NYPD_Shooting_Basic_Exploration

```r
library(ggplot2)
library(dplyr)
library(shiny)
library(zoo)
library(chron)
```

Cursory visual and summary examination:

```r
df = read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
head(df, 1)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME   BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00 QUEENS                        105
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                    false
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX VIC_RACE X_COORD_CD
## 1                                           18-24       M    BLACK    1058925
##   Y_COORD_CD Latitude Longitude                                        Lon_Lat
## 1     180924 40.66296 -73.73084 POINT (-73.73083868899994 40.662964620000025)
```

```r
str(df)
```

```
## 'data.frame':    27312 obs. of  21 variables:
##  $ INCIDENT_KEY           : int  228798151 137471050 147998800 146837977 58921844 219559682 85295722
##  $ OCCUR_DATE             : chr  "05/27/2021" "06/27/2014" "11/21/2015" "10/09/2015" ...
##  $ OCCUR_TIME             : chr  "21:30:00" "17:40:00" "03:56:00" "18:30:00" ...
##  $ BORO                   : chr  "QUEENS" "BRONX" "QUEENS" "BRONX" ...
##  $ LOC_OF_OCCUR_DESC      : chr  "" "" "" "" ...
##  $ PRECINCT               : int  105 40 108 44 47 81 114 81 105 101 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LOC_CLASSFCTN_DESC     : chr  "" "" "" "" ...
##  $ LOCATION_DESC          : chr  "" "" "" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "true" "false" ...
##  $ PERP_AGE_GROUP         : chr  "" "" "" "" ...
##  $ PERP_SEX               : chr  "" "" "" "" ...
##  $ PERP_RACE              : chr  "" "" "" "" ...
##  $ VIC_AGE_GROUP          : chr  "18-24" "18-24" "25-44" "<18" ...
##  $ VIC_SEX                : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "WHITE" "WHITE HISPANIC" ...
##  $ X_COORD_CD             : num  1058925 1005028 1007668 1006537 1024922 ...
##  $ Y_COORD_CD             : num  180924 234516 209837 244511 262189 ...
##  $ Latitude               : num  40.7 40.8 40.7 40.8 40.9 ...
##  $ Longitude              : num  -73.7 -73.9 -73.9 -73.9 -73.9 ...
##  $ Lon_Lat                : chr  "POINT (-73.73083868899994 40.662964620000025)" "POINT (-73.9249423
```

```r
summary(df)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME             BORO
```

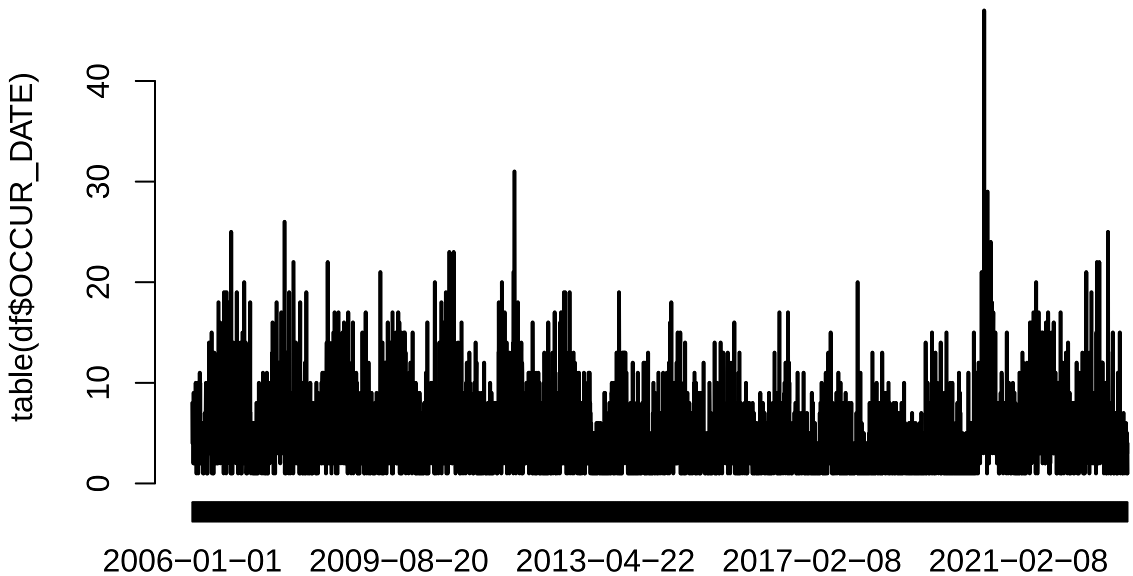
```
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  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       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        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.:1000028   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
```

Single variable examination

Date

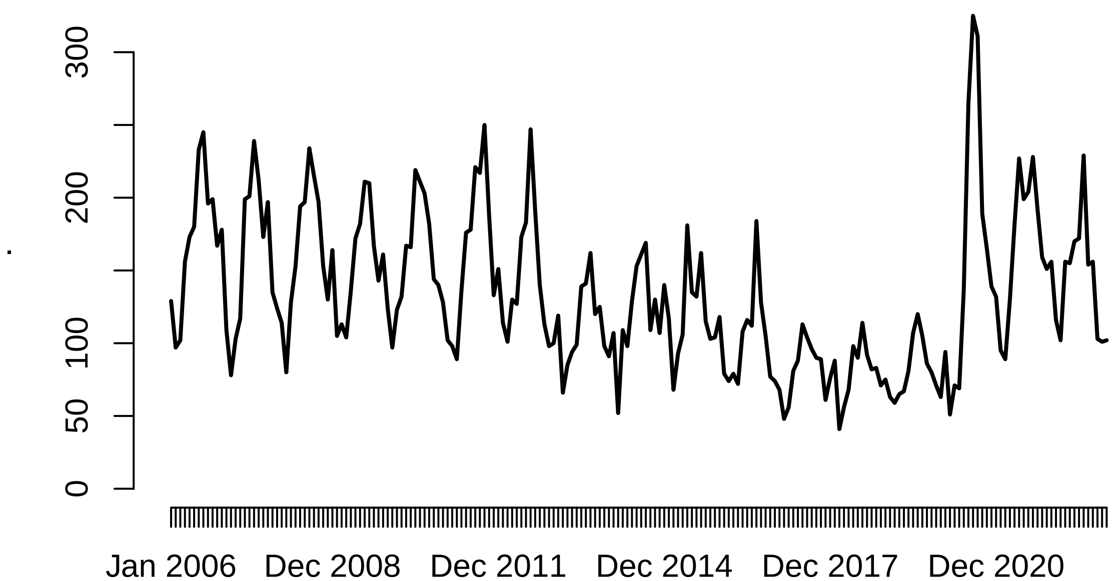
```
#parse OCCUR_DATE to date format
df$OCCUR_DATE = as.Date(df$OCCUR_DATE, format = "%m/%d/%Y")
```

```r
plot(table(df$OCCUR_DATE), type = 'l')
```



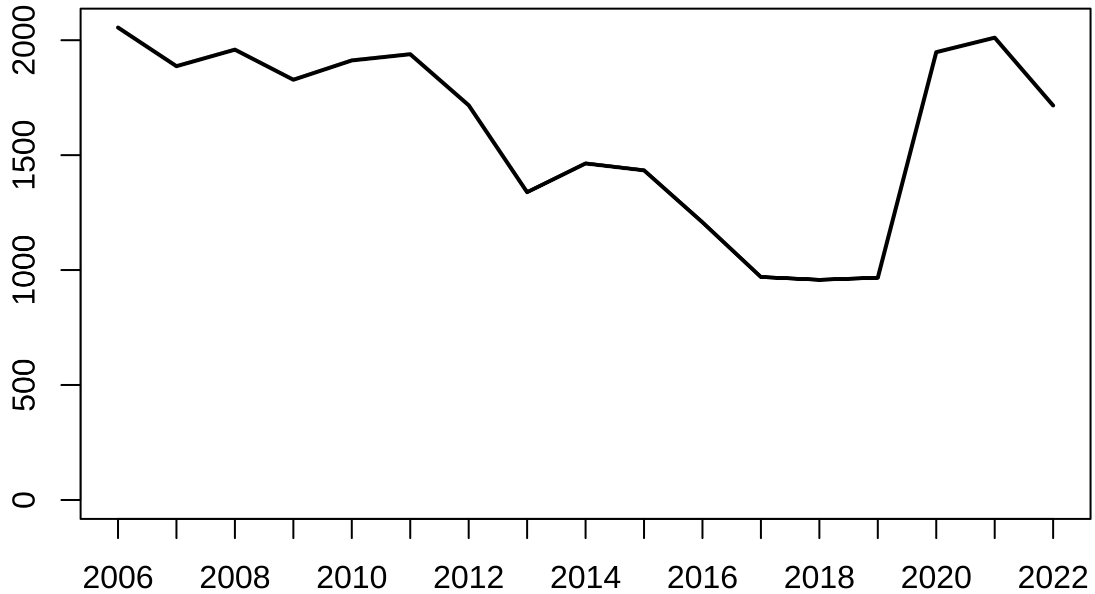

```r
#this chart is messy but we can see a few things:
#-seasonal peaks and lulls in shootings - presumably lower in the winter and higher in the summer
#-an overall drop in shooting incidents from the beginning of the data in 2006 until 2020
#-a large spike around the period of unrest following the killing of George Floyd with overall levels s
#let's bin the dates into individual months to produce a clearer chart
```

```r
df$Month = df$OCCUR_DATE %>% as.yearmon()
df$Month %>% table() %>% plot(type = 'l')
```




```r
#this monthly plot is better, but it would be nice to plot each year separately and have the x-axis be
#as well as to simply aggregate by year
```

```r
df$Year = df$OCCUR_DATE %>% format("%Y")
df$Year %>% table() %>% plot(type = "l")
```




```r
# with this yearly plot we can see a sizeable reduction in shootings - almost 50% over about 10 years,
# that has persisted until the end of the dataset in 2022
```

```r
df$Year = df$OCCUR_DATE %>% format("%Y") %>% as.integer()
df$Month = df$OCCUR_DATE %>% format("%m") %>% as.integer()
yearmon_df = df %>% group_by(Year, Month) %>% summarise(Count = n()) %>% as.data.frame()
```

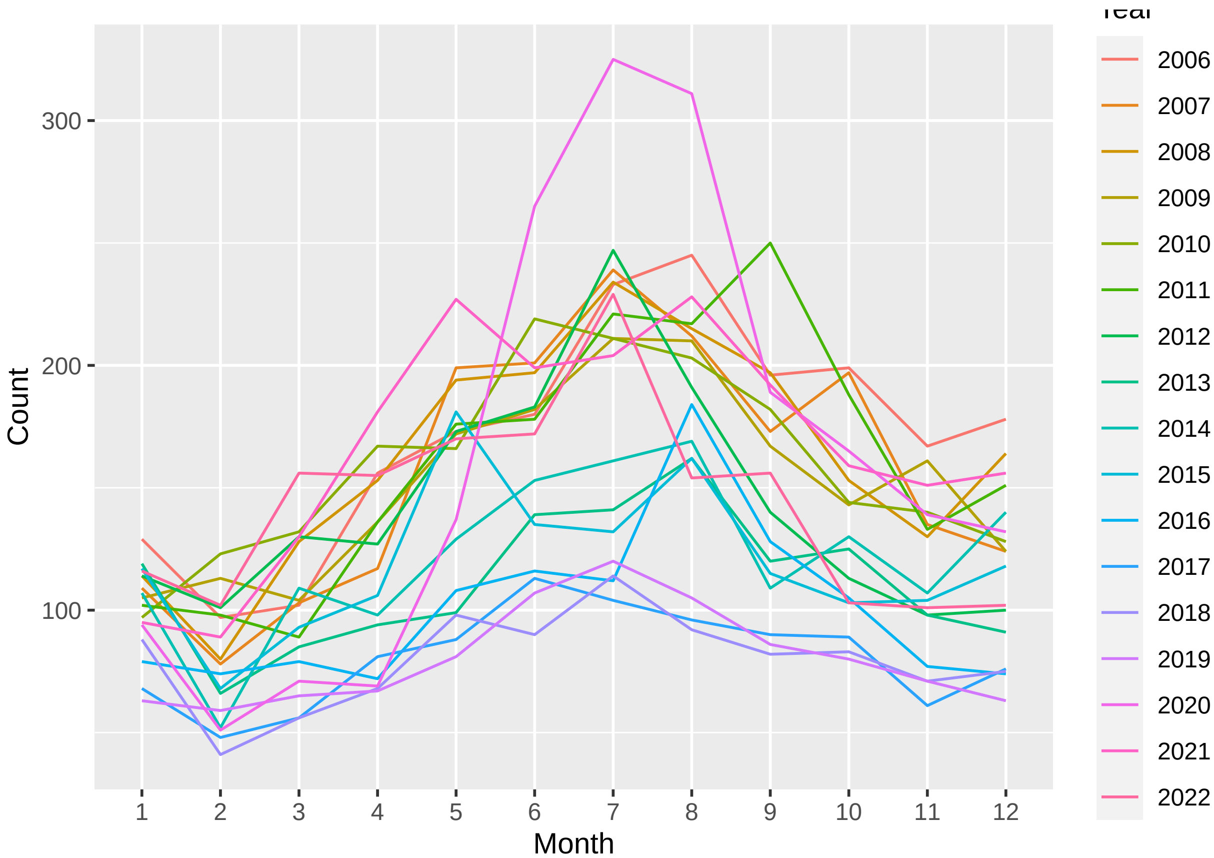
```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```r
yearmon_df$Year = yearmon_df$Year %>% as.factor()
yearmon_df$Month = yearmon_df$Month %>% as.factor()
```

```r
yearmon_df %>% head(3)
```

```
##   Year Month Count
## 1 2006     1   129
## 2 2006     2    97
## 3 2006     3   102
```

```r
ggplot(yearmon_df, aes(x = Month, y = Count, group = Year, col = Year)) + geom_line()
```




```
#here we can confirm that shootings tend to peak in the summer.
#we can also see the surge in shootings in the summer of 2020 after the killing of George Floyd
```

Time

```r
df$OCCUR_TIME %>% unique() %>% head(20)
```

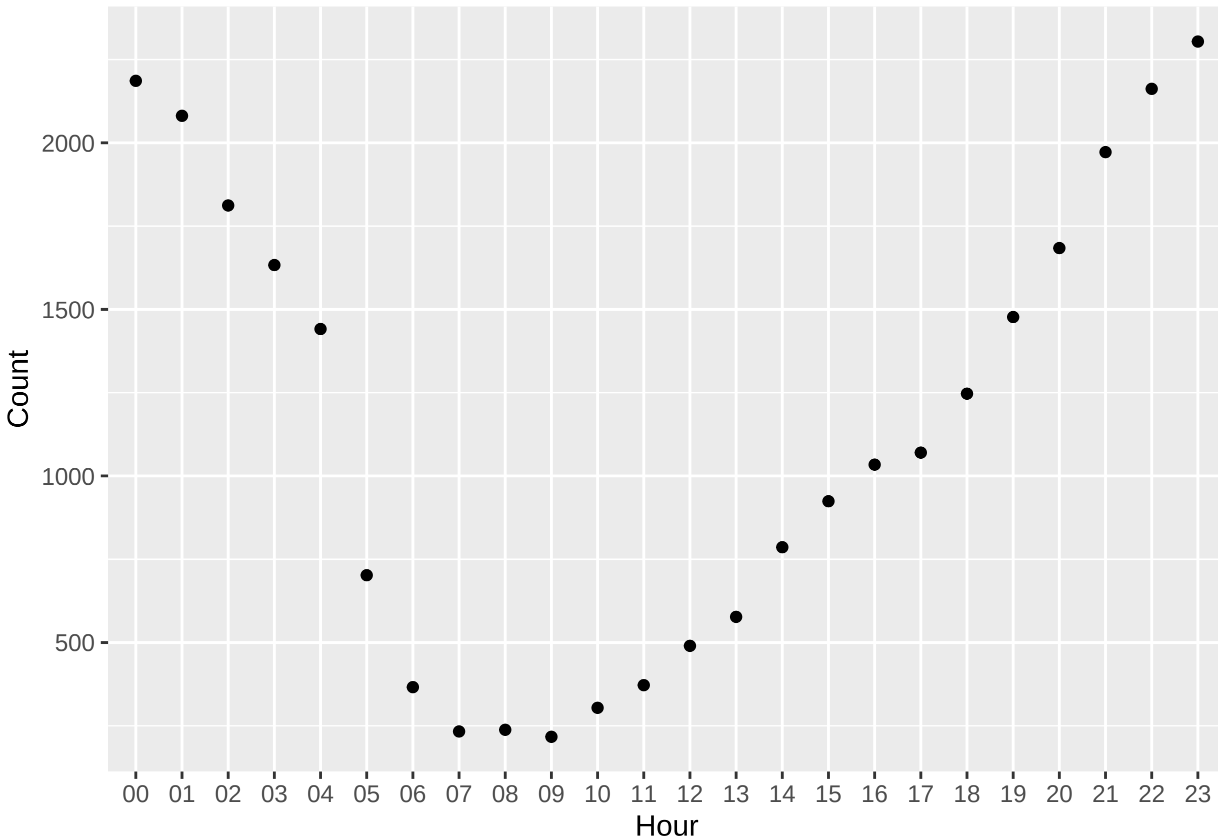
```
##  [1] "21:30:00" "17:40:00" "03:56:00" "18:30:00" "22:58:00" "21:36:00"
##  [7] "22:47:00" "19:41:00" "05:45:00" "01:10:00" "03:21:00" "01:27:00"
## [13] "20:17:00" "21:58:00" "20:13:00" "02:22:00" "21:07:00" "02:44:00"
## [19] "21:17:00" "23:16:00"
```

```r
df$OCCUR_TIME %>% table() %>% sort(decreasing = TRUE) %>% head(20)
```

```
## .
## 23:30:00 00:30:00 01:30:00 02:00:00 21:00:00 22:30:00 01:00:00 04:00:00
##      179      156      153      148      145      140      133      130
## 23:00:00 21:30:00 22:00:00 02:30:00 00:50:00 03:30:00 01:15:00 03:00:00
##      130      125      121      108      105      104      103      100
## 04:30:00 00:15:00 20:00:00 23:50:00
##       99       98       97       96
```

```
#unfortunately while there are time bins included down to the minute, many have been apparently categor
# we will group into uniform bins in order to get a decent visualization of the time distribution
```

```r
df$Hour = df$OCCUR_TIME %>% substr(0,2)
hour_df = df$Hour %>% table() %>% as.data.frame() %>% setNames(c("Hour", "Count"))
ggplot(hour_df, aes(x = Hour, y = Count)) + geom_point()
```




```
# are lowest in the morning and then rise throughout the day, peaking late at night
```

Borough

```
table(df$BORO)
```

```
##
##         BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND
##          7937         10933          3572          4094           776
```

```
# straightforward categories with complete data
```

"LOC_OF_OCCUR_DESC"

```
table(df$LOC_OF_OCCUR_DESC)
```

```
##
##           INSIDE OUTSIDE
##   25596      242    1474
```

```
# not particularly helpful with the vast majority of values missing
```

Precinct

```
table(df$PRECINCT)
```

```
##
##     1     5     6     7     9    10    13    14    17    18    19    20    22    23    24    25
```

```
##    25    58    28   109   109    73    60    56    10    34    20    40     1   487   105   461
##    26    28    30    32    33    34    40    41    42    43    44    45    46    47    48    49
##   149   343   229   634   225   316   908   494   850   758  1020   182   895   953   787   353
##    50    52    60    61    62    63    66    67    68    69    70    71    72    73    75    76
##   154   583   372   153    70   282    46  1216    32   466   459   579   109  1452  1557   167
##    77    78    79    81    83    84    88    90    94   100   101   102   103   104   105   106
##   795    62  1012   799   500   124   280   315    86   170   489   210   593   102   479   224
##   107   108   109   110   111   112   113   114   115   120   121   122   123
##   101    67   115   160    11    23   802   369   179   572   112    61    31
```

```
#clearly there is large variance in shooting incidents between different precincts. Some have none whil
#obviously needs to be mapped to be meaningful
```

```
df$PRECINCT %>% table() %>% sum()
```

```
## [1] 27312
```

```
#confirming that all incidents are placed in a precinct - no missing values
```

Jurisdiction Code

```
table(df$JURISDICTION_CODE)
```

```
##
##     0     1     2
## 22809    74  4427
```

```
#according to NYC's data website, 0=Patrol, 1=Transit, 2=Housing
```

"LOC_CLASSFCTN_DESC"

```
table(df$LOC_CLASSFCTN_DESC)
```

```
##
##                COMMERCIAL    DWELLING     HOUSING       OTHER PARKING LOT
##       25596           100         127         280          31           7
##  PLAYGROUND        STREET     TRANSIT     VEHICLE
##          30          1103          15          23
```

```
#vast majority have missing value
```

"LOCATION_DESC"

```
table(df$LOCATION_DESC)
```

```
##
##                                                (null)                       ATM
##                     14977                         977                         1
##                      BANK              BAR/NIGHT CLUB        BEAUTY/NAIL SALON
##                         3                         628                       112
##               CANDY STORE                 CHAIN STORE                CHECK CASH
##                         7                           5                         1
##        CLOTHING BOUTIQUE             COMMERCIAL BLDG                DEPT STORE
##                        14                         292                         9
##            DOCTOR/DENTIST                  DRUG STORE      DRY CLEANER/LAUNDRY
##                         1                          14                        31
```

```
##         FACTORY/WAREHOUSE                  FAST FOOD               GAS STATION
##                         8                        104                        71
##            GROCERY/BODEGA       GYM/FITNESS FACILITY                  HOSPITAL
##                       694                          3                        65
##               HOTEL/MOTEL              JEWELRY STORE              LIQUOR STORE
##                        35                         12                        41
##              LOAN COMPANY   MULTI DWELL - APT BUILD MULTI DWELL - PUBLIC HOUS
##                         1                       2835                      4832
##                      NONE          PHOTO/COPY STORE                 PVT HOUSE
##                       175                          1                       951
##          RESTAURANT/DINER                     SCHOOL                SHOE STORE
##                       204                          1                        10
##            SMALL MERCHANT SOCIAL CLUB/POLICY LOCATI          STORAGE FACILITY
##                        37                         72                         1
##        STORE UNCLASSIFIED                SUPERMARKET           TELECOMM. STORE
##                        36                         21                        11
##             VARIETY STORE                VIDEO STORE
##                        11                          8
```

```r
#interesting categories here but more than half still have missing value
```

STATISTICAL_MURDER_FLAG

```r
table(df$STATISTICAL_MURDER_FLAG)
```

```
##
## false  true
## 22046  5266
```

```r
# from NYC's data website: "Shooting resulted in the victim's death which would be counted as a murder"
```

Shooter Age/Sex/Race

```r
table(df$PERP_AGE_GROUP)
```

```
##
##          (null)     <18    1020   18-24     224   25-44   45-64     65+     940
##    9344     640    1591       1    6222       1    5687     617      60       1
## UNKNOWN
##    3148
```

```r
table(df$PERP_SEX)
```

```
##
##        (null)       F       M       U
##   9310    640     424   15439    1499
```

```r
table(df$PERP_RACE)
```

```
##
##                                                             (null)
##                           9310                                 640
## AMERICAN INDIAN/ALASKAN NATIVE       ASIAN / PACIFIC ISLANDER
##                              2                             154
##                          BLACK                  BLACK HISPANIC
##                          11432                            1314
```

```
##                          UNKNOWN                           WHITE
##                             1836                             283
##                   WHITE HISPANIC
##                             2341
```

```r
#naturally there is a substantial proportion of missing values. Presumably police can't necessarily eve
#typical profile of categorized shooter is young, male, black/hispanic
#based on number of null/missing values it looks like a perp description (i.e. these columns in a singl
```

Victim Age/Sex/Race

```r
table(df$VIC_AGE_GROUP)
```

```
##
##     <18    1022   18-24   25-44   45-64     65+ UNKNOWN
##    2839       1   10086   12281    1863     181      61
```

```r
table(df$VIC_SEX)
```

```
##
##     F     M     U
##  2615 24686    11
```

```r
table(df$VIC_RACE)
```

```
##
## AMERICAN INDIAN/ALASKAN NATIVE       ASIAN / PACIFIC ISLANDER
##                             10                            404
##                          BLACK                 BLACK HISPANIC
##                          19439                           2646
##                        UNKNOWN                          WHITE
##                             66                            698
##                 WHITE HISPANIC
##                           4049
```

```r
#naturally victims have many fewer missing values (they got shot, much easier to find)
#profile again is young, male, black/hispanic
```

Geolocation Data - [needs to be visualized with geographical package]

Modeling fatality proportion vs precinct number of shootings

```r
#create a dataframe with each princinct's shooting count, fatality count, and proportion of shootings t
df$Fatal = df$STATISTICAL_MURDER_FLAG %>% recode('true' = 1, 'false' = 0)
fatality_prop_df = df %>% group_by(PRECINCT) %>% summarise(Count = n(), Fatalities = sum(Fatal), Fatal_
fatality_prop_df %>% head(5)
```

A typical question that arises from examining crime data is whether police/emergency resources are being fairly distributed throughout a jurisdiction. While we don't have any sort of deployment or response time data here for NYPD we can check to see if there is any relationship between the number of shootings in a precinct and the proportion that are fatal as a sort of proxy for the speed/efficacy of emergency response in general.

```
## # A tibble: 5 x 4
##   PRECINCT Count Fatalities Fatal_prop
##      <int> <int>      <dbl>      <dbl>
## 1        1    25          7      0.28
```

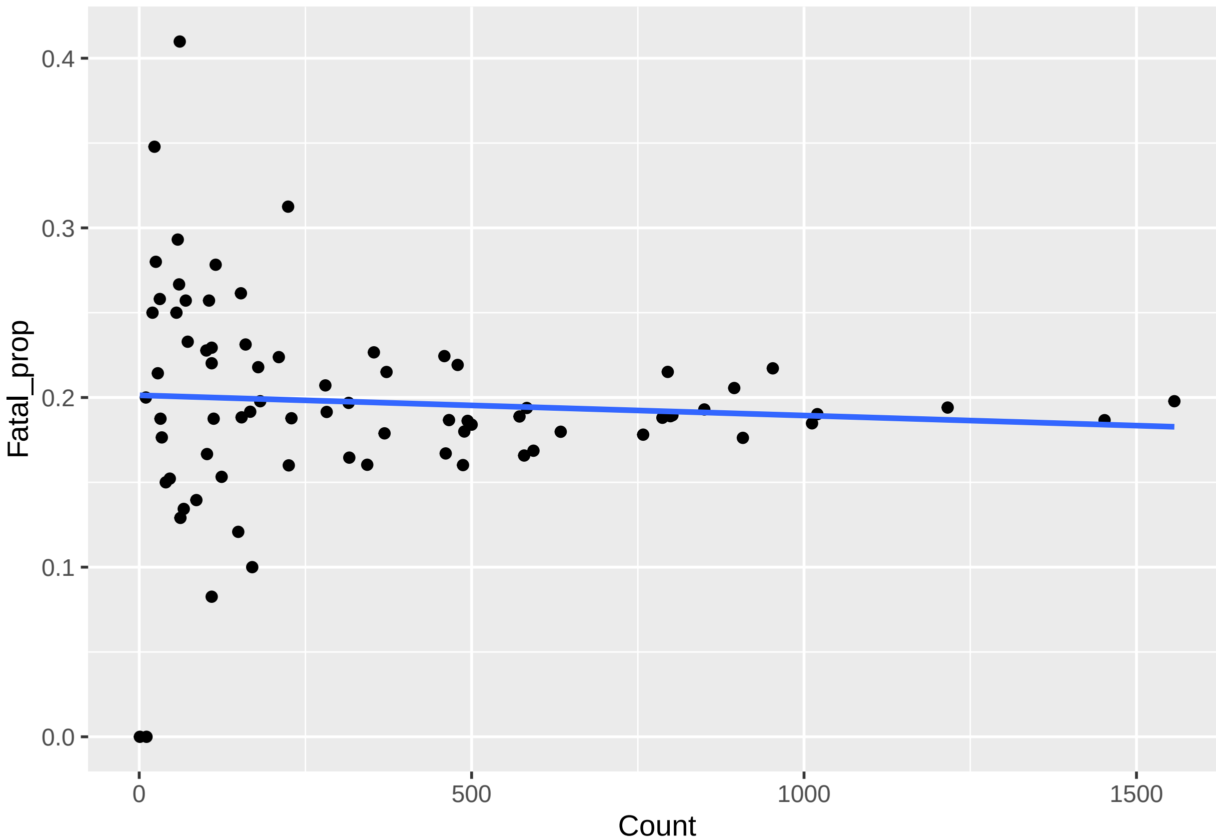
```
## 2         5    58          17      0.293
## 3         6    28           6      0.214
## 4         7   109           9      0.0826
## 5         9   109          25      0.229
```

```r
#Calculate linear model with x=number of shootings in a precinct and y=proportion of shootings that are
#As we can see from the model and the graph below there is essentially no correlation,
#so there is no suggestion *in this data* that more dangerous precincts are experiencing a generally wo
#emergency response.
model = lm(formula = Fatal_prop ~ Count, data = fatality_prop_df, )
summary(model)
```

```
##
## Call:
## lm(formula = Fatal_prop ~ Count, data = fatality_prop_df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.201298 -0.024434 -0.002226  0.027257  0.209253
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.013e-01  9.720e-03  20.711   <2e-16 ***
## Count       -1.192e-05  1.932e-05  -0.617    0.539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06047 on 75 degrees of freedom
## Multiple R-squared:  0.005046,   Adjusted R-squared:  -0.00822
## F-statistic: 0.3804 on 1 and 75 DF,  p-value: 0.5393
```

```r
ggplot(fatality_prop_df, aes(x = Count, y = Fatal_prop)) + geom_point() + geom_smooth(method = 'lm', se
```

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



Data Bias and Quality Discussion

- It is not clear whether this data includes instances where people literally were hit with a bullet or if there are also incidents where a victim was just shot at; either way there are presumably more 'shots fired' incidents not included in this data set which have different feature distributions from this dataset
- A lot of the location description columns are missing so many values that they are not particularly useful
- Perpetrator description columns may be subject to direct bias as they may be garnered from witness statements which can be faulty
- Victim description columns should be better since it is easier to actually locate and confirm a shooting victim