NYPD Shooting by Age and Location

2023-02-08

Purpose and Problem Statement

In this analysis we will take a look at the nypd shooting data and attempt to ascertain what age groups are more responsible for shootings and if factors such as location descriptions, age group, and boro are a significant predictor of a shooting being a murder.

This information could potentially be used to impact policy designed to reduce shootings / murders.

Below are the list of packages required to run this file.

library(“tidyverse”) library(“lubridate”) library(“dplyr”) library(“ggplot2”) library(“scales”)

```
knitr::opts_chunk$set(echo = FALSE)
options(width = 120)
library("tidyverse")
```

```
## ── Attaching core tidyverse packages ──────────────────────────────────────────────────────── tidyvers
e 2.0.0 ──
## ✔ dplyr     1.1.0     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.1     ✔ tibble    3.1.8
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ─────────────────────────────────────────────────────────────────────────── tidyverse_conf
licts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the  ]8;;http://conflicted.r-lib.org/ conflicted package ]8;;  to force all conflicts to become errors
```

```
library("lubridate")
library("dplyr")
library("ggplot2")
library("scales")
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

Downloading the data.

Reading the csv file from the above link.

```
nypd<-read_csv(url_in[1])
```

```
## Rows: 25596 Columns: 19
## ── Column specification ────────────────────────────────────────────────────────────────────────────────
────────
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_R
ACE, L...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude
## 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.
```

Cleaning Data.

```
#Tidy data and give summary
#Remove columns not needed
nypd<-select(nypd, -c(JURISDICTION_CODE, Lon_Lat, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, BORO, PRECINCT, PE
RP_RACE, ))

#Format OCCUR_DATE Column
nypd$OCCUR_DATE<-as.Date(nypd$OCCUR_DATE, "%m/%d/%Y")

#Omit all NA values for locations
nypd.l<-subset(nypd, !is.na(nypd$LOCATION_DESC))
```

```
summary(nypd.l)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME       LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  9953245   Min.   :2006-01-01   Length:10619      Length:10619       Mode :logical
##  1st Qu.: 52550582   1st Qu.:2008-10-27   Class1:hms        Class :character   FALSE:8200
##  Median : 80989164   Median :2011-09-28   Class2:difftime   Mode  :character   TRUE :2419
##  Mean   :101844573   Mean   :2012-09-18   Mode  :numeric
##  3rd Qu.:150705404   3rd Qu.:2016-02-25
##  Max.   :238490103   Max.   :2021-12-31
##  PERP_AGE_GROUP       PERP_SEX          VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:10619       Length:10619       Length:10619       Length:10619       Length:10619
##  Class :character   Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Ages of Shooters and Victims

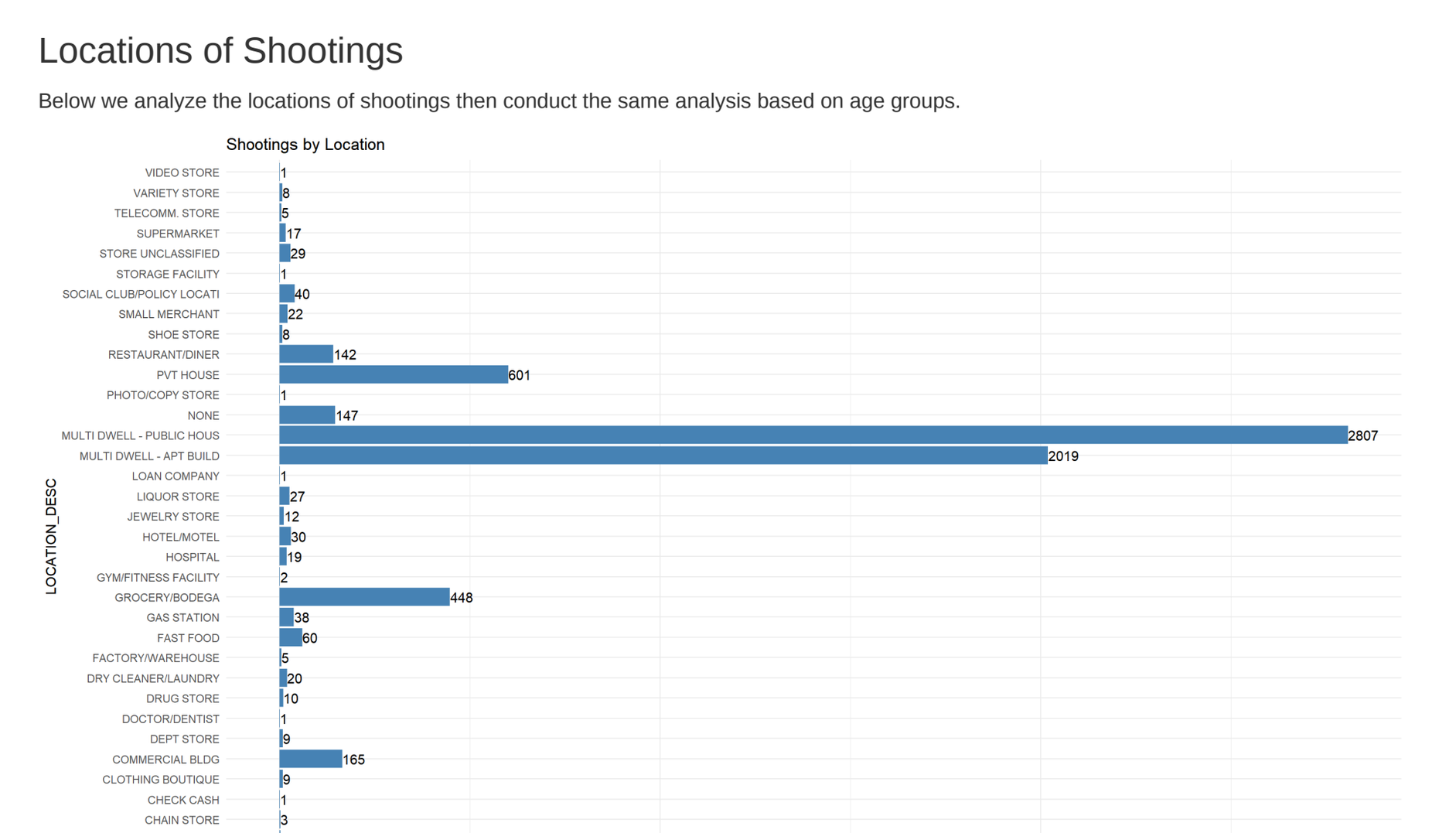
```
##       age count.perp count.vic
## 1     <18        601       768
## 2   18-24       2588      2650
## 3   25-44       2388      3175
## 4   45-64        257       559
## 5     65+         39        62
## 6 UNKNOWN       1368        27
```

The above table shows that the majority of known perps are between the ages of 18 and 44.

The table also shows that the majority of known victims are also between the ages of 18 and 44 with a larger amount in the 25-44 year old group when compared to perpetrators.

Locations of Shootings

Below we analyze the locations of shootings then conduct the same analysis based on age groups.

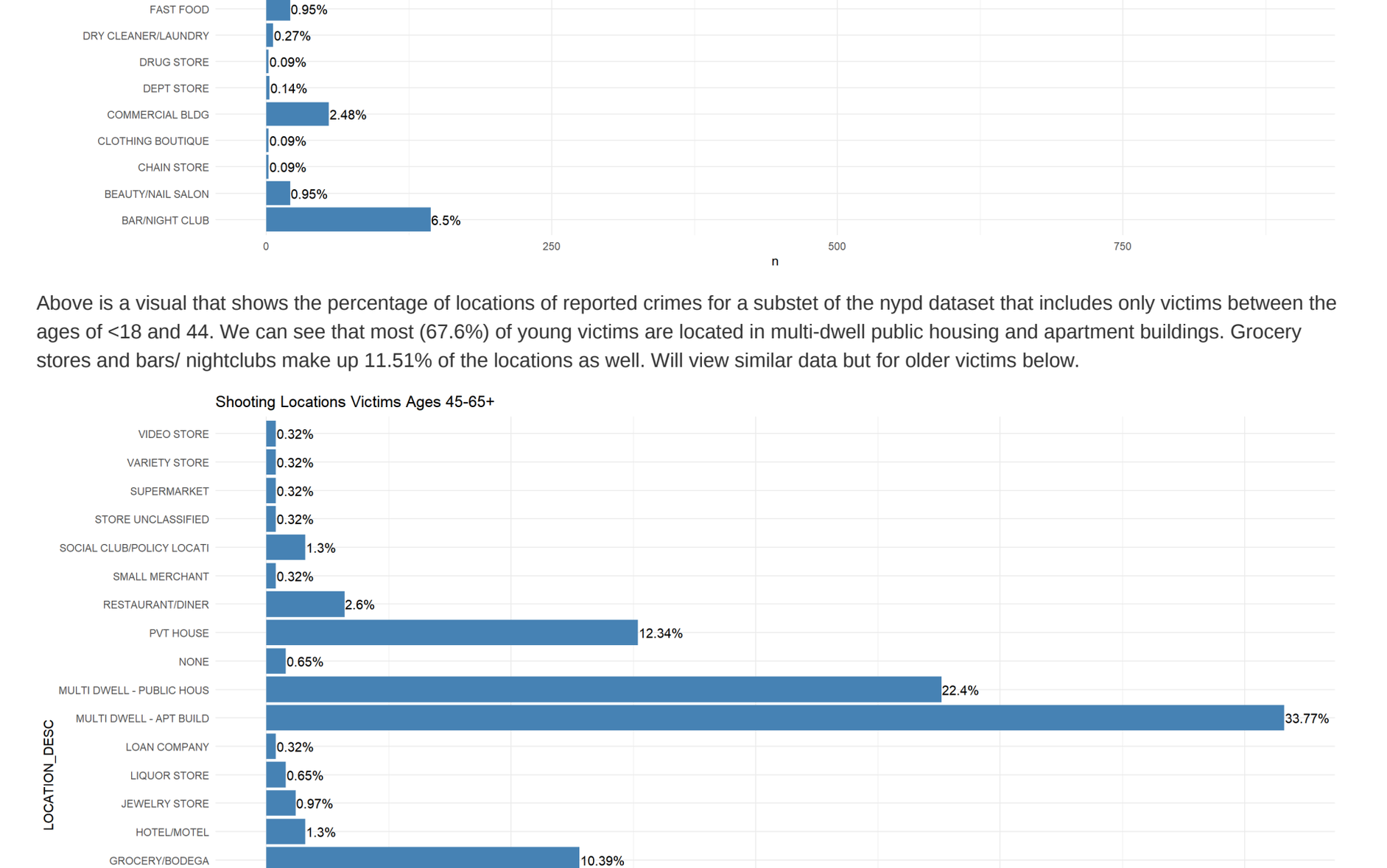


Above is a brief view into the different locations where shootings occurs. My goal is to take this insight and analyze the locations in relation to victim and perpetrator ages. Essentially, where do young people commit shootings / are victims and where do older people commit shootings / are victims?

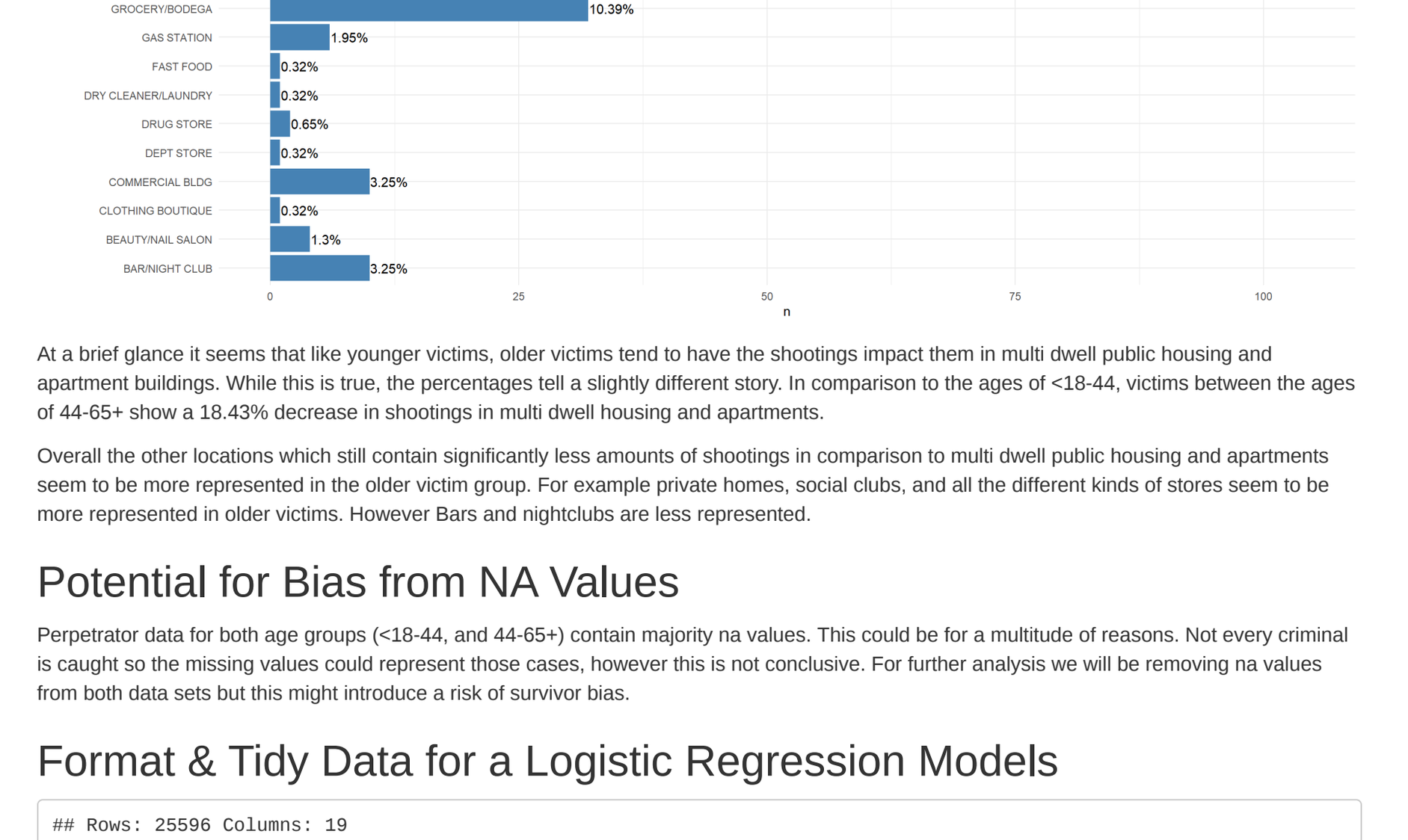
Age Based Analysis of Shooting Locations

```
## Warning in nypd.young.vic$VIC_AGE_GROUP == c("<18", "18-24", "25-44"): longer object length is not a multiple
of
## shorter object length
```

```
## Warning in nypd.old.vic$VIC_AGE_GROUP == c("45-64", "65+"): longer object length is not a multiple of shorter
object
## length
```



Above is a visual that shows the percentage of locations of reported crimes for a subset of the nypd dataset that includes only victims between the ages of <18 and 44. We can see that most (67.6%) of young victims are located in multi-dwell public housing and apartment buildings. Grocery stores and bars/ nightclubs make up 11.51% of the locations as well. Will view similar data but for older victims below.



At a brief glance it seems that like younger victims, older victims tend to have the shootings impact them in multi dwell public housing and apartment buildings. While this is true, the percentages tell a slightly different story. In comparison to the ages of <18-44, victims between the ages of 44-65+ show a 18.43% decrease in shootings in multi dwell housing and apartments.

Overall the other locations which still contain significantly less amounts of shootings in comparison to multi dwell public housing and apartments seem to be more represented in the older victim group. For example private homes, social clubs, and all the different kinds of stores seem to be more represented in older victims. However Bars and nightclubs are less represented.

Potential for Bias from NA Values

Perpetrator data for both age groups (<18-44, and 44-65+) contain majority na values. This could be for a multitude of reasons. Not every criminal is caught so the missing values could represent those cases, however this is not conclusive. For further analysis we will be removing na values from both data sets but this might introduce a risk of survivor bias.

Format & Tidy Data for a Logistic Regression Models

```
## Rows: 25596 Columns: 19
## ── Column specification ────────────────────────────────────────────────────────────────────────────────
────────
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_R
ACE, L...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude
## 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.
```

The below model attempts to show if location description is significant predictor
of a shooting being a murder

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ LOCATION_DESC, family = "binomial",
##     data = nypd.m)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0168  -1.4106   0.7216   0.8691   1.1774
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             1.457e+01  6.242e+02   0.023    0.981
## LOCATION_DESCBAR/NIGHT CLUB            -1.359e+01  6.242e+02  -0.022    0.983
## LOCATION_DESCBEAUTY/NAIL SALON         -1.342e+01  6.242e+02  -0.022    0.983
## LOCATION_DESCCANDY STORE                1.340e-06  8.827e+02   0.000    1.000
## LOCATION_DESCCHAIN STORE               -2.913e+01  1.081e+03  -0.027    0.979
## LOCATION_DESCCHECK CASH                -2.913e+01  1.081e+03  -0.027    0.979
## LOCATION_DESCCLOTHING BOUTIQUE         -1.428e+01  6.242e+02  -0.023    0.982
## LOCATION_DESCCOMMERCIAL BLDG           -1.360e+01  6.242e+02  -0.022    0.983
## LOCATION_DESCDEPT STORE                -1.457e+01  6.242e+02  -0.023    0.981
## LOCATION_DESCDOCTOR/DENTIST             1.341e-06  1.081e+03   0.000    1.000
## LOCATION_DESCDRUG STORE                -1.387e+01  6.242e+02  -0.022    0.982
## LOCATION_DESCDRY CLEANER/LAUNDRY        1.340e-06  6.838e+02   0.000    1.000
## LOCATION_DESCFACTORY/WAREHOUSE         -1.347e+01  6.242e+02  -0.022    0.983
## LOCATION_DESCFAST FOOD                 -1.285e+01  6.242e+02  -0.021    0.984
## LOCATION_DESCGAS STATION               -1.393e+01  6.242e+02  -0.022    0.982
## LOCATION_DESCGROCERY/BODEGA            -1.337e+01  6.242e+02  -0.021    0.983
## LOCATION_DESCGYM/FITNESS FACILITY      -2.913e+01  1.081e+03  -0.027    0.979
## LOCATION_DESCHOSPITAL                  -1.387e+01  6.242e+02  -0.022    0.982
## LOCATION_DESCHOTEL/MOTEL               -1.457e+01  6.242e+02  -0.023    0.981
## LOCATION_DESCJEWELRY STORE             -1.318e+01  6.242e+02  -0.021    0.983
## LOCATION_DESCLIQUOR STORE              -1.457e+01  6.242e+02  -0.023    0.981
## LOCATION_DESCLOAN COMPANY              -2.913e+01  1.081e+03  -0.027    0.979
## LOCATION_DESCMULTI DWELL - APT BUILD   -1.379e+01  6.242e+02  -0.022    0.982
## LOCATION_DESCMULTI DWELL - PUBLIC HOUS -1.335e+01  6.242e+02  -0.021    0.983
## LOCATION_DESCNONE                      -1.342e+01  6.242e+02  -0.021    0.983
## LOCATION_DESCPHOTO/COPY STORE           1.340e-06  1.081e+03   0.000    1.000
## LOCATION_DESCPVT HOUSE                 -1.403e+01  6.242e+02  -0.022    0.982
## LOCATION_DESCRESTAURANT/DINER          -1.267e+01  6.242e+02  -0.020    0.984
## LOCATION_DESCSHOE STORE                -1.406e+01  6.242e+02  -0.023    0.982
## LOCATION_DESCSMALL MERCHANT            -1.416e+01  6.242e+02  -0.023    0.982
## LOCATION_DESCSOCIAL CLUB/POLICY LOCATI -1.325e+01  6.242e+02  -0.021    0.983
## LOCATION_DESCSTORAGE FACILITY           1.340e-06  1.081e+03   0.000    1.000
## LOCATION_DESCSTORE UNCLASSIFIED        -1.359e+01  6.242e+02  -0.022    0.983
## LOCATION_DESCSUPERMARKET               -1.387e+01  6.242e+02  -0.022    0.982
## LOCATION_DESCTELECOMM. STORE           -1.416e+01  6.242e+02  -0.023    0.982
## LOCATION_DESCVARIETY STORE             -1.347e+01  6.242e+02  -0.022    0.983
## LOCATION_DESCVIDEO STORE                1.340e-06  1.081e+03   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6861.7  on 5863  degrees of freedom
## Residual deviance: 6740.7  on 5827  degrees of freedom
##   (7204 observations deleted due to missingness)
## AIC: 6814.7
##
## Number of Fisher Scoring iterations: 13
```

The above model shows that no particular location description is statistically significant when trying to predict if a shooting was a murder or not. P values are too high.

The below model attempts to show if perp age group is significant predictor
of a shooting being a murder

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_AGE_GROUP, family = "binomial",
##     data = nypd.m)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8454   0.6342   0.6857   0.7970   1.0455
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.50157    0.06780  22.147  < 2e-16 ***
## PERP_AGE_GROUP18-24 -0.17371    0.07506  -2.314   0.0206 *
## PERP_AGE_GROUP25-44 -0.51756    0.07463  -6.935 4.05e-12 ***
## PERP_AGE_GROUP45-64 -0.88869    0.11313  -7.856 3.98e-15 ***
## PERP_AGE_GROUP65+   -1.18311    0.27671  -4.276 1.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14344  on 13067  degrees of freedom
## Residual deviance: 14215  on 13063  degrees of freedom
## AIC: 14225
##
## Number of Fisher Scoring iterations: 4
```

According to the model, perp age group is a significant predictor of a shooting not being a murder. The model shows that they older the age group, the less likely the shooting is to be a murder.

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP, family = "binomial",
##     data = nypd.m)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8627   0.6231   0.7069   0.7744   0.9057
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.54072    0.06887  22.371  < 2e-16 ***
## VIC_AGE_GROUP18-24   -0.28146    0.07740  -3.636 0.000276 ***
## VIC_AGE_GROUP25-44   -0.49001    0.07508  -6.527 6.71e-11 ***
## VIC_AGE_GROUP45-64   -0.56412    0.09949  -5.670 1.43e-08 ***
## VIC_AGE_GROUP65+     -0.86156    0.21588  -3.991 6.58e-05 ***
## VIC_AGE_GROUPUNKNOWN -0.44211    0.34037  -1.299 0.193977
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14344  on 13067  degrees of freedom
## Residual deviance: 14279  on 13062  degrees of freedom
## AIC: 14291
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ BORO, family = "binomial",
##     data = nypd.m)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7256   0.7151   0.7227   0.7434   0.7623
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.08707    0.03626  29.979   <2e-16 ***
## BOROBROOKLYN       0.12216    0.05017   2.435   0.0149 *
## BOROMANHATTAN      0.14618    0.06659   2.195   0.0281 *
## BOROQUEENS         0.05777    0.06409   0.901   0.3674
## BOROSTATEN ISLAND  0.09638    0.10733   0.898   0.3692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14344  on 13067  degrees of freedom
## Residual deviance: 14337  on 13063  degrees of freedom
## AIC: 14347
##
## Number of Fisher Scoring iterations: 4
```

The above data model shows the Bronx, Brooklyn, and Manhattan are statistically significant when predicting murders.

Potential Bias in Analysis

When setting out on this analysis I initially assumed that younger folks were more likely to be shooters than older folks and while this might be the case for ages 65+, this was not necessarily the case for ages 25-44. A look at the numbers mitigated these biases. That being said people ages 65+ were significantly less likely to be shooters than those in the younger age groups.

As mentioned before another potential source of bias is resulting from from the omission of a large amount of NA values when analyzing the perp data. Perps who were not caught could potentially be over represented in a certain age group which has the potential to skew the data one way or another. In this event, we cannot be as conclusive with perp data as we are with victim data but the analysis was still worthwhile.

Conclusion

Given our analysis above it seems that shooting location and age possibly tell a story of proximity. Younger shooters and victims seem to both be primarily be in public housing and apartment buildings. This could imply that younger shooters tend to shoot younger victims more than they do older victims.

We also used 3 logistic regression modes to show the potential impact of location description, perp/victim, and boro on the shooting being classified as a murder.

The models indicated that location description is not a good predictor of murder and that perp/vic age group, and boro are statistically significant factors.

People in the <18-44 age group are more likely to be shot and murdered, and shootings that occur in the Bronx, Brooklyn, and Manhattan are more likely to be murders.