NYDP Shooting Incident Data

Khoa Bui

03/05/2024

Dataset Description

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity. Please refer to the attached data footnotes for additional information about this dataset.

Source: NYPD Shooting Incident

Step 0: Install And/Or Import Libraries

(Optional): If you don't have any of these packages installed yet, uncomment these lines below and run it

(Required): I used extra library called "dplyr", so please make sure you at least install that packages before knit my file.

```
# install.packages("tidyverse")
# install.packages("lubridate")
# install.packages("ggplot2")
# install.packages("dplyr")
library(dplyr)
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Step 1: Import Dataset

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

# read_cvs is used because it being used to read comma seperated values file.
shootingData = read_csv(dataUrl)
```

```
## 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.
```

Step 1.5: How to read dataset

Eliminated some column because didn't need to use. The definition of each in-use column is included but please feel free to explore the rest using the NYPD Shooting Incident website provided above, if needed.

Row Description

- Each **row** in this dataset is presenting unique **shooting incident**.

Column Description

- **INCIDENT_KEY**: (dbl) Unique incident ID assigned for each incident
- **OCCUR_DATE**: (chr) Date of shooting incident in mm/dd/yyyy
- **OCCUR_TIME**: (time) Time of the shooting incident in hh/mm/ss using 24hours system
- **BORO**: (chr) Borough where the shooting incident occurred
- **STATISTICAL_MURDER_FLAG**: (lgl) True/ False if the system became a murder
- **PERP_AGE_GROUP**: (chr) Perpetrator's age group
- **PERP_SEX**: (chr) Perpetrator's sex identification
- **PERP_RACE**: (chr) Perpetrator's race identification
- **VIC_AGE_GROUP**: (chr) Victim's age group
- **VIC_SEX**: (chr) Victim's sex identification
- **VIC_RACE**: (chr) Victim's race

Step 2: Tidy and Transform Data

Eliminating these column from the dataset: LOC_OF_OCCUR_DESC, PRECINCT, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat.

```
# Remove unused column from dataset
shootingData <- shootingData %>% select(-c(
                      LOC_OF_OCCUR_DESC,
                      PRECINCT,
                      JURISDICTION_CODE,
                      LOC_CLASSFCTN_DESC,
                      LOCATION_DESC,
                      X_COORD_CD,
                      Y_COORD_CD,
                      Latitude,
                      Longitude,
                      Lon_Lat
```

```
                      ))
glimpse(shootingData) # Print dataset after removed unused column
```

```
## Rows: 27,312
## Columns: 11
## $ INCIDENT_KEY            <dbl> 228798151, 137471050, 147998800, 146837977, 58~
## $ OCCUR_DATE              <chr> "05/27/2021", "06/27/2014", "11/21/2015", "10/~
## $ OCCUR_TIME              <time> 21:30:00, 17:40:00, 03:56:00, 18:30:00, 22:58~
## $ BORO                    <chr> "QUEENS", "BRONX", "QUEENS", "BRONX", "BRONX",~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, ~
## $ PERP_AGE_GROUP          <chr> NA, NA, NA, NA, "25-44", NA, NA, NA, NA, "25-4~
## $ PERP_SEX                <chr> NA, NA, NA, NA, "M", NA, NA, NA, NA, "M", NA, ~
## $ PERP_RACE               <chr> NA, NA, NA, NA, "BLACK", NA, NA, NA, NA, "BLAC~
## $ VIC_AGE_GROUP           <chr> "18-24", "18-24", "25-44", "<18", "45-64", "25~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "WHITE", "WHITE HISPANIC", "~
```

Since **OCCUR_DATE**: (Char) Date of shooting incident in mm/dd/yyyy, we should convert it to data data type instead

```
shootingData$OCCUR_DATE <- mdy(shootingData$OCCUR_DATE)
```

We know that **PERP_AGE_GROUP** and **VIC_AGE_GROUP** is having data type of "chr", but we also want to remove all data that doesn't make senes, so we have to manually cross checking with the actual excel file to remove it for now

```
# Remove error values in dataset
shootingData = subset(shootingData, PERP_AGE_GROUP!="1020" & PERP_AGE_GROUP!="224" & PERP_AGE_GROUP!="9
```

Key observations on data type conversion are:

- **OCCUR_DATE**: Will be used to get Year.
- **BORO**: Will be treated as a factor.
- **PERP_AGE_GROUP**: Will be treated as a factor.
- **PERP_SEX**: Will be treated as a factor.
- **PERP_RACE**: Will be treated as a factor.
- **VIC_AGE_GROUP**: Will be treated as a factor.
- **VIC_SEX**: Will be treated as a factor.
- **VIC_RACE**: Will be treated as a factor.
- **STATISTICAL_MURDER_FLAG**: Will be treated as a factor

Unknown/ Missing Value

The reason behind there is missing value in **PERP_AGE_GROUP**, **PERP_SEX**, **PERP_RACE** is because these case is till a cold case/ unsolved case. In other to ensure our data is correct, I will avoid using these column, by avoid doing study on PERP AGE, SEX, and RACE since to me, it is not enough data to study. Instead, I will study how many unsolved case, how long the case have been unsolved for, and all the related things. To do so, I will treat all "NA" or "UNKNOWN" as "Unknown"

```r
# Tidy and transform data
shootingData = shootingData %>%
  replace_na(list(PERP_AGE_GROUP = "Unknown", PERP_SEX = "Unknown", PERP_RACE = "Unknown"))
shootingData <- shootingData %>%
  mutate_at(c("PERP_AGE_GROUP","PERP_SEX","PERP_RACE"), list(~ifelse(. == "(null)", "Unknown", .)))
shootingData <- shootingData %>%
  mutate_at(c("PERP_AGE_GROUP","PERP_RACE","VIC_AGE_GROUP",
              "VIC_RACE"), list(~ifelse(. == "UNKNOWN", "Unknown", .)))
shootingData <- shootingData %>%
  mutate_at(c("PERP_SEX","VIC_SEX"), list(~ifelse(. == "U", "Unknown", .)))

# Add OCCUR_YEAR to the dataset
shootingData$OCCUR_YEAR <- year(shootingData$OCCUR_DATE)

# Transform to factor
shootingData$BORO = as.factor(shootingData$BORO)
shootingData$PERP_AGE_GROUP = as.factor(shootingData$PERP_AGE_GROUP)
shootingData$PERP_SEX = as.factor(shootingData$PERP_SEX)
shootingData$PERP_RACE = as.factor(shootingData$PERP_RACE)
shootingData$VIC_AGE_GROUP = as.factor(shootingData$VIC_AGE_GROUP)
shootingData$VIC_SEX = as.factor(shootingData$VIC_SEX)
shootingData$VIC_RACE = as.factor(shootingData$VIC_RACE)
shootingData$STATISTICAL_MURDER_FLAG <- factor(shootingData$STATISTICAL_MURDER_FLAG)
```

Summary Statistic

```r
summary(shootingData)
```

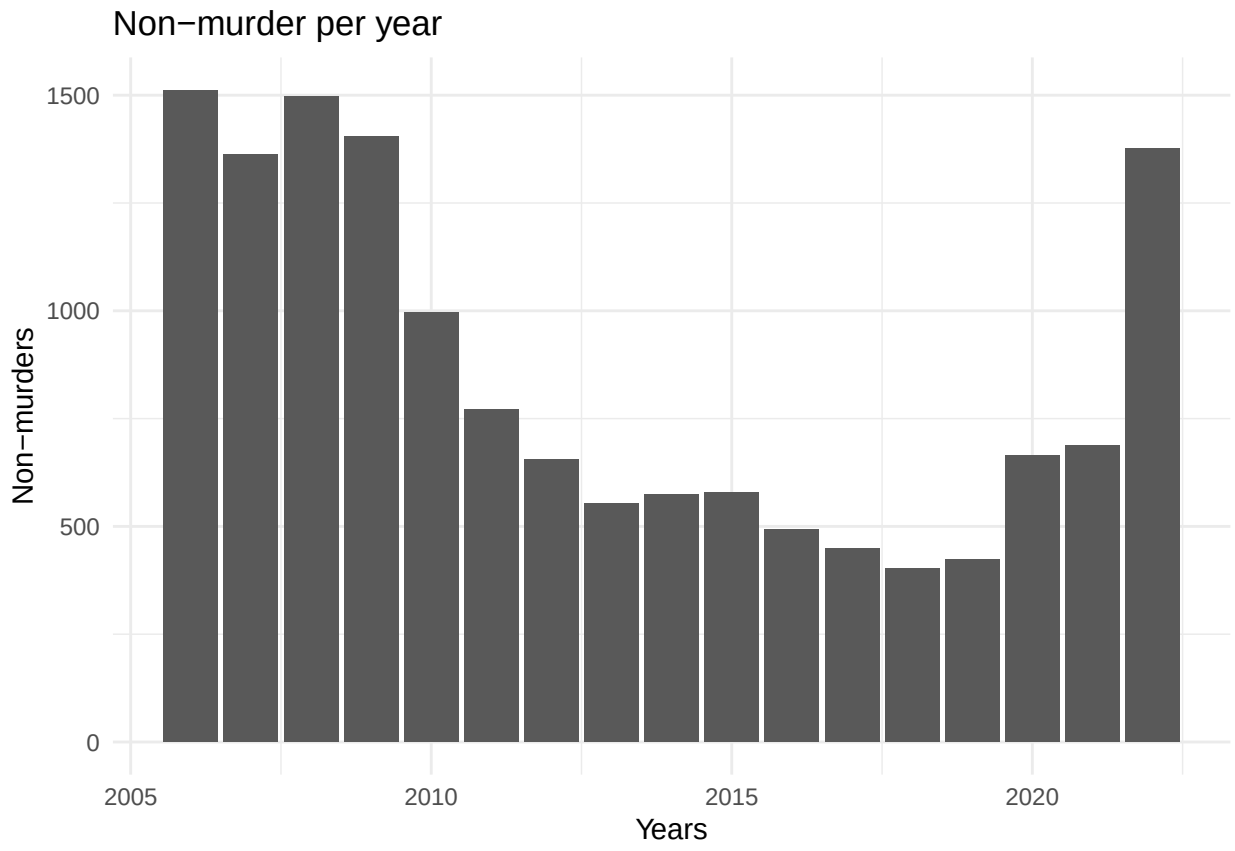
```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:17964
##  1st Qu.: 49856480   1st Qu.:2008-08-05   Class1:hms
##  Median : 81781918   Median :2011-11-18   Class2:difftime
##  Mean   :112646564   Mean   :2013-05-11   Mode  :numeric
##  3rd Qu.:178651739   3rd Qu.:2018-04-26
##  Max.   :261190187   Max.   :2022-12-31
##
##             BORO      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP    PERP_SEX
##  BRONX        :5423   FALSE:14404             <18    :1591   F       :  424
##  BROOKLYN     :6641   TRUE : 3560             18-24  :6221   M       :15435
##  MANHATTAN    :2541                           25-44  :5687   Unknown: 2105
##  QUEENS       :2728                           45-64  : 617
##  STATEN ISLAND: 631                           65+    :  60
##                                               Unknown:3788
##
##                            PERP_RACE     VIC_AGE_GROUP     VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    :2027   F       : 1922
##  ASIAN / PACIFIC ISLANDER      :  154   18-24  :6517   M       :16034
##  BLACK                         :11430   25-44  :7937   Unknown:    8
##  BLACK HISPANIC                : 1314   45-64  :1290
##  Unknown                       : 2442   65+    : 137
##  WHITE                         :  283   Unknown:  56
##  WHITE HISPANIC                : 2339
##                            VIC_RACE       OCCUR_YEAR
```

```
##  AMERICAN INDIAN/ALASKAN NATIVE:    8   Min.   :2006
##  ASIAN / PACIFIC ISLANDER      :  307   1st Qu.:2008
##  BLACK                         :12250   Median :2011
##  BLACK HISPANIC                : 1800   Mean   :2013
##  Unknown                       :   48   3rd Qu.:2018
##  WHITE                         :  552   Max.   :2022
##  WHITE HISPANIC                : 2999
```

Step 3: Visualizations and Analysis

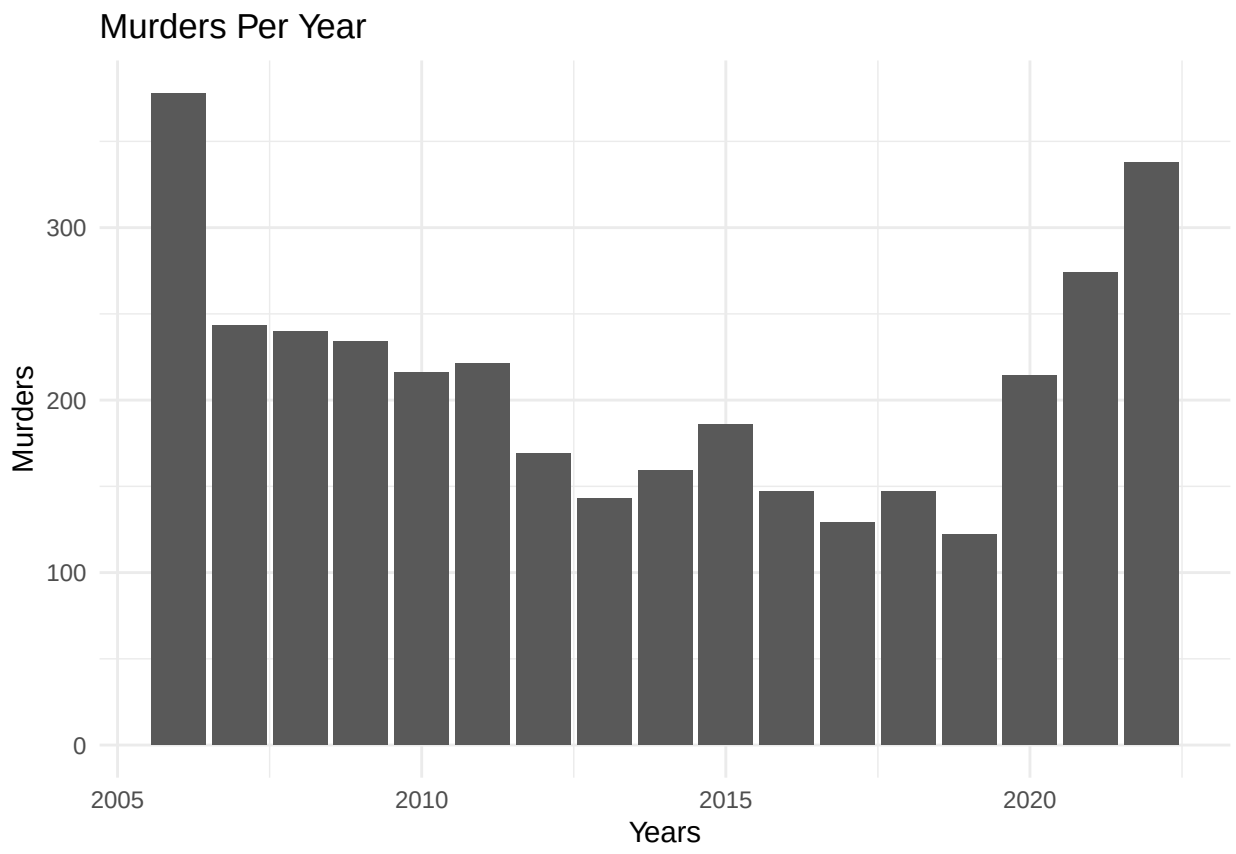
1. Non-murder by year and murders by year

```
shootingData %>%
  filter(STATISTICAL_MURDER_FLAG == FALSE) %>%
  ggplot(aes(x = OCCUR_YEAR)) +
  geom_bar() +
  labs(title = "Non-murder per year",
       x = "Years",
       y = "Non-murders") +
  theme_minimal()
```




```
shootingData %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  ggplot(aes(x = OCCUR_YEAR)) +
```

```
  geom_bar() +
  labs(title = "Murders Per Year",
       x = "Years",
       y = "Murders") +
  theme_minimal()
```




```
table(shootingData$OCCUR_YEAR, shootingData$STATISTICAL_MURDER_FLAG)
```

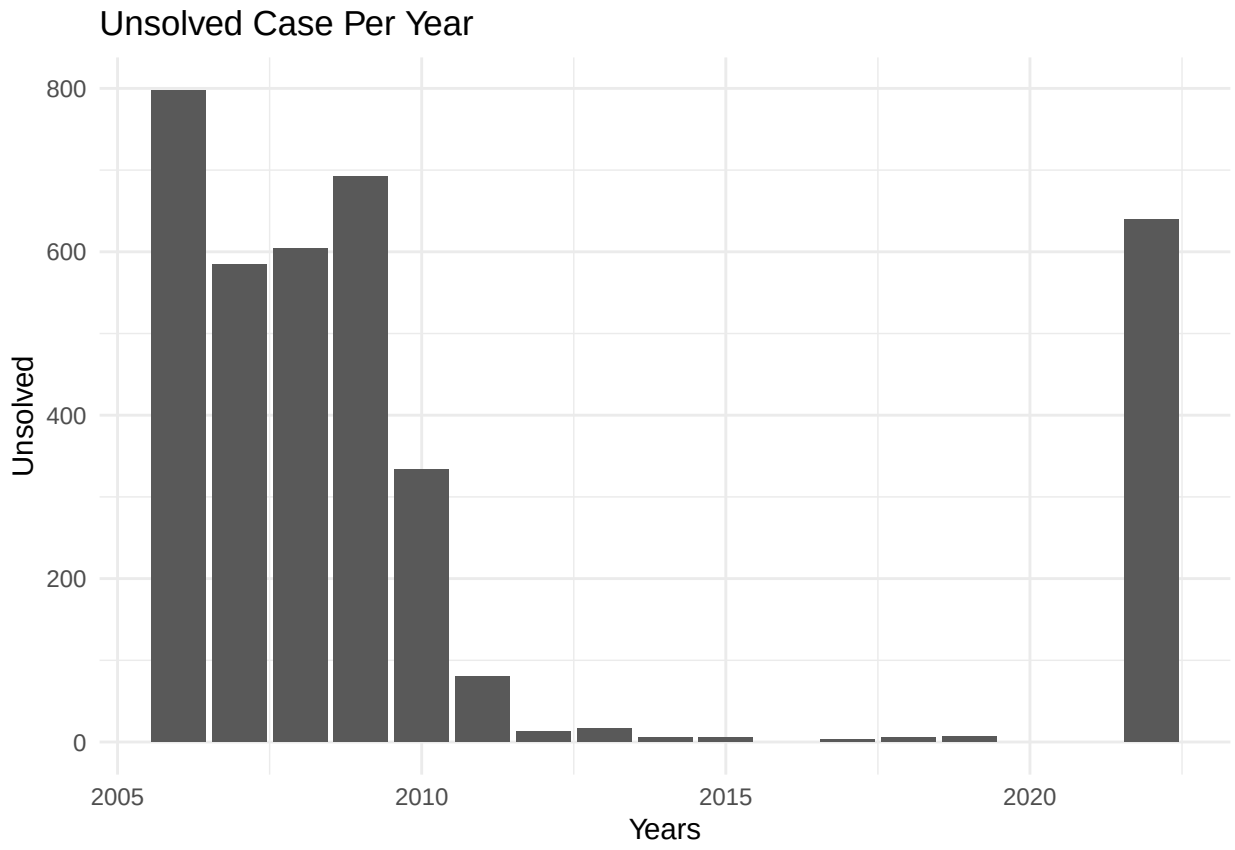
```
##
##        FALSE TRUE
##   2006  1512  378
##   2007  1363  243
##   2008  1497  240
##   2009  1405  234
##   2010   997  216
##   2011   771  221
##   2012   656  169
##   2013   554  143
##   2014   574  159
##   2015   579  186
##   2016   493  147
##   2017   449  129
##   2018   402  147
##   2019   423  122
##   2020   664  214
```

```
##    2021   688  274
##    2022  1377  338
```

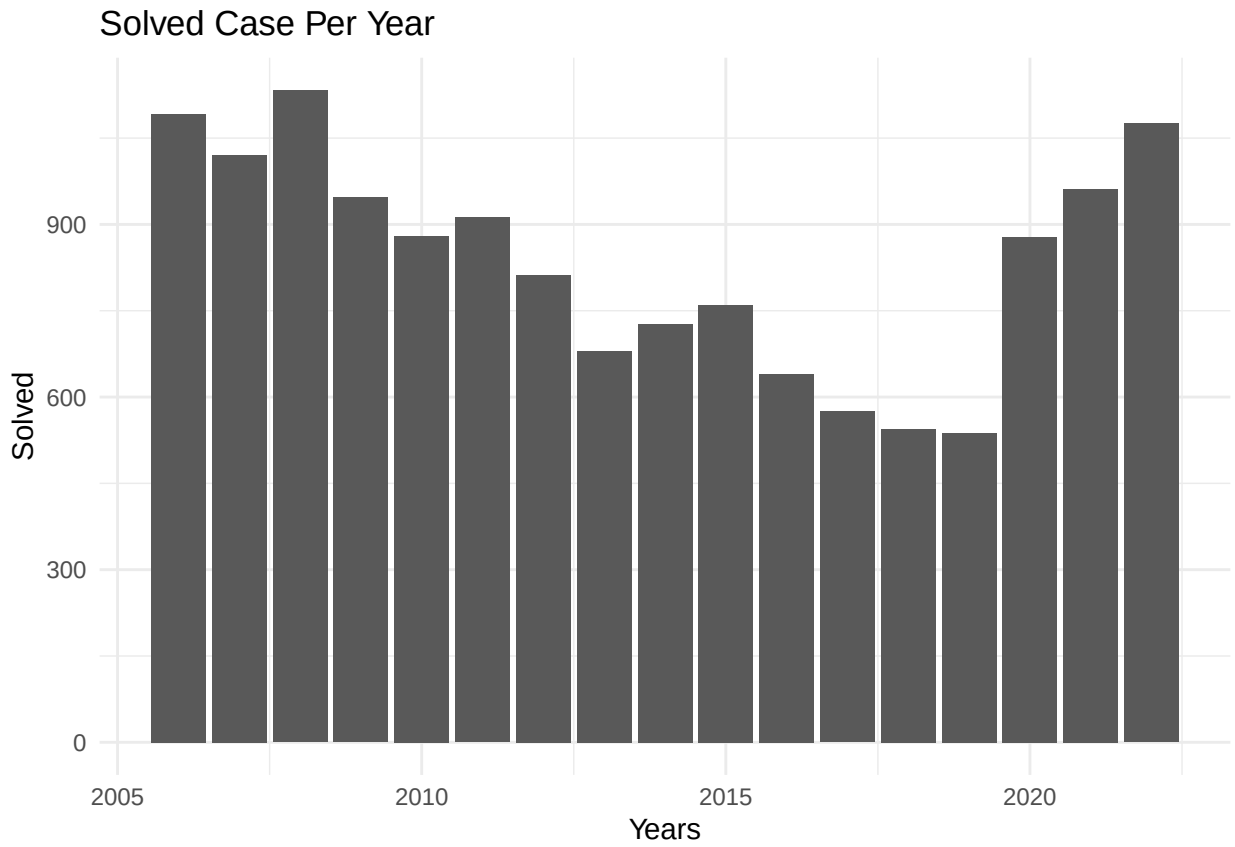
Based on graph and data table above, since murder is way less than non-murder, the rate of change is a bit different, but in the safe zone to say it is align with each other.

2. Solved vs Unsolved Case by Year Unsolved Case

```
shootingData %>%
  filter(PERP_AGE_GROUP == "Unknown") %>%
  ggplot(aes(x = OCCUR_YEAR)) +
  geom_bar() +
  labs(title = "Unsolved Case Per Year",
       x = "Years",
       y = "Unsolved") +
  theme_minimal()
```




```
shootingData %>%
  filter(PERP_AGE_GROUP != "Unknown") %>%
  ggplot(aes(x = OCCUR_YEAR)) +
  geom_bar() +
  labs(title = "Solved Case Per Year",
       x = "Years",
       y = "Solved") +
  theme_minimal()
```




```
# I used PERP_AGE_GROUP because if the value is Unknown
# that mean the case is not solved/ missing data
table(shootingData$OCCUR_YEAR, shootingData$PERP_AGE_GROUP)
```

```
##
##        <18 18-24 25-44 45-64 65+ Unknown
##   2006 138   498   425    29   2     798
##   2007 134   479   372    26  10     585
##   2008 166   582   346    36   3     604
##   2009 102   461   357    25   2     692
##   2010 107   426   310    34   3     333
##   2011 136   476   256    43   1      80
##   2012  85   392   305    24   6      13
##   2013  58   342   255    20   5      17
##   2014  78   347   274    22   6       6
##   2015  85   334   314    21   6       5
##   2016  53   261   292    33   1       0
##   2017  47   220   270    36   2       3
##   2018  49   201   256    37   1       5
##   2019  53   191   254    37   3       7
##   2020  77   298   428    72   3       0
##   2021  95   336   488    40   3       0
##   2022 128   377   485    82   3     640
```

Based on these data above, is it safe to say that, the further the year, the more case didn't solved, isn't it? Maybe. But I also have another theory. Maybe the data for these years was before when we actually collect

data, so these data wasn't collected properly. This is something we need to dive deep to ensure we known our dataset inside out.

3. Building logistic regression model to predict if the victim will be survied Logistic regression is an instance of classification technique that you can use to predict a qualitative response. I will use logistic regression models to estimate the probability that a murder case belongs to a particular victim's profile.

The output shows the coefficients, their standard errors, the z-values, and the associated p-values.The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

```
# Logistics Regression
glm.fit <- glm(STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX,
               data = shootingData, family = binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP + VIC_RACE +
##     VIC_SEX, family = binomial, data = shootingData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9986  -0.6893  -0.6158  -0.5350   2.1157
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -12.76308  114.10228  -0.112  0.91094
## VIC_AGE_GROUP18-24                0.30495    0.07224   4.221 2.43e-05 ***
## VIC_AGE_GROUP25-44                0.55537    0.07006   7.927 2.25e-15 ***
## VIC_AGE_GROUP45-64                0.66478    0.09183   7.239 4.51e-13 ***
## VIC_AGE_GROUP65+                  0.90917    0.19774   4.598 4.27e-06 ***
## VIC_AGE_GROUPUnknown              0.57580    0.34918   1.649  0.09915 .
## VIC_RACEASIAN / PACIFIC ISLANDER 11.36796  114.10233   0.100  0.92064
## VIC_RACEBLACK                    11.05406  114.10226   0.097  0.92282
## VIC_RACEBLACK HISPANIC           10.90933  114.10227   0.096  0.92383
## VIC_RACEUnknown                  10.49549  114.10321   0.092  0.92671
## VIC_RACEWHITE                    11.41747  114.10230   0.100  0.92029
## VIC_RACEWHITE HISPANIC           11.21451  114.10227   0.098  0.92171
## VIC_SEXM                         -0.16254    0.05909  -2.751  0.00595 **
## VIC_SEXUnknown                   -0.32960    1.12749  -0.292  0.77003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17887  on 17963  degrees of freedom
## Residual deviance: 17723  on 17950  degrees of freedom
## AIC: 17751
##
## Number of Fisher Scoring iterations: 11
```

Notable Findings: The age bracket of the **affected individual** appears pivotal in predicting their chances of surviving a shooting incident. Notably, individuals within the **< 18** and **18-24** age categories exhibit the

highest likelihood of surviving such incidents. Conversely, the probability of survival steadily **diminishes** across successive age groups. Particularly grim are the outcomes for shootings involving individuals **aged 65 and above**, where fatalities are prevalent.

Area to dive deep

- Are there additional variables besides age that could serve as indicators for the fatality of a shooting incident, such as the victim's sex?

Step 4: Conclusion && Identify Bias

Conclusion

We explored how many unsolved case, among with if the shot is fatal what is the varibles to make that shot fatal.

Room for bias

One of the most common bias is that if the victim is in one race, normally the perp will also be in the same Race. There is also another bias that if there is a gun shot, it should be fatal. For the scope of this project, I didn't touch perp race for that reason, but dive deep into if the shot fatal or not.