NYPD Shooting Incidents - Data Exploration

23/11/2021

Import necessary libraries and the NYPD shooting csv file:

```
library(tidyverse)
library(ggplot2)
library(lubridate) # For use in converting date/time variables
```

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

Identification of bias

Before starting with data exploration, I want to first discuss the reliability of the data source and identify potential sources of bias. This data set only contains information on *reported* cases of shooting incidents that took place in New York City. Therefore, the analysis in this document is performed on a subset of all potential shooting incidents. It excludes shootings that were not reported nor captured by the NYPD. As a result, the real number of total cases may differ from the official records.

I believe the data source is reliable because the data comes from police reports. Most of the data categories don't contain bias because it is based on facts, for example, date, time of shooting, location, and victim details.

However, there are three fields that could lead to bias. These are the fields relating to the age, sex and race of the perpetrator. These three perpetrator fields contain information on potential suspects, which leads me to think that most of the data on the perpetrator comes from eye-witness reports from the surviving victims or nearby witnesses. This means that the information from these three columns cannot be considered as hard facts, and therefore can contain bias. For example, it can be difficult to tell the age of a person you don't know. If the shooter is fully clothed and wearing a mask, then it can be difficult to tell the sex or race of the person. The victims could default to describing the shooter as a particular race due to the victim's societal bias.

Initial look at data

First lets see what types of data the csv file contains. The subheadings are as follows:

```
colnames(shooting_csv)
```

```
##  [1] "INCIDENT_KEY"            "OCCUR_DATE"
##  [3] "OCCUR_TIME"              "BORO"
##  [5] "PRECINCT"                "JURISDICTION_CODE"
##  [7] "LOCATION_DESC"           "STATISTICAL_MURDER_FLAG"
##  [9] "PERP_AGE_GROUP"          "PERP_SEX"
## [11] "PERP_RACE"               "VIC_AGE_GROUP"
## [13] "VIC_SEX"                 "VIC_RACE"
## [15] "X_COORD_CD"              "Y_COORD_CD"
## [17] "Latitude"                "Longitude"
```

```
## [19] "Lon_Lat"
```

There are 19 different field names. Out of these, there are 8 headings that contains information that I don't know how to utilize (since I'm not from the US), or not useful for analysis. They are:
1. Incident_key
2. Precinct
3. Jurisdiction_code
4. x_coord_cd
5. y_coord_cd
6. Latitude
7. Longitude
8. Lon_Lat

I will remove the above 8 columns from the data table.

```
shooting <- shooting_csv %>%
  select(-c(INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, Latitude,
            Longitude, Lon_Lat))
```

Now check for missing data:

```
sum(is.na(shooting))
```

```
## [1] 38398
```

```
sum(is.na(shooting$OCCUR_DATE))
```

```
## [1] 0
```

```
sum(is.na(shooting$OCCUR_TIME))
```

```
## [1] 0
```

```
sum(is.na(shooting$BORO))
```

```
## [1] 0
```

```
sum(is.na(shooting$LOCATION_DESC))
```

```
## [1] 13581
```

```
sum(is.na(shooting$STATISTICAL_MURDER_FLAG))
```

```
## [1] 0
```

```
sum(is.na(shooting$PERP_AGE_GROUP))
```

```
## [1] 8295
```

```
sum(is.na(shooting$PERP_SEX))
```

```
## [1] 8261
```

```
sum(is.na(shooting$PERP_RACE))
```

```
## [1] 8261
```

```
sum(is.na(shooting$VIC_AGE_GROUP))
```

```
## [1] 0
```

```
sum(is.na(shooting$VIC_SEX))
```

```
## [1] 0
```

```
sum(is.na(shooting$VIC_RACE))
```

```
## [1] 0
```

This NYPD shooting dataset contains some missing data regarding two main categories: location description and information regarding the perpetrator (particularly age, sex and race). This is to be expected. The police do not always know who the perpetrators are, and the location description appears to only refer to general locations such as hotel, bar, supermarket, etc. Shooting incidents that occur in an alley probably doesn't have a suitable description, hence the missing data.

Data Exploration

Lets do some basic data exploration to get a better understanding of the data. I will mainly focus on the victims in this analysis, because there are many missing data points regarding the shooter. I will use questions or comparisons as subheadings to organize this section.

How many shootings resulted in death?

The statistical_murder_flag variable indicates whether or not the shooting incident resulted in the victim's death, which would indicate murder.

```
# Length of this column also indicates total cases because it contains no missing data
length(shooting$STATISTICAL_MURDER_FLAG)
```

```
## [1] 23585
```

```
murders <- sum(shooting$STATISTICAL_MURDER_FLAG == TRUE)
murders
```

```
## [1] 4500
```

```
ratio <- sum(shooting$STATISTICAL_MURDER_FLAG == TRUE)/length(shooting$STATISTICAL_MURDER_FLAG)
ratio
```

```
## [1] 0.1907992
```

This dataset contains 23585 cases of shootings, out of which 4500 are considered to be cases of murder. According to the footnotes from the data source, the STATISTICAL_MURDER_FLAG variable is defined to be shootings that resulted in the victim's death and would be counted as a murder. Using this definition, the result would indicate that only 19% of shootings result in death, and that the survival rate for shootings in New York is 81%. This is a very interesting result to me, because I expected the death rate for shooting incidents to be higher than 1 in 5 cases. My initial uneducated guess would have been a death rate of around 40—50% because of how dangerous gunshot wounds can be.

Does race affect the rate of shooting incidents?

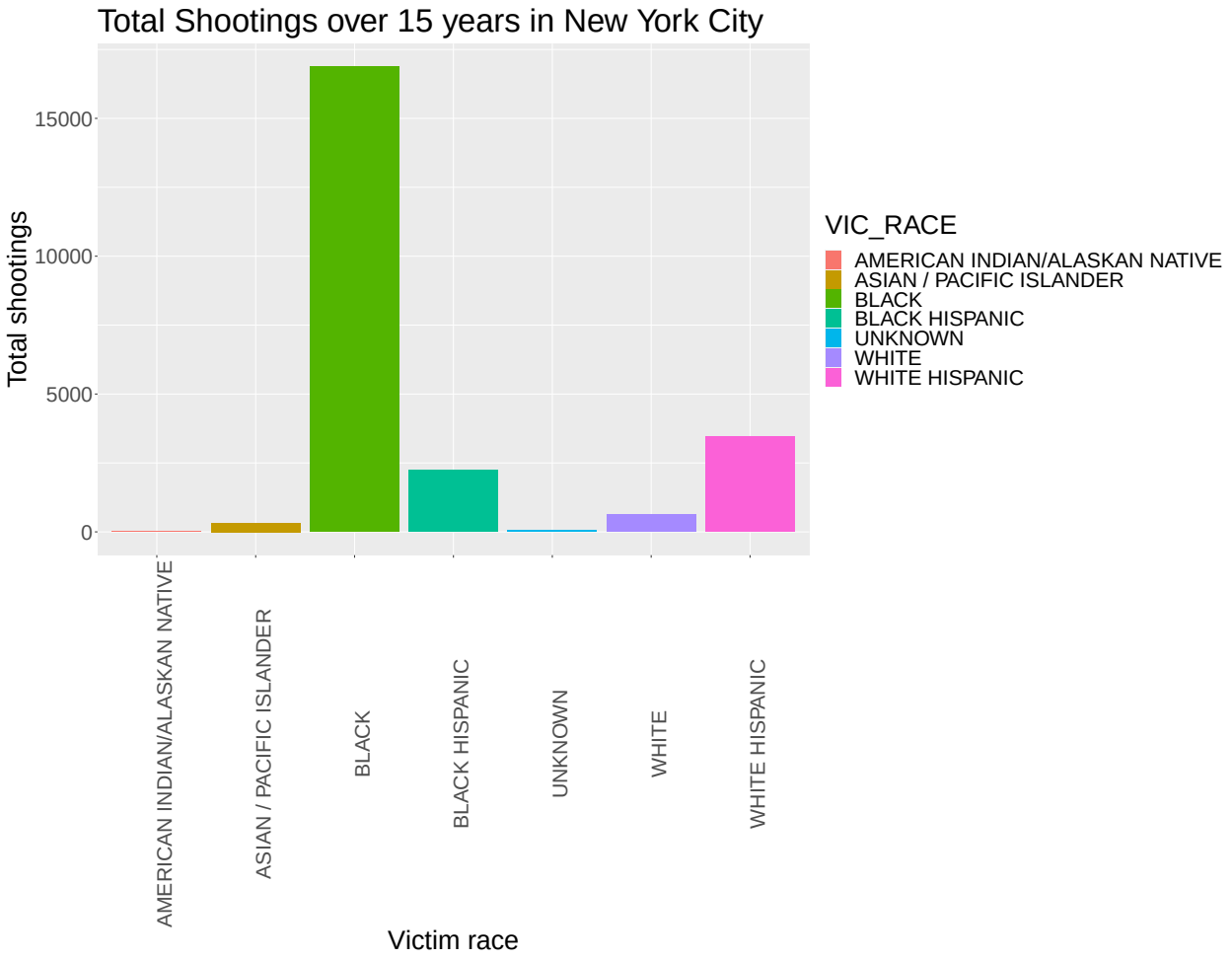
Lets see the racial distribution of all the shooting incidents:

```
sort(table(shooting$VIC_RACE), decreasing = TRUE)
```

```
##
##                          BLACK                 WHITE HISPANIC
##                          16869                          3450
##                 BLACK HISPANIC                          WHITE
##                           2245                           620
##       ASIAN / PACIFIC ISLANDER                        UNKNOWN
##                            327                            65
## AMERICAN INDIAN/ALASKAN NATIVE
```

```
##                                 9
```

```
shooting %>%
  ggplot(aes(x=VIC_RACE, fill = VIC_RACE)) +
  geom_bar() +
  theme(text=element_text(size=28), axis.text.x = element_text(angle = 90)) +
  labs(x = "Victim race", y = "Total shootings", title = "Total Shootings over 15 years in New York City
```




```
total_cases <- length(shooting$VIC_RACE)
total_cases
```

```
## [1] 23585
```

```
black_cases <- sum(shooting$VIC_RACE == "BLACK") / total_cases * 100
cat(black_cases, '%')
```

```
## 71.52427 %
```

```
white_hisp_cases <- sum(shooting$VIC_RACE == "WHITE HISPANIC") / total_cases * 100
cat(white_hisp_cases, '%')
```

```
## 14.62794 %
```

```
black_hisp_cases <- sum(shooting$VIC_RACE == "BLACK HISPANIC") / total_cases * 100
cat(black_hisp_cases, '%')
```

```
## 9.518762 %
```

```
top_3 <- black_cases + white_hisp_cases + black_hisp_cases
cat(top_3, '%')
```

```
## 95.67098 %
```

```
white_cases <- sum(shooting$VIC_RACE == "WHITE") / total_cases * 100
cat(white_cases, '%')
```

```
## 2.628789 %
```

The analysis shows that the top three racial groups make up a staggering 95.7% of all shooting victims in New York since 2006. Black people were exposed to the most incidents at 71.5%, followed by white and black Hispanic people at 14.6% and 9.5%, respectively. White Americans only make up 2.6% of all shooting victims.

There is a large disparity between the victim races. Shootings are disproportionate against black people, which is to be expected because from news reports I am aware that racism is still a big problem in America.

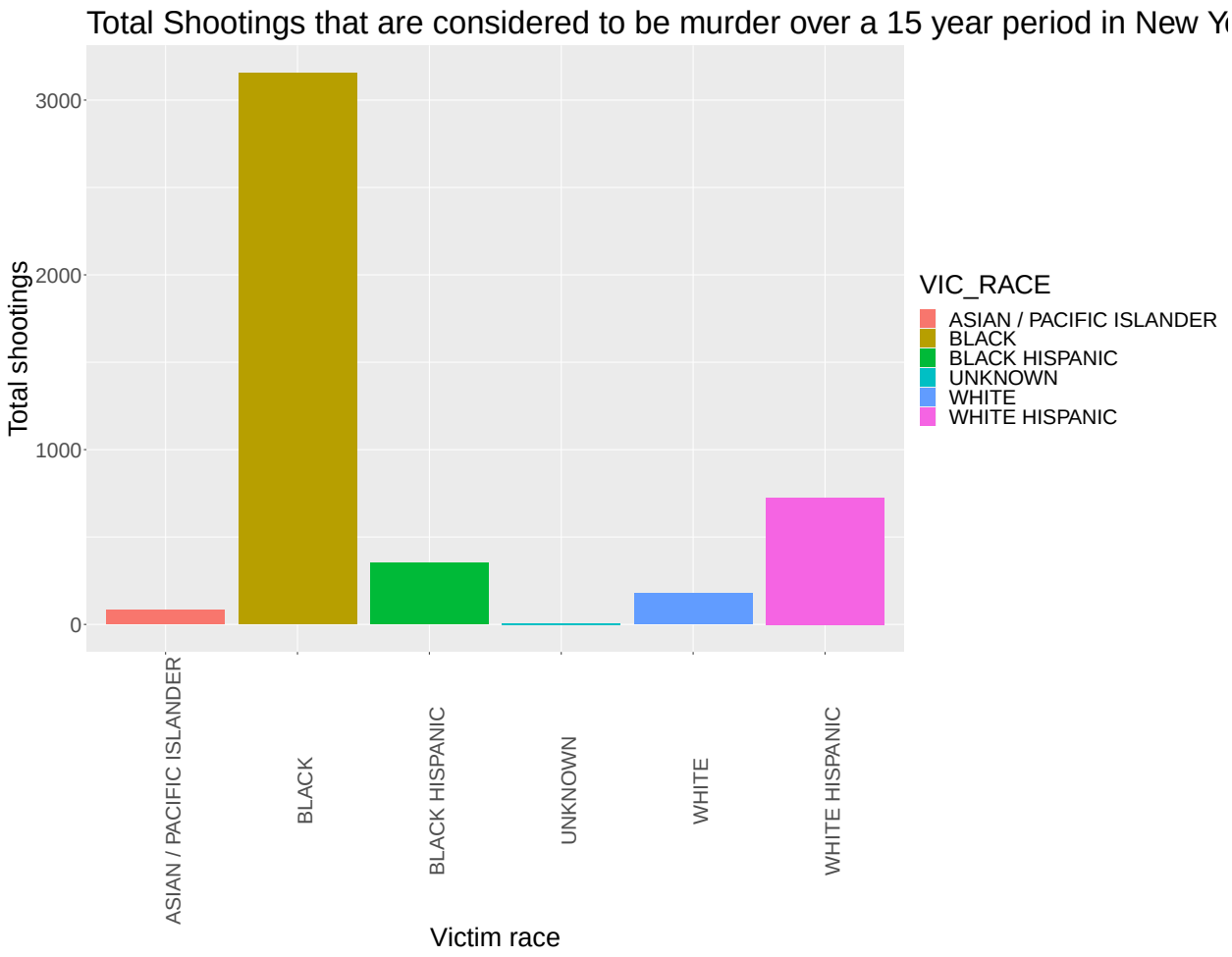
Does race affect death rate?

The graph above illustrated all shootings, but lets take a closer look at the cases that are considered to be murders to see if the same pattern can be seen.

```
race_death<- shooting %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE)
sort(table(race_death$VIC_RACE), decreasing = TRUE)
```

```
##
##                    BLACK           WHITE HISPANIC           BLACK HISPANIC
##                     3155                      725                      352
##                    WHITE ASIAN / PACIFIC ISLANDER                  UNKNOWN
##                      178                       83                        7
```

```
race_death %>%
  ggplot(aes(x=VIC_RACE, fill = VIC_RACE)) +
  geom_bar() +
  theme(text=element_text(size=28), axis.text.x = element_text(angle = 90)) +
  labs(x = "Victim race", y = "Total shootings", title = "Total Shootings that are considered to be murd
```




```
total_cases <- length(race_death$VIC_RACE)
total_cases
```

```
## [1] 4500
```

```
black_cases <- sum(race_death$VIC_RACE == "BLACK") / total_cases * 100
cat(black_cases, '%')
```

```
## 70.11111 %
```

```
white_hisp_cases <- sum(race_death$VIC_RACE == "WHITE HISPANIC") / total_cases * 100
cat(white_hisp_cases, '%')
```

```
## 16.11111 %
```

```
black_hisp_cases <- sum(race_death$VIC_RACE == "BLACK HISPANIC") / total_cases * 100
cat(black_hisp_cases, '%')
```

```
## 7.822222 %
```

```
white_cases <- sum(race_death$VIC_RACE == "WHITE") / total_cases * 100
cat(white_cases, '%')
```

```
## 3.955556 %
```

The top three racial groups that experience the most death is the same as in the total shootings, with black people at the top, followed by Hispanic white and black people. This is to be expected, since exposure to

more shooting incidents greatly increases the number of deaths.

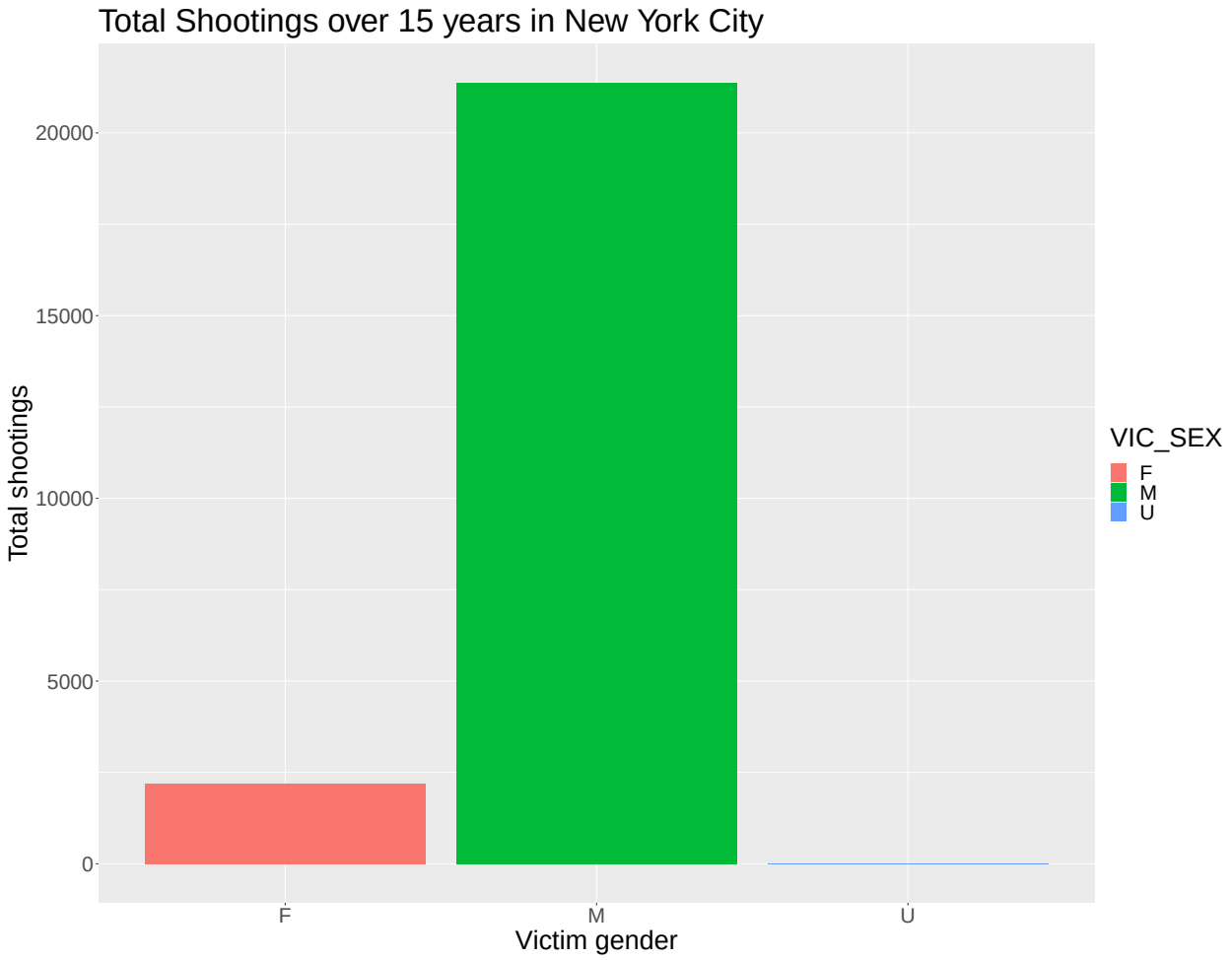
Does gender affect the rate of shooting incidents?

Lets see if there is any difference between the amount of male and female victims:

```
table(shooting$VIC_SEX)
```

```
##
##     F     M     U
##  2204 21370    11
```

```
shooting %>%
  ggplot(aes(x=VIC_SEX, fill = VIC_SEX)) +
  geom_bar() +
  theme(text=element_text(size=28)) +
  labs(x = "Victim gender", y = "Total shootings", title = "Total Shootings over 15 years in New York C
```




```
total_cases <- length(shooting$VIC_SEX)
total_cases
```

```
## [1] 23585
```

```
male_cases <- sum(shooting$VIC_SEX == "M") / total_cases * 100
cat(male_cases, '%')
```

```
## 90.60844 %
```

```
female_cases <- sum(shooting$VIC_SEX == "F") / total_cases * 100
cat(female_cases, '%')
```

```
## 9.344923 %
```

There is a surprisingly large gap between male and female victims. The data shows that men are ten times more likely to get shot than women. This is a huge surprise for me and I can't come up with a hypothesis on why that is the case.

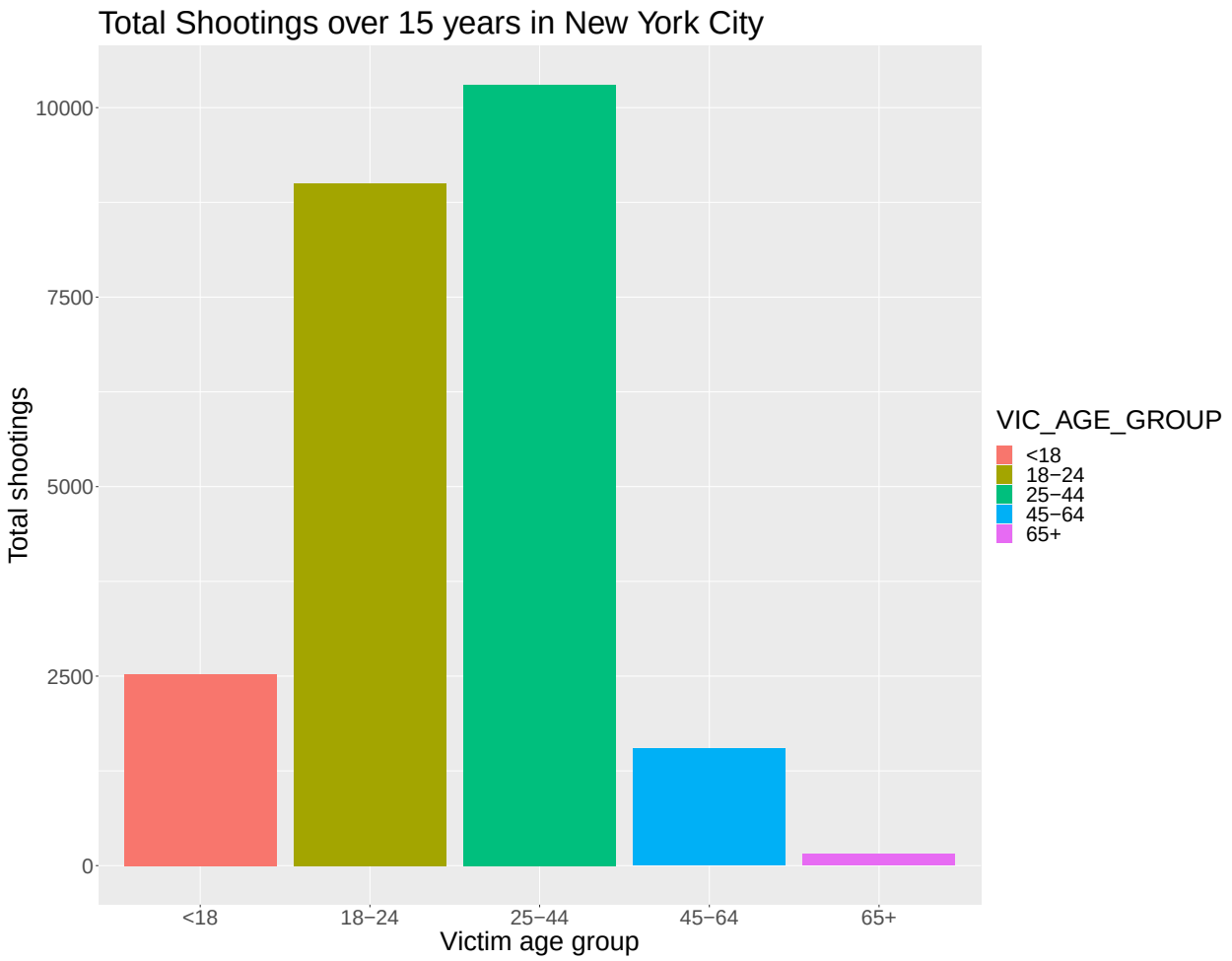
What is the age group distribution of shooting victims?

Lets see which age groups contain the most victims:

```
age_data <- shooting %>%
  filter(VIC_AGE_GROUP != 'UNKNOWN') # Drop data points with unknown age
table(age_data$VIC_AGE_GROUP)
```

```
##
##   <18 18-24 25-44 45-64   65+
##  2525  9003 10303  1541   154
```

```
age_data %>%
  ggplot(aes(x=VIC_AGE_GROUP, fill = VIC_AGE_GROUP)) +
  geom_bar() +
  theme(text=element_text(size=28)) +
  labs(x = "Victim age group", y = "Total shootings", title = "Total Shootings over 15 years in New Yor
```




```
combined <- shooting %>%
  filter(VIC_AGE_GROUP=='18-24' | VIC_AGE_GROUP=='25-44')
ratio <- length((combined$VIC_AGE_GROUP)) / total_cases * 100
ratio
```

```
## [1] 81.85711
```

Two age groups make up the large majority, which are the 18-24 and 25-44 age groups. Together, they make up 81.9% of shooting victims. These two age groups are young adults and working class citizens, which means that they are also the most active age groups in the community. This result lines up with my expectations.

Are there any potential insights between solved and unsolved cases?

I am interested to see if I can spot any interesting insights between solved and unsolved cases. The data doesn't indicate whether the shooter was arrested or not, but for the current analysis, I will assume that if the police have information regarding the age, sex and race of the perpetrator, then that would mean that the police at least has an idea who was responsible.

If the perpetrator's information is missing, then that means that the police have no information on the shooter. For analysis, I will define the number of unsolved cases to be the cases where the police have no leads on the shooter. I'm aware that simply because the police has general information on the shooter doesn't mean that they have been directly identified or arrested. But without more data, I can only assume that cases with missing perpetrator data are definitely unsolved, because having no information on the shooter also means that the shooter has escaped capture and thus have gotten away with the crime. I will therefore

label the two groups as potentially solved and definitely unsolved.

Organize data

I will split the data into two tables: the first table will contain cases where the police has general information on the shooter (i.e. there is data on the age, sex and race of the shooter), and a second section that contains cases that has no information on the perpetrator.

```
PotentiallySolved <- shooting %>% filter(!is.na(shooting$PERP_AGE_GROUP))
Unsolved <- shooting %>% filter(is.na(shooting$PERP_AGE_GROUP))

total_PotentiallySolved <- length(PotentiallySolved$OCCUR_DATE) # Total cases of potentially solved sho
total_PotentiallySolved
```

```
## [1] 15290
```

```
total_Unsolved <- length(Unsolved$OCCUR_DATE) # Total cases of definitely unsolved shootings
total_Unsolved
```

```
## [1] 8295
```

After re-organizing, the PotentiallySolved data set contains 15290 cases, while the Unsolved data set contains 8295 cases.

Comparison between solved and unsolved cases

Out of 23585 total cases of shootings, 8295 cases contain no information regarding the perpetrator. This shows that at least a third (or 35.2%) of all shooting incidents in New York are unsolved. Lets see what the ratio of murder cases are unsolved, and compare it to unsolved cases where the victims survived.

```
total_murders <- sum(shooting$STATISTICAL_MURDER_FLAG == TRUE) # Total cases of all murders

unsolved_murders <- sum(Unsolved$STATISTICAL_MURDER_FLAG == TRUE) # Number of unsolved murders
unsolved_murders_ratio <- unsolved_murders / total_murders
unsolved_murders_ratio
```

```
## [1] 0.3448889
```

```
survivors <- sum(shooting$STATISTICAL_MURDER_FLAG == FALSE) # Shootings that did not result in death
unsolved_shootings <- sum(Unsolved$STATISTICAL_MURDER_FLAG == FALSE)
unsolved_shootings_ratio <- unsolved_shootings / survivors
unsolved_shootings_ratio
```

```
## [1] 0.3533141
```

This is another interesting statistic, because both ratios are around 35%. It appears that the ratio of unsolved murder cases is almost identical to the ratio of unsolved shootings where the victim survived. This is also the same ratio of all unsolved cases. This surprised me, because I thought that the murder category would contain a much larger percentage of unsolved cases since it is more likely to identify the shooter if the victim survived.

Do shootings increase over time?

```
over_time <- shooting %>%
  mutate(val = 1) %>%
  mutate(date = mdy(OCCUR_DATE)) %>%
  mutate(month = month(date)) %>%
  mutate(year = year(date)) %>%
```

```
  group_by(month, year) %>%
  summarise(total = sum(val))

unique(over_time$month)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12
```

```
unique(over_time$year)
```

```
##  [1] 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
```

```
length(over_time$total)
```

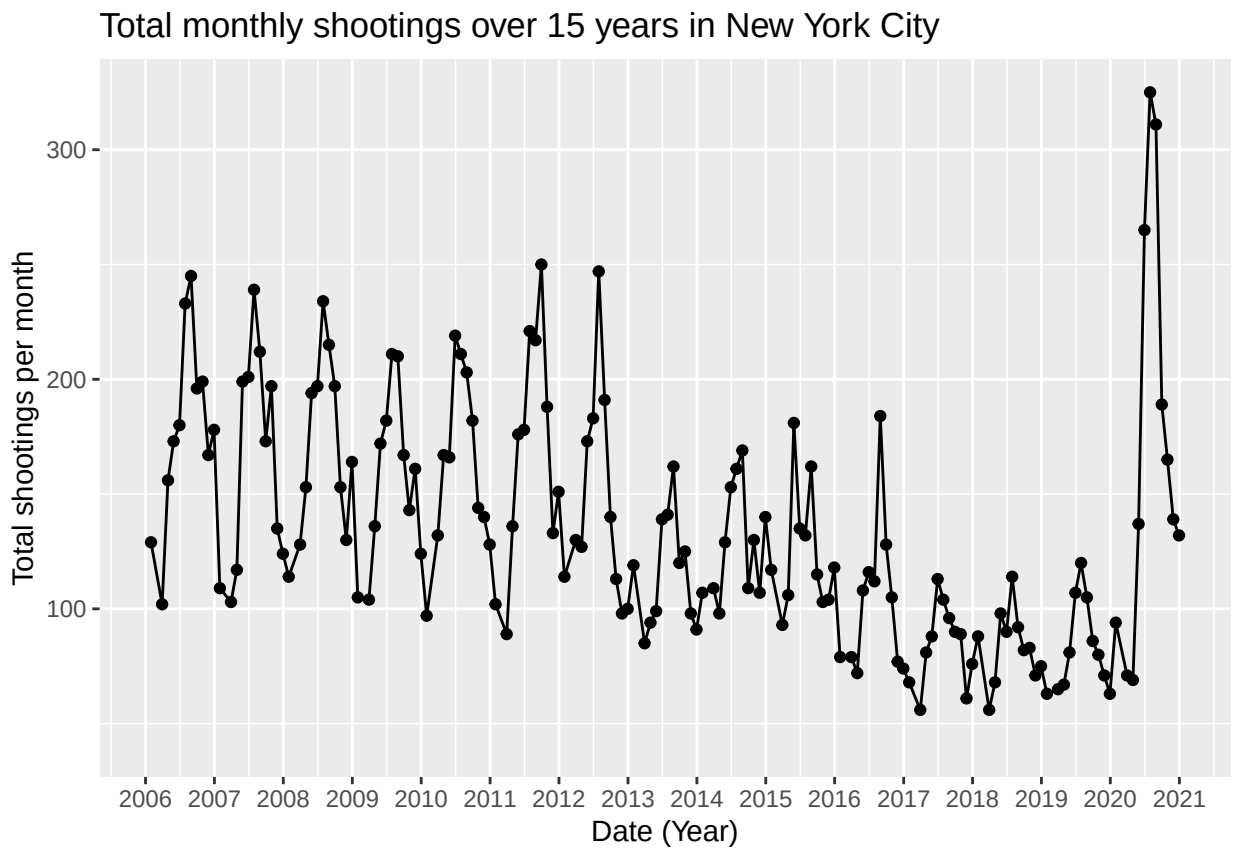
```
## [1] 180
```

```
over_time
```

```
## # A tibble: 180 x 3
## # Groups:   month [12]
##    month  year total
##    <dbl> <dbl> <dbl>
##  1     1  2006   129
##  2     1  2007   109
##  3     1  2008   114
##  4     1  2009   105
##  5     1  2010    97
##  6     1  2011   102
##  7     1  2012   114
##  8     1  2013   119
##  9     1  2014   107
## 10     1  2015   117
## # ... with 170 more rows
```

The new table consists of 12 unique months, and 15 unique years. I am expecting to have 12*15=180 total data points, which is the same size as the total column, which stores the cumulative shootings for the respective month. This leads me to believe that the data has been successfully grouped by month and year, so lets take a look at the plot of this table:

```
over_time %>%
  mutate(YearMonth = as.Date(paste0("30-", month,"-", year),"%d-%m-%Y")) %>%
  ggplot(aes(x=YearMonth, y=total, group = 1)) +
  geom_line() +
  geom_point() +
  scale_x_date(date_labels = "%Y", date_breaks = '1 year') +
  labs(x = "Date (Year)", y = "Total shootings per month", title = "Total monthly shootings over 15 yea
```



The graph shows that there is a cyclic sinusoidal pattern of total monthly shootings. It starts off low at the beginning of the year, but then rises sharply over the next few months and peaks during mid-year, around the July/August period. The total shootings then starts to drop over the next six months just as quickly as it rose, until it is back to its lowest rate in January. This result is interesting to me because and I did not expect that there would be any consistent patterns regarding monthly shooting data.

The large cyclic pattern was very consistent for 7 years from 2006 to 2013. From 2013 onwards, the number of shootings started to decrease compared to previous years. When the Covid19 pandemic started in early 2020, the total number of shooting cases rose exponentially, potentially because of increased tensions due to the virus and lock down, especially from those who opposed the mask mandate and vaccines.

Due to this insight regarding a consistent annual cyclic pattern, my hypothesis is that the time of year has a big influence on total shootings. Lets find out if a model tuned to the month can make an accurate prediction of total monthly shootings.

Model

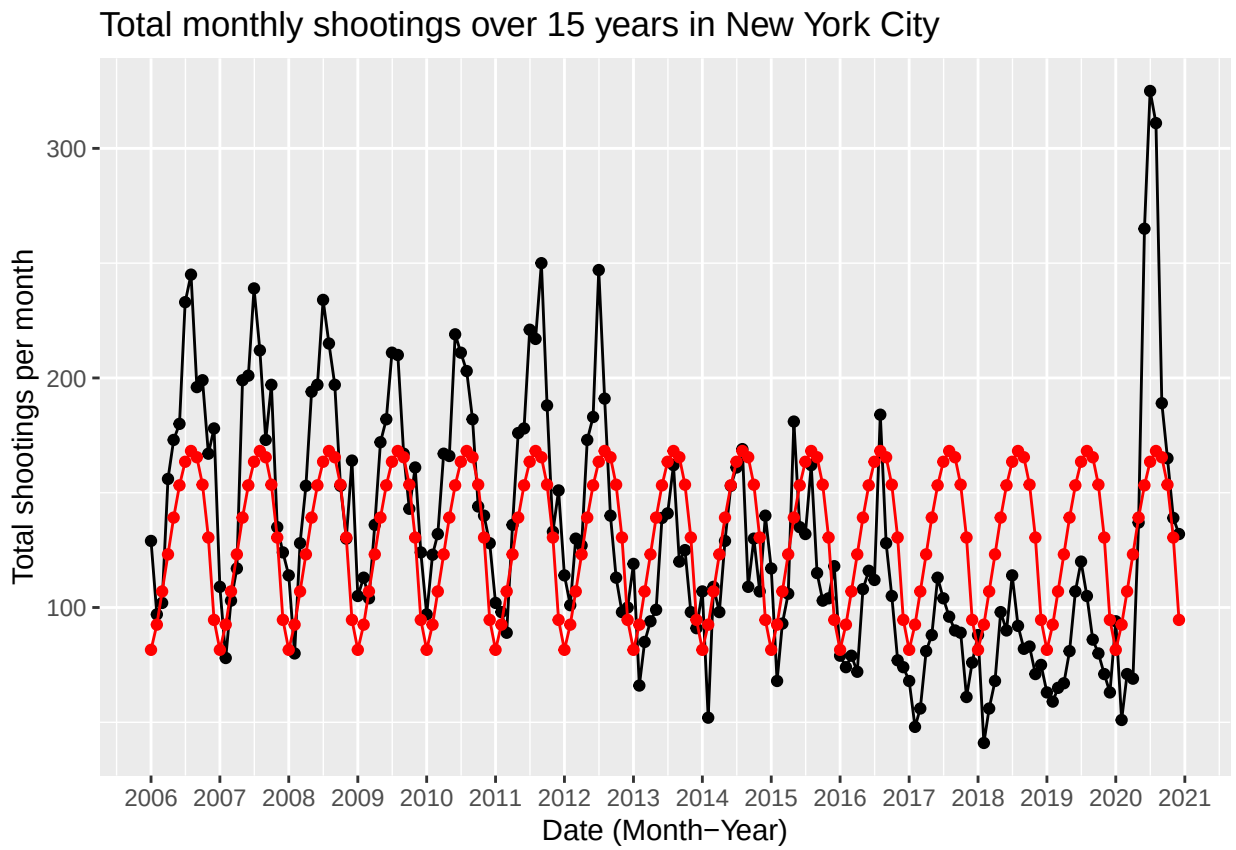
I don't have any experience modelling data, so I will only present a simple model for this dataset. I decided to use a cubic polynomial equation to relate the month to the total shooting cases.

```
mod <- lm(total ~ month + I(month^2) + I(month^3), data = over_time)
summary(mod)
```

```
##
## Call:
## lm(formula = total ~ month + I(month^2) + I(month^3), data = over_time)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -83.452 -34.151  -0.485  30.798 161.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  75.9091    18.6612   4.068 7.15e-05 ***
## month         2.3886    11.9330   0.200  0.84158
## I(month^2)    3.5680     2.0901   1.707  0.08956 .
## I(month^3)   -0.3031     0.1060  -2.860  0.00475 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.18 on 176 degrees of freedom
## Multiple R-squared:  0.316,  Adjusted R-squared:  0.3043
## F-statistic:  27.1 on 3 and 176 DF,  p-value: 1.848e-14
```

```
over_time$pred <- predict(mod)
over_time %>%
  mutate(YearMonth = as.Date(paste0("01-", month,"-", year),"%d-%m-%Y")) %>%
  ggplot() +
  geom_line(aes(x=YearMonth, y=total)) +
  geom_point(aes(x=YearMonth, y=total)) +
  geom_line(aes(x=YearMonth, y=pred), color = 'red') +
  geom_point(aes(x=YearMonth, y=pred), color = 'red') +
  scale_x_date(date_labels = "%Y", date_breaks = '1 year') +
  labs(x = "Date (Month-Year)", y = "Total shootings per month", title = "Total monthly shootings over
```



There is still room for improvement for this model. The sinusoidal pattern repeats every year until the beginning of 2020, which is when the Covid pandemic started. The data during this lock down period would not follow the same trend as previous years. Therefore, I chose to optimize the model a little bit by considering the data after January 2020 to be outliers, to see if excluding this portion of the dataset could lead to a better model.

```
exclude_outliers <- over_time %>%
  filter(year < 2020)

print('Double check that 1 year or 12 months has been taken out:')
```

```
## [1] "Double check that 1 year or 12 months has been taken out:"
```

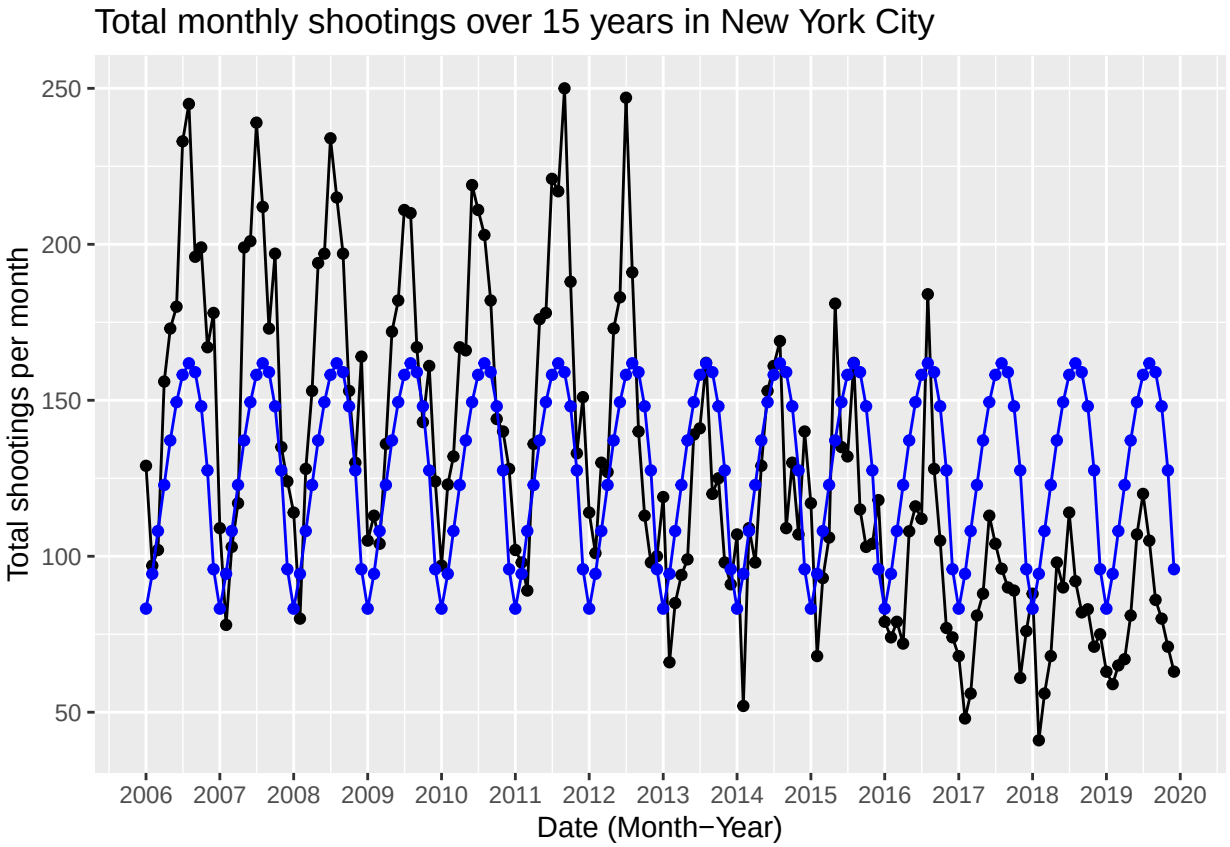
```
length(over_time$year) - length(exclude_outliers$year)
```

```
## [1] 12
```

```
mod2 <- lm(total ~ month + I(month^2) + I(month^3), data = exclude_outliers)
summary(mod2)
```

```
##
## Call:
## lm(formula = total ~ month + I(month^2) + I(month^3), data = exclude_outliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.012 -32.984   0.522  32.266  90.988
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  76.1061    17.8649   4.260 3.43e-05 ***
## month         4.5414    11.4237   0.398   0.6915
## I(month^2)    2.7999     2.0009   1.399   0.1636
## I(month^3)   -0.2534     0.1015  -2.498   0.0135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.86 on 164 degrees of freedom
## Multiple R-squared:  0.3043, Adjusted R-squared:  0.2916
## F-statistic: 23.91 on 3 and 164 DF,  p-value: 6.896e-13
```

```
exclude_outliers$pred2 <- predict(mod2)
exclude_outliers %>%
  mutate(YearMonth = as.Date(paste0("01-", month,"-", year),"%d-%m-%Y")) %>%
  ggplot() +
  geom_line(aes(x=YearMonth, y=total)) +
  geom_point(aes(x=YearMonth, y=total)) +
  geom_line(aes(x=YearMonth, y=pred2), color = 'blue') +
  geom_point(aes(x=YearMonth, y=pred2), color = 'blue') +
  scale_x_date(date_labels = "%Y", date_breaks = '1 year') +
  labs(x = "Date (Month-Year)", y = "Total shootings per month", title = "Total monthly shootings over
```



Conclusions

The improved model still looks the same as before and is unable to predict the peaks of the cycles, and also failed to predict the reduced total shootings that occured after 2017. The adjusted R-squared value of this model is only 0.29, so the model is really bad. A better model would be one that is more complex than a linear model, as well as incorporating more variables, such as victim race, age group and location data.

I'm not familiar with events that occur in the United States leading up to this period, so I don't have any personal insights into why shootings would increase until August, and then sharply drop. Looking at the data, I notice that the majority of the victims are in the 18-24 and 25-44 age groups. I consider people in the 18 to 30 year age group as young adults with most of them either attending college or just started at their first jobs. I don't have data to support this statement, so it's just a personal assumption. I also know that summer in the US starts in July, so students have their summer breaks in July and August. My thought is that the period leading up to summer break has many more young adults roaming the streets and causing trouble, which might have the effect of increasing monthly shootings. From a financial point of view, white Americans tend to have more wealth than black and Hispanic Americans, so my guess is that young white students tend to vacation in safer neighborhoods, or even overseas, while black and Hispanic students stay where they are. This is perhaps why when a shooting does occur, these racial groups are more likely to get shot.

Updated edits (after first submission):

Do shootings increase over time?

The graph of shootings over time when shown in months appear to be sinusoidal in nature, and I don't yet know how to model something that complicated. Therefore, to make the modelling section a little better, I will condense it to annual total shootings:

```r
annual_shootings <- shooting %>%
  mutate(val = 1) %>%  ## Set val=1 so that there is something numeric to sum when grouping by year
  mutate(date = mdy(OCCUR_DATE)) %>%
  mutate(year = year(date)) %>%
  group_by(year) %>%
  summarise(total = sum(val))

unique(annual_shootings$year)
```

```
##  [1] 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
```

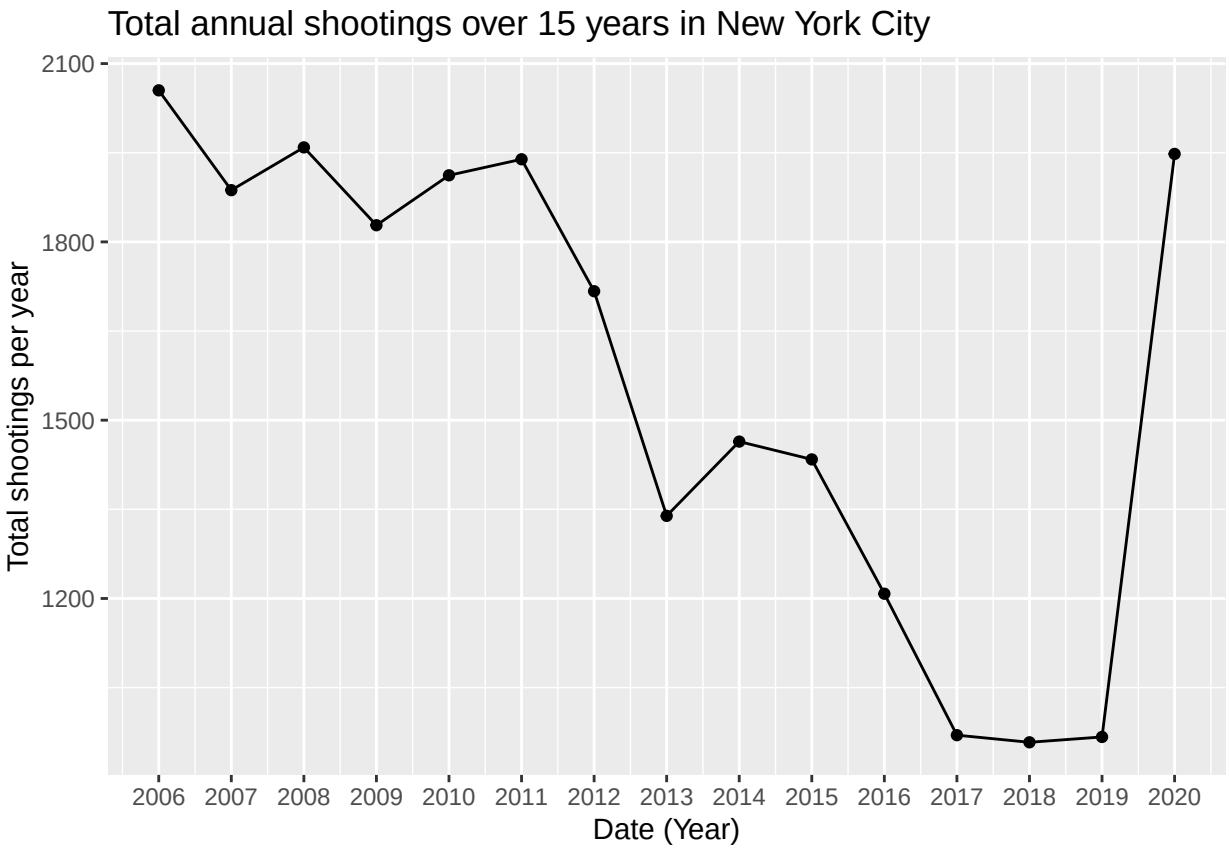
```r
length(annual_shootings$total)
```

```
## [1] 15
```

```r
annual_shootings
```

```
## # A tibble: 15 x 2
##     year total
##    <dbl> <dbl>
##  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
```

```r
annual_shootings %>%
  #mutate(YearMonth = as.Date(paste0("01-", month,"-", year),"%d-%m-%Y")) %>%
  ggplot(aes(x=year, y=total, group = 1)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(n.breaks = 15) +
  labs(x = "Date (Year)", y = "Total shootings per year", title = "Total annual shootings over 15 years
```



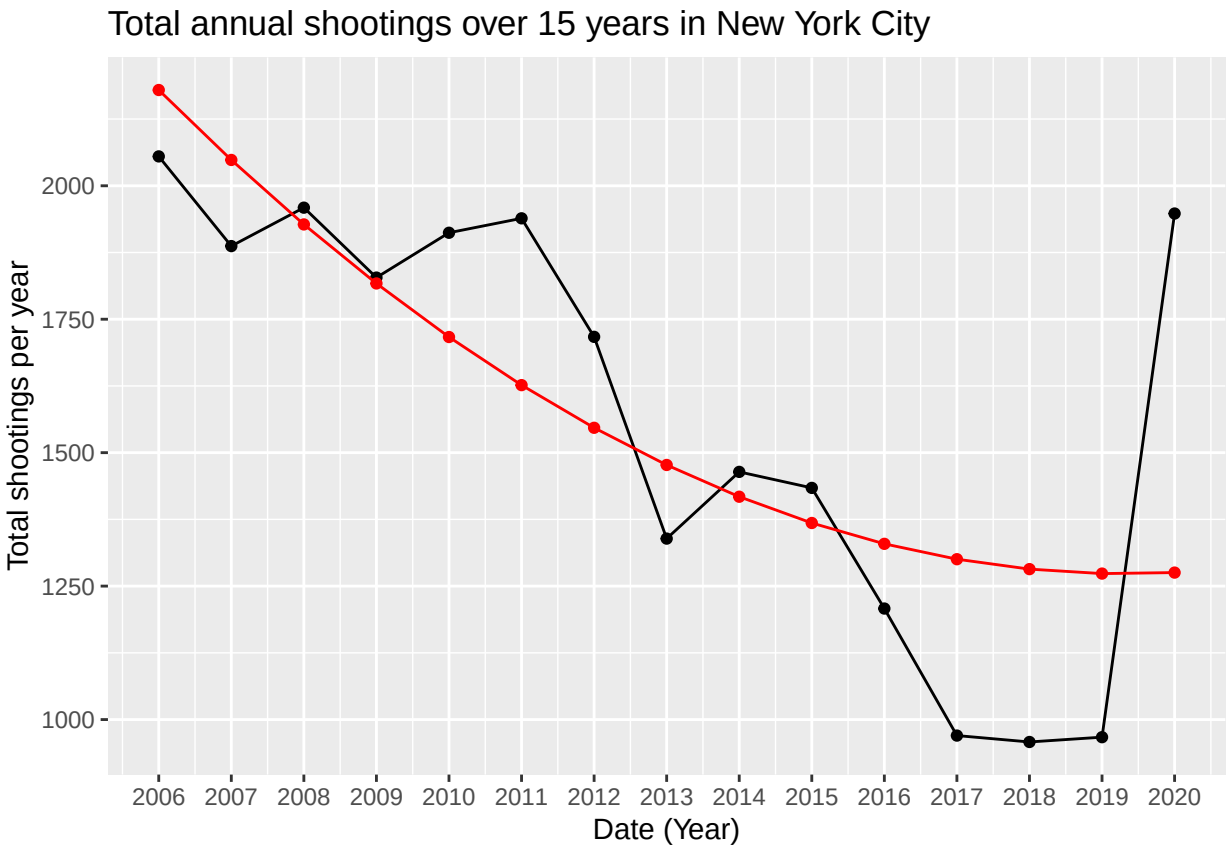
Model

```
mod <- lm(total ~ year + I(year^2) + I(year^3), data = annual_shootings)
summary(mod)
```

```
##
## Call:
## lm(formula = total ~ year + I(year^2) + I(year^3), data = annual_shootings)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -330.4 -149.7   11.0  118.1  672.7
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.084e+07  1.829e+07   1.140    0.277
## year        -2.064e+04  1.817e+04  -1.136    0.278
## I(year^2)    5.111e+00  4.513e+00   1.133    0.280
## I(year^3)           NA         NA      NA       NA
##
## Residual standard error: 289.8 on 12 degrees of freedom
## Multiple R-squared:  0.5585, Adjusted R-squared:  0.4849
## F-statistic: 7.591 on 2 and 12 DF,  p-value: 0.007404
```

```
annual_shootings$pred <- predict(mod)  ## Add predictions to a new column
annual_shootings %>%
```

```
  #mutate(YearMonth = as.Date(paste0("01-", month,"-", year),"%d-%m-%Y")) %>%
  ggplot() +
  geom_line(aes(x=year, y=total)) +
  geom_point(aes(x=year, y=total)) +
  geom_line(aes(x=year, y=pred), color = 'red') +
  geom_point(aes(x=year, y=pred), color = 'red') +
  scale_x_continuous(n.breaks = 15) +
  labs(x = "Date (Year)", y = "Total shootings per year", title = "Total annual shootings over 15 years
```



This is the model that I made to predict annual shootings. I chose to use a cubic polynomial equation to relate the year to total annual shootings, because I wanted my model to be able to curve. The recent spike in shootings in 2020 is a large outlier that skews my prediction, so I chose to improve my model by excluding the 2020 data.

```
exclude_outliers <- annual_shootings %>%
  filter(year < 2020)

print('Double check that 1 year or 12 months has been taken out:')
```

```
## [1] "Double check that 1 year or 12 months has been taken out:"
```

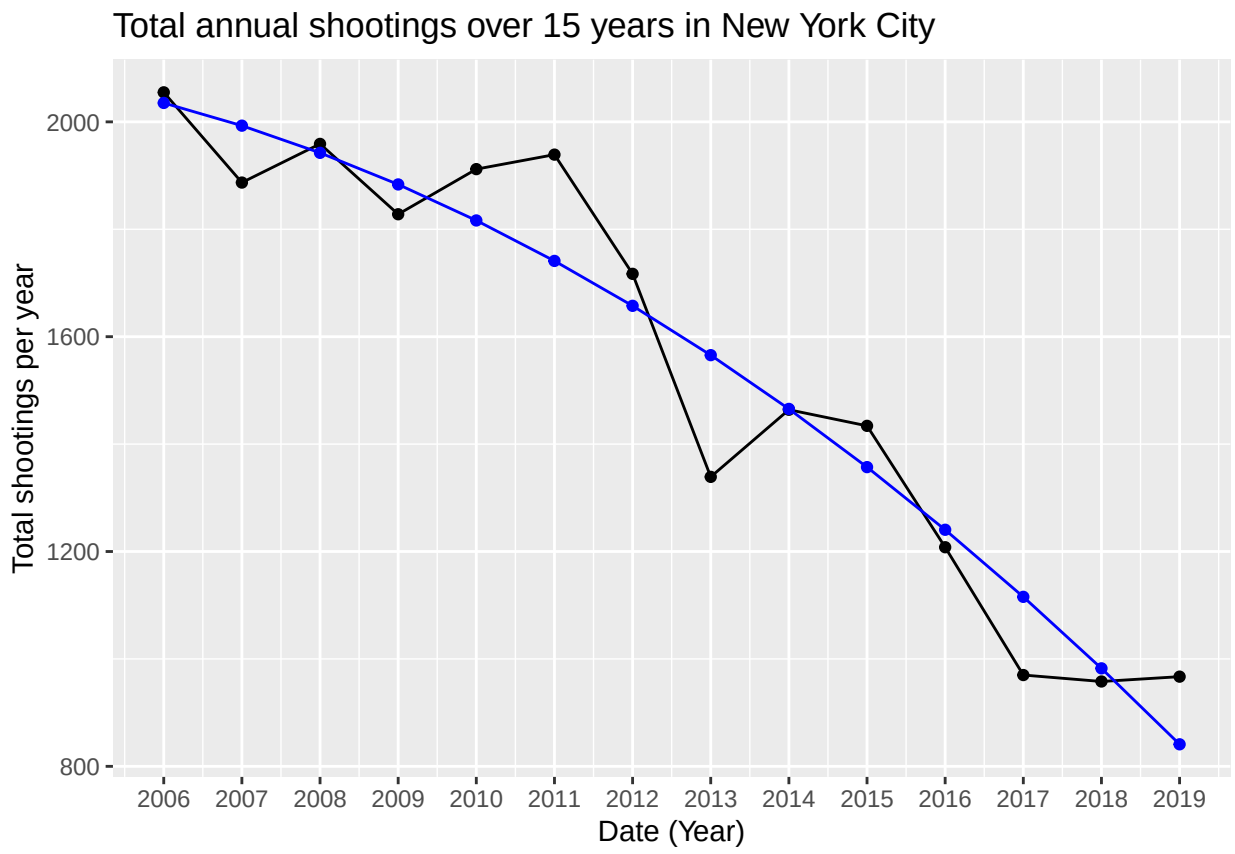
```
length(annual_shootings$year) - length(exclude_outliers$year)
```

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

```
mod2 <- lm(total ~ year + I(year^2) + I(year^3), data = exclude_outliers)
summary(mod2)
```

```
##
## Call:
## lm(formula = total ~ year + I(year^2) + I(year^3), data = exclude_outliers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -226.635  -49.774    7.562   72.523  197.900
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.654e+07  9.131e+06  -1.811   0.0975 .
## year         1.653e+04  9.074e+03   1.821   0.0958 .
## I(year^2)   -4.129e+00  2.254e+00  -1.832   0.0942 .
## I(year^3)           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121.7 on 11 degrees of freedom
## Multiple R-squared:  0.9236, Adjusted R-squared:  0.9098
## F-statistic: 66.53 on 2 and 11 DF,  p-value: 7.171e-07
```

```r
exclude_outliers$pred2 <- predict(mod2)
exclude_outliers %>%
  ggplot() +
  geom_line(aes(x=year, y=total)) +
  geom_point(aes(x=year, y=total)) +
  geom_line(aes(x=year, y=pred2), color = 'blue') +
  geom_point(aes(x=year, y=pred2), color = 'blue') +
  scale_x_continuous(n.breaks = 15) +
  labs(x = "Date (Year)", y = "Total shootings per year", title = "Total annual shootings over 15 years
```




```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_South Africa.1252  LC_CTYPE=English_South Africa.1252
## [3] LC_MONETARY=English_South Africa.1252 LC_NUMERIC=C
## [5] LC_TIME=English_South Africa.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.7.10 forcats_0.5.1    stringr_1.4.0    dplyr_1.0.7
##  [5] purrr_0.3.4      readr_2.0.2      tidyr_1.1.4      tibble_3.1.4
##  [9] ggplot2_3.3.5    tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.7       assertthat_0.2.1 digest_0.6.27    utf8_1.2.2
##  [5] R6_2.5.1         cellranger_1.1.0 backports_1.2.1  reprex_2.0.1
##  [9] evaluate_0.14    highr_0.9        httr_1.4.2       pillar_1.6.3
## [13] rlang_0.4.11     curl_4.3.2       readxl_1.3.1     rstudioapi_0.13
```

```
## [17] rmarkdown_2.11   labeling_0.4.2   bit_4.0.4        munsell_0.5.0
## [21] broom_0.7.9      compiler_4.1.1   modelr_0.1.8     xfun_0.26
## [25] pkgconfig_2.0.3  htmltools_0.5.2  tidyselect_1.1.1 fansi_0.5.0
## [29] crayon_1.4.1     tzdb_0.1.2       dbplyr_2.1.1     withr_2.4.2
## [33] grid_4.1.1       jsonlite_1.7.2   gtable_0.3.0     lifecycle_1.0.1
## [37] DBI_1.1.1        magrittr_2.0.1   scales_1.1.1     cli_3.0.1
## [41] stringi_1.7.4    vroom_1.5.5      farver_2.1.0     fs_1.5.0
## [45] xml2_1.3.2       ellipsis_0.3.2   generics_0.1.0   vctrs_0.3.8
## [49] tools_4.1.1      bit64_4.0.5      glue_1.4.2       hms_1.1.1
## [53] parallel_4.1.1   fastmap_1.1.0    yaml_2.2.1       colorspace_2.0-2
## [57] rvest_1.0.1      knitr_1.34       haven_2.4.3
```