Week 3 Project - NYPD Shooting Incident Data (Historic)

Oliviu Lazar

2024-04-16

Import libraries

Import necessary libraries

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
```

Data Description

The NYPD Shooting Incident Data data set offers comprehensive details on shooting incidents in New York City, encompassing dates, times, locations, perpetrator and victim demographics, and other pertinent information. We'll import and analyze this data set to glean insights into citywide shooting incidents. The data set spans every shooting incident in NYC from 2006 to the end of 2022. It undergoes manual extraction quarterly and is vetted by the Office of Management Analysis and Planning before publication on the NYPD website. Each record represents a shooting incident in NYC, providing event specifics, occurrence details, and demographics of suspects and victims. This data serves as a valuable resource for public exploration of shooting and criminal activities.

Importing the Dataset

Read the CSV file from the URL and store it in a data frame.

```
url_nypd <-
  "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

nypd_raw_data <- read_csv(url_nypd)
```

```
## 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.
```

Clean the raw data

Before conducting the analysis, the data set undergoes cleaning to standardize column names, convert data types (e.g., dates to DATE), handle missing values, and resolve inconsistencies.

```
nypd_clean_data <- nypd_raw_data %>%
  select(c(
    "DATE" = "OCCUR_DATE",
    "TIME" = "OCCUR_TIME",
    "BOROUGH" = "BORO",
    "PRECINCT",
    "MURDER_FLAG" = "STATISTICAL_MURDER_FLAG",
    "PERP_AGE_GROUP",
    "PERP_SEX",
    "PERP_RACE",
    "VIC_AGE_GROUP",
    "VIC_SEX",
    "VIC_RACE"
  )) %>%
  mutate(
    DATE = mdy(DATE),
    PRECINCT = as.integer(PRECINCT),
    YEAR = year(DATE),
    MURDER_FLAG = as.integer(MURDER_FLAG)
  )

head(nypd_clean_data)
```

```
## # A tibble: 6 x 12
##   DATE       TIME   BOROUGH  PRECINCT MURDER_FLAG PERP_AGE_GROUP PERP_SEX
##   <date>     <time> <chr>       <int>       <int> <chr>          <chr>
## 1 2021-05-27 21:30  QUEENS        105           0 <NA>           <NA>
## 2 2014-06-27 17:40  BRONX          40           0 <NA>           <NA>
## 3 2015-11-21 03:56  QUEENS        108           1 <NA>           <NA>
## 4 2015-10-09 18:30  BRONX          44           0 <NA>           <NA>
## 5 2009-02-19 22:58  BRONX          47           1 25-44          M
## 6 2020-10-21 21:36  BROOKLYN       81           1 <NA>           <NA>
## # i 5 more variables: PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, YEAR <dbl>
```

Summary statistics of the dataset

Compute summary statistics for each variable (column) in the data frame and display the count of missing values for each column. Missing values will be addressed individually for subsequent analyses and visualizations.

```
summary(nypd_clean_data)
```

```
##       DATE                TIME            BOROUGH             PRECINCT
##  Min.   :2006-01-01   Length:27312      Length:27312       Min.   :  1.00
##  1st Qu.:2009-07-18   Class1:hms        Class :character   1st Qu.: 44.00
##  Median :2013-04-29   Class2:difftime   Mode  :character   Median : 68.00
##  Mean   :2014-01-06   Mode  :numeric                       Mean   : 65.64
##  3rd Qu.:2018-10-15                                        3rd Qu.: 81.00
##  Max.   :2022-12-31                                        Max.   :123.00
##   MURDER_FLAG     PERP_AGE_GROUP       PERP_SEX          PERP_RACE
##  Min.   :0.0000   Length:27312       Length:27312       Length:27312
##  1st Qu.:0.0000   Class :character   Class :character   Class :character
##  Median :0.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.1928
##  3rd Qu.:0.0000
##  Max.   :1.0000
##  VIC_AGE_GROUP        VIC_SEX            VIC_RACE             YEAR
##  Length:27312       Length:27312       Length:27312       Min.   :2006
##  Class :character   Class :character   Class :character   1st Qu.:2009
##  Mode  :character   Mode  :character   Mode  :character   Median :2013
##                                                           Mean   :2013
##                                                           3rd Qu.:2018
##                                                           Max.   :2022
```

```
missing_counts <- colSums(is.na(nypd_clean_data))
missing_counts
```

```
##           DATE           TIME        BOROUGH       PRECINCT    MURDER_FLAG
##              0              0              0              0              0
## PERP_AGE_GROUP       PERP_SEX      PERP_RACE  VIC_AGE_GROUP        VIC_SEX
##           9344           9310           9310              0              0
##       VIC_RACE           YEAR
##              0              0
```

Analysis and Visualizations

Murders by year

We calculate the total number of murders reported by the NYPD for each year using the nypd_clean_data data set. The data is grouped by year, and the sum of murders is calculated using the summarize() function. Subsequently, the first few rows of the resulting data set named murders_by_year are displayed using the head() function. Since there is no missing data in the columns needed for this analysis, we utilize the entire data set.

The results are visualized in a line plot to illustrate the trend of the total number of murders by year.

```
murders_by_year <- nypd_clean_data %>%
  group_by(YEAR) %>%
  summarize(Murders = sum(MURDER_FLAG))

head(murders_by_year)
```

```
## # A tibble: 6 x 2
##    YEAR Murders
##   <dbl>   <int>
## 1  2006     445
## 2  2007     373
## 3  2008     362
## 4  2009     348
## 5  2010     405
## 6  2011     373
```

```r
ggplot(murders_by_year, aes(x = YEAR, y = Murders)) +
  geom_line(color = "blue") +  # Use geom_line for line graph
  labs(title = "NYPD Murders by Year",
       x = "Year",
       y = "Number of Murders") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

The results indicate a downward trend from 2006 until 2019, followed by an increase since 2019. This analysis lays the groundwork for comprehending long-term trends, pinpointing areas of concern, and guiding strategic interventions to tackle public safety challenges.

Regression Modeling of murders by year

Fit and visualize the relationship between the time and the number of murders.

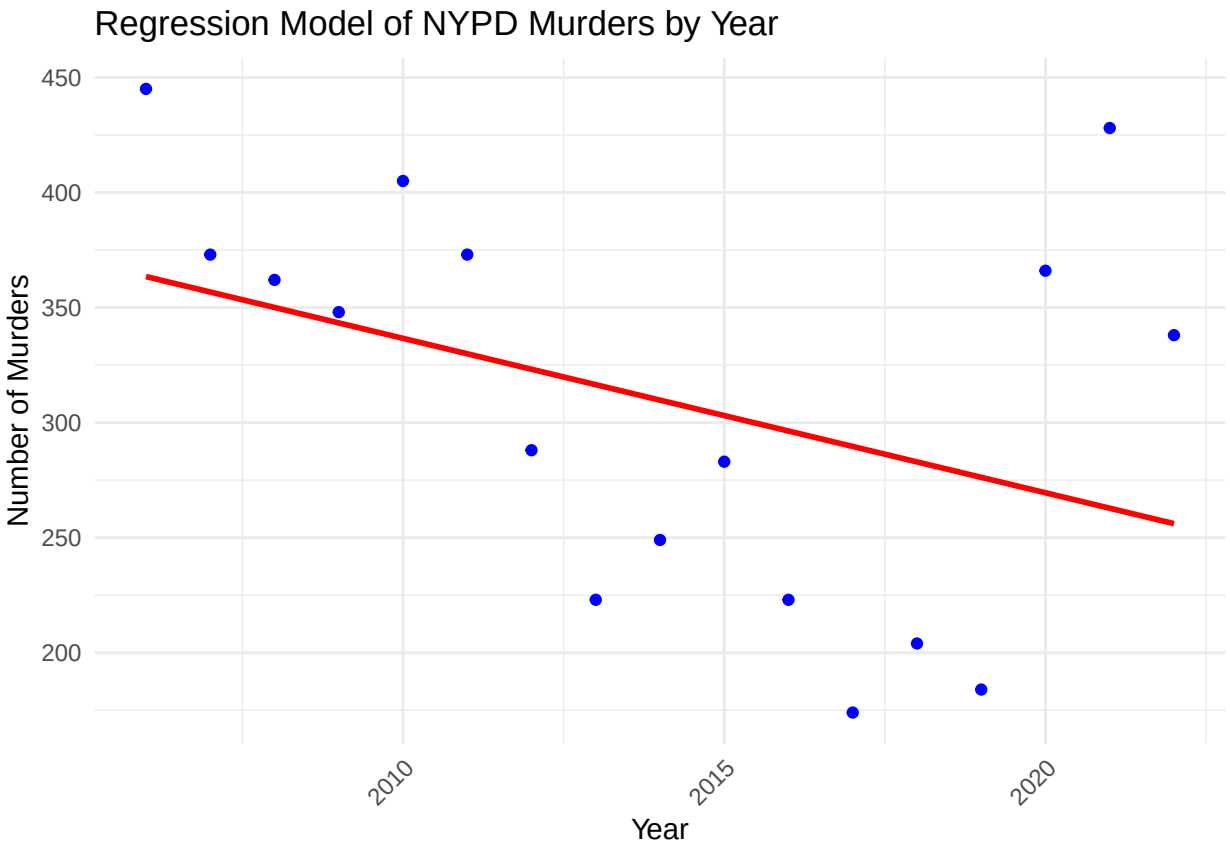
```r
model <- lm(Murders ~ YEAR, data = murders_by_year)

summary(model)
```

```
##
## Call:
## lm(formula = Murders ~ YEAR, data = murders_by_year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -115.632  -73.343    4.681   68.392  165.211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13825.284   8321.696   1.661    0.117
## YEAR           -6.711      4.132  -1.624    0.125
##
## Residual standard error: 83.46 on 15 degrees of freedom
## Multiple R-squared:  0.1496, Adjusted R-squared:  0.09286
## F-statistic: 2.638 on 1 and 15 DF,  p-value: 0.1252
```

```r
ggplot(murders_by_year, aes(x = YEAR, y = Murders)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Regression Model of NYPD Murders by Year",
       x = "Year",
       y = "Number of Murders") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



We can refine and enhance the linear regression model by exploring options such as feature engineering and/or polynomial regression. These approaches aim to better capture the underlying relationships in the data and improve the model's predictive performance.

Race-on-race Shootings

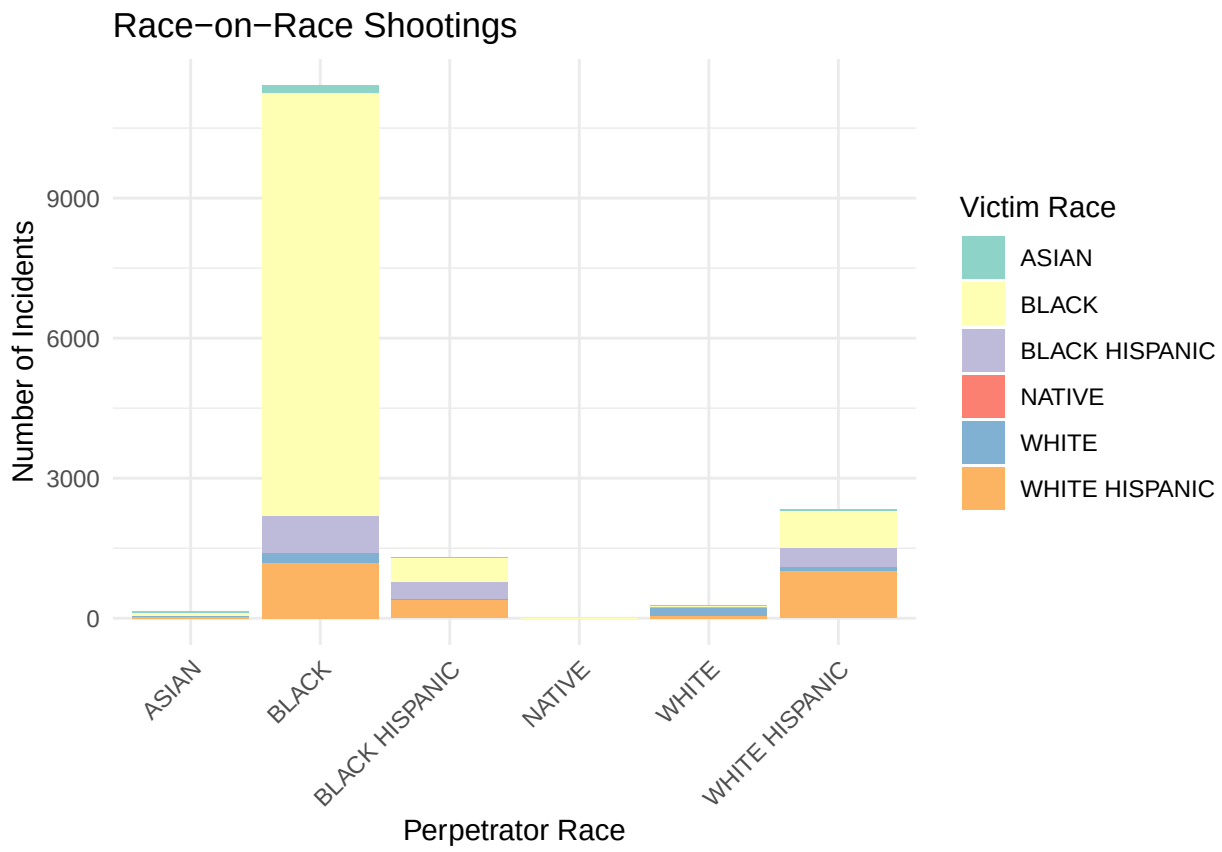
Performing a race-on-race shootings analysis necessitates the removal of missing and unknown race rows from the shooting data to uphold data integrity and ensure accurate analysis. Rows lacking race information can compromise data set reliability, leading to biased or misleading results. Including such rows in analysis may distort statistical measures and obscure patterns or trends in race-related shooting incidents. Removal of these rows enhances the accuracy of statistical analyses and visualizations, elevating the overall data set quality. This ensures that insights or conclusions drawn from the data are grounded in complete and reliable information, fostering more robust and trustworthy findings.

```
nypd_clean_data <- nypd_clean_data %>%
  filter(PERP_RACE != 'U' &
           VIC_RACE != 'U' &
           PERP_RACE != '(null)' &
           VIC_RACE != '(null)' &
           PERP_RACE != 'UNKNOWN' &
           VIC_RACE != 'UNKNOWN')

nypd_clean_data$PERP_RACE <-
  replace(nypd_clean_data$PERP_RACE,
          nypd_clean_data$PERP_RACE == "AMERICAN INDIAN/ALASKAN NATIVE", "NATIVE")
```

```r
nypd_clean_data$PERP_RACE <-
  replace(nypd_clean_data$PERP_RACE,
          nypd_clean_data$PERP_RACE == "ASIAN / PACIFIC ISLANDER", "ASIAN")
nypd_clean_data$VIC_RACE <-
  replace(nypd_clean_data$VIC_RACE,
          nypd_clean_data$VIC_RACE == "AMERICAN INDIAN/ALASKAN NATIVE", "NATIVE")
nypd_clean_data$VIC_RACE <-
  replace(nypd_clean_data$VIC_RACE,
          nypd_clean_data$VIC_RACE == "ASIAN / PACIFIC ISLANDER", "ASIAN")

ggplot(nypd_clean_data, aes(x = PERP_RACE, fill = VIC_RACE)) +
  geom_bar() +
  labs(title = "Race-on-Race Shootings",
       x = "Perpetrator Race",
       y = "Number of Incidents",
       fill = "Victim Race") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set3")  # Using a different color for each victim race
```



The analysis highlights disparities and patterns in the occurrence of shootings between individuals of the same race (in particular BLACK on BLACK shootings). Factors such as socioeconomic status, community dynamics, and historical context likely play significant roles in shaping these patterns. Furthermore, the analysis underscores the importance of addressing systemic issues such as racism, inequality, and access to resources to mitigate the prevalence of shootings within racial groups.

Sources of Bias

When performing the analysis above, we must consider any potential sources of bias such as:

Underreporting: Not all shooting incidents may be reported, leading to an underestimation of the true number of shootings and biased datasets.

Selection Bias: The data set may not represent all demographics or areas equally, skewing conclusions about the prevalence or distribution of shootings.

Sampling Bias: Non-random sampling methods or incomplete data collection can bias analyses by excluding certain incidents or time periods.

Data Collection Methods: Inaccurate or incomplete reporting methods can introduce biases, especially in demographic information like race.

Data Quality Issues: Inaccuracies, inconsistencies, or missing data can bias analyses and affect conclusions.

Contextual Factors: Social, economic, and political factors can influence both the occurrence of shootings and the recording of data, leading to bias.

Personal Bias: In conducting the analysis, I must be aware of personal biases which can stem from various factors such as personal experiences, cultural background, upbringing, education, social environment, among others.

Awareness of these biases and rigorous evaluation of data set limitations are crucial for interpreting results accurately in analyses of NYPD shooting data

Conclusion

The analysis of murders by year and race-on-race shootings provides valuable insights into the dynamics of violent crime and racial disparities in New York City.

Regarding murders by year, the data reveals fluctuations in homicide rates over time, with periods of decline followed by recent increases. This analysis underscores the importance of ongoing monitoring and proactive measures to address emerging trends and mitigate the impact of violent crime on communities.

Similarly, the race-on-race shootings analysis sheds light on disparities in the prevalence of shootings within racial groups. By examining patterns of shootings between individuals of the same race, this analysis highlights the complex interplay of socioeconomic factors, community dynamics, and systemic inequalities.