NYPD Shooting Incident Data

Student

2023-02-19

Questions of interest: 1.) Have shooting been increasing or decreasing over time. 2.) What are the demographics of the victims? 3.) Can the demographics of the victims be used to model to number victim in the dataset? 4.) Can the age range of the victim be used to predict the likelihood of dies from the shooting?

First, I am going to declare my libraries.

```
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Below is a link to the in data for NYPD Shooting Incidents. This data includes information on each shooting incident from 2006 to 2021. It includes information like date, time, and location of the shooting, along with basic demographic information on both the suspect and the victim. The demographic information includes age range, race, and sex. There is also an indicator on weather or not the shooting resulted in a death.

```
url_NYPD <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shooting_data <- read_csv(url_NYPD)
```

It is usually a good idea to a quick exploration of the data using the summary function in R.

```
summary(shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  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:25596       Length:25596       Length:25596       Length:25596
##  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:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  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
##
```

A lot of the key fields, including OCCUR_DATE, are in character format. So, there really isn't much to glean from the summary.

The next step is to use the summary function to look for missing values.

```
summary(is.na(shooting_data))
```

```
##  INCIDENT_KEY    OCCUR_DATE      OCCUR_TIME         BORO
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:25596     FALSE:25596     FALSE:25596     FALSE:25596
##
##   PRECINCT       JURISDICTION_CODE LOCATION_DESC   STATISTICAL_MURDER_FLAG
##  Mode :logical   Mode :logical     Mode :logical   Mode :logical
##  FALSE:25596     FALSE:25594       FALSE:10619     FALSE:25596
##                  TRUE :2           TRUE :14977
##  PERP_AGE_GROUP   PERP_SEX       PERP_RACE       VIC_AGE_GROUP
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:16252     FALSE:16286     FALSE:16286     FALSE:25596
##  TRUE :9344      TRUE :9310      TRUE :9310
##   VIC_SEX         VIC_RACE       X_COORD_CD      Y_COORD_CD
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:25596     FALSE:25596     FALSE:25596     FALSE:25596
##
##   Latitude       Longitude        Lon_Lat
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:25596     FALSE:25596     FALSE:25596
##
```

From this, we can see that typically over a third of the demographic information is missing for the suspect. Also, it is more likely that we will find racial bias in the suspect information. So, for this analysis, we will focus on the demographics of the victim.

Next, let's check the incident key to make sure it was unique to each row. Based on the query below it is not. However, this is mentioned in the footnotes attached to the landing page for the data: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8

"A shooting incident can have multiple victims involved and as a result duplicate INCIDENT_KEY's are produced. Each INCIDENT_KEY represents a victim but similar duplicate keys are counted as one incident"

```
dup_key <- shooting_data %>%
  dplyr::group_by (INCIDENT_KEY) %>%
  dplyr::summarise(cnt = n()) %>%
  dplyr::filter(cnt > 1)

head(dup_key)
```

```
## # A tibble: 6 x 2
##   INCIDENT_KEY   cnt
##          <dbl> <int>
## 1      9953250     2
## 2      9953255     2
## 3     10038637     2
## 4     10137408     2
## 5     10137411     3
## 6     10137412     2
```

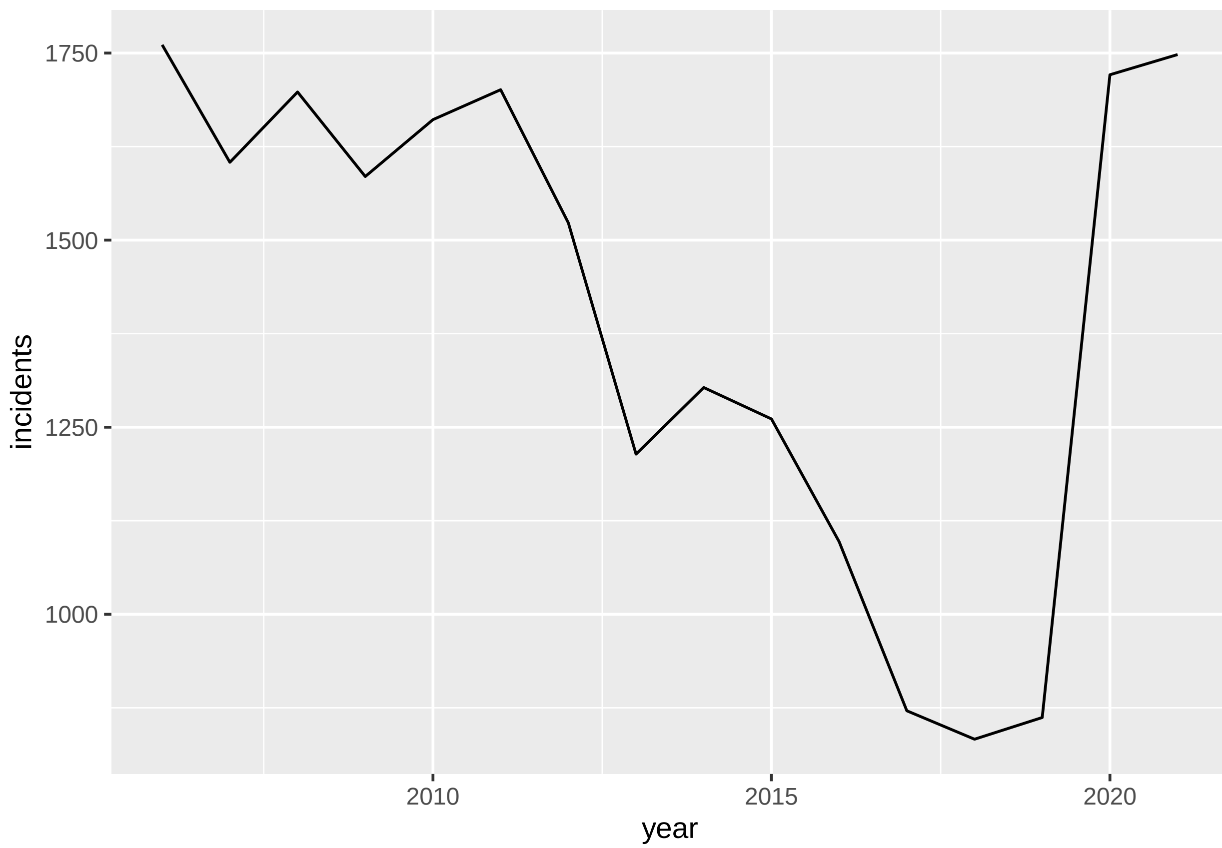
For this analysis, when referring to "incidents", it implies counting unique "INCIDENT_KEY"s, and when referring to "victims", it implies counting distinct rows. Typically, when looking at frequencies over time, the measure will be the count of "incidents", and when looking at demographics, like age or race, the count will be "victims".

Next, lets summarized the data to group by "INCIDENT_KEY", and the needed columns for the exploratory analysis to create a column named "VICTIMS" that it the count of distinct rows. Also, the "OCCUR_DATE" field was not in a date format, so let's change it to a date format for easier manipulation.

```
shooting_summary <- shooting_data %>%
  dplyr::group_by (INCIDENT_KEY,OCCUR_DATE,BORO,
              PRECINCT,VIC_AGE_GROUP,VIC_SEX,
              VIC_RACE) %>%
  dplyr::summarise(victim_cnt = n()) %>%
  dplyr::mutate(OCCUR_DATE=mdy(OCCUR_DATE))
```

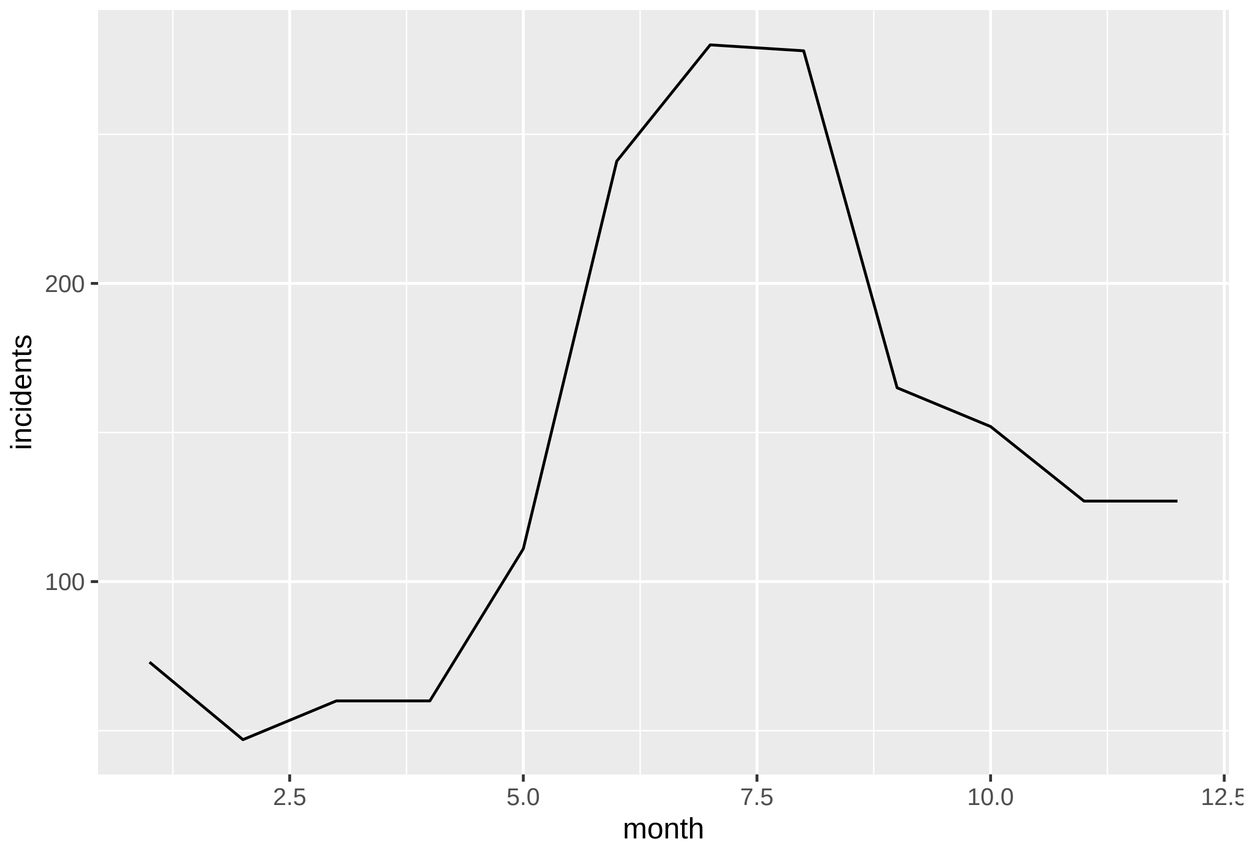
Now, let's see if shootings have been increasing over the years.

```
incidents_year <- shooting_summary %>%
  dplyr::group_by(year=year(OCCUR_DATE)) %>%
  dplyr::summarise(incidents = n())

ggplot(data=incidents_year, aes(x=year, y=incidents)) +
  geom_line()
```



It looks like shooting incidents were decreasing over the years but saw a major increase in 2020 and 2021. Now, let's see what months in 2020 saw the most shootings. Note, the footnotes states that the data only includes, "valid shooting incidents resulting in an injured victim". This can affect the measure of true shootings over time. Also, these are just raw numbers. For a future analysis, it would be best to adjust these numbers by population size.

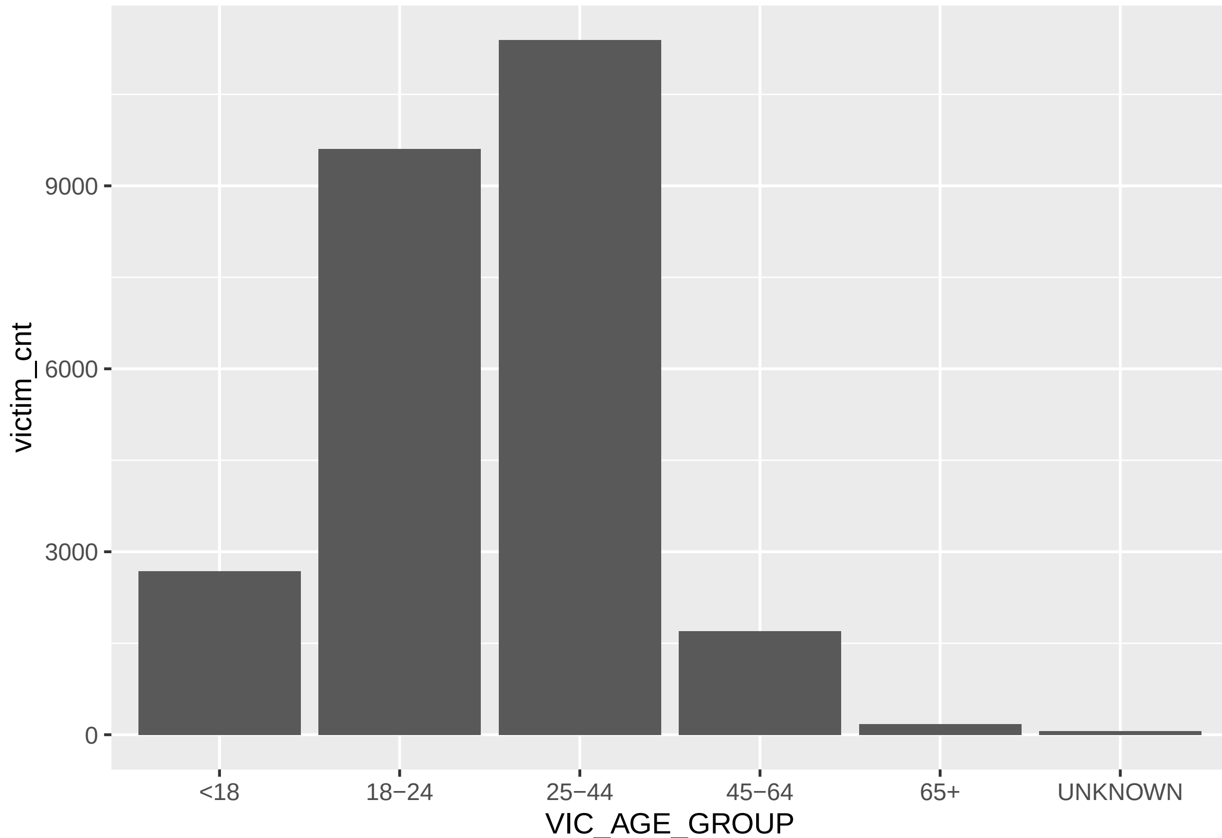
```
incidents_2020 <- shooting_summary %>%
  dplyr::group_by(year=year(OCCUR_DATE),month=month(OCCUR_DATE)) %>%
  dplyr::summarise(incidents = n()) %>%
  dplyr::filter(year == 2020)

ggplot(data=incidents_2020, aes(x=month, y=incidents)) +
  geom_line()
```



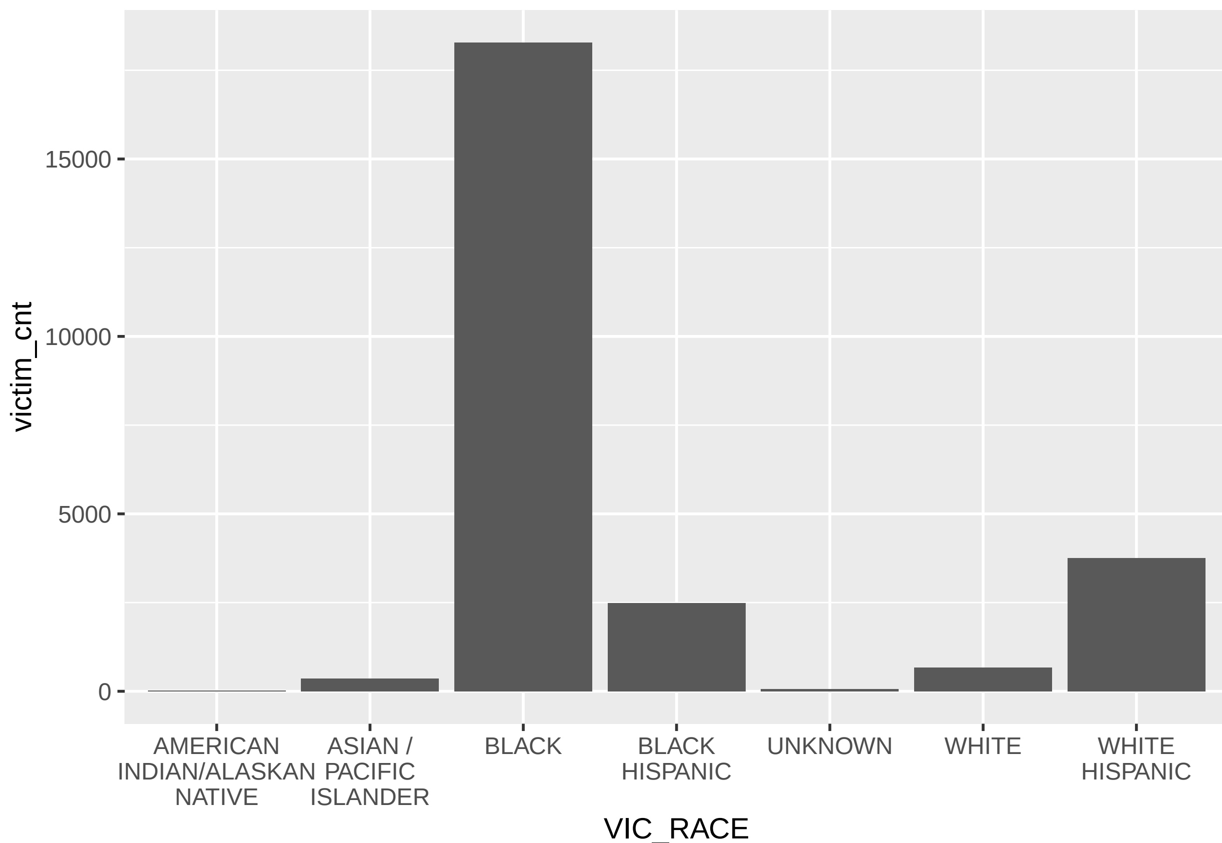
It looks like there was a spike in shootings during the summer months.

Now, let's explore some of the demographics of the victims.

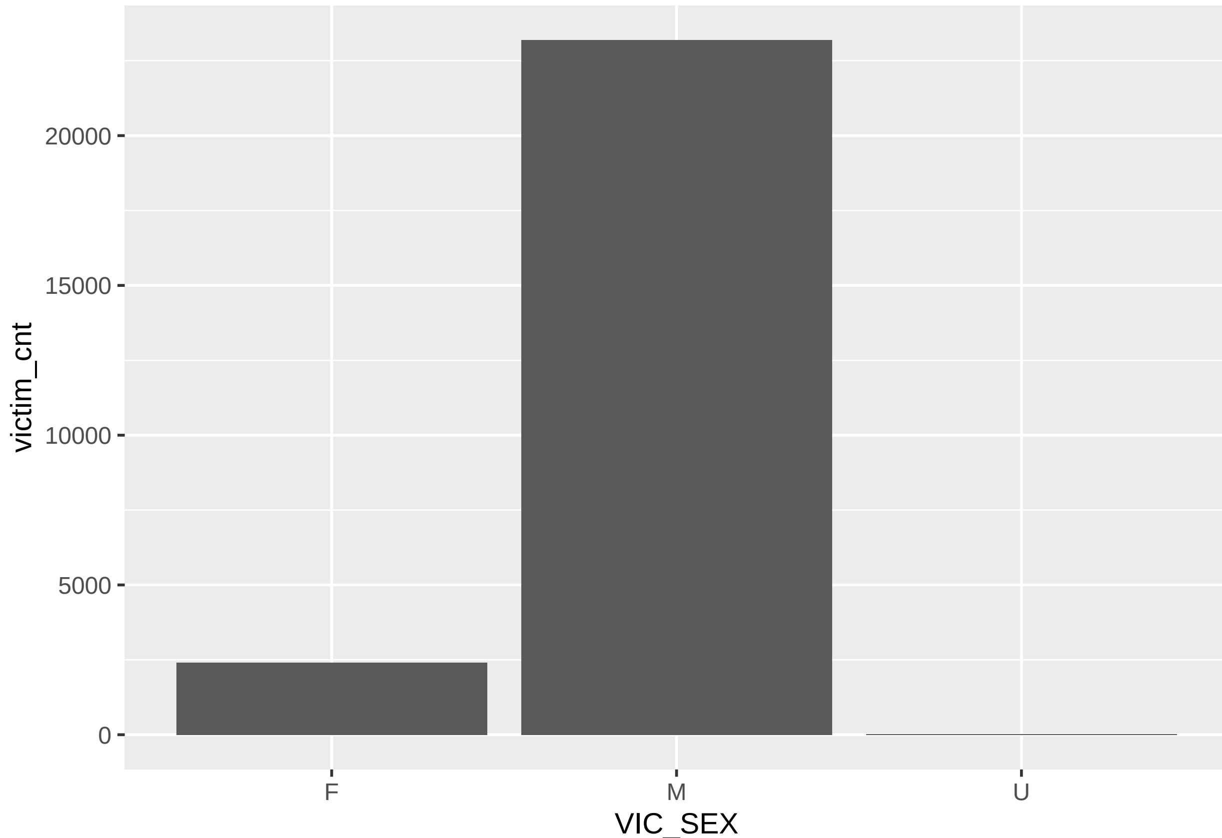
```
## Victim Count by Age Group
ggplot(data=shooting_summary, aes(x=VIC_AGE_GROUP, y=victim_cnt)) +
  geom_bar(stat="identity")
```



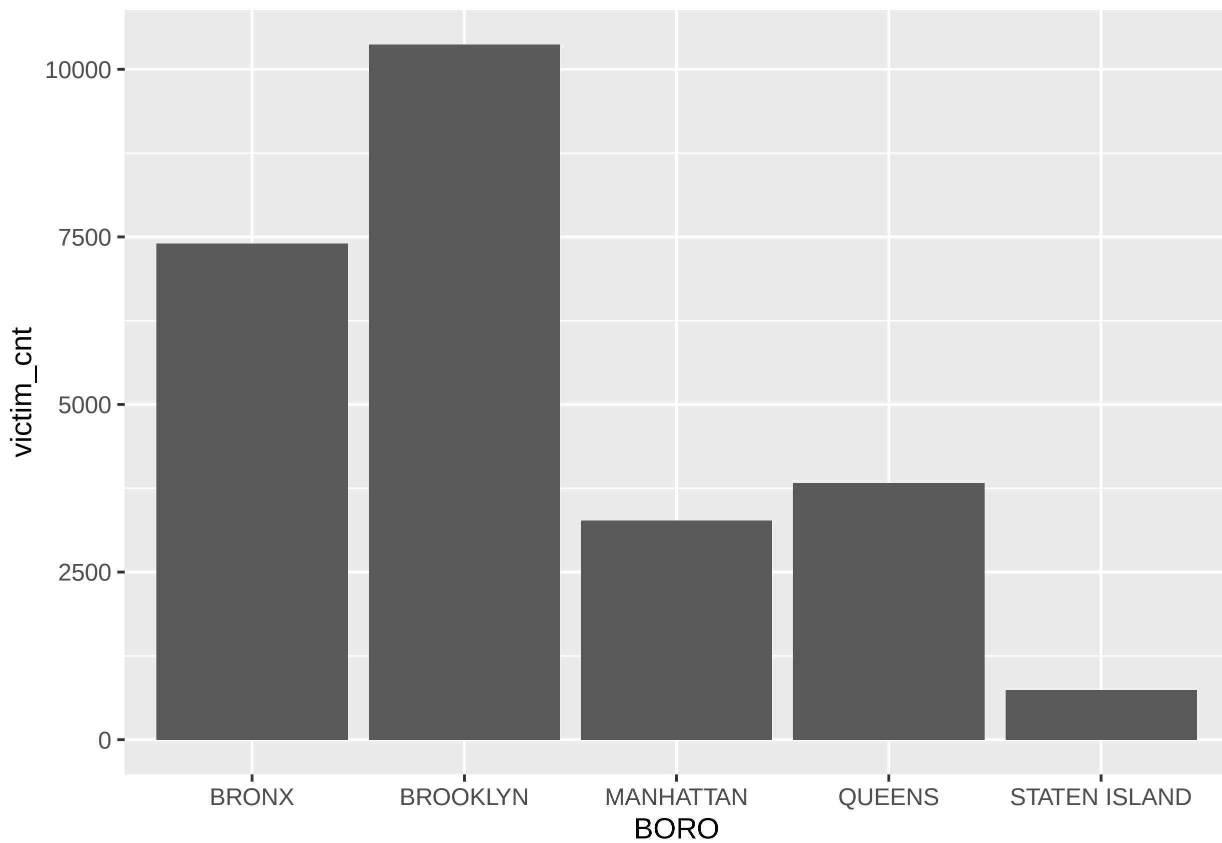

```
## Victim Count by Race
ggplot(data=shooting_summary, aes(x=VIC_RACE, y=victim_cnt)) +
  geom_bar(stat="identity") +
  scale_x_discrete(labels = function(x) str_wrap(x, width = 10))
```




```
## Victim Count by Sex
ggplot(data=shooting_summary, aes(x=VIC_SEX, y=victim_cnt)) +
  geom_bar(stat="identity")
```




```r
## Victim Count by Boro
ggplot(data=shooting_summary, aes(x=BORO, y=victim_cnt)) +
  geom_bar(stat="identity")
```



The majority of the shooting victims in the NYPD data appear to be black males between the ages of 18 and 44. The boros with the highest number of shootings are Brooklyn and Bronx.

Let's run a linear regression model with the victim count as the response and the demographic fields as our predictors.

```
shooting_summary_model <- shooting_summary %>%
  dplyr::group_by(VIC_AGE_GROUP,VIC_RACE,VIC_SEX,BORO) %>%
  dplyr::summarise(victim_cnt = sum(victim_cnt))

lm_shooting = lm(victim_cnt ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX + BORO, data = shooting_summary_model)
summary(lm_shooting)
```

```
##
## Call:
## lm(formula = victim_cnt ~ VIC_AGE_GROUP + VIC_RACE + VIC_SEX +
##     BORO, data = shooting_summary_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -361.35 -132.94  -34.54   66.25 2904.81
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -263.071    138.675  -1.897 0.058962 .
## VIC_AGE_GROUP18-24                  132.861     60.039   2.213 0.027796 *
```

```
## VIC_AGE_GROUP25-44                184.854     59.900   3.086 0.002254 **
## VIC_AGE_GROUP45-64                 -9.613     61.520  -0.156 0.875955
## VIC_AGE_GROUP65+                  -60.989     65.460  -0.932 0.352377
## VIC_AGE_GROUPUNKNOWN             -102.187     78.626  -1.300 0.194900
## VIC_RACEASIAN / PACIFIC ISLANDER  146.025    135.517   1.078 0.282262
## VIC_RACEBLACK                     502.072    131.984   3.804 0.000179 ***
## VIC_RACEBLACK HISPANIC            211.635    132.869   1.593 0.112452
## VIC_RACEUNKNOWN                   114.060    143.377   0.796 0.427054
## VIC_RACEWHITE                     190.460    132.428   1.438 0.151608
## VIC_RACEWHITE HISPANIC            249.201    132.447   1.882 0.061051 .
## VIC_SEXM                          170.525     38.814   4.393 1.64e-05 ***
## VIC_SEXU                         -103.489    127.761  -0.810 0.418689
## BOROBROOKLYN                       43.811     54.767   0.800 0.424484
## BOROMANHATTAN                     -80.888     58.187  -1.390 0.165708
## BOROQUEENS                        -60.558     56.697  -1.068 0.286496
## BOROSTATEN ISLAND                -158.361     63.202  -2.506 0.012852 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 300.7 on 253 degrees of freedom
## Multiple R-squared:  0.2781, Adjusted R-squared:  0.2296
## F-statistic: 5.734 on 17 and 253 DF,  p-value: 5.291e-11
```

This model is statistically significant and has an R-squared of ~.28. It looks like we may be able to improve the model adjusting the demographic fields or by creating new fields from our existing fields since several individual values of the categorical fields are not statistically significant. Also, there are likely other predictors in the dataset that could be added to improve the model. My bias is reflected in the predictors that were chosen for the model.

Finally, lets see if the victim's age group is useful in predicting the likelihood of dying from the shooting. In many instances, logistic regression is a useful machine learning algorithm for determining the probability of an event occurring given the predictor variables in the data.

```r
## Add 'death' field to the data.  This will be the response in our model.
shooting_data_model <- shooting_data %>%
  dplyr::mutate(death = if_else(STATISTICAL_MURDER_FLAG == "TRUE",1,0))

summary(shooting_data_model$death)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1925  0.0000  1.0000
```

```r
## Create training and testing datasets
set.seed(54)
randomize.rows = sample(nrow(shooting_data_model))
shooting.data = shooting_data_model[randomize.rows, ]

n = floor(0.8 * nrow(shooting.data))
index = sample(seq_len(nrow(shooting.data)), size = n)

shooting.train = shooting.data[index, ]
shooting.test = shooting.data[-index, ]

## Train the logistic regresssion model
```

```r
logit_shooting <- glm(death ~ VIC_AGE_GROUP, data = shooting.train, family = "binomial")
summary(logit_shooting)
```

```
##
## Call:
## glm(formula = death ~ VIC_AGE_GROUP, family = "binomial", data = shooting.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9246  -0.6966  -0.5972  -0.5344   2.0085
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.87425    0.06384 -29.357  < 2e-16 ***
## VIC_AGE_GROUP18-24    0.24060    0.07086   3.396 0.000685 ***
## VIC_AGE_GROUP25-44    0.58170    0.06877   8.458  < 2e-16 ***
## VIC_AGE_GROUP45-64    0.83263    0.08881   9.376  < 2e-16 ***
## VIC_AGE_GROUP65+      1.24564    0.18979   6.563 5.27e-11 ***
## VIC_AGE_GROUPUNKNOWN  0.64558    0.33436   1.931 0.053511 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19977  on 20475  degrees of freedom
## Residual deviance: 19794  on 20470  degrees of freedom
## AIC: 19806
##
## Number of Fisher Scoring iterations: 4
```

```r
prob <- predict(logit_shooting, newdata=shooting.test, type="response")
summary(prob)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1331  0.1633  0.2154  0.1913  0.2154  0.3478
```

It appears that the victim's age group is useful in predicting the likelihood of dying from the shooting. The coefficients are all statistically significant and follow the expected pattern. However, the coefficient of the intercept is approximately -1.87 and the largest coefficient in our model is about 1.24. Also, the max value of our predicted response is $< .5$. So, the model is no better than predicting that each observation will result in the patient living. More predictors will be needed to make a useful model.

Some of my bias appears in what predictors are used in the model. I only included the age predictor, but I could have included other predictors from this dataset or from another dataset. Also, some of the age groups are label 'UNKNOWN'. There are several options for dealing with missing data, but I chose to do nothing in this case.