MSDS - Datascience as a Field - Week 3

Marco Tulio Teixeira

2023-08-30

Project Summary

Import, tidy and analyze the NYPD Shooting Incident dataset 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.

Project Goals / Rationale

Validate the number of domestic incidents by year, sex and age range.

Note: This document/presentation doesn't contain any political or any type of discrimination. It also should not be considered a proper analysis due the limited amount of information.

Project Steps/Processes

1. Dependencies Installation
2. Data Collection and Understanding
3. Data Cleaning
4. Data Aggregation
5. Visualization
6. Analysis
7. Modeling
8. Answers and Conclusion

Dependencies Installation

Installing and adding required libraries

```r
# Importing useful libraries
library(ggmap)
library(lubridate)
library(dplyr)
library(tidyverse)
library(patchwork)
```

Data Collection

Downloading NYPD Shooting Incident Dataset

```
url_base = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
nypd_shooting_data_raw = read.csv(url_base, sep=",")
```

Data Cleaning

Steps considered on data cleaning process

- Filtering JURISDICTION_CODE where is equals to 2 (Housing);
- Filtering VIC_SEX different the U (UNKNOW);
- Filtering PERP_AGE_GROUP different the 1020 (ERROR DATA);
- Create OCURR_DATE_TIME combining OCCUR_DATE and OCCUR_TIME
- Convert STATISTICAL_MURDER_FLAG to integer (0, 1)
- Removing unuseful columns: INCIDENT_KEY, OCCUR_TIME, PRECINCT, JURISDICTION_CODE, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Lon_Lat

```
# Cleaning process
nypd_shooting_cleaned = nypd_shooting_data_raw %>%
    filter(JURISDICTION_CODE == 2, VIC_SEX != "U", PERP_AGE_GROUP != "1020") %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
    unite(
      "OCCUR_DATE_TIME", c(OCCUR_DATE, OCCUR_TIME),
      sep = " ", na.rm = TRUE, remove = FALSE
    ) %>%
    mutate(
      OCCUR_DATE_TIME = as_datetime(OCCUR_DATE_TIME),
      OCCUR_YEAR = year(OCCUR_DATE)
    ) %>%
    mutate(
      STATISTICAL_MURDER_FLAG = as.integer(as.logical(STATISTICAL_MURDER_FLAG))
    ) %>%
    select(
      -c(INCIDENT_KEY, OCCUR_TIME, PRECINCT, JURISDICTION_CODE,
        LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION_DESC,
        X_COORD_CD, Y_COORD_CD, Lon_Lat)
    )
```

Data Aggregation

Splitting data by sex:

```r
nypd_shooting_by_sex = nypd_shooting_cleaned %>%
  group_by(VIC_SEX) %>%
  summarise(count_shoots = n(), .groups = 'drop') %>%
  mutate(perc = (count_shoots / sum(count_shoots))*100) %>%
  mutate(percentage = round(perc, 2))

# Split data to validate if there is a murder or not
nypd_shooting_by_sex_with_murder = nypd_shooting_cleaned %>%
  filter(STATISTICAL_MURDER_FLAG == 1) %>%
  group_by(VIC_SEX) %>%
  summarise(count_shoots = n(), .groups = 'drop') %>%
  mutate(perc = (count_shoots / sum(count_shoots))*100) %>%
  mutate(percentage = round(perc, 2))

nypd_shooting_by_sex_with_no_murder = nypd_shooting_cleaned %>%
  filter(STATISTICAL_MURDER_FLAG == 0) %>%
  group_by(VIC_SEX) %>%
  summarise(count_shoots = n(), .groups = 'drop') %>%
  mutate(perc = (count_shoots / sum(count_shoots))*100) %>%
  mutate(percentage = round(perc, 2))
```

Splitting data by sex and age range:

```r
nypd_shooting_by_male_age_range = nypd_shooting_cleaned %>%
  filter(!is.na(PERP_AGE_GROUP), PERP_AGE_GROUP != "", PERP_AGE_GROUP != "(null)" ) %>%
  filter(VIC_SEX == "M") %>%
  group_by(PERP_AGE_GROUP) %>%
  summarise(count_shoots = n(), .groups = 'drop') %>%
  mutate(perc = (count_shoots / sum(count_shoots))*100) %>%
  mutate(percentage = round(perc, 2))

nypd_shooting_by_female_age_range = nypd_shooting_cleaned %>%
  filter(!is.na(PERP_AGE_GROUP), PERP_AGE_GROUP != "", PERP_AGE_GROUP != "(null)" ) %>%
  filter(VIC_SEX == "F") %>%
  group_by(PERP_AGE_GROUP) %>%
  summarise(count_shoots = n(), .groups = 'drop') %>%
  mutate(perc = (count_shoots / sum(count_shoots))*100) %>%
  mutate(percentage = round(perc, 2))
```

Splitting data by sex and race:

```
nypd_shooting_male_by_race = nypd_shooting_cleaned %>%
  filter(!is.na(PERP_RACE), PERP_RACE != "", PERP_RACE != "(null)" ) %>%
  filter(VIC_SEX == "M") %>%
  group_by(PERP_RACE) %>%
  summarise(count_shoots = n(), .groups = 'drop') %>%
  mutate(perc = (count_shoots / sum(count_shoots))*100) %>%
  mutate(percentage = round(perc, 2))

nypd_shooting_female_by_race = nypd_shooting_cleaned %>%
  filter(!is.na(PERP_RACE), PERP_RACE != "", PERP_RACE != "(null)" ) %>%
  filter(VIC_SEX == "F") %>%
  group_by(PERP_RACE) %>%
  summarise(count_shoots = n(), .groups = 'drop') %>%
  mutate(perc = (count_shoots / sum(count_shoots))*100) %>%
  mutate(percentage = round(perc, 2))
```

Data Visualization

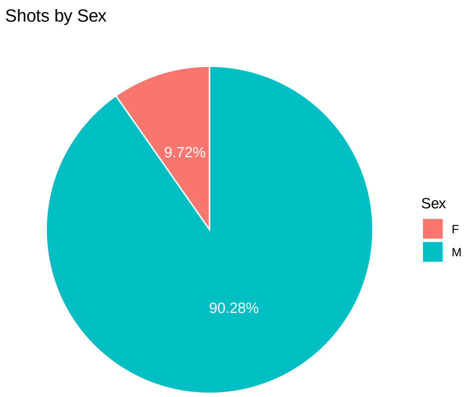
Spatial Analysis

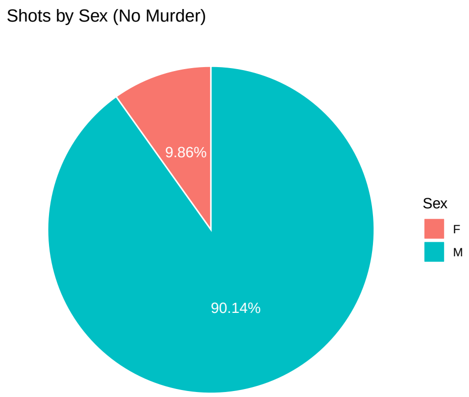
```r
# Show occurrencies on a map
nyc_map <- get_map(location = "New York", maptype = "roadmap", )
ggmap::register_google(key = api_key, write = TRUE)
p = ggmap(nyc_map)
p + geom_point(
  data=nypd_shooting_cleaned,
  aes(x=Longitude, y=Latitude),
  color="red",
  alpha=0.5
)
```

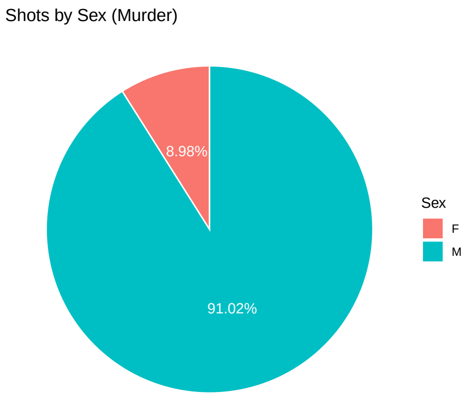
Demographic Analysis

Victims by Sex

- At the first sigh we can conclude that the number of domestic shots incidents are greater for victims are categorized as the **Sex M (Male)**

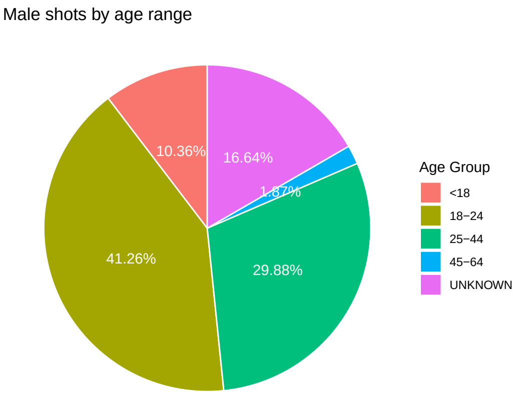

Shots by Sex

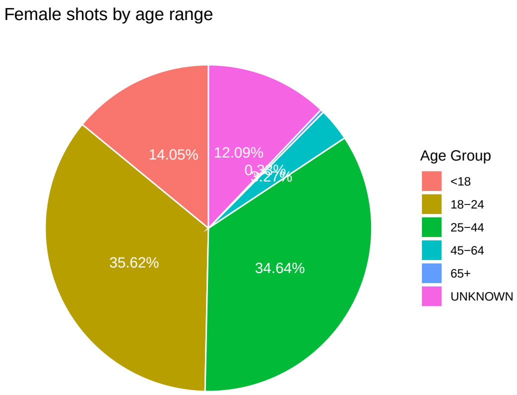

Shots by Sex (No Murder)


Shots by Sex (Murder)

Victims by Age

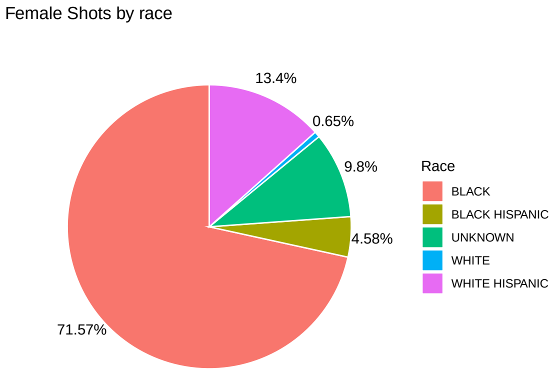
- The number of incidents are higher in the age ranges of 18-24 and 25-44 for both **Female** and **Male**

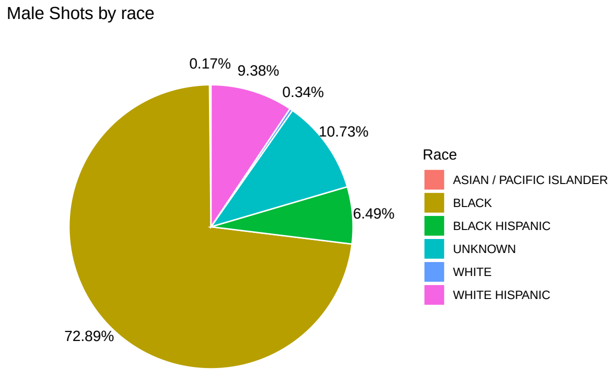

Male shots by age range


Female shots by age range

Victims by race

- Considering only the information provided at this moment we can assume that there's a huge difference between victims by race.
- Could be interesting to have a comparison between social classes, education level and neighbors to achieve a better understanding.





Analysis and Conclusion

- The initial year provided by the dataset is 2006;
- Dataset contains a sort of biases - Newest incidents are well reported than old ones;
- The usage of other dataset to aggregate more information should be considered to a better understanding of this data - Datasets like Social Classes, poverty, political situations and others.