NYPD Incident Project

Richard Ebersole

2023-11-30

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: ##Import Data

```
# Set the Url variable to the link of your CSV file
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"

# Read in the CSV file using the read.csv() function
nypd_shooting <- 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 of data (Pre-cleaning)

```
summary(nypd_shooting)
```

```
##   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
```

##Clean Data

```r
#Set date, select desired rows
nypd_clean <- nypd_shooting %>%
  select(c("OCCUR_DATE", "PERP_AGE_GROUP","VIC_AGE_GROUP", "BORO"))%>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))

#Get rid of rows with empty values, and odd values (no one is 940 years old in NYC)
nypd_cleaner <- nypd_clean[complete.cases(nypd_clean), ]
nypd_cleaner <- subset(nypd_cleaner, !(PERP_AGE_GROUP %in% c("1020", "940", "224", "UNKNOWN", "(null)")
nypd_cleaner <- subset(nypd_cleaner, !(VIC_AGE_GROUP %in% c("1022", "UNKNOWN")))
```

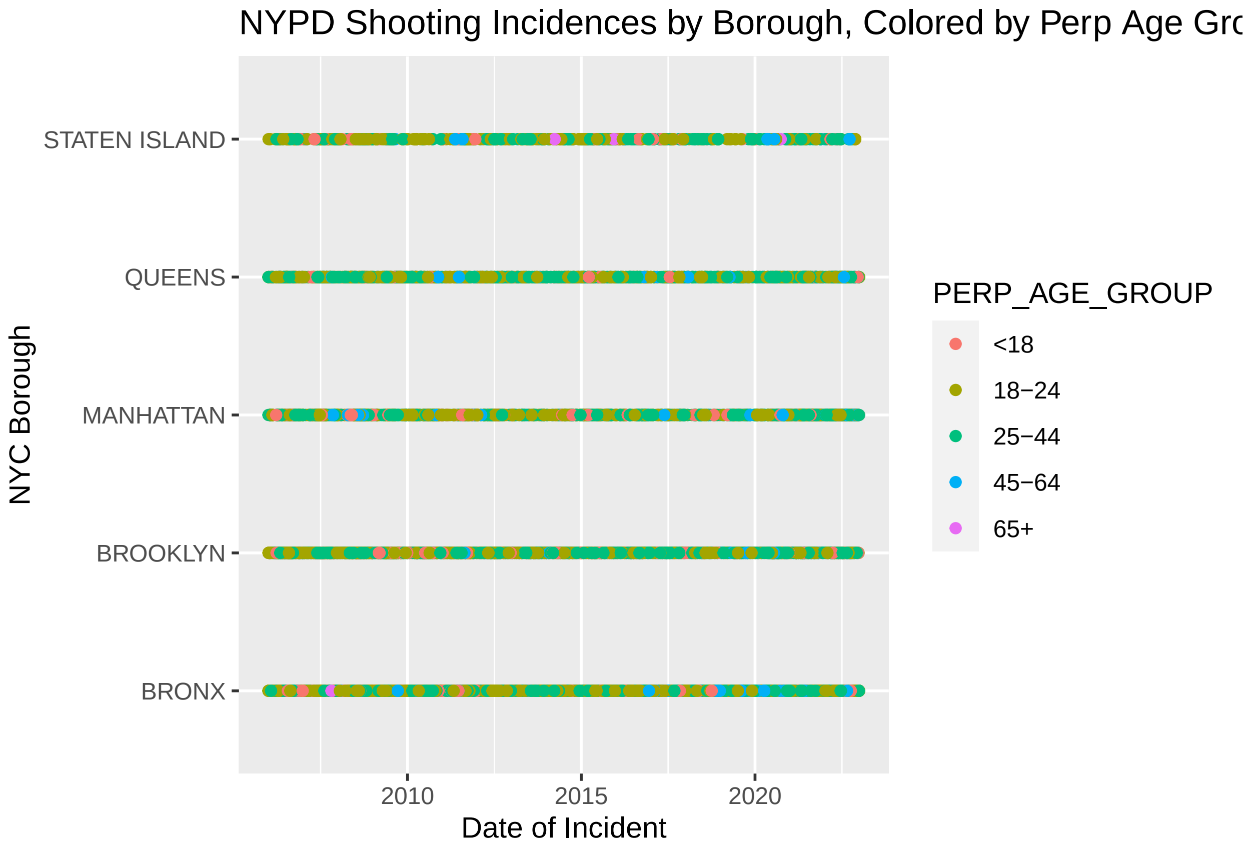
##Summary of data (Post-cleaning)

```
summary(nypd_cleaner)
```

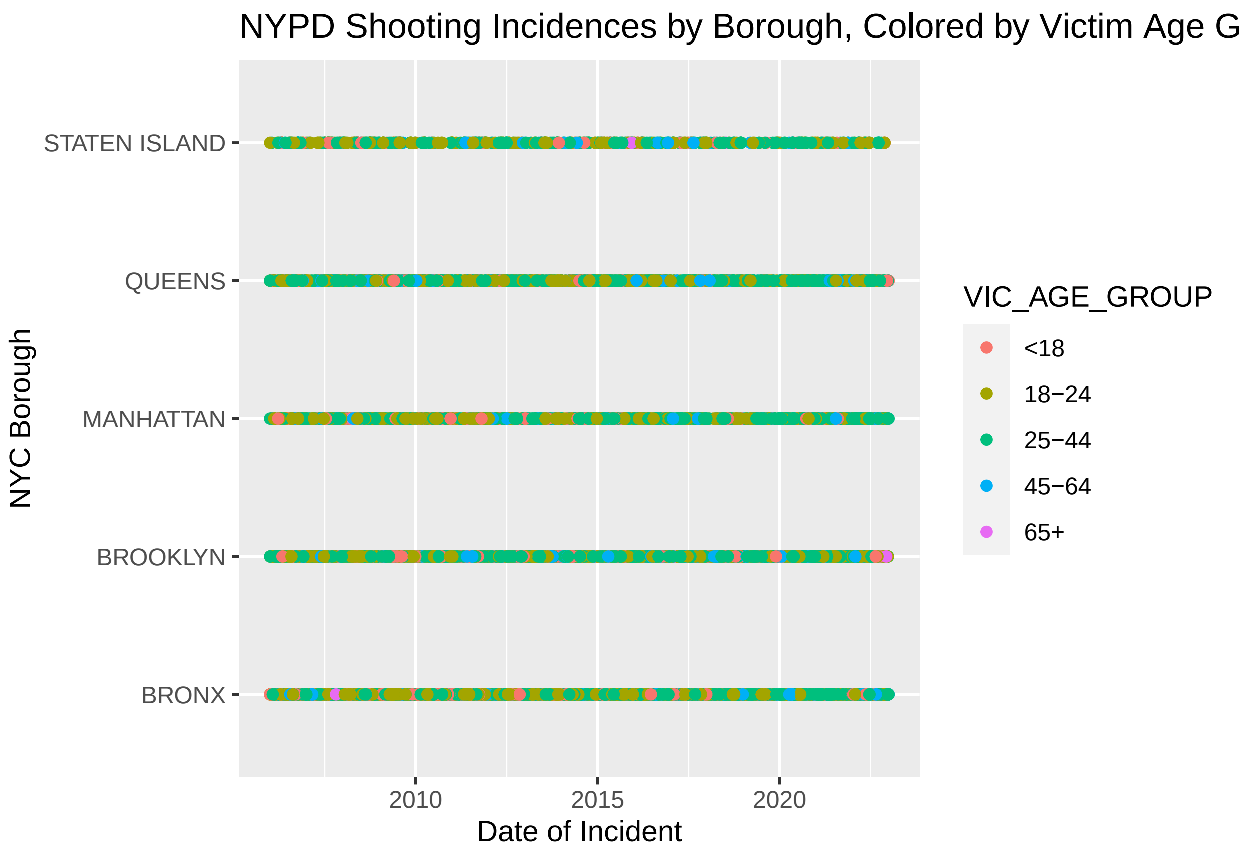
```
##    OCCUR_DATE         PERP_AGE_GROUP     VIC_AGE_GROUP          BORO
##  Min.   :2006-01-01   Length:14122       Length:14122       Length:14122
##  1st Qu.:2009-05-14   Class :character   Class :character   Class :character
##  Median :2013-06-25   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2014-01-14
##  3rd Qu.:2018-11-01
##  Max.   :2022-12-31
```

##Plot

```
#Plot Incidents by Borough Over Time, colored by Perpetrator Age Group
ggplot(nypd_cleaner, aes(x = OCCUR_DATE, y = BORO, color = PERP_AGE_GROUP)) + labs(title = "NYPD Shootin
```



```
#Plot Incidents by Borough over time., colored by Victim's Age Group
ggplot(nypd_cleaner, aes(x = OCCUR_DATE, y = BORO, color = VIC_AGE_GROUP)) + labs(title = "NYPD Shooting
```



##Analysis

The two visualizations show shooting incidences in NYC, separated by borough, and color coded to either the victim's or the perpetrator's age group.

##Model

```
# Create a data frame with character variables
df <- data.frame(VIC_AGE_GROUP = na.omit(nypd_cleaner$VIC_AGE_GROUP), PERP_AGE_GROUP = na.omit(nypd_clea

# Convert the character variables to numeric variable, get rid of NA
df$VIC_AGE_GROUP <- as.numeric(na.omit((gsub("-", ".", df$VIC_AGE_GROUP))))
```

```
## Warning: NAs introduced by coercion
```

```
#Could use this instead of victim, if you wanted a model focused on perpetrator
#df$PERP_AGE_GROUP <- na.omit(as.numeric(gsub("-", ".", df$PERP_AGE_GROUP)))

# Create a linear model with the numeric variables
model <- lm(VIC_AGE_GROUP ~ OCCUR_DATE, data = df)
# Print the model summary
summary(model)
```

```
##
## Call:
## lm(formula = VIC_AGE_GROUP ~ OCCUR_DATE, data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2855 -5.6020  0.2289  1.5730 22.4265
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.830e+01  5.629e-01   32.51   <2e-16 ***
## OCCUR_DATE  3.734e-04  3.464e-05   10.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.383 on 12450 degrees of freedom
##   (1670 observations deleted due to missingness)
## Multiple R-squared:  0.009249,   Adjusted R-squared:  0.009169
## F-statistic: 116.2 on 1 and 12450 DF,  p-value: < 2.2e-16
```

##Statement of Bias One major potential source of bias for me is that while I am looking at victim and perpetrator ages, I am in the 18-24 age group. If asked before this assignment, I would have presumed my age group was the most consistently the perpetrator and not the victim. I did not take direct measures to mitigate bias, but made sure after the fact that I gave each age group a fair view. I also recognize that the age groups vary greatly in size, as each range in age is not the same. Perhaps that is bias from the source of the data collection.