

NYPD INCIDENT REPORT

Hamid Aboutaher

-|-|-|-|-|-|-|-|Step 1|-|-|-|-|-|-|-|-|

Introduction

This project is about details of every shooting incident in NYC from 2006 to the end of the last calendar year, precisely extracted quarterly and reviewed by the NYPD's Office of Management Analysis and Planning. It includes data on the incidents, locations, timings, as well as suspect and victim demographics, serving as a resource for the public to examine shooting - criminal activities.

Installing libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.4
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2    3.4.4      v tibble    3.2.1
## v lubridate  1.9.3      v tidyr     1.3.0
## v purrr      1.0.2
## -- 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(readr)
library(lubridate)
```

There are 21 columns in dataset. First we check the datatype then we remove unwanted columns.

```
# list datatype of each column
str(df)
```

```
## spc_tbl_ [27,312 x 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ INCIDENT_KEY      : num [1:27312] 2.29e+08 1.37e+08 1.48e+08 1.47e+08 5.89e+07 ...
## $ OCCUR_DATE         : chr [1:27312] "05/27/2021" "06/27/2014" "11/21/2015" "10/09/2015" ...
## $ OCCUR_TIME         : 'hms' num [1:27312] 21:30:00 17:40:00 03:56:00 18:30:00 ...
## $ ..- attr(*, "units")= chr "secs"
## $ BORO               : chr [1:27312] "QUEENS" "BRONX" "QUEENS" "BRONX" ...
## $ LOC_OF_OCCUR_DESC   : chr [1:27312] NA NA NA NA ...
## $ PRECINCT           : num [1:27312] 105 40 108 44 47 81 114 81 105 101 ...
## $ JURISDICTION_CODE   : num [1:27312] 0 0 0 0 0 0 0 0 0 0 ...
```

```
## $ LOC_CLASSFCTN_DESC      : chr [1:27312] NA NA NA NA ...
## $ LOCATION_DESC           : chr [1:27312] NA NA NA NA ...
## $ STATISTICAL_MURDER_FLAG: logi [1:27312] FALSE FALSE TRUE FALSE TRUE TRUE ...
## $ PERP_AGE_GROUP          : chr [1:27312] NA NA NA NA ...
## $ PERP_SEX                : chr [1:27312] NA NA NA NA ...
## $ PERP_RACE               : chr [1:27312] NA NA NA NA ...
## $ VIC_AGE_GROUP           : chr [1:27312] "18-24" "18-24" "25-44" "<18" ...
## $ VIC_SEX                 : chr [1:27312] "M" "M" "M" "M" ...
## $ VIC_RACE                : chr [1:27312] "BLACK" "BLACK" "WHITE" "WHITE HISPANIC" ...
## $ X_COORD_CD              : num [1:27312] 1058925 1005028 1007668 1006537 1024922 ...
## $ Y_COORD_CD              : num [1:27312] 180924 234516 209837 244511 262189 ...
## $ Latitude                : num [1:27312] 40.7 40.8 40.7 40.8 40.9 ...
## $ Longitude               : num [1:27312] -73.7 -73.9 -73.9 -73.9 -73.9 ...
## $ Lon_Lat                 : chr [1:27312] "POINT (-73.73083868899994 40.662964620000025)" "POINT (-73.73083868899994 40.662964620000025)" ...
## - attr(*, "spec")=
## .. 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()
## .. )
## - attr(*, "problems")=<externalptr>
```

Let's remove [X_COORD_CD,Y_COORD_CD, Latitude, Longitude, Lon_Lat] columns from the dataset.

```
# Remove specified columns
df <- df[, -which(names(df) %in% c("X_COORD_CD", "Y_COORD_CD", "Latitude", "Longitude", "Lon_Lat"))]
```

Removing columns with missing values

```
df <- df[,colSums(is.na(df))==0]
colnames(df)
```

```
## [1] "INCIDENT_KEY"      "OCCUR_DATE"
## [3] "OCCUR_TIME"        "BORO"
## [5] "PRECINCT"          "STATISTICAL_MURDER_FLAG"
## [7] "VIC_AGE_GROUP"      "VIC_SEX"
## [9] "VIC_RACE"
```

Converting data types

```
library(dplyr)
df <- df %>%
  mutate(INCIDENT_KEY = as.factor(INCIDENT_KEY),
         OCCUR_DATE = as.Date(OCCUR_DATE, format = "%Y-%m-%d"), # Specify format
         OCCUR_TIME = as.POSIXct(OCCUR_TIME, format = ""),
         BORO = as.factor(BORO),
         PRECINCT = as.factor(PRECINCT),
         STATISTICAL_MURDER_FLAG = as.logical(STATISTICAL_MURDER_FLAG),
         VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP),
         VIC_SEX = as.factor(VIC_SEX),
         VIC_RACE = as.factor(VIC_RACE))
```

Summary of data

```
summary(df)
```

```
##      INCIDENT_KEY      OCCUR_DATE      OCCUR_TIME
## 173354054: 18 Min. :NA      Min. :1970-01-01 00:00:00.0000
## 23749375 : 12 1st Qu.:NA      1st Qu.:1970-01-01 03:27:00.0000
## 24717013 : 12 Median :NA      Median :1970-01-01 15:11:00.0000
## 33478089 : 12 Mean :NaN      Mean :1970-01-01 12:41:31.7091
## 33706902 : 12 3rd Qu.:NA      3rd Qu.:1970-01-01 20:45:00.0000
## 35803777 : 12 Max. :NA      Max. :1970-01-01 23:59:00.0000
## (Other) :27234 NA's :27312
##      BORO      PRECINCT      STATISTICAL_MURDER_FLAG VIC_AGE_GROUP
## BRONX : 7937 75 : 1557 Mode :logical <18 : 2839
## BROOKLYN :10933 73 : 1452 FALSE:22046 1022 : 1
## MANHATTAN : 3572 67 : 1216 TRUE :5266 18-24 :10086
## QUEENS : 4094 44 : 1020 25-44 :12281
## STATEN ISLAND: 776 79 : 1012 45-64 : 1863
## 47 : 953 65+ : 181
## (Other):20102 UNKNOWN: 61
## VIC_SEX      VIC_RACE
```

```
## F: 2615 AMERICAN INDIAN/ALASKAN NATIVE: 10
## M:24686 ASIAN / PACIFIC ISLANDER : 404
## U: 11 BLACK :19439
## BLACK HISPANIC : 2646
## UNKNOWN : 66
## WHITE : 698
## WHITE HISPANIC : 4049
```

-|-|-|-|-|-|-|-|-|-|-| **Step 3** -|-|-|-|-|-|-|-|-|-|-|

```
library(dplyr)
library(ggplot2)

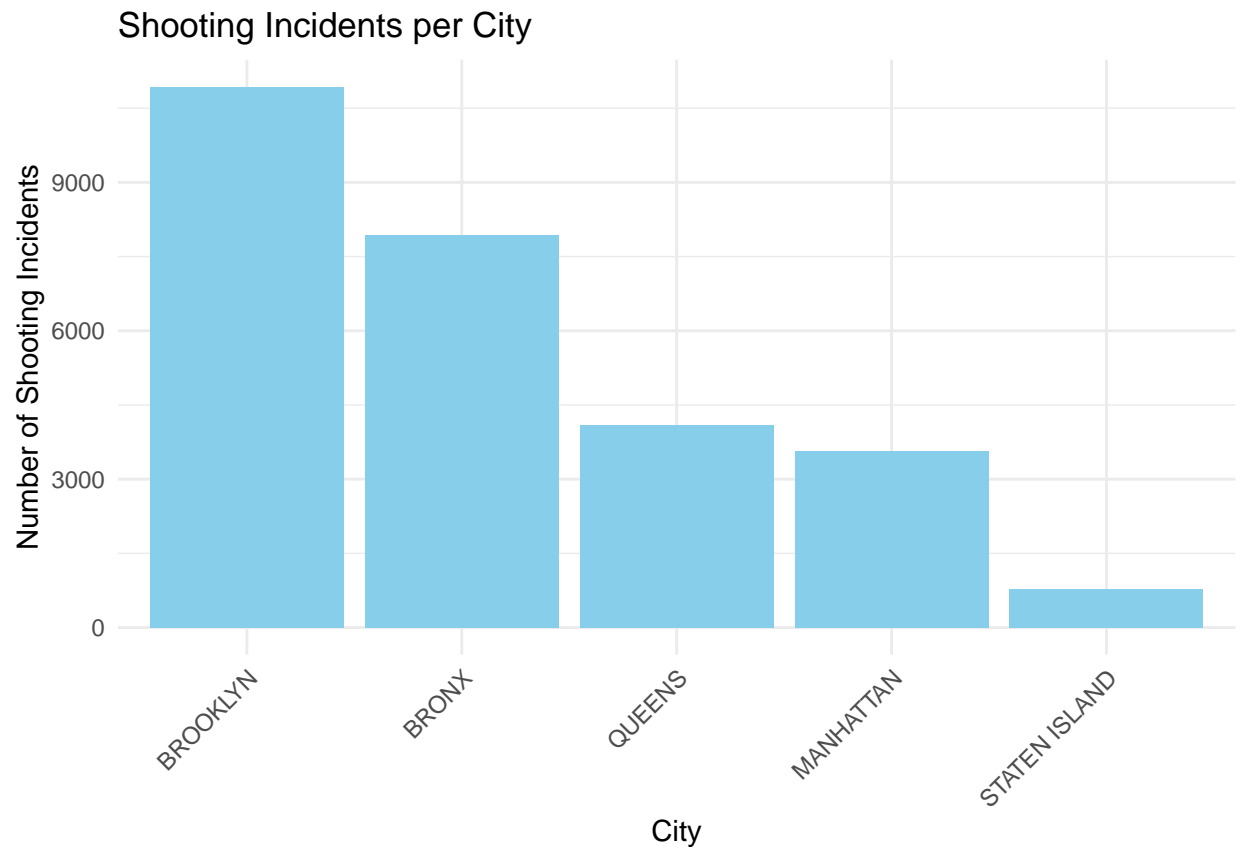
# Convert OCCUR_DATE to Date type
df$OCCUR_DATE <- as.Date(df$OCCUR_DATE, format = "%Y-%m-%d")

# Extract year from OCCUR_DATE
df$YEAR <- lubridate::year(df$OCCUR_DATE)

# Count shooting incidents per city
city_shootings <- df %>%
  group_by(BORO) %>%
  summarise(shootings = n())

# Count shooting incidents per year
yearly_shootings <- df %>%
  group_by(YEAR) %>%
  summarise(shootings = n())

# Visualization 1: Shooting incidents per city
ggplot(city_shootings, aes(x = reorder(BORO, -shootings), y = shootings)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Shooting Incidents per City",
       x = "City",
       y = "Number of Shooting Incidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
library(ggplot2)

# Count shooting incidents per victim age group
age_group_shootings <- df %>%
  group_by(VIC_AGE_GROUP) %>%
  summarise(shootings = n())

# Create bar plot for shooting incidents by victim age group
ggplot(age_group_shootings, aes(x = VIC_AGE_GROUP, y = shootings)) +
  geom_bar(stat = "identity", fill = "Green") +
  labs(title = "Shooting Incidents by Victim Age Group",
       x = "Victim Age Group",
       y = "Number of Shooting Incidents") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

