NYPD Shooting Incidents

Thomas Bohn

2023-04-20

Introduction

Data Science Process

The following report follows the Data Science Process from beginning to end, ensuring there is a discussion on the following areas in the flow:

- Import
- Tidy
- Transform
- Visualize
- Model
- Communicate

Overview of Report Structure

The following report will contain the following sections:

- **Background**: Why should I care?
- **Data Source**: Where is your data from?
- **Tidying and Transform the Data**: How has the data been cleaned and transformed?
- **Analysis and Visualizations**: What does it tell you?
- **Models & Conclusions**: What do you conclude?
- **Review of Bias**: How could you be wrong?

By including comprehensive details in a well structured document, the results and findings of this analysis should be reproducible for any user.

R Libraries Utilized

The analysis in this report will utilize the following libraries in R for Data Analysis:

```
library(tidyverse)
library(lubridate)
library(tinytex)
```

Background

What is an NYPD Shooting Incident

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.

Each incident is described with the following attributes:

- **INCIDENT_KEY**: Randomly generated persistent ID for each arrest
- **OCCUR_DATE**: Exact date of the shooting incident
- **OCCUR_TIME**: Exact time of the shooting incident
- **BORO**: Borough where the shooting incident occurred
- **PRECINCT**: Precinct where the shooting incident occurred
- **JURISDICTION_CODE**: Jurisdiction where the shooting incident occurred. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions
- **LOCATION_DESC**: Location of the shooting incident
- **STATISTICAL_MURDER_FLAG**: Shooting resulted in the victim's death which would be counted as a murder
- **PERP_AGE_GROUP**: Perpetrator's age within a category
- **PERP_SEX**: Perpetrator's sex description
- **PERP_RACE**: Perpetrator's race description
- **VIC_AGE_GROUP**: Victim's age within a category
- **VIC_SEX**: Victim's sex description
- **VIC_RACE**: Victim's race description
- **X_COORD_CD**: Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
- **Y_COORD_CD**: Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
- **Latitude**: Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
- **Longitude**: Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
- **Lon_Lat**: Longitude and Latitude Coordinates for mapping

Data Source

Source of Data

The data used for the analysis is sourced from https://catalog.data.gov/dataset and provided by **NYC OpenData**. The data source is described as:

> List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.
>
> 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.

It can be found the following the following Github URL: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

Import the Data to R

```
#Build URLs to access the data from the web
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
```

```
#Read in the data to datasets
nypd <- read_csv(url)
```

```
#Preview the dataset
head(nypd)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
## 1    228798151 05/27/2021 21:30      QUEENS   <NA>                   105
## 2    137471050 06/27/2014 17:40      BRONX    <NA>                    40
## 3    147998800 11/21/2015 03:56      QUEENS   <NA>                   108
## 4    146837977 10/09/2015 18:30      BRONX    <NA>                    44
## 5     58921844 02/19/2009 22:58      BRONX    <NA>                    47
## 6    219559682 10/21/2020 21:36      BROOKLYN <NA>                    81
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

Tidying and Transform the Data

The following outlines how the data was modified to be tidy and transformed to contain variables for further analysis. This section contains:

- A summary of the data
- Clean up of the dataset by changing appropriate variables to factors, updating date types, and getting rid of any columns not needed
- Transforming the data to add useful variables and derived elements
- Summary of the data to be sure there is no missing data

Data Summerization

```
#Preview the data set in R
glimpse(nypd)
```

```
## Rows: 27,312
## Columns: 21
## $ 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",~
## $ LOC_OF_OCCUR_DESC       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ PRECINCT                <dbl> 105, 40, 108, 44, 47, 81, 114, 81, 105, 101, 2~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 2~
## $ LOC_CLASSFCTN_DESC      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ LOCATION_DESC           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "MULTI DWE~
## $ 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", "~
## $ X_COORD_CD              <dbl> 1058925.0, 1005028.0, 1007667.9, 1006537.4, 10~
## $ Y_COORD_CD              <dbl> 180924.0, 234516.0, 209836.5, 244511.1, 262189~
## $ Latitude                <dbl> 40.66296, 40.81035, 40.74261, 40.83778, 40.886~
## $ Longitude               <dbl> -73.73084, -73.92494, -73.91549, -73.91946, -7~
## $ Lon_Lat                 <chr> "POINT (-73.73083868899994 40.662964620000025)~
```

```
#Summary of the NYPD Shootings Incidents Dataset
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
```

```
#Show the column names of the columns in NYPD Shootings Incidents data sets
nypd_cols <- colnames(nypd)
nypd_cols <- str_to_lower(nypd_cols)
nypd_cols
```

```
##  [1] "incident_key"            "occur_date"
##  [3] "occur_time"              "boro"
##  [5] "loc_of_occur_desc"       "precinct"
##  [7] "jurisdiction_code"       "loc_classfctn_desc"
##  [9] "location_desc"           "statistical_murder_flag"
## [11] "perp_age_group"          "perp_sex"
## [13] "perp_race"               "vic_age_group"
## [15] "vic_sex"                 "vic_race"
## [17] "x_coord_cd"              "y_coord_cd"
## [19] "latitude"                "longitude"
## [21] "lon_lat"
```

Profile the Data

```
#Profile some base data variables in the data set
nypd %>% count(BORO)
```

```
## # 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
```

```
nypd %>% count(JURISDICTION_CODE)
```

```
## # A tibble: 4 x 2
##   JURISDICTION_CODE     n
##               <dbl> <int>
## 1                 0 22809
## 2                 1    74
## 3                 2  4427
## 4                NA     2
```

```
nypd %>% count(STATISTICAL_MURDER_FLAG)
```

```
## # A tibble: 2 x 2
##   STATISTICAL_MURDER_FLAG     n
##   <lgl>                   <int>
## 1 FALSE                   22046
## 2 TRUE                     5266
```

```
nypd %>% count(PERP_AGE_GROUP)
```

```
## # A tibble: 11 x 2
##    PERP_AGE_GROUP     n
##    <chr>          <int>
##  1 (null)           640
##  2 1020               1
##  3 18-24           6222
##  4 224                1
##  5 25-44           5687
##  6 45-64            617
##  7 65+               60
##  8 940                1
##  9 <18             1591
## 10 UNKNOWN         3148
## 11 <NA>            9344
```

```
nypd %>% count(PERP_SEX)
```

```
## # A tibble: 5 x 2
##   PERP_SEX     n
##   <chr>    <int>
## 1 (null)     640
## 2 F          424
## 3 M        15439
## 4 U         1499
## 5 <NA>      9310
```

```
nypd %>% count(PERP_RACE)
```

```
## # A tibble: 9 x 2
##   PERP_RACE                          n
##   <chr>                          <int>
## 1 (null)                           640
## 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
## 9 <NA>                            9310
```

```
nypd %>% count(VIC_AGE_GROUP)
```

```
## # A tibble: 7 x 2
##   VIC_AGE_GROUP     n
##   <chr>         <int>
## 1 1022              1
## 2 18-24         10086
## 3 25-44         12281
## 4 45-64          1863
## 5 65+             181
## 6 <18            2839
## 7 UNKNOWN          61
```

```
nypd %>% count(VIC_SEX)
```

```
## # A tibble: 3 x 2
##   VIC_SEX     n
##   <chr>   <int>
## 1 F        2615
## 2 M       24686
## 3 U          11
```

```
nypd %>% count(VIC_RACE)
```

```
## # A tibble: 7 x 2
##   VIC_RACE                           n
##   <chr>                          <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE    10
## 2 ASIAN / PACIFIC ISLANDER         404
## 3 BLACK                          19439
## 4 BLACK HISPANIC                  2646
## 5 UNKNOWN                           66
## 6 WHITE                            698
## 7 WHITE HISPANIC                  4049
```

```
nypd %>% count(LOCATION_DESC)
```

```
## # A tibble: 41 x 2
##    LOCATION_DESC         n
##    <chr>             <int>
##  1 (null)             977
##  2 ATM                  1
##  3 BANK                 3
##  4 BAR/NIGHT CLUB     628
##  5 BEAUTY/NAIL SALON  112
##  6 CANDY STORE          7
##  7 CHAIN STORE          5
##  8 CHECK CASH           1
##  9 CLOTHING BOUTIQUE   14
## 10 COMMERCIAL BLDG    292
## # i 31 more rows
```

Scope for Initial Tidy

List of initial tidy adjustments to make:

- Adjust Header names to be lower case
- INCIDENT_KEY cast as Int
- OCCUR_DATE parsed as Date and OCCUR_TIME parsed as time
- Mixed Case for Boro and set as Factor
- Map JURISDICTION_CODE to Factor Values
- Map PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE as Factors
- Derive year, month, hour, minute data Drop LOCATION_DESC and Geo information

Data Quality Isseus

The following data issues are observed but will not be changed in the data source:

- 9,344 perp age groups missing (NA)
- 9,310 perp sex missing (NA)
- 9,310 perp race missing (NA)

Since the above values are related to the perpetrator of the shooting, and are relatively similar in size, it can be assumed that NA indicates not perpetrator was identified. These values should still remain in the dataset as they are still valid incidents. They are also different than unknown, where a perpetrator was identified, but the witness or victim could not identify the demographic details.

The following items will remain as NA but will be outside of the factor levels:

- 2 missing jurisdiction codes (NA)
- Perp Age Groups out of domain (value of 1020, 224, 940)

These records will remain, but will be mapped to NA as there is no logic mapping for them.

The following variable will be dropped from the dataset:

- 14,977 location description missing (NA)

There are too many missing values from this attribute to be useful for analysis. It is possible it could be used in the future, if further relationships can be identified that justify why there are so many NA values.

Tidy the NYPD Shootings Incidents Data

```r
#Define the factor levels to use in the tify process
f_sex = c("M", "F", "U")
f_age_group = c("<18", "18-24", "25-44", "45-64", "65+", "Unknown")
f_boro = c("Manhattan", "Brooklyn", "Queens",
           "Bronx", "Staten Island", "Unknown")
f_race = c("White", "Black", "American Indian/Alaskan Native",
           "Asian / Pacific Islander", "Unknown")
f_ethn = c("Hispanic or Latino", "Not Hispanic or Latino", "Unknown")
f_jur = c("Patrol", "Transit", "Housing", "Other")
f_month = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
            "Aug", "Sep", "Oct", "Nov", "Dec")
```

```r
#Tidy the dataset
tidy_nypd <- nypd %>%
  #Adjust Header names to be lower case
  rename_with(tolower) %>%
  #incident_key and precinct cast as int
  mutate(incident_key = as.integer(incident_key)) %>%
  mutate(precinct = as.integer(precinct)) %>%
  #occur_date parsed as date and occur_time parsed as time
  mutate(occur_date = parse_date(occur_date, format = "%m/%d/%Y")) %>%
  mutate(occur_time = parse_time(as.character(occur_time)))  %>%
  #mixed case for boro and set as Factor
  mutate(boro = str_to_title(boro)) %>%
  mutate(boro = factor(boro, levels = f_boro)) %>%
  #Map jurisdiction_code to Factor Values
  #0(Patrol), 1(Transit) and 2(Housing) represent NYPD
  #whilst codes 3 and more represent non NYPD jurisdictions
  mutate(jurisdiction = fct_recode(as.character(jurisdiction_code),
                                    "Patrol" = "0",
                                    "Transit" = "1",
                                    "Housing" = "2"
                                   ),
         .after = jurisdiction_code
   ) %>%
  mutate(jurisdiction = factor(jurisdiction, levels = f_jur)) %>%
  #Map perp_age_group, perp_sex, perp_race, vic_age_group,
  #vic_sex, vic_race as Factors
  mutate(perp_sex = factor(perp_sex, levels = f_sex)) %>%
  mutate(vic_sex = factor(vic_sex, levels = f_sex)) %>%
  mutate(perp_age_group = factor(str_to_title(perp_age_group),
                                 levels = f_age_group)) %>%
  mutate(vic_age_group = factor(str_to_title(vic_age_group),
                                levels = f_age_group)) %>%
  #Drop columns not needed for analysis
  select(-c("location_desc", "x_coord_cd", "y_coord_cd",
            "latitude", "longitude", "lon_lat"))
```

Transform the NYPD Shootings Incidents Data

```
tidy_nypd <- tidy_nypd %>%
  #Map month, year, hour
  mutate(occur_month = month(occur_date, label = TRUE, abbr = TRUE),
         .after = occur_date) %>%
  mutate(occur_year = as.integer(year(occur_date)), .after = occur_date) %>%
  mutate(occur_year_month = format(as.Date(occur_date), "%Y-%m"),
         .after = occur_time) %>%
  mutate(occur_hour = hour(occur_time), .after = occur_time) %>%
  #Map perp_race to mixed case
  mutate(perp_race = str_to_title(perp_race)) %>%
  #Derive perp ethnicity from perp race by
  #consolidating to Hispanic and not Hispanic
  mutate(perp_ethn = fct_collapse(perp_race,
            "Unknown" = c("Unknown"),
            "Not Hispanic or Latino" = c("White", "Black",
                                         "American Indian/Alaskan Native",
                                         "Asian / Pacific Islander"),
            "Hispanic or Latino" = c("White Hispanic", "Black Hispanic")
          ),
         .after = perp_race
   ) %>%
  #Turn perp ethnicity into a factor with levels
  mutate(perp_ethn = factor(perp_ethn, levels = f_ethn)) %>%
  #Remove Hispanic from perp race
  mutate(perp_race2 = fct_collapse(perp_race,
            "White" = c("White", "White Hispanic"),
            "Black" = c("Black", "Black Hispanic")
          ),
         .after = perp_race
   ) %>%
  #Turn perp race into a factor with levels
  mutate(perp_race2 = factor(perp_race2, levels = f_race)) %>%
  #Rename original field as diversity group
  rename(perp_diversity_group = perp_race, perp_race = perp_race2) %>%
  #Map vic_race to mixed case
  mutate(vic_race = str_to_title(vic_race)) %>%
  #Derive victim ethnicity from victim race by
  #consolidating to Hispanic and not Hispanic
  mutate(vic_ethn = fct_collapse(vic_race,
            "Unknown" = c("Unknown"),
            "Not Hispanic or Latino" = c("White", "Black",
                                         "American Indian/Alaskan Native",
                                         "Asian / Pacific Islander"),
            "Hispanic or Latino" = c("White Hispanic", "Black Hispanic")
          ),
         .after = vic_race
   ) %>%
  #Turn victim ethnicity into a factor with levels
  mutate(vic_ethn = factor(vic_ethn, levels = f_ethn)) %>%
  #Remove Hispanic from victim race
  mutate(vic_race2 = fct_collapse(vic_race,
            "White" = c("White", "White Hispanic"),
            "Black" = c("Black", "Black Hispanic")
          ),
         .after = vic_race
   ) %>%
  #Turn victim race into a factor with levels
  mutate(vic_race2 = factor(vic_race2, levels = f_race)) %>%
  #Rename original field as diversity group
  rename(vic_diversity_group = vic_race, vic_race = vic_race2)
```

Summary of the Tidyed and Transformed Data

```
#Review the overall table after the tidy function
head(tidy_nypd, n = 5)
```

```
## # A tibble: 5 x 24
##   incident_key occur_date occur_year occur_month occur_time occur_hour
##          <int> <date>          <int> <ord>       <time>          <int>
## 1    228798151 2021-05-27       2021 May         21:30              21
## 2    137471050 2014-06-27       2014 Jun         17:40              17
## 3    147998800 2015-11-21       2015 Nov         03:56               3
## 4    146837977 2015-10-09       2015 Oct         18:30              18
## 5     58921844 2009-02-19       2009 Feb         22:58              22
## # i 18 more variables: occur_year_month <chr>, boro <fct>,
## #   loc_of_occur_desc <chr>, precinct <int>, jurisdiction_code <dbl>,
## #   jurisdiction <fct>, loc_classfctn_desc <chr>,
## #   statistical_murder_flag <lgl>, perp_age_group <fct>, perp_sex <fct>,
## #   perp_diversity_group <chr>, perp_race <fct>, perp_ethn <fct>,
## #   vic_age_group <fct>, vic_sex <fct>, vic_diversity_group <chr>,
## #   vic_race <fct>, vic_ethn <fct>
```

```
#Summary of the NYPD Shootings Incidents Dataset
summary(tidy_nypd)
```

```
##   incident_key         occur_date           occur_year    occur_month   
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :2006   Jul    : 3238  
##  1st Qu.: 63860880   1st Qu.:2009-07-18   1st Qu.:2009   Aug    : 3156  
##  Median : 90372218   Median :2013-04-29   Median :2013   Jun    : 2829  
##  Mean   :120860536   Mean   :2014-01-06   Mean   :2013   Sep    : 2572  
##  3rd Qu.:188810230   3rd Qu.:2018-10-15   3rd Qu.:2018   May    : 2571  
##  Max.   :261190187   Max.   :2022-12-31   Max.   :2022   Oct    : 2279  
##                                                          (Other):10667  
##   occur_time         occur_hour    occur_year_month              boro     
##  Length:27312      Min.   : 0.00   Length:27312       Manhattan    : 3572  
##  Class1:hms        1st Qu.: 3.00   Class :character   Brooklyn     :10933  
##  Class2:difftime   Median :15.00   Mode  :character   Queens       : 4094  
##  Mode  :numeric    Mean   :12.22                      Bronx        : 7937  
##                    3rd Qu.:20.00                      Staten Island:  776  
##                    Max.   :23.00                      Unknown      :    0  
##
```

```
##  loc_of_occur_desc     precinct      jurisdiction_code  jurisdiction
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Patrol :22809
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Transit:   74
##  Mode  :character   Median : 68.00   Median :0.0000    Housing: 4427
##                     Mean   : 65.64   Mean   :0.3269    Other  :    0
##                     3rd Qu.: 81.00   3rd Qu.:0.0000    NA's   :    2
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  loc_classfctn_desc statistical_murder_flag perp_age_group perp_sex
##  Length:27312       Mode :logical           <18    :1591   M   :15439
##  Class :character   FALSE:22046             18-24  :6222   F   :  424
##  Mode  :character   TRUE :5266              25-44  :5687   U   : 1499
##                                             45-64  : 617   NA's: 9950
##                                             65+    :  60
##                                             Unknown:3148
##                                             NA's   :9987
##  perp_diversity_group                           perp_race
##  Length:27312         White                           : 2624
##  Class :character     Black                           :12746
##  Mode  :character     American Indian/Alaskan Native:    2
##                       Asian / Pacific Islander      :  154
##                       Unknown                       : 1836
##                       NA's                          : 9950
##
##                  perp_ethn     vic_age_group   vic_sex   vic_diversity_group
##  Hispanic or Latino    : 3655   <18    : 2839   M:24686   Length:27312
##  Not Hispanic or Latino:11871   18-24  :10086   F: 2615   Class :character
##  Unknown               : 1836   25-44  :12281   U:   11   Mode  :character
##  NA's                  : 9950   45-64  : 1863
##                                 65+    :  181
##                                 Unknown:   61
##                                 NA's   :    1
##                           vic_race                         vic_ethn
##  White                         : 4747   Hispanic or Latino    : 6695
##  Black                         :22085   Not Hispanic or Latino:20551
##  American Indian/Alaskan Native:   10   Unknown               :   66
##  Asian / Pacific Islander      :  404
##  Unknown                       :   66
##
##
```

```
#validate the mapping of diversity group, race, and ethnicity for perps
tidy_nypd %>% count(perp_diversity_group, perp_race, perp_ethn)
```

```
## # A tibble: 9 x 4
##   perp_diversity_group           perp_race                      perp_ethn      n
##   <chr>                          <fct>                          <fct>      <int>
## 1 (Null)                         <NA>                           <NA>         640
## 2 American Indian/Alaskan Native American Indian/Alaskan Native Not Hispa~     2
## 3 Asian / Pacific Islander       Asian / Pacific Islander       Not Hispa~   154
## 4 Black                          Black                          Not Hispa~ 11432
## 5 Black Hispanic                 Black                          Hispanic ~  1314
## 6 Unknown                        Unknown                        Unknown     1836
## 7 White                          White                          Not Hispa~   283
## 8 White Hispanic                 White                          Hispanic ~  2341
## 9 <NA>                           <NA>                           <NA>        9310
```

```
#validate the mapping of diversity group, race, and ethnicity for victims
tidy_nypd %>% count(vic_diversity_group, vic_race, vic_ethn)
```

```
## # A tibble: 7 x 4
##   vic_diversity_group            vic_race                       vic_ethn       n
##   <chr>                          <fct>                          <fct>      <int>
## 1 American Indian/Alaskan Native American Indian/Alaskan Native Not Hispa~    10
## 2 Asian / Pacific Islander       Asian / Pacific Islander       Not Hispa~   404
## 3 Black                          Black                          Not Hispa~ 19439
## 4 Black Hispanic                 Black                          Hispanic ~  2646
## 5 Unknown                        Unknown                        Unknown       66
## 6 White                          White                          Not Hispa~   698
## 7 White Hispanic                 White                          Hispanic ~  4049
```

```
#validate the mapping of jurisdiction code and jurisdiction
tidy_nypd %>% count(jurisdiction_code, jurisdiction)
```

```
## # A tibble: 4 x 3
##   jurisdiction_code jurisdiction     n
##               <dbl> <fct>        <int>
## 1                 0 Patrol       22809
## 2                 1 Transit         74
## 3                 2 Housing       4427
## 4                NA <NA>             2
```

```
#validate the mapping of year
tidy_nypd %>% count(occur_year)
```

```
## # A tibble: 17 x 2
##    occur_year     n
##         <int> <int>
##  1       2006  2055
##  2       2007  1887
##  3       2008  1959
##  4       2009  1828
##  5       2010  1912
##  6       2011  1939
##  7       2012  1717
##  8       2013  1339
##  9       2014  1464
## 10       2015  1434
## 11       2016  1208
## 12       2017   970
## 13       2018   958
## 14       2019   967
## 15       2020  1948
## 16       2021  2011
## 17       2022  1716
```

```
#validate the mapping of month
tidy_nypd %>% count(occur_month)
```

```
## # A tibble: 12 x 2
##    occur_month     n
##    <ord>       <int>
##  1 Jan          1716
##  2 Feb          1340
##  3 Mar          1688
##  4 Apr          1983
##  5 May          2571
##  6 Jun          2829
##  7 Jul          3238
##  8 Aug          3156
##  9 Sep          2572
## 10 Oct          2279
## 11 Nov          1944
## 12 Dec          1996
```

```
#validate the mapping of date, year, month
tidy_nypd %>% count(occur_date, occur_year, occur_month)
```

```
## # A tibble: 5,761 x 4
##    occur_date occur_year occur_month     n
##    <date>          <int> <ord>       <int>
##  1 2006-01-01       2006 Jan             8
##  2 2006-01-02       2006 Jan             4
##  3 2006-01-03       2006 Jan             4
##  4 2006-01-04       2006 Jan             4
##  5 2006-01-05       2006 Jan             4
##  6 2006-01-06       2006 Jan             4
##  7 2006-01-07       2006 Jan             2
##  8 2006-01-08       2006 Jan             4
##  9 2006-01-09       2006 Jan             9
## 10 2006-01-10       2006 Jan             5
## # i 5,751 more rows
```

```
#validate the mapping of hour
tidy_nypd %>% count(occur_hour)
```

```
## # A tibble: 24 x 2
##    occur_hour     n
##         <int> <int>
##  1          0  2186
##  2          1  2081
##  3          2  1812
##  4          3  1633
##  5          4  1441
##  6          5   702
##  7          6   366
##  8          7   233
##  9          8   238
## 10          9   217
## # i 14 more rows
```

```
#validate the mapping of perp and victim age groups
tidy_nypd %>% count(perp_age_group)
```

```
## # A tibble: 7 x 2
##   perp_age_group     n
##   <fct>          <int>
## 1 <18             1591
## 2 18-24           6222
## 3 25-44           5687
## 4 45-64            617
## 5 65+               60
## 6 Unknown         3148
## 7 <NA>            9987
```

```
tidy_nypd %>% count(vic_age_group)
```

```
## # A tibble: 7 x 2
##   vic_age_group     n
##   <fct>         <int>
## 1 <18            2839
## 2 18-24         10086
## 3 25-44         12281
## 4 45-64          1863
## 5 65+             181
## 6 Unknown          61
## 7 <NA>              1
```

Analysis and Visualizations

Through analysis and visualization, I would like to look at factors and trends that influnce shootings and murders in NYC based on the shooting incident report data source. In order to better understand the conditions for shootings, I'd like to do some analysis around the following areas:

- How many shootings occur per day?
- What is the trend of shootings over time?
- What borough has the most shootings?
- What time of day has the most shootings?
- what month has the most shootings?
- What is the most deadly borough?
- What age group shoots what age group?

How Many Shootings Occur Each Day in NYC?

```
#Summary of shootings per year
shootings_per_year <- tidy_nypd %>%
  group_by(occur_year) %>%
  summarize(
    shootings = n(),
    murders = sum(statistical_murder_flag == TRUE)
  )

shootings_per_year
```

```
#Calculate metrics for shootings per day and murders per day for the year 2021
shootings_per_day <- tidy_nypd %>%
  filter(occur_year == 2021) %>%
  group_by(occur_date, occur_year, occur_month) %>%
  summarize(
    shootings = n(),
    murders = sum(statistical_murder_flag == TRUE)
  ) %>%
  group_by(occur_year) %>%
  summarize(
    shootings = sum(shootings),
    murders = sum(murders),
    shootings_per_day = sum(shootings) / n(),
    murders_per_day = sum(murders) / n()
  )
```

```
## ‘summarise()‘ has grouped output by ’occur_date’, ’occur_year’. You can
## override using the ‘.groups‘ argument.
```

```
shootings_per_day
```

```
## # A tibble: 1 x 5
##   occur_year shootings murders shootings_per_day murders_per_day
##        <int>     <int>   <int>             <dbl>           <dbl>
## 1       2021      2011     428              5.66            1.21
```

Conclusion: On average in 2021 there were 5 shooting incidents each day in New York City and that they resulted in at least 1 or more murders every day.

What Do Shootings and Murders Look Like Over Time?

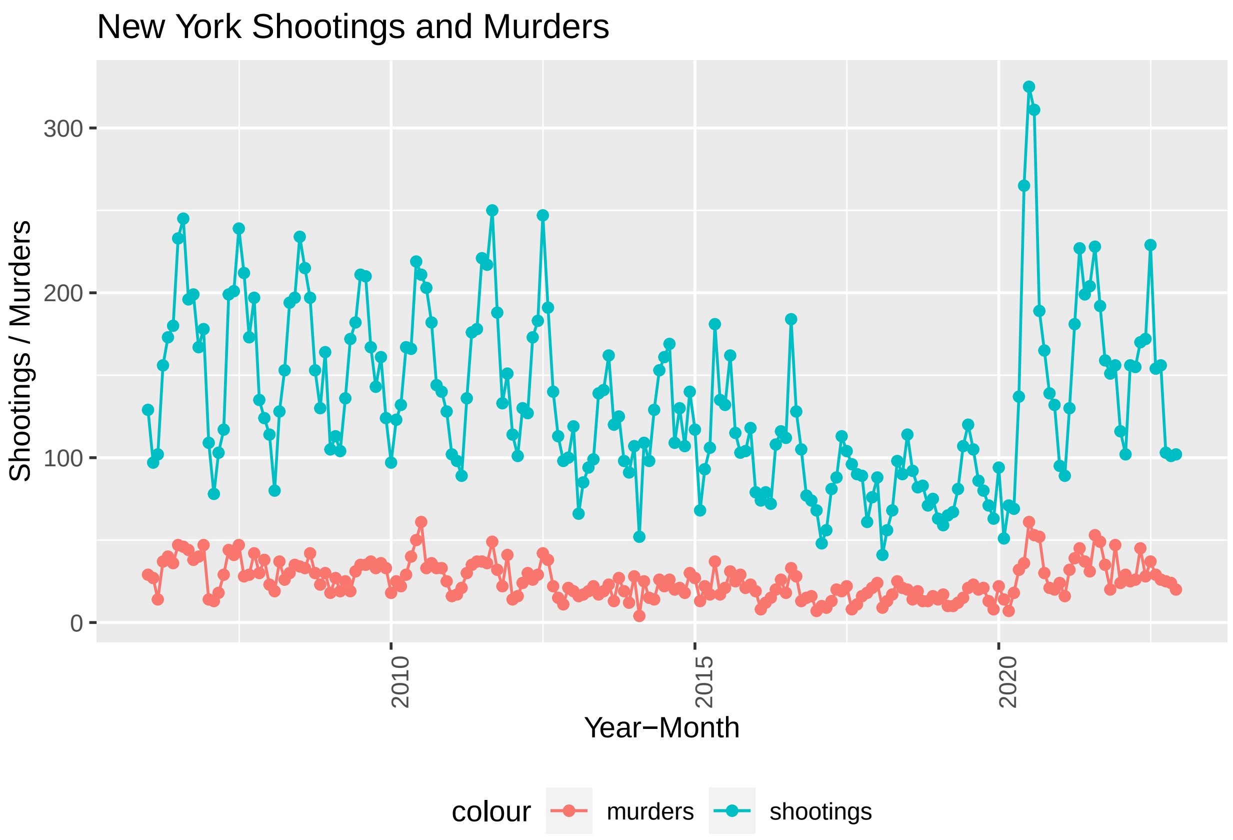
```
#Create a summarized time series of shootings and murders
nypd_over_time <- tidy_nypd %>%
  mutate(occur_year_month = as.Date(paste(occur_year_month, "-01", sep=""))) %>%
  group_by(occur_year_month, occur_year, occur_month) %>%
  summarize(
    shootings = n(),
    murders = sum(statistical_murder_flag == TRUE),
    pct_murder = sum(statistical_murder_flag == TRUE) / n()
  )
```

```
## ‘summarise()‘ has grouped output by ’occur_year_month’, ’occur_year’. You can
## override using the ‘.groups‘ argument.
```

```
nypd_over_time
```

```
## # A tibble: 204 x 6
## # Groups:   occur_year_month, occur_year [204]
##    occur_year_month occur_year occur_month shootings murders pct_murder
##    <date>                <int> <ord>           <int>   <int>      <dbl>
##  1 2006-01-01             2006 Jan               129      29      0.225
##  2 2006-02-01             2006 Feb                97      27      0.278
##  3 2006-03-01             2006 Mar               102      14      0.137
##  4 2006-04-01             2006 Apr               156      37      0.237
##  5 2006-05-01             2006 May               173      40      0.231
##  6 2006-06-01             2006 Jun               180      36      0.2
##  7 2006-07-01             2006 Jul               233      47      0.202
##  8 2006-08-01             2006 Aug               245      46      0.188
##  9 2006-09-01             2006 Sep               196      44      0.224
## 10 2006-10-01             2006 Oct               199      38      0.191
## # i 194 more rows
```

```
#Plot the time series
nypd_over_time %>%
  ggplot(aes(x = occur_year_month, y = shootings)) +
  geom_line(aes(color = "shootings")) +
  geom_point(aes(color = "shootings")) +
  geom_line(aes(y = murders, color = "murders")) +
  geom_point(aes(y = murders, color = "murders")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "New York Shootings and Murders", y = NULL) +
  xlab("Year-Month") +
  ylab("Shootings / Murders")
```



Conclusion: After 2000, there was a noticeable increase in shootings on a monthly basis (pulling up murders as well).

How Do Shootings in Boroughs Change Over Time?

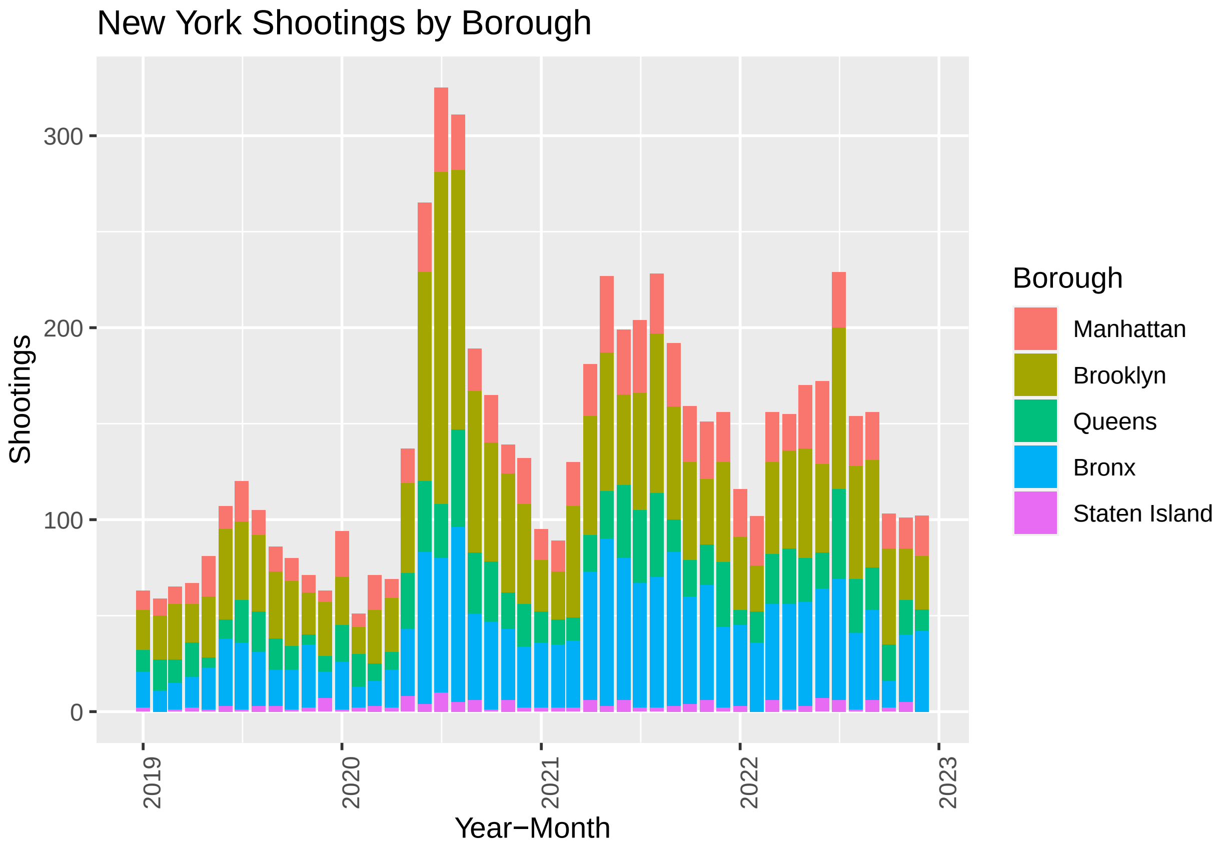
```
#Create a summarized time series of shootings broken out by borough
boro_over_time <- tidy_nypd %>%
  filter(occur_year >= 2019) %>%
  mutate(occur_year_month = as.Date(paste(occur_year_month, "-01", sep=""))) %>%
  group_by(occur_year, occur_year_month, boro) %>%
  summarize(
    shootings = n(),
    murders = sum(statistical_murder_flag == TRUE),
    pct_murder = sum(statistical_murder_flag == TRUE) / n()
  )
```

```
## ‘summarise()‘ has grouped output by ’occur_year’, ’occur_year_month’. You can
## override using the ‘.groups‘ argument.
```

```
boro_over_time
```

```
## # A tibble: 237 x 6
## # Groups:   occur_year, occur_year_month [48]
##    occur_year occur_year_month boro          shootings murders pct_murder
##         <int> <date>           <fct>            <int>   <int>      <dbl>
##   1      2019 2019-01-01       Manhattan           10       3      0.3
##   2      2019 2019-01-01       Brooklyn            21       5      0.238
##   3      2019 2019-01-01       Queens              11       2      0.182
##   4      2019 2019-01-01       Bronx               19       4      0.211
##   5      2019 2019-01-01       Staten Island        2       0      0
##   6      2019 2019-02-01       Manhattan            9       0      0
##   7      2019 2019-02-01       Brooklyn            23       9      0.391
##   8      2019 2019-02-01       Queens              16       6      0.375
##   9      2019 2019-02-01       Bronx               11       2      0.182
## 10       2019 2019-03-01       Manhattan            9       2      0.222
## # i 227 more rows
```

```
#Plot the time series
boro_over_time %>%
  ggplot(aes(x = occur_year_month, y = shootings, fill = boro)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "right",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "New York Shootings by Borough", y = NULL,
       fill = "Borough") +
  xlab("Year-Month") +
  ylab("Shootings")
```



Conclusion: Generally it looks like Brooklyn and the Bronx have the majority of shootings each month, but a different visualization would likely show this more clearly. Also, in the middle of 2000,there was a clear spike in shoots that were driven largely by changes in Brooklyn.

What Boroughs Have the Most Shootings After 2020?

```
#Create a dataset of shootings in each borough after 2020
boro_shootings <- tidy_nypd %>%
  filter(occur_year >= 2020) %>%
  group_by(boro) %>%
  summarize(
    shootings = n(),
    murders = sum(statistical_murder_flag == TRUE),
    pct_murder = round(sum(murders) / sum(shootings) * 100, digits = 1)
  )
boro_shootings
```

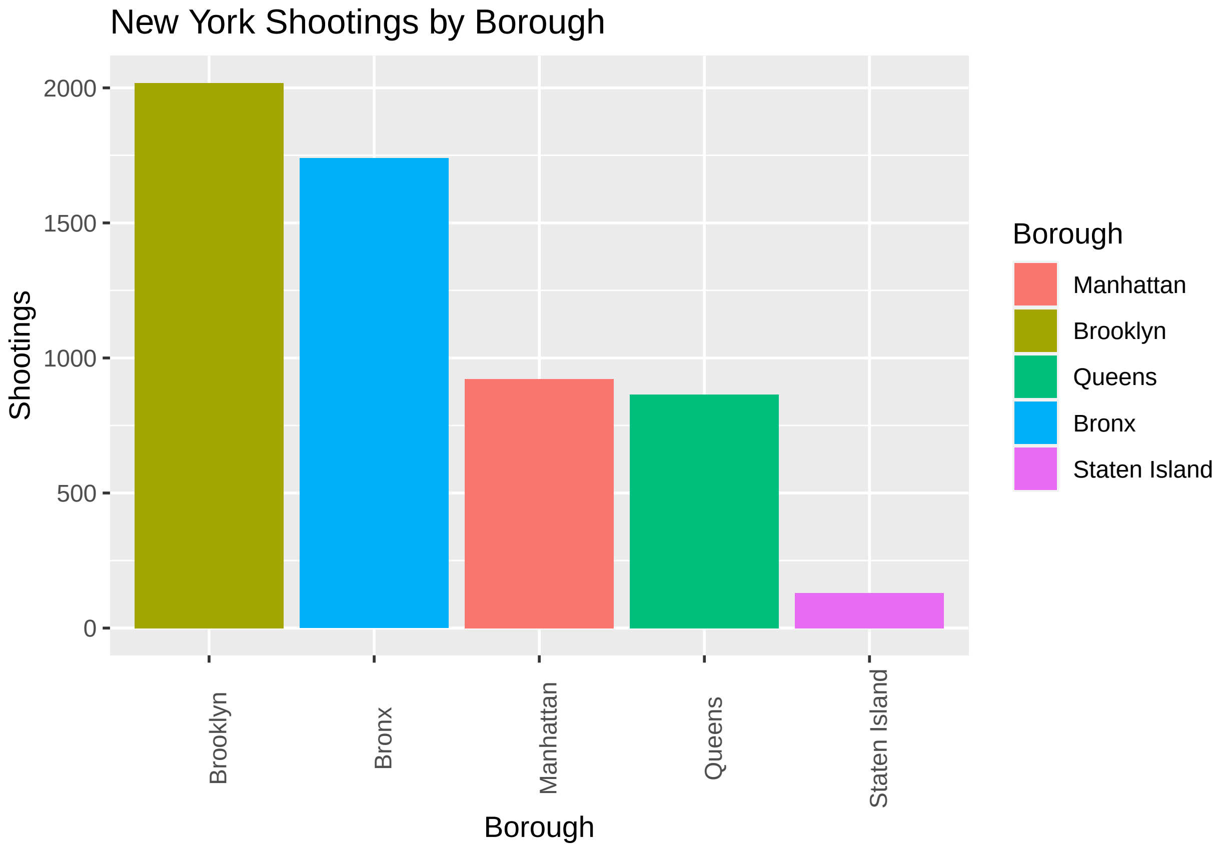
```
## # A tibble: 5 x 4
##   boro          shootings murders pct_murder
##   <fct>             <int>   <int>      <dbl>
## 1 Manhattan           922     162       17.6
## 2 Brooklyn           2018     385       19.1
## 3 Queens              865     169       19.5
## 4 Bronx              1740     381       21.9
## 5 Staten Island       130      35       26.9
```

```
#Write the summary of the where shootings occur
b_b_shootings = boro_shootings[boro_shootings$boro == "Brooklyn", ]$shootings +
  boro_shootings[boro_shootings$boro == "Bronx", ]$shootings
all_shootings = sum(boro_shootings$shootings)
pct_shootings = round(b_b_shootings / all_shootings * 100, 0)

print(paste0("There where ", b_b_shootings ,
             " shootings in Brooklyn and the Bronx, which account for ",
             pct_shootings, "% of the overall ", all_shootings,
             " shootings that occured in NYC."))
```

```
## [1] "There where 3758 shootings in Brooklyn and the Bronx, which account for 66% of the overall 5675
```

```
#Plot the Bar Chart
boro_shootings %>%
  ggplot(aes(x = reorder(boro, -shootings), y = shootings, fill = boro)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "right",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "New York Shootings by Borough", y = NULL,
       fill = "Borough") +
  xlab("Borough") +
  ylab("Shootings")
```



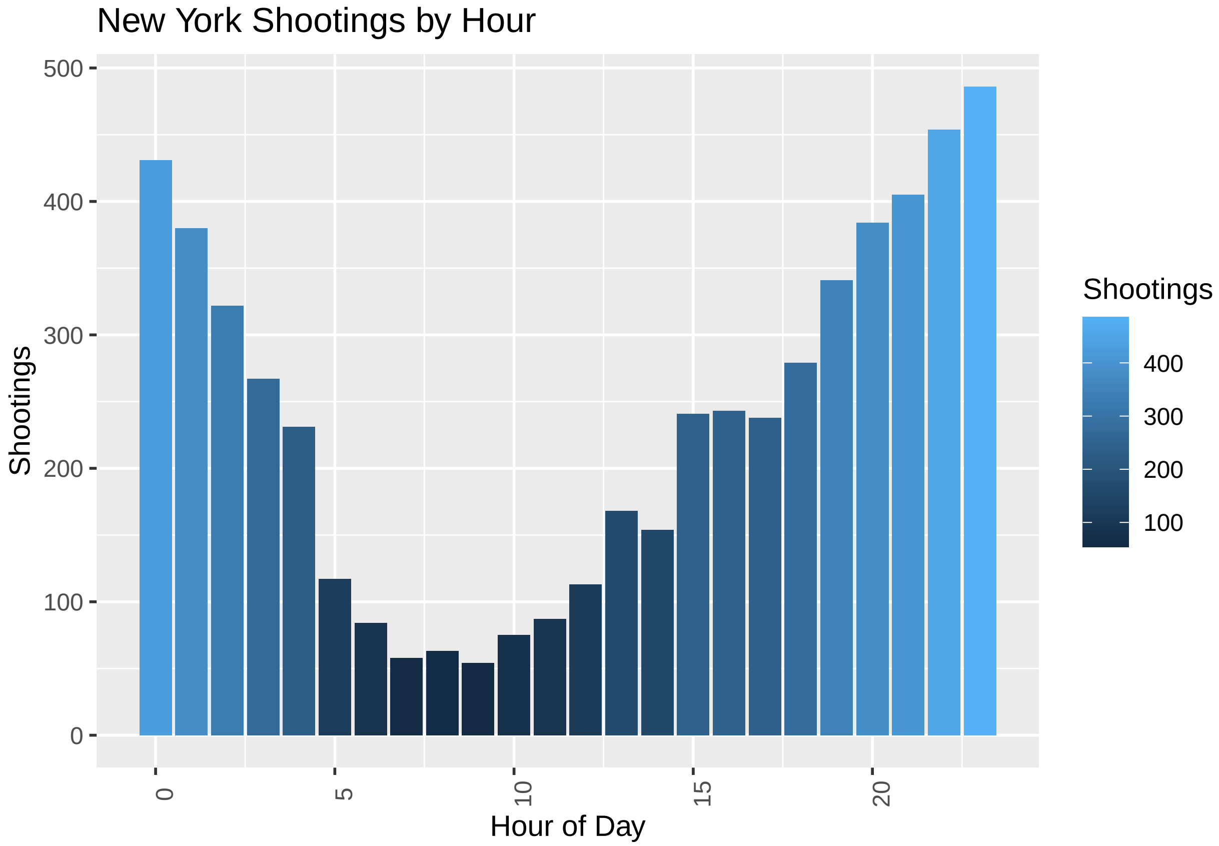
Conclusion: The Bronx and Brooklyn tend to account for the majority of the shooting incidents in NYC.

What is the Worst Time of Day for Shootings After 2020?

```
#Create a dataset of shootings in each hour after 2020
hour_shootings <- tidy_nypd %>%
  filter(occur_year >= 2020) %>%
  group_by(occur_hour) %>%
  summarize(
    shootings = n(),
    murders = sum(statistical_murder_flag == TRUE),
    pct_murder = round(sum(murders) / sum(shootings) * 100, digits = 1)
  )
hour_shootings
```

```
## # A tibble: 24 x 4
##    occur_hour shootings murders pct_murder
##         <int>     <int>   <int>      <dbl>
##  1          0       431      96       22.3
##  2          1       380      68       17.9
##  3          2       322      52       16.1
##  4          3       267      38       14.2
##  5          4       231      52       22.5
##  6          5       117      19       16.2
##  7          6        84       20        23.8
##  8          7        58        9        15.5
##  9          8        63       17        27
## 10          9        54       13        24.1
## # i 14 more rows
```

```
#Plot the Bar Chart
hour_shootings %>%
  ggplot(aes(x = occur_hour, y = shootings, fill = shootings)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "right",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "New York Shootings by Hour", y = NULL,
       fill = "Shootings") +
  xlab("Hour of Day") +
  ylab("Shootings")
```



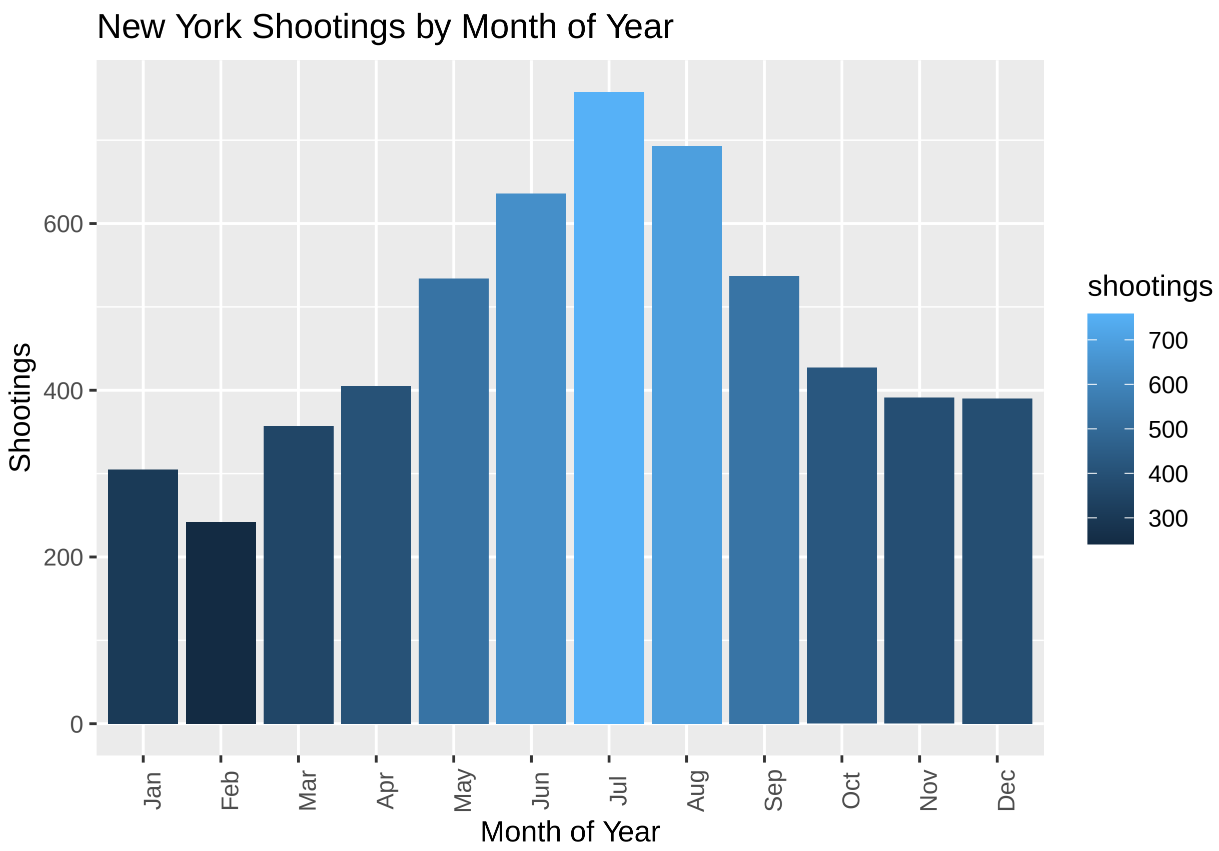
Conclusion: The 3 worst hours of the day for shootings are 10pm, 11pm, and 12pm.

What is the Worst Month of the Year for Shootings After 2020?

```
#Create a dataset of shootings in each hour after 2020
month_shootings <- tidy_nypd %>%
  filter(occur_year >= 2020) %>%
  group_by(occur_month) %>%
  summarize(
    shootings = n(),
    murders = sum(statistical_murder_flag == TRUE),
    pct_murder = round(sum(murders) / sum(shootings) * 100, digits = 1)
  )
month_shootings
```

```
## # A tibble: 12 x 4
##    occur_month shootings murders pct_murder
##    <ord>           <int>   <int>      <dbl>
##  1 Jan               305      70       23  
##  2 Feb               242      59       24.4
##  3 Mar               357      64       17.9
##  4 Apr               405      83       20.5
##  5 May               534     122       22.8
##  6 Jun               636     101       15.9
##  7 Jul               758     129       17  
##  8 Aug               693     135       19.5
##  9 Sep               537     127       23.6
## 10 Oct               427      90       21.1
## 11 Nov               391      65       16.6
## 12 Dec               390      87       22.3
```

```
#Plot the Bar Chart
month_shootings %>%
  ggplot(aes(x = occur_month, y = shootings, fill = shootings)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "right",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "New York Shootings by Month of Year", y = NULL) +
  xlab("Month of Year") +
  ylab("Shootings")
```



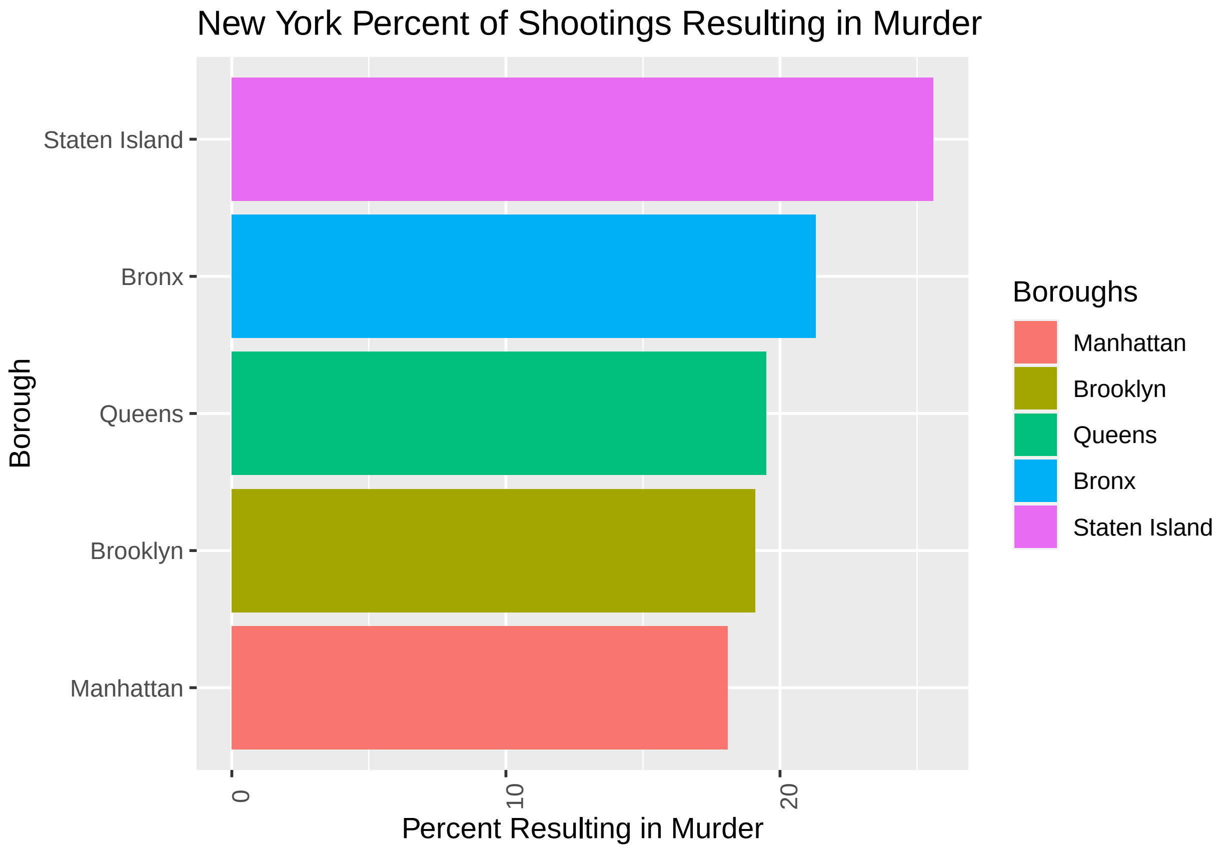
Conclusion: The 3 worst months for shootings are June, July, and August.

What is the Most Deadly Boroughs for Shootings for the Past 3 Years?

```
#Create a summarized dataset
top_boro_last_3 <- boro_over_time %>%
  filter(occur_year >= 2019) %>%
  group_by(boro) %>%
  summarize(
    shootings = sum(shootings),
    murders = sum(murders),
    pct_murder = round(sum(murders) / sum(shootings) * 100, digits = 1)
  )
top_boro_last_3
```

```
## # A tibble: 5 x 4
##   boro          shootings murders pct_murder
##   <fct>             <int>   <int>      <dbl>
## 1 Manhattan          1068     193       18.1
## 2 Brooklyn           2390     457       19.1
## 3 Queens             1021     199       19.5
## 4 Bronx              2007     427       21.3
## 5 Staten Island       156      40       25.6
```

```
#Plot the bar chart
top_boro_last_3 %>%
  ggplot(aes(x = pct_murder, y = boro, fill = boro)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "right",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "New York Percent of Shootings Resulting in Murder", y = NULL,
       fill = "Boroughs") +
  xlab("Percent Resulting in Murder") +
  ylab("Borough")
```



Conclusion: Staten Island in the most deadly place to be part of a shooting (results in the highest murders per shooting).

What pattern of shootings do we see for age groups?

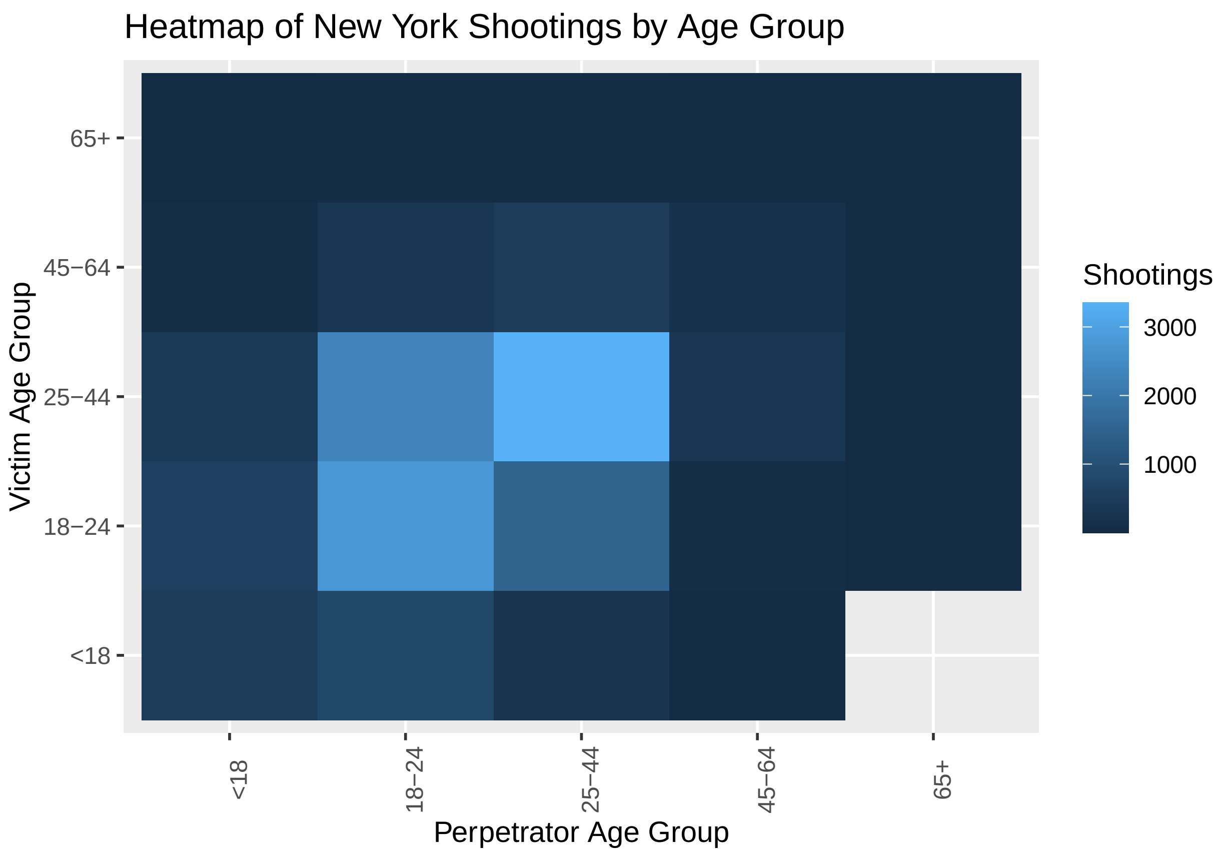
```
#create an aggregate data set of perp age group and victim age groop
nypd_age_group <- tidy_nypd %>%
  group_by(perp_age_group, vic_age_group) %>%
  summarize(
    shootings = n(),
    murders = sum(statistical_murder_flag == TRUE),
    pct_murder = sum(statistical_murder_flag == TRUE) / n()
  )
```

```
## ‘summarise()‘ has grouped output by ’perp_age_group’. You can override using
## the ‘.groups‘ argument.
```

```
nypd_age_group
```

```
## # A tibble: 41 x 5
## # Groups:   perp_age_group [7]
##    perp_age_group vic_age_group shootings murders pct_murder
##    <fct>          <fct>             <int>   <int>      <dbl>
##  1 <18            <18                 484      69      0.143
##  2 <18            18-24               621     110      0.177
##  3 <18            25-44               397      94      0.237
##  4 <18            45-64                77      13      0.169
##  5 <18            65+                  10       1      0.1
##  6 <18            Unknown               2       0      0
##  7 18-24          <18                 788     143      0.181
##  8 18-24          18-24              2758     570      0.207
##  9 18-24          25-44              2294     505      0.220
## 10 18-24          45-64               329      75      0.228
## # i 31 more rows
```

```
#create the heatmap
nypd_age_group %>%
  filter(!is.na(perp_age_group)
         & !is.na(vic_age_group)
         & perp_age_group != "Unknown"
         & vic_age_group != "Unknown"
        ) %>%
  ggplot(aes(x = perp_age_group, y = vic_age_group)) +
  geom_raster(aes(fill = shootings)) +
  theme(legend.position = "right",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Heatmap of New York Shootings by Age Group", y = NULL,
       fill = "Shootings") +
  xlab("Perpetrator Age Group") +
  ylab("Victim Age Group")
```



Conclusion: Most shootings happen with your own age group, for example, the highest rate of shootings are committed by people aged 25-44 against victims who are also 25-44.

Additional Questions to Explore and Investigate

After completing some initial analysis and visualization, there is much more to explore and investigate. The following is a list of some potential questions to answer:

1. Is there any interaction or correlation between perpetrator gender and the victims?
2. Is there any interaction or correlation between perpetrator gender and the victims?
3. What month are you most likely to observe a shooting?
4. Can we predict the level of shootings based on the current data? how accurately?
5. What time of day are you most likely to observe a shooting? Does it vary by jurisdiction?
6. Does temperature affect how many shootings occur?

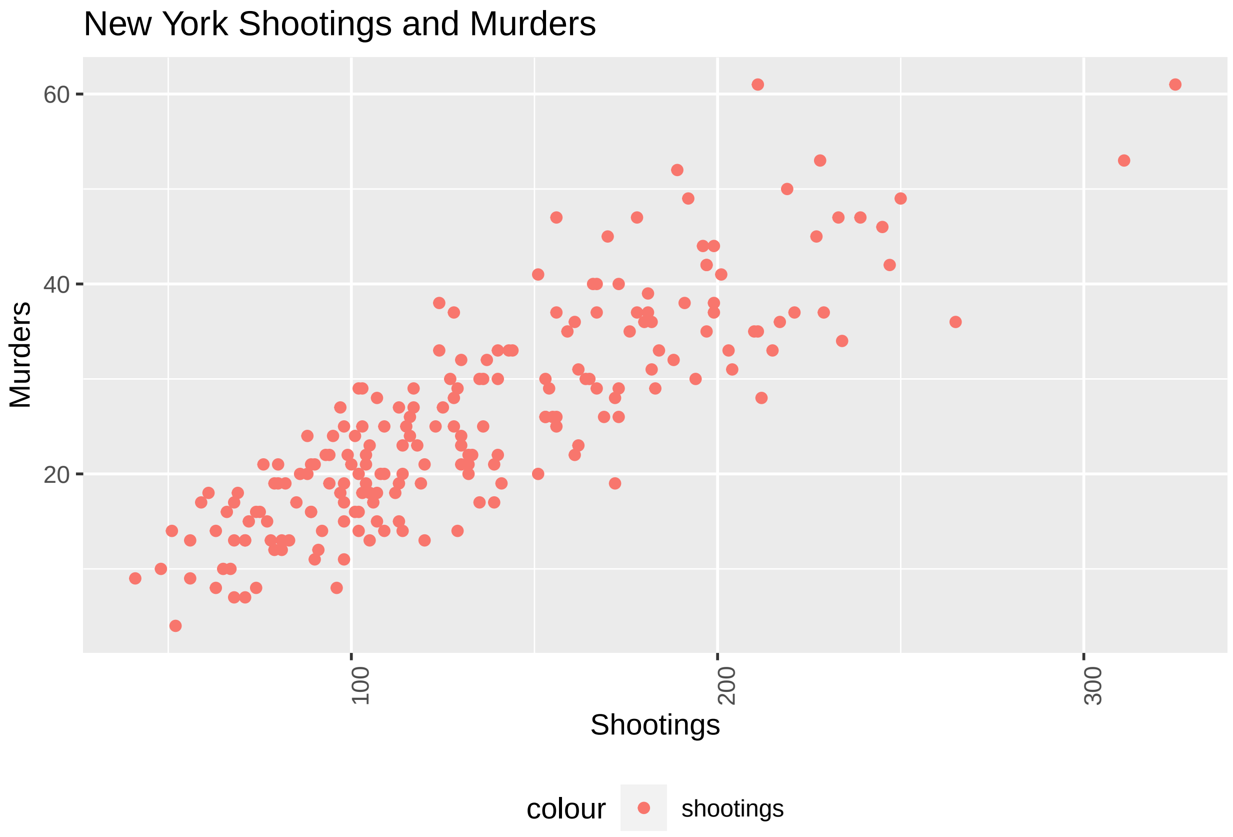
Models & Conclusions

Create Linear Prediction Models

```
#show the data source to be used in the model
head(nypd_over_time, n = 3)
```

```
## # A tibble: 3 x 6
## # Groups:   occur_year_month, occur_year [3]
##   occur_year_month occur_year occur_month shootings murders pct_murder
##   <date>                <int> <ord>           <int>   <int>      <dbl>
## 1 2006-01-01             2006 Jan               129      29      0.225
## 2 2006-02-01             2006 Feb                97      27      0.278
## 3 2006-03-01             2006 Mar               102      14      0.137
```

```
#build a Scatter Plot of shootings vs. murders
nypd_over_time %>%
  ggplot(aes(x = shootings, y = murders)) +
  geom_point(aes(color = "shootings")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "New York Shootings and Murders", y = NULL) +
  xlab("Shootings") +
  ylab("Murders")
```



```
#create the Prediction Model
mod <- lm(murders ~ shootings, data = nypd_over_time)
summary(mod)
```

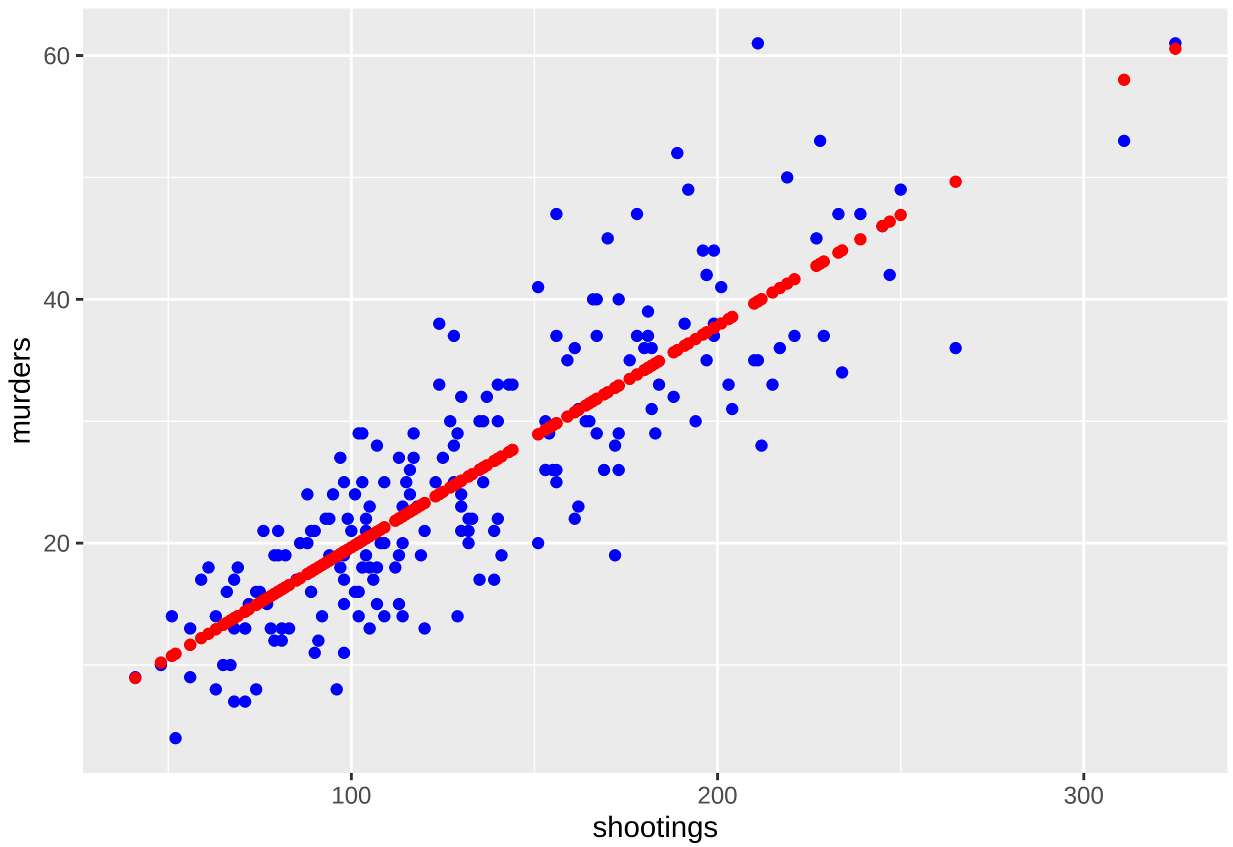
```
##
## Call:
## lm(formula = murders ~ shootings, data = nypd_over_time)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7430  -4.1080  -0.0157   3.6410  21.1672
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.475585   1.132499   1.303    0.194
## shootings   0.181788   0.007878  23.075   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.89 on 202 degrees of freedom
## Multiple R-squared:  0.725,  Adjusted R-squared:  0.7236
## F-statistic: 532.4 on 1 and 202 DF,  p-value: < 2.2e-16
```

```r
#Add the predictions to a data frame
nypd_over_time_w_pred <- nypd_over_time %>%
  modelr::add_predictions(mod)
nypd_over_time_w_pred
```

```
## # A tibble: 204 x 7
## # Groups:   occur_year_month, occur_year [204]
##    occur_year_month occur_year occur_month shootings murders pct_murder  pred
##    <date>                <int> <ord>           <int>   <int>      <dbl> <dbl>
##  1 2006-01-01             2006 Jan               129      29      0.225  24.9
##  2 2006-02-01             2006 Feb                97      27      0.278  19.1
##  3 2006-03-01             2006 Mar               102      14      0.137  20.0
##  4 2006-04-01             2006 Apr               156      37      0.237  29.8
##  5 2006-05-01             2006 May               173      40      0.231  32.9
##  6 2006-06-01             2006 Jun               180      36      0.2    34.2
##  7 2006-07-01             2006 Jul               233      47      0.202  43.8
##  8 2006-08-01             2006 Aug               245      46      0.188  46.0
##  9 2006-09-01             2006 Sep               196      44      0.224  37.1
## 10 2006-10-01             2006 Oct               199      38      0.191  37.7
## # i 194 more rows
```

```r
#plot the actual values and predictions
nypd_over_time_w_pred %>% ggplot() +
  geom_point(aes(x = shootings, y = murders), color = "blue") +
  geom_point(aes(x = shootings, y = pred), color = "red")
```



Conclusion: While there is a relationship and some correlation between shootings and murders, its not as strong as one might think. Further analysis is required to identify other strong predictors of murders.

Conclusions

After completing the analysis of data, visualization, and modeling, we can conclude the following:

Question	Conclusion
How many shootings occur per day?	On average in 2021 there were 5 shooting incidents each day in New York City and that they resulted in at least 1 or more murders every day.
What is the trend of shootings over time?	After 2000, there was a noticeable increase in shootings on a monthly basis (pulling up murders as well).
What borough has the most shootings?	The Bronx and Brooklyn tend to account for the majority of the shooting incidents in NYC.
What time of day has the most shootings?	The 3 worst hours of the day for shootings are 10pm, 11pm, and 12pm.
what month has the most shootings?	The 3 worst months for shootings are June, July, and August.
What is the most deadly borough?	Stanten Island in the most deadly place to be part of a shooting (results in the highest murders per shooting).

Question	Conclusion
What age group shoots what age group?	Most shootings happen with your own age group, for example, the highest rate of shootings are committed by people aged 25-44 against victims who are also 25-44.
Is there a strong correlation between shootings and murders?	While there is a relationship and some correlation between shootings and murders, its not as strong as one might think. Further analysis is required to identify other strong predictors of murder.

Review of Bias

Considering Bias, I would place it into 3 categories:

1. Who is providing the data
2. Who is collecting the data
3. Who is analyzing the data

Provider: As the data is based on incident reports and manually extracted from witnesses and victoms, the bias of people is included in the data. Details abvout the suspect are observations and likely concrete data. The shooting incident was observed through bias eyes and bias factors were likely collected.

Collector: The incident reports that sit behind the data are collected by police officers. What is document is shaped by their thoughts, opinions, and bias as a police officer. It is unclear whether all the factors are provided by victim or are assessed by the officer. How the data is collected at the point of the incident and potentially at the point of interpretation when the data set is built. All of this can shape what is in the data source and include individuals bias in the collection process.

Analyzer: As the analyst, I bring my own biases to the data. I am an urban resident, so in some ways I may thing I understand the dynamics of city life and city crime. On the other hand, I am a white male who has little exposure to gun violence and no exposure to policing. I am not a subject matter expert in this area and may not interpret the details or factors correctly. I also bring to the analysis my own assumptions and bias about sex, age, race, and ethnicity.

Session Summary

```
sessionInfo()
```

```
## R version 4.2.3 (2023-03-15 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22621)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
```

```
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] tinytex_0.45    lubridate_1.9.2 forcats_1.0.0   stringr_1.5.0
##  [5] dplyr_1.1.1     purrr_1.0.1     readr_2.1.4     tidyr_1.3.0
##  [9] tibble_3.2.1    ggplot2_3.4.2   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] highr_0.10       pillar_1.9.0     compiler_4.2.3   tools_4.2.3
##  [5] bit_4.0.5        digest_0.6.31    timechange_0.2.0 evaluate_0.20
##  [9] lifecycle_1.0.3  gtable_0.3.3     pkgconfig_2.0.3  rlang_1.1.0
## [13] cli_3.6.1        rstudioapi_0.14  curl_5.0.0       parallel_4.2.3
## [17] yaml_2.3.7       xfun_0.38        fastmap_1.1.1    withr_2.5.0
## [21] knitr_1.42       generics_0.1.3   vctrs_0.6.1      hms_1.1.3
## [25] bit64_4.0.5      grid_4.2.3       tidyselect_1.2.0 glue_1.6.2
## [29] R6_2.5.1         fansi_1.0.4      vroom_1.6.1      rmarkdown_2.21
## [33] modelr_0.1.11    farver_2.1.1     tzdb_0.3.0       magrittr_2.0.3
## [37] backports_1.4.1  scales_1.2.1     htmltools_0.5.5  colorspace_2.1-0
## [41] labeling_0.4.2   utf8_1.2.3       stringi_1.7.12   munsell_0.5.0
## [45] broom_1.0.4      crayon_1.5.2
```