Week 3 Project

Anonymized for Project Submission

2023-09-12

Intro

In this document, I will clean and analyze the NYPD Shooting Incident dataset. The dataset is available in the link below. https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

Libraries

The libraries used for this assignment are: ggplot2, dplyr, knitr, rmarkdown, readr, tidyverse, and lubridate.

Importing the data

```
URL <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
NYPD <- readr::read_csv(URL)
```

```
## 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.
```

Summary Stats

Now, let's take a brief look at the summary statistics of the dataset. Right now, this means nothing since it is uncleaned and unverified. This step helps with a quick look at the data, but there could be outliers or incorrectly input data. This summary will show the barebones of the statistics: min/max/quartiles for numeric columns and length/datatype for the string-based columns. This summary should only be used for a very loose idea of the data, such as how large the dataframe is and what the rough ranges are.

```
summary(NYPD)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME           BORO          
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312      
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character  
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character  
##  Mean   :120860536                      Mode  :numeric                      
##  3rd Qu.:188810230                                                          
##  Max.   :261190187                                                          
##                                                                             
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312      
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character  
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character  
##                     Mean   : 65.64   Mean   :0.3269                      
##                     3rd Qu.: 81.00   3rd Qu.:0.0000                      
##                     Max.   :123.00   Max.   :2.0000                      
##                                      NA's   :2                           
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP    
##  Length:27312       Mode :logical           Length:27312      
##  Class :character   FALSE:22046             Class :character  
##  Mode  :character   TRUE :5266              Mode  :character  
##                                                               
##                                                               
##                                                               
##                                                               
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX         
##  Length:27312       Length:27312       Length:27312       Length:27312      
##  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:27312       Min.   : 914928   Min.   :125757   Min.   :40.51  
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67  
##  Mode  :character   Median :1007731   Median :194487   Median :40.70  
##                     Mean   :1009449   Mean   :208127   Mean   :40.74  
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82  
##                     Max.   :1066815   Max.   :271128   Max.   :40.91  
##                                                        NA's   :10     
##    Longitude        Lon_Lat         
##  Min.   :-74.25   Length:27312      
##  1st Qu.:-73.94   Class :character  
##  Median :-73.92   Mode  :character  
##  Mean   :-73.91                     
##  3rd Qu.:-73.88                     
##  Max.   :-73.70                     
##  NA's   :10
```

Fixing the data

My first step is to convert the OCCUR_DATE column to a datetime data type. The lubridate library function make this an easy, one liner task. Having the date in proper format is important for more accurately reading and manipulating it in future code.

```r
NYPD$OCCUR_DATE <- lubridate::mdy(NYPD$OCCUR_DATE)
```

The following shows that there are a lot of NA values in 3 columns (10,000+) as well as a smaller amount (but still a lot) of NA values in other columns. The first three columns will be removed as there is not enough data in them to justify keeping them around. This cell will count NA and then remove the 3 columns of LOC_CLASSFCTN_DESC, LOCATION_DESC, and LOC_OF_OCCUR_DESC.

```r
sapply(NYPD, function(x) sum(is.na(x)))
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO       LOC_OF_OCCUR_DESC                PRECINCT
##                       0                   25596                       0
##       JURISDICTION_CODE      LOC_CLASSFCTN_DESC           LOCATION_DESC
##                       2                   25596                   14977
## STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP                PERP_SEX
##                       0                    9344                    9310
##               PERP_RACE           VIC_AGE_GROUP                 VIC_SEX
##                    9310                       0                       0
##                VIC_RACE              X_COORD_CD              Y_COORD_CD
##                       0                       0                       0
##                Latitude               Longitude                 Lon_Lat
##                      10                      10                      10
```

```r
NYPD <- subset(NYPD, select = -c(LOC_CLASSFCTN_DESC, LOCATION_DESC, LOC_OF_OCCUR_DESC))
```

Now it is time to turn the Victim and Perp related columns into categorical variables. From a brief glance of the dataset, it is clear that the age should be categorical as there are a lot of repeat values that appear within ranges rather than a specific number. Categorical is the most logical for the Victim/Perp Sex as there is a finite amount of options. The same concept holds true for race.

```r
# Perps
NYPD$PERP_AGE_GROUP <- as.factor(NYPD$PERP_AGE_GROUP)
NYPD$PERP_SEX <- as.factor(NYPD$PERP_SEX)
NYPD$PERP_RACE <- as.factor(NYPD$PERP_RACE)
# Victims
NYPD$VIC_AGE_GROUP <- as.factor(NYPD$VIC_AGE_GROUP)
NYPD$VIC_RACE <- as.factor(NYPD$VIC_RACE)
NYPD$VIC_SEX <- as.factor(NYPD$VIC_SEX)
```

With the columns categorized properly, it is time to inspect and verify the data. Anything that does not make sense, such as a negative age bracket or an invalid sex should be investigated further and removed from the dataset. After inspection, there are a few invalid inputs, such as 1020 for age (nobody is immortal). Race and sex seem fine enough. The number of "Unknown" in the Age/Sex/Race column can indicate that the shooter may have gotten away with their crime (upon investigation) or the victim survived and did not further report the incident.

```r
# Perp levels inspection
levels(NYPD$PERP_AGE_GROUP)
```

```
##  [1] "(null)"  "<18"     "1020"    "18-24"   "224"     "25-44"   "45-64"  
##  [8] "65+"     "940"     "UNKNOWN"
```

```r
levels(NYPD$PERP_SEX)
```

```
## [1] "(null)" "F"      "M"      "U"
```

```r
levels(NYPD$PERP_RACE)
```

```
## [1] "(null)"                         "AMERICAN INDIAN/ALASKAN NATIVE"
## [3] "ASIAN / PACIFIC ISLANDER"       "BLACK"                         
## [5] "BLACK HISPANIC"                 "UNKNOWN"                       
## [7] "WHITE"                          "WHITE HISPANIC"
```

```r
# Victim levels inspection
levels(NYPD$VIC_AGE_GROUP)
```

```
## [1] "<18"     "1022"    "18-24"   "25-44"   "45-64"   "65+"     "UNKNOWN"
```

```r
levels(NYPD$VIC_SEX)
```

```
## [1] "F" "M" "U"
```

```r
levels(NYPD$VIC_RACE)
```

```
## [1] "AMERICAN INDIAN/ALASKAN NATIVE" "ASIAN / PACIFIC ISLANDER"      
## [3] "BLACK"                          "BLACK HISPANIC"                
## [5] "UNKNOWN"                        "WHITE"                         
## [7] "WHITE HISPANIC"
```

Going through each of the columns in the dataframe to audit the data clearly and check to see if there are any other columns that do not contribute significant amounts of information to the overall dataset. By using the 'aggregate' command, we can get a quick count of the number of times a given value appears. Through this, jurisdiction code seems to be cleaned or clean enough. These aggregates have shown that the data is cleaned and it is time to analyze it.

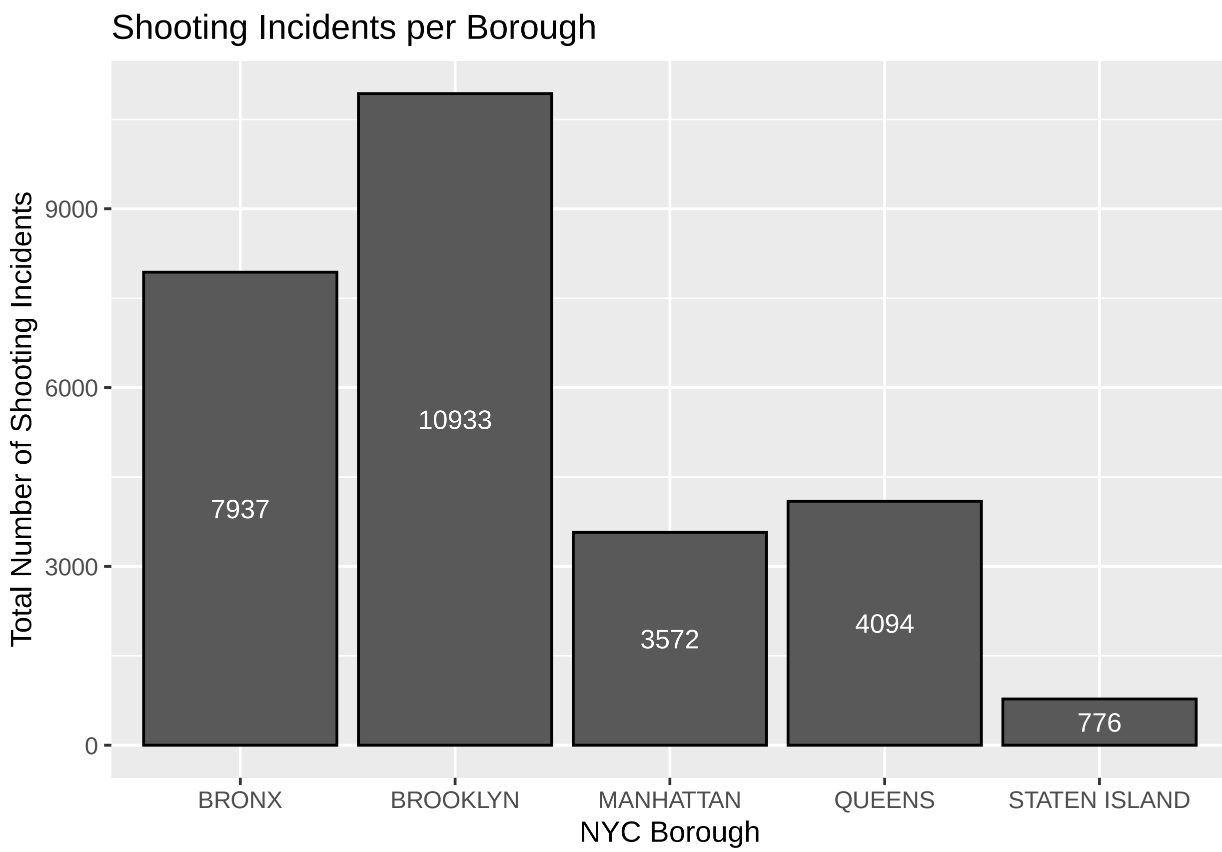
```r
boro_count <- NYPD %>% count(BORO)
boro_count
```

```
## # A tibble: 5 x 2
##   BORO              n
##   <chr>         <int>
## 1 BRONX          7937
## 2 BROOKLYN      10933
## 3 MANHATTAN      3572
## 4 QUEENS         4094
## 5 STATEN ISLAND   776
```

Plotting the data

This plot shows the number of firearm incidents in each NYC borough. This graph shows that the highest concentration of firearm incidents is within Brooklyn and the Bronx. After looking up the population of NYC's boroughs, this is unusual as the Bronx and Manhattan have a similar number of people however there is a stark contrast between the number of firearm incidents. Brooklyn and Queens have similar populations but there is a large difference in incidents. Source: https://en.wikipedia.org/wiki/Boroughs_of_New_York_City#Background. Wikipedia's source pulls directly from the US Census data.

```
ggplot(NYPD, aes(x = BORO)) +
  geom_bar(color = "black") +
  stat_count(geom="text", colour = "white", size = 3.5,
             aes(label = after_stat(count)), position = position_stack(vjust = 0.5)) +
  ggtitle("Shooting Incidents per Borough") +
  xlab("NYC Borough") + ylab("Total Number of Shooting Incidents")
```



This chunk shows the number of shootings committed by each race in the dataset. It is also where I did some minor cleaning to remove NA values, which are different from "Unknown" values as NA indicates that there was nothing entered versus unknown representing that the police were unable to identify any usable information from witnesses or investigation.

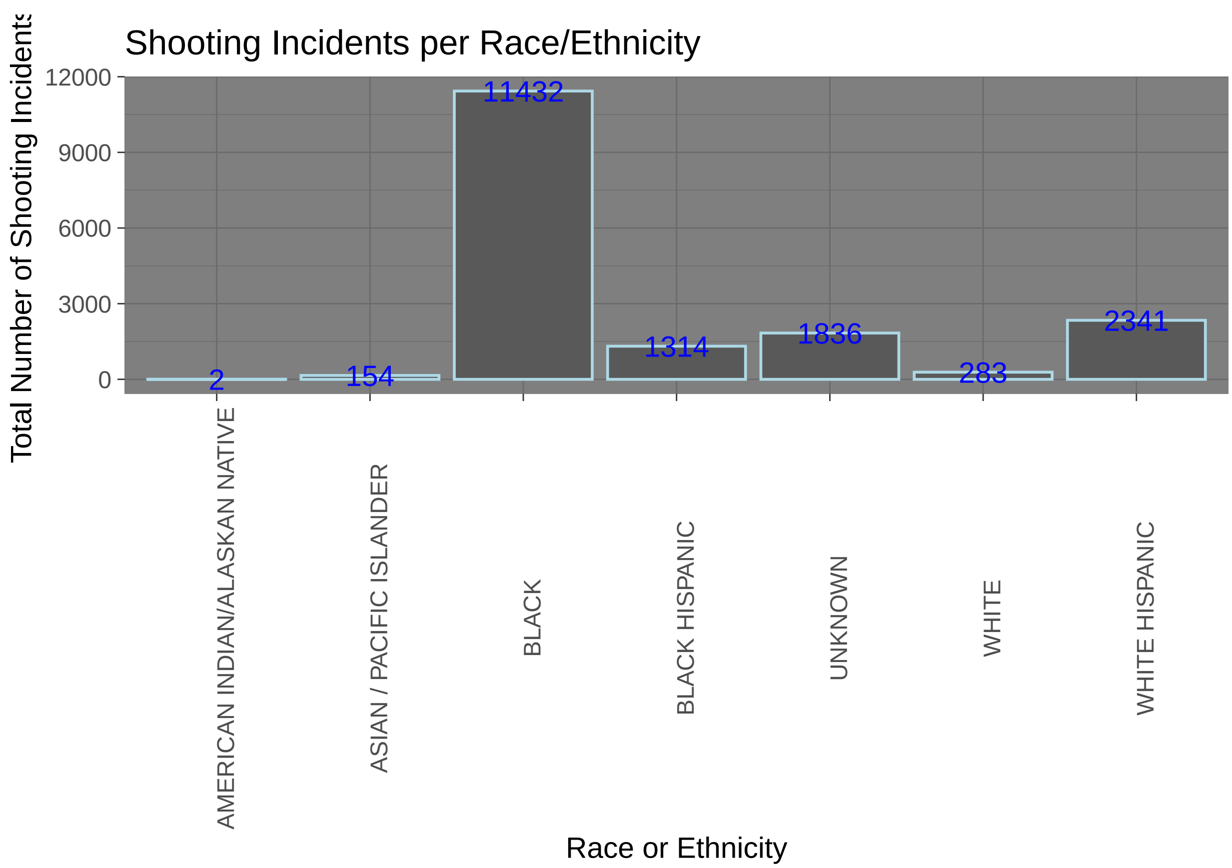
```
race <- as.data.frame(table(NYPD$PERP_RACE))
race <- race[-1,]
names(race)[names(race) == "Var1"] <- "Ethnicity"
names(race)[names(race) == "Freq"] <- "Incidents"

race
```

```
##                           Ethnicity Incidents
## 2 AMERICAN INDIAN/ALASKAN NATIVE         2
## 3       ASIAN / PACIFIC ISLANDER       154
## 4                          BLACK     11432
## 5                 BLACK HISPANIC      1314
## 6                        UNKNOWN      1836
## 7                          WHITE       283
## 8                 WHITE HISPANIC      2341
```

This graph shows the number of shooting incidents committed by each race/ethnicity. There is a large discrepancy between the incidents by black NYC residents and all other races. The next closest group are both of the Hispanic demographics and unknown. The unknown could be that the shooter was not able to be identified through police investigation and likely escaped or there were no witnesses around to give a description.

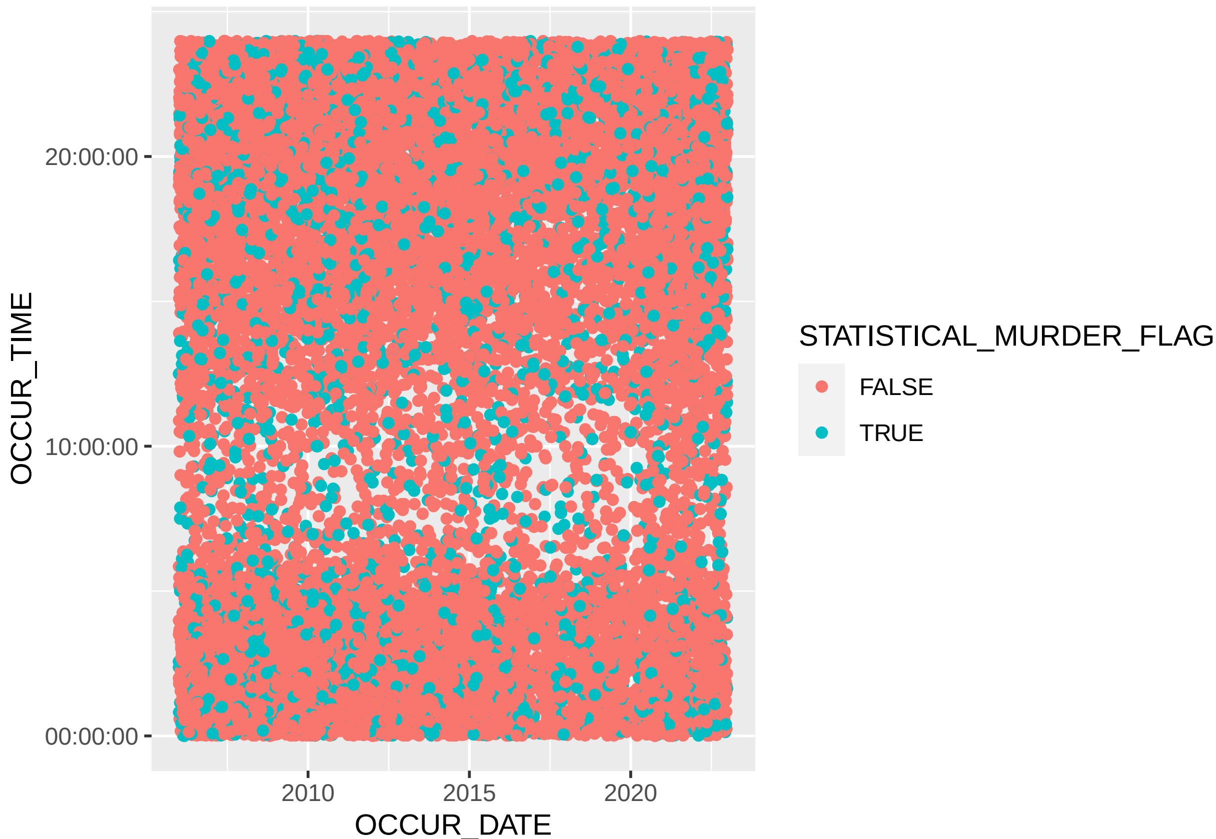
```
ggplot(race, aes(x = Ethnicity, y =Incidents)) + geom_col(color = "lightblue")  +
  theme_dark() +
  ggtitle("Shooting Incidents per Race/Ethnicity") +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Race or Ethnicity") + ylab("Total Number of Shooting Incidents") +
  geom_text(aes(label = Incidents), vjust = 0.5, colour = "blue")
```



Modeling the Data

Here, I will take a simple linear model of the data to loosely predict if a shooting incident is a homicide based on the time and date. First is a plot comparing the time of day to whether or not an incident was a statistical murder, based on the original column from the dataset rather than statistical analysis within this document.

```
ggplot(NYPD, aes(x = OCCUR_DATE, y = OCCUR_TIME, color = STATISTICAL_MURDER_FLAG)) +
  geom_point()
```



In this model, the p value for the date and time are both outside of the statistically significant values of >= 0.95 or <= 0.05. With the p-value in mind, the date and time are not good predictors of if a shooting incident will be a statistical murder. Even when re-modeling using only date or only time, the p-value is still outside of the statistically significant range. The R-squared value approaches 0, which means that the model is not a good fit and that the variation in statistical murder is unaffected by time of day or date.

```
model <- lm(STATISTICAL_MURDER_FLAG ~ OCCUR_DATE + OCCUR_TIME, data = NYPD)
summary(model)$coeff
```

```
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  1.951441e-01 2.074065e-02  9.4087711 5.400914e-21
## OCCUR_DATE  -2.177427e-07 1.271571e-06 -0.1712391 8.640370e-01
## OCCUR_TIME   2.551004e-08 7.807089e-08  0.3267548 7.438559e-01
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = STATISTICAL_MURDER_FLAG ~ OCCUR_DATE + OCCUR_TIME,
##     data = NYPD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1945 -0.1934 -0.1924 -0.1915  0.8091
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.951e-01  2.074e-02   9.409   <2e-16 ***
## OCCUR_DATE  -2.177e-07  1.272e-06  -0.171    0.864
## OCCUR_TIME   2.551e-08  7.807e-08   0.327    0.744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3945 on 27309 degrees of freedom
## Multiple R-squared:  4.819e-06,  Adjusted R-squared:  -6.842e-05
## F-statistic: 0.0658 on 2 and 27309 DF,  p-value: 0.9363
```

Biases

Going in to this assignment, I had a loose idea of what I could expect out of the racial distribution of shooting incidents, with Black Americans at the upper end and Asian Americans at the lower end. While the data shows that it fits the stereotype, race alone is not enough of a reason for it to happen. There are many variables in play that lead to this which are not shown within the data, such as socioeconomic standings and cultural differences. Another bias within my pre-assignment thoughts was actually a lack of a thought: I forgot about the indigenous peoples as well as did not think about splitting up the Latino race into white and black. The best way to mitigate either forms of these bias is to only think about it within the context of the data. By ignoring my pre-assignment thoughts, I can draw a more objective conclusion using the data with potential evidence rather than random anecdotes.