Data Science As A Field NYC Shootings Project

CU Student

5/21/2021

Data Science as a Field NYC Shooting Project

This is the week three assignment for the Data Science as a Field course. We'll examine and make a repeatable report about data from shootings in New York City. The data is read in and summarized before any transformations.

```
df <-read.csv('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD')

summary(df)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class :character   Class :character
##  Median : 83365370   Mode  :character   Mode  :character   Mode  :character
##  Mean   :102218616
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Length:23568
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  Mean   : 66.21   Mean   :0.3323
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:23568       Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE          X_COORD_CD         Y_COORD_CD
##  Length:23568       Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23568
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Cleaning the dataset by making all column names lowercase, and making appropriate columns dates or factors. Age group columns could be converted to numeric by taking the average of the range but for this analysis I've decided to convert them to factors. Then summarizing after changes.

```r
names(df) <- tolower(names(df))

factor_cols = c('boro', 'precinct', 'jurisdiction_code', 'location_desc', 'perp_sex', 'perp_race', 'vic

df[factor_cols] <- lapply(df[factor_cols], as.factor)

df <- df %>%
  mutate(occur_date = mdy(occur_date))

summary(df)
```

```
##   incident_key         occur_date          occur_time
##  Min.   :  9953245   Min.   :2006-01-01   Length:23568
##  1st Qu.: 55317014   1st Qu.:2008-12-30   Class :character
##  Median : 83365370   Median :2012-02-26   Mode  :character
##  Mean   :102218616   Mean   :2012-10-03
##  3rd Qu.:150772442   3rd Qu.:2016-02-28
##  Max.   :222473262   Max.   :2020-12-31
##
##             boro         precinct     jurisdiction_code
##  BRONX        :6700   75     : 1367   0   :19624
##  BROOKLYN     :9722   73     : 1282   1   :   54
##  MANHATTAN    :2921   67     : 1102   2   : 3888
##  QUEENS       :3527   79     :  920   NA’s:    2
##  STATEN ISLAND: 698   44     :  842
##                       47     :  815
##                       (Other):17240
##                    location_desc   statistical_murder_flag perp_age_group
##                           :13581   Length:23568                   :8459
##  MULTI DWELL - PUBLIC HOUS: 4230   Class :character        18-24  :5448
##  MULTI DWELL - APT BUILD  : 2551   Mode  :character        25-44  :4613
##  PVT HOUSE                :  858                           UNKNOWN:3156
##  GROCERY/BODEGA           :  572                           <18    :1354
##  BAR/NIGHT CLUB           :  558                           45-64  : 481
##  (Other)                  : 1218                           (Other):  57
##  perp_sex           perp_race     vic_age_group   vic_sex
##   : 8425   BLACK         :9855   <18     : 2525   F: 2195
##  F:  334                   :8425   18-24  : 9000   M:21353
##  M:13305   WHITE HISPANIC:1961   25-44  :10287   U:   20
##  U: 1504   UNKNOWN       :1869   45-64  : 1536
##            BLACK HISPANIC:1081   65+    :  155
##            WHITE         : 255   UNKNOWN:   65
##            (Other)       : 122
##                              vic_race      x_coord_cd         y_coord_cd
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Length:23568       Length:23568
##  ASIAN / PACIFIC ISLANDER      :  320   Class :character   Class :character
##  BLACK                         :16846   Mode  :character   Mode  :character
##  BLACK HISPANIC                : 2244
##  UNKNOWN                       :  102
##  WHITE                         :  615
##  WHITE HISPANIC                : 3432
##     latitude       longitude        lon_lat
##  Min.   :40.51   Min.   :-74.25   Length:23568
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Many of the features describing the perpetrator and the victim have missing data, represented by an empty string, and a value denoting unknown. I plan to fill missing data with the appropriate unknown value for each column and to allow that unknown category to remain a factor.

Borough Over the Years Analysis

For an initial analysis, I'd like to understand how the shootings are spread over the years and boroughs.

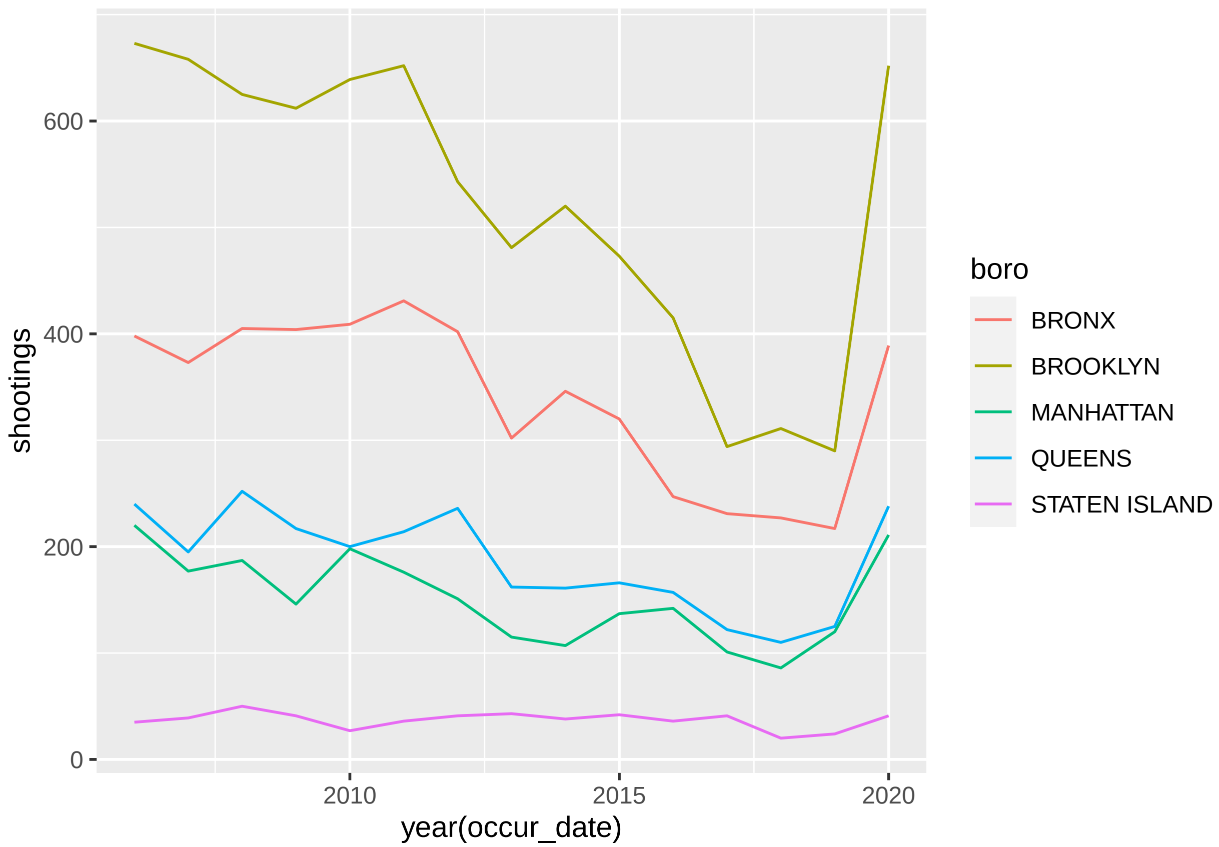
```
boro_year_total <- df %>%
  group_by(boro, year(occur_date)) %>%
  summarize(shootings = n_distinct(incident_key),) %>%
  ungroup()
```

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

```
summary(boro_year_total)
```

```
##             boro    year(occur_date)   shootings
##  BRONX        :15   Min.   :2006     Min.   : 20.0
##  BROOKLYN     :15   1st Qu.:2009     1st Qu.:112.5
##  MANHATTAN    :15   Median :2013     Median :211.0
##  QUEENS       :15   Mean   :2013     Mean   :247.5
##  STATEN ISLAND:15   3rd Qu.:2017     3rd Qu.:381.0
##                     Max.   :2020     Max.   :673.0
```

```
boro_year_total %>%
  ggplot(aes(x=`year(occur_date)`, y=shootings, group=boro, color=boro)) +
    geom_line()
```



2020 Marked a sharp increase in shootings for all the boroughs except Staten Island. This trend is certainly one for further analysis although detecting the cause is probably beyond the scope of this dataset.

The borough and year seem to go a long way in explaining the yearly shootings, so I've prepared a first model using those two variables to predict the number of shootings in a year.

```
mod <- lm(shootings ~ `year(occur_date)` + boro, data=boro_year_total)
summary(mod)
```

```
##
## Call:
## lm(formula = shootings ~ ‘year(occur_date)‘ + boro, data = boro_year_total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -191.502  -31.991   -1.818   26.871  194.272
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        18976.133   3499.565   5.422 8.16e-07 ***
## ‘year(occur_date)‘    -9.258      1.738  -5.325 1.19e-06 ***
## boroBROOKLYN         182.467     23.752   7.682 7.62e-11 ***
## boroMANHATTAN       -188.467     23.752  -7.935 2.63e-11 ***
## boroQUEENS          -153.733     23.752  -6.472 1.19e-08 ***
## boroSTATEN ISLAND   -303.133     23.752 -12.762  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 65.05 on 69 degrees of freedom
## Multiple R-squared:  0.8848, Adjusted R-squared:  0.8765
## F-statistic:   106 on 5 and 69 DF,  p-value: < 2.2e-16
```

Even with a simple linear model, the borough and year are highly effective in predicting the yearly shootings.

Location Type Analysis

I'd also like to understand more about the types of locations where the shootings are occuring and if the year over year trend is as apparent when slicing the data that way.

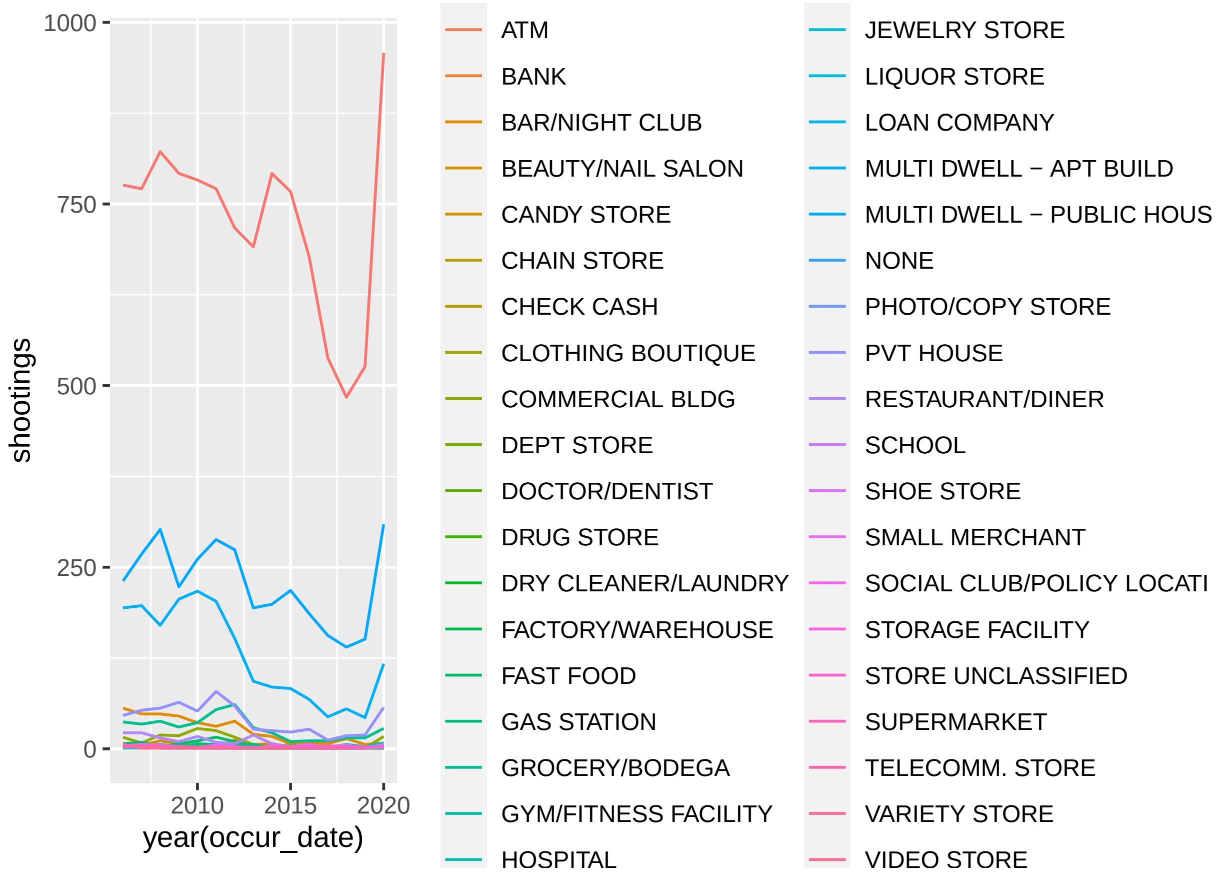
```
location_total <- df %>%
  group_by(location_desc, year(occur_date)) %>%
  summarize(shootings = n_distinct(incident_key),) %>%
  ungroup()
```

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

```
summary(location_total)
```

```
##                    location_desc year(occur_date)   shootings
##                           : 15   Min.   :2006     Min.   :  1.00
##  BAR/NIGHT CLUB           : 15   1st Qu.:2009     1st Qu.:  1.00
##  COMMERCIAL BLDG          : 15   Median :2012     Median :  4.00
##  GROCERY/BODEGA           : 15   Mean   :2012     Mean   : 61.26
##  MULTI DWELL - APT BUILD  : 15   3rd Qu.:2016     3rd Qu.: 22.50
##  MULTI DWELL - PUBLIC HOUS: 15   Max.   :2020     Max.   :958.00
##  (Other)                  :213
```

```
location_total %>%
  ggplot(aes(x=`year(occur_date)`, y=shootings, group =location_desc, color=location_desc)) +
    geom_line()
```



A few locations have vastly more shootings than the others. Which are those in 2020?

```
location_total %>% filter(shootings >25 & `year(occur_date)` == 2020)
```

```
## # A tibble: 5 x 3
##   location_desc                 ‘year(occur_date)‘ shootings
##   <fct>                                      <dbl>     <int>
## 1 ""                                          2020       958
## 2 "GROCERY/BODEGA"                            2020        28
## 3 "MULTI DWELL - APT BUILD"                   2020       117
## 4 "MULTI DWELL - PUBLIC HOUS"                 2020       309
## 5 "PVT HOUSE"                                 2020        57
```

Conclusion

This preliminary analysis shows a large increase in shootings in 2020 and the types of locations and boros with the most shootings. Looking more closely at residential shootings seems to be a promising area for future analysis.

A large source of bias in this analysis is the fact that the count of shootings does not take into account the populations and demographics of each of these boroughs. The boroughs are quite different in these respects so presenting total counts of shootings rather than rates is not indicative of the full picture. With the high number of rows missing data, it is also reasonable to wonder if there is reporting bias affecting this data even before this analysis.

The population/demographic bias could be addressed by supplementing this dataset with general information about the boroughs and analyzing the shootings in the context of the population and population density.

Addressing the possible reporting bias would likely involve a lot of in-person investigation to fill in the missing data or an effort to supplement this data with some collected by an outside organization.

Regarding personal bias, I'm neither an expert on shootings nor on New York so I have a large number of blind spots regarding this data and likely do not understand the nuance or context. I believe I have presented this report as an initial overview of a narrow slice of the data and not attempted to make it seem more authoritative or exhaustive than is warranted. Finally, I have noted that even a more thorough analysis will likely not provide us with causal information about these shootings.