# Final Project 1 - NYPD Shooting Incident Data

2022-09-26

Hi there!

This notebook contains an analysis of the NYPD data and the required steps for the Final Project 1 submission.

## The Data

The dataset 'NYPD Shooting Incident Data (Historic)' contains a list of every shooting incident in New York City from 2006 until the end of 2021.

For every incident there is a wealth of additional information available.

All data about the incidents is first reviewed before it is made public and added to the dataset.

## My Questions

Based on the data set I have the following questions I would like to answer.

1. What is the overall trend in shootings and murders?
2. Is there a specific pattern we can find and what does it tell us?

## Download Data

First we download the data and show a summary of the raw and unprocessed data.

```
# Import libraries
library(dplyr)
library(lubridate)
library(ggplot2)
library(readr)

# Data URL
nypddata_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

# Download Data
raw_data = read_csv(nypddata_url)

# Show Data Summary
summary(raw_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms         Class :character
##  Median : 86437258   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX             VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

## Tidy Data

As a next step we will perform the following basic data processing steps:

- Change date types when necessary
- Remove columns not needed
- Rename columns to display friendly names
- Remove any rows with missing data if needed.
- Show Summary

Since I don't posses the domain knowledge about the NYPD dataset I personally think I will not be able to make accurate decisions about the way to substitute the missing data with other values.

In my analysis I will not remove or modify any of the missing values as this will severely impact the information shown.

```
# Tidy Data
nypd_data <- raw_data %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
    rename("IncidentID" = INCIDENT_KEY,
           "IncidentDate" = OCCUR_DATE,
           "IncidentTime" = OCCUR_TIME,
           "LocationDescription" = LOCATION_DESC,
           "IncidentWasMurder" = STATISTICAL_MURDER_FLAG
           ) %>%
    select(-c(BORO, PRECINCT, JURISDICTION_CODE, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, Latitude, Longitu

# Show Summary
summary(nypd_data)
```

```
##    IncidentID         IncidentDate        IncidentTime      LocationDescription
##  Min.   :  9953245   Min.   :2006-01-01   Length:25596      Length:25596
##  1st Qu.: 61593633   1st Qu.:2009-05-10   Class1:hms        Class :character
##  Median : 86437258   Median :2012-08-26   Class2:difftime   Mode  :character
##  Mean   :112382648   Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:166660833   3rd Qu.:2017-07-01
##  Max.   :238490103   Max.   :2021-12-31
##  IncidentWasMurder
##  Mode :logical
##  FALSE:20668
##  TRUE :4928
##
##
##
```

## Visualizations and Analysis

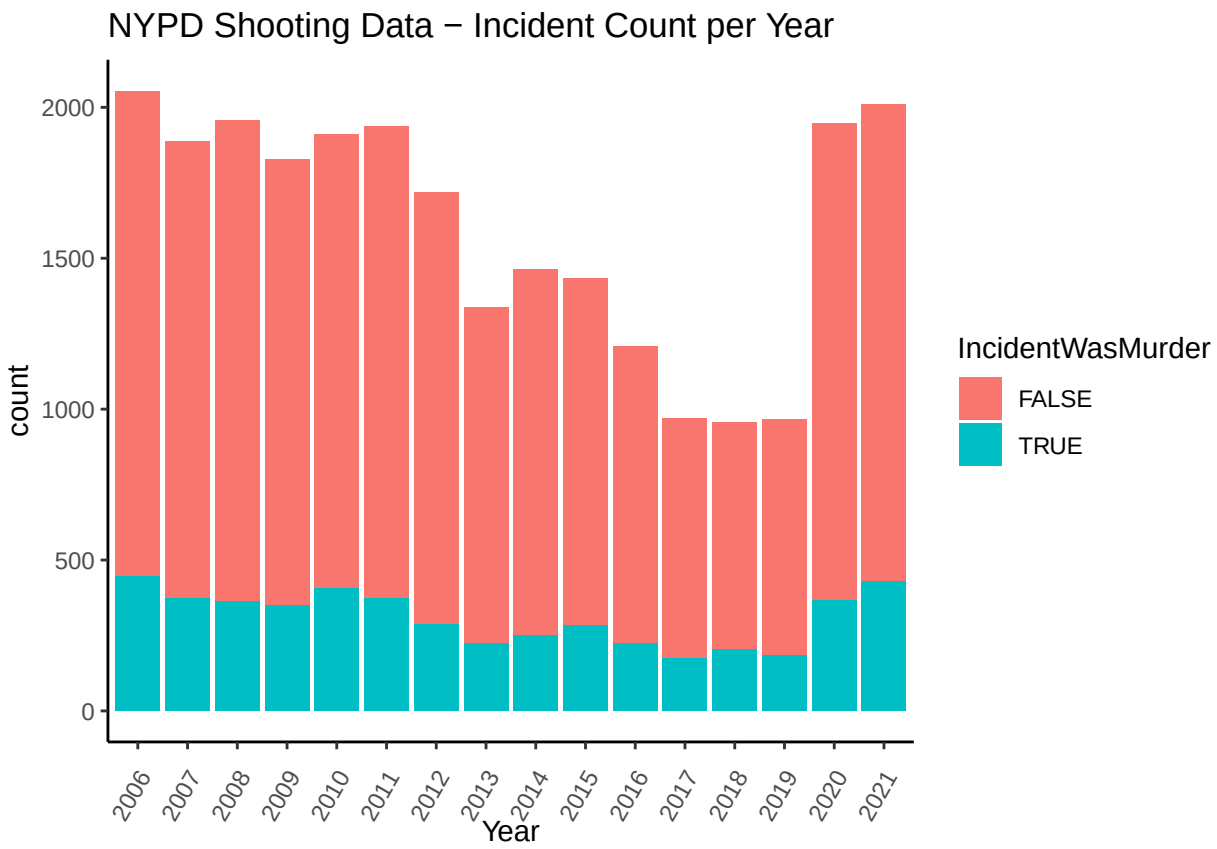
Let's create a first simple but really interresting visualization and analysis.

What I would like to know is how much incidents there are each year, how much of those are murders and what is the general trend for these 2?

Lets create the visualization first.

```
# Process Data
plot1_data <- nypd_data %>%
    mutate(month = format(IncidentDate, "%m"), Year = format(IncidentDate, "%Y")) %>%
    group_by(Year)

# Plot Data
g1 <- ggplot(plot1_data, aes(x = Year, fill = IncidentWasMurder))
g1 +  geom_bar() +
      theme_classic() +
      theme(axis.text.x = element_text(angle = 60, vjust = 0.5, hjust = 0.75)) +
      labs(title = "NYPD Shooting Data - Incident Count per Year")
```




The plot shows some interesting information. We see that every year there are many incidents of which only a small part are actual murders.

We can see that from 2006 until 2017 there is in general a downward trend in incidents and incidents that are considered murders.

From 2017 until 2019 this downward trend seems to have stopped and remained stable. Without additional data and/or research we can only speculate why the downward trend stopped and stabilized.

But the most interesting part happens in 2020. There is suddenly a large spike in the number of incidents and murders. It almost doubles compared to 2019. The same-thing happens in 2021 a further increase of incidents and murders compared to 2020.

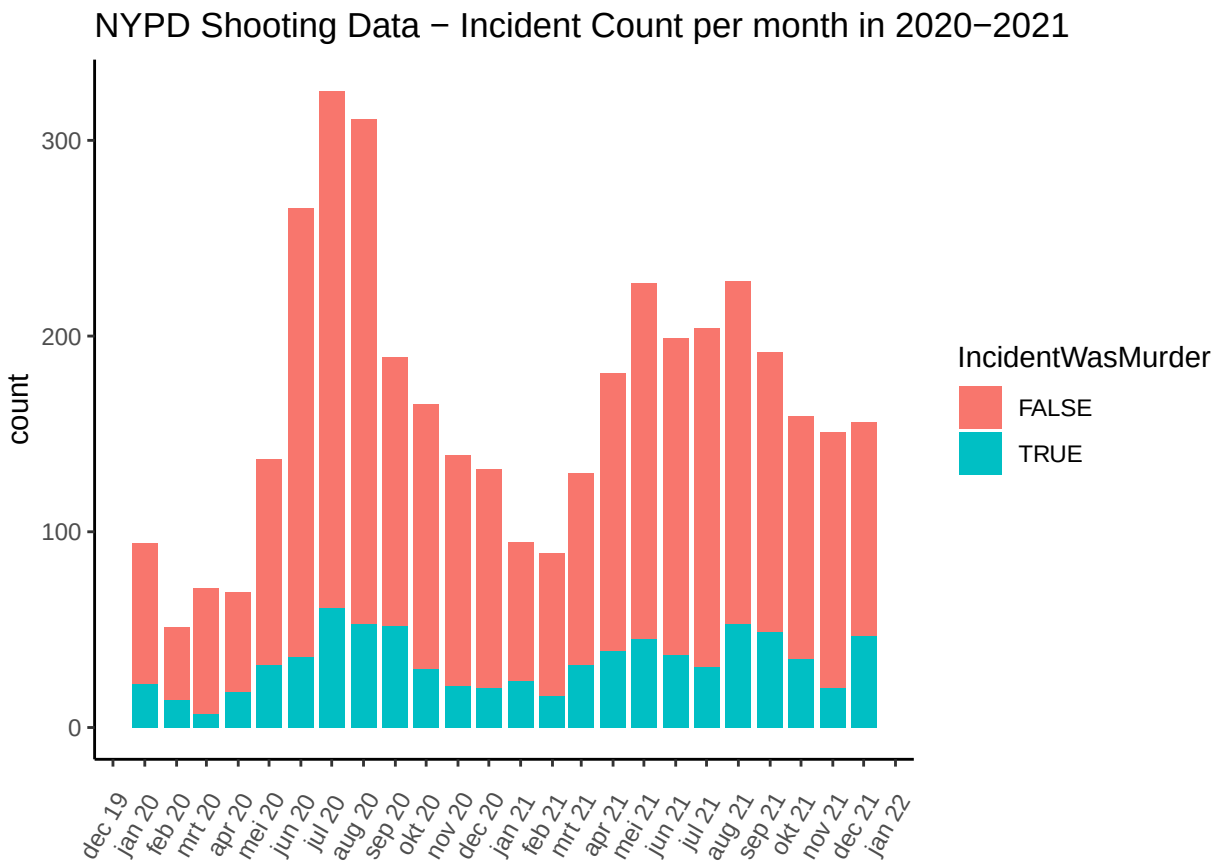
While I can't prove this but I think this is very likely due to the outbreak of COVID in early 2020. Many people had to stay at home..this can be a cause for serious stress I presume. Especially when the living area is smaller and you have to share it with people. That might cause some people to reach a certain tipping point. Another reason could be that a lot of people lost their jobs..this could also cause some serious levels of stress.

Lets take a more detailed look at the numbers per month for only 2020 and 2021.

```
# Process Data
plot2_data <- nypd_data %>%
    mutate(Year = format(IncidentDate, "%Y")) %>%
    group_by(month = floor_date(IncidentDate, unit = "month")) %>%
    filter(Year > 2019)

# Plot Data
g2 <- ggplot(plot2_data, aes(x = month, fill = IncidentWasMurder))
```

```
g2 +  geom_bar() +
      scale_x_date(NULL, date_labels = "%b %y", breaks = "month") +
      theme_classic() +
      theme(axis.text.x = element_text(angle = 60, vjust = 0.5, hjust = 0.75)) +
      labs(title = "NYPD Shooting Data - Incident Count per month in 2020-2021")
```

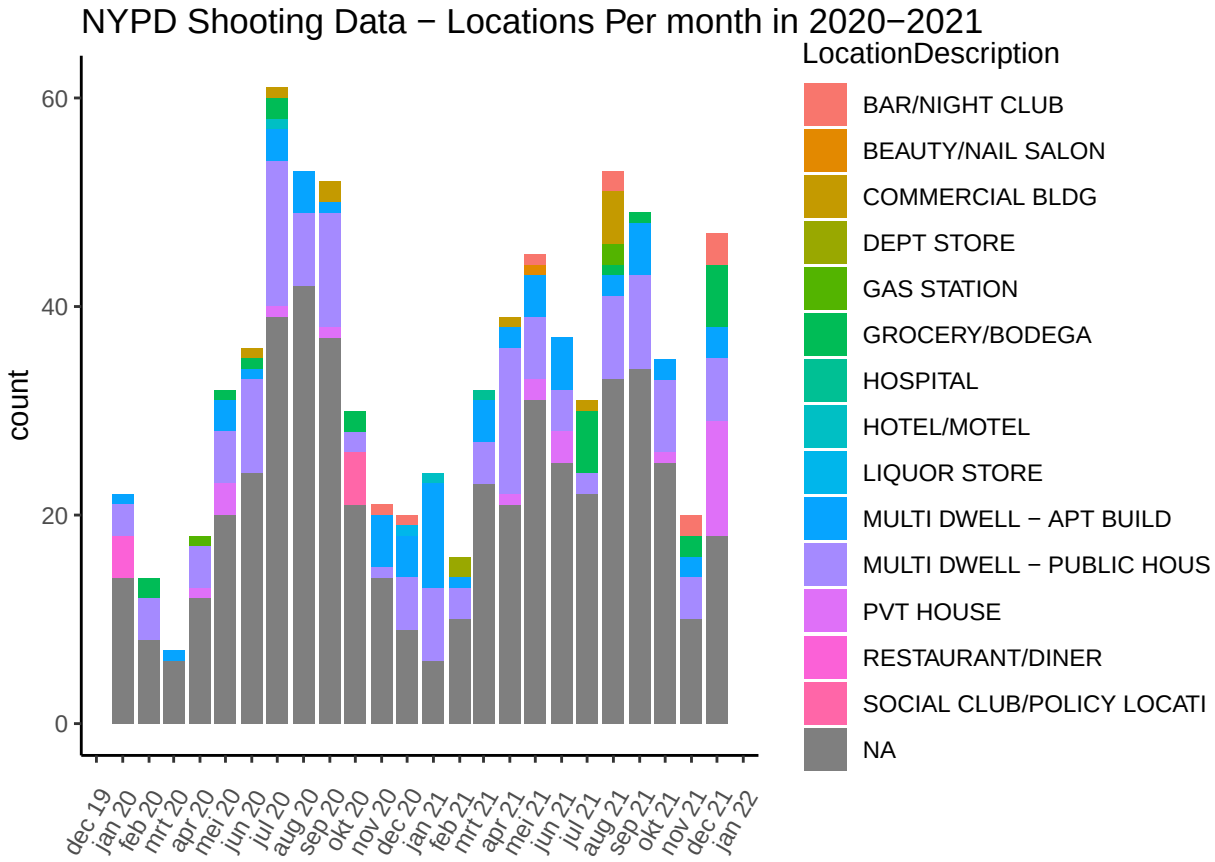



This provides an interesting view. Already in May there is a sharp increase in incidents and murders visible. From September on wards there is a sharp drop but early in the year 2021 the numbers again start to rise sharply.

As a next analysis and visualization it would be very interesting to see at what type of location the largest increase in murders would have occurred.

```
# Process Data
plot3_data <- nypd_data %>%
    mutate(Year = format(IncidentDate, "%Y")) %>%
    group_by(month = floor_date(IncidentDate, unit = "month")) %>%
    filter(Year > 2019 & IncidentWasMurder == TRUE)

# Plot Data
g3 <- ggplot(plot3_data, aes(x = month, fill = LocationDescription))
g3 +  geom_bar() +
      scale_x_date(NULL, date_labels = "%b %y", breaks = "month") +
      theme_classic() +
      theme(axis.text.x = element_text(angle = 60, vjust = 0.5, hjust = 0.75)) +
      labs(title = "NYPD Shooting Data - Locations Per month in 2020-2021")
```




The plot shows that there is a large increase in murders in the Location Description - Multi Dwelling Public Housing.

The largest increase is however in the missing data. Unfortunately we can't draw any conclusions based on missing values.

## Conclusion

In this project report we looked at the NYPD Shooting Incidents dataset. We created multiple analysis and plots and looked particularly at the number of incidents per year and per month. We also analysed the years 2020 and 2021 in detail.

I didn't discover any bias related to the specific columns of the data I used.

As far as personal bias is concerned... this might be the case. Being from Europe we usually get to see only the 'bad moments' of american police when someone is shot during his or her arrest. This might have an effect on objectivity.

Write the conclusion to your project report and include any possible sources of bias. Be sure to identify what your personal bias might be and how you have mitigated that.