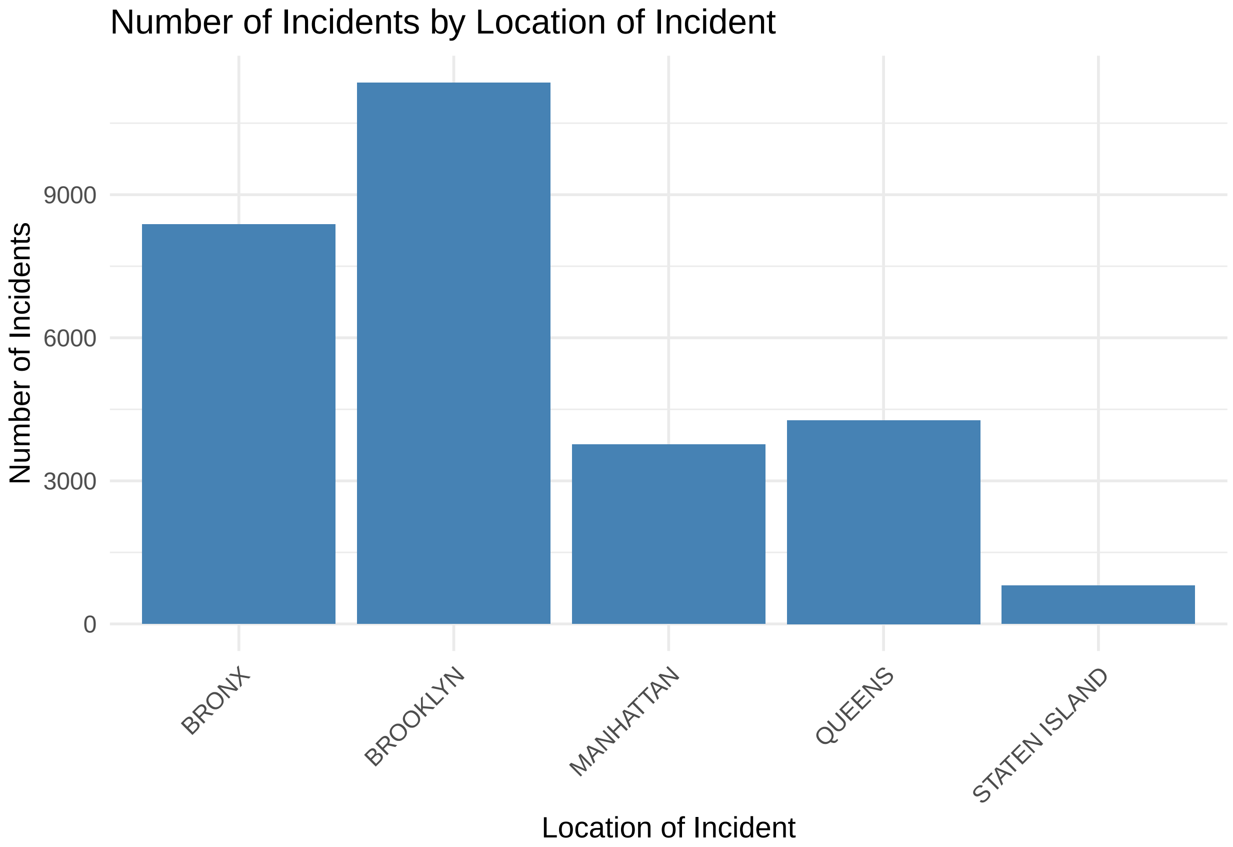
```r
# Calculate the residuals from the linear model
cases_per_year <- cases_per_year %>%
  mutate(RESIDUALS = residuals(lm_model))

ggplot(cases_per_year, aes(x = as.numeric(YEAR), y = RESIDUALS)) +
  geom_point(size = 3) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Residual Plot: Actual vs Predicted NYPD Cases per Year",
    x = "Year",
    y = "Residuals"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



The residual plot provides insights into the linear model. If the residuals are randomly scattered around zero, it indicates that the model fits the data well. However, if there are deviations or patterns in the residuals, this may suggest that other factors are affecting the data that the model does not account for.

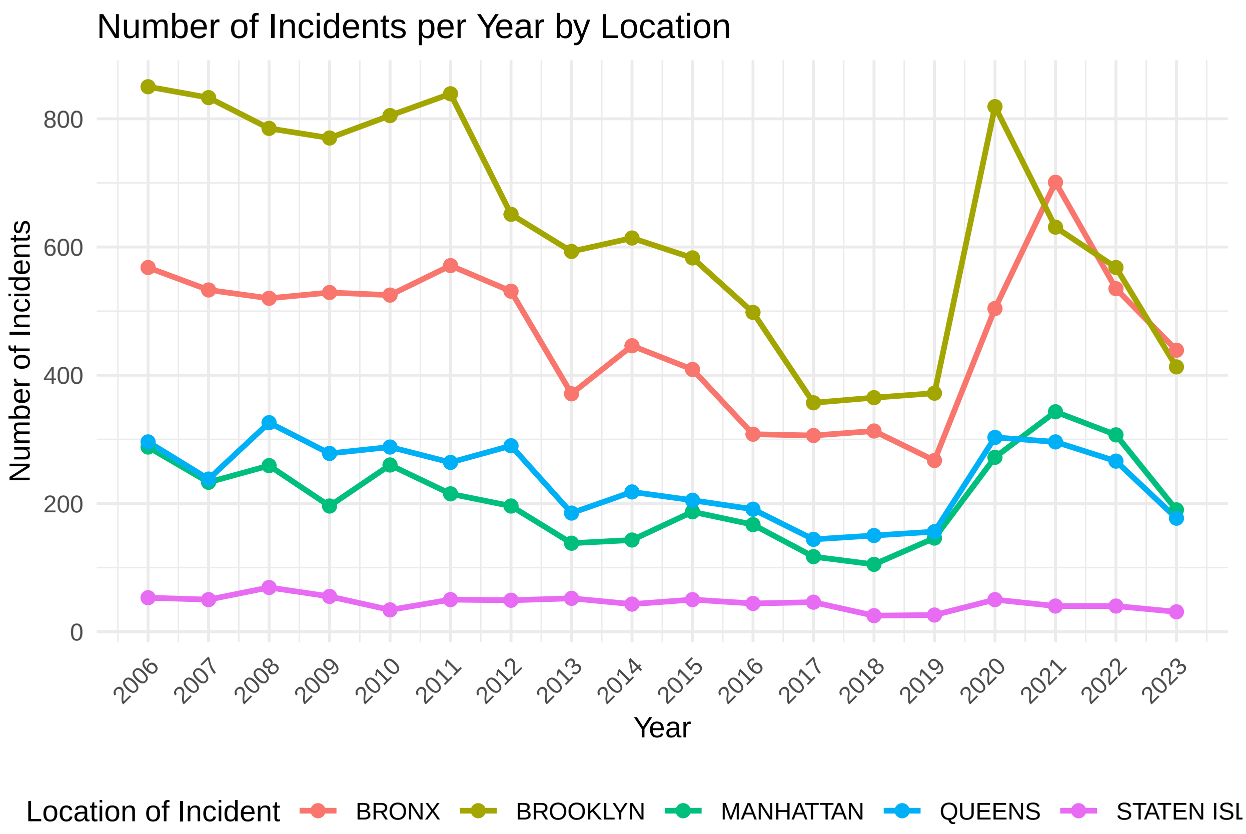
Visualization i was interested in.

```
incidents_by_location <- nypd_cases_1 %>%
    count(Location_Of_Incident)

# Create the bar plot
ggplot(incidents_by_location, aes(x = Location_Of_Incident, y = n)) +
    geom_bar(stat = "identity", fill = "steelblue") +
    labs(
        title = "Number of Incidents by Location of Incident",
        x = "Location of Incident",
        y = "Number of Incidents"
    ) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Visualization

```
nypd_cases_1 <- nypd_cases_1 %>%
    mutate(Year = year(Occur_Date))

incidents_per_year_location <- nypd_cases_1 %>%
    group_by(Year, Location_Of_Incident) %>%
    summarize(Incident_Count = n(), .groups = 'drop')

ggplot(incidents_per_year_location, aes(x = Year, y = Incident_Count, color = Location_Of_Incident)) +
    geom_line(linewidth = 1) +
    geom_point(size = 2) +
    labs(
        title = "Number of Incidents per Year by Location",
        x = "Year",
        y = "Number of Incidents",
        color = "Location of Incident"
    ) +
    theme_minimal() +
    scale_x_continuous(breaks = seq(min(incidents_per_year_location$Year), max(incidents_per_year_locati
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    theme(legend.position = "bottom")
```



I wanted to see what the visualization would look like with the number of cases made based on location and in the years followed. since the last visualization was with all locations added.