2019 NYPD Shooting Incident Data

Downloading packages

We have downloaded **tidyverse** and **lubridate**

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Importing Data

The data contains a breakdown every shooting in NYC between 2006 and 2020. I want learn: What factors affect the number of shootings and is there away to minimize the number of shootings in the future? Here is a link of the CSV for the data: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic/resource/c564b578-fd8a-4005-8365-34150d306cc4

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

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
```

```
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_double(),
##   Y_COORD_CD = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

Cleaning Data

We first want to remove any columns that will not be used in the analysis.

```
NYPD_shooting <- NYPD_shooting %>% select(-c(X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lon_Lat,
INCIDENT_KEY,VIC_RACE,PERP_RACE,OCCUR_TIME,STATISTICAL_MURDER_FLAG,JURISDICTION_CODE,PRECINCT,
PERP_AGE_GROUP,VIC_AGE_GROUP))
```

We changed the name of **OCCUR_DATE** to **Date** because it is a simpler title and changed the data type of date

```
NYPD_shooting <- NYPD_shooting %>% select(-c(LOCATION_DESC)) %>% rename(Date = 'OCCUR_DATE',
Borough = "BORO", Perpetrator_Sex = 'PERP_SEX', Victim_Sex = 'VIC_SEX') %>%
mutate(Date= mdy(Date))
```

Coverting to a dataframe

```
NYPD <- as.data.frame(NYPD_shooting)
```

Convert columns that are characters to factors with levels

```
NYPD$Perpetrator_Sex <-factor(NYPD$Perpetrator_Sex, levels = c("M", "F", "U" ), labels =
c("Male", "Female", "Unidentified"))
NYPD$Victim_Sex <-factor(NYPD$Victim_Sex, levels = c("M", "F", "U" ), labels =
c("Male", "Female", "Unidentified"))
NYPD$Borough = factor(NYPD$Borough)
```

I am adding a row called shootings so I can use the sum function to find number of shootings based off different dates, Boroughs, Perpetrator_Sex, and Victim_Sex.

```
NYPD$"Shootings" <- 1
```

Analysis of data

Creating data table for shootings in each Borough

```
shootings_by_Borough <- NYPD %>% group_by(Date,Borough) %>% summarize(Shootings = sum(Shootings))
```

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

Finding summary data

```
summary(shootings_by_Borough)
```

```
##       Date                     Borough       Shootings
##  Min.   :2006-01-01   BRONX        :3084   Min.   : 1.000
##  1st Qu.:2009-05-06   BROOKLYN     :3845   1st Qu.: 1.000
##  Median :2012-08-29   MANHATTAN    :1785   Median : 1.000
##  Mean   :2013-01-23   QUEENS       :2075   Mean   : 2.085
```

```
##  3rd Qu.:2016-08-11   STATEN ISLAND: 521   3rd Qu.: 2.000
##  Max.   :2020-12-31                        Max.   :19.000
```

Create data table for shootings for every Date

```
shootings_by_day <- NYPD%>% group_by(Date) %>% summarize(Shootings= sum(Shootings))
```

Create data table for shootings for every Month

```
shootings_by_month <- NYPD %>% group_by(date_month=floor_date(Date,"month"))%>%
summarize(Shootings = sum(Shootings))
shootings_by_month<-as.data.frame(shootings_by_month)

head(shootings_by_month)
```

```
##   date_month Shootings
## 1 2006-01-01       129
## 2 2006-02-01        97
## 3 2006-03-01       102
## 4 2006-04-01       156
## 5 2006-05-01       173
## 6 2006-06-01       180
```

Create data table for shootings by the Perpetrator's Sex

```
shootings_by_perpetrator_sex<- NYPD %>% group_by(Date,Perpetrator_Sex) %>%
summarize(Shootings= sum(Shootings))
```

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

Create table for shootings by the Perpetrator's Sex

```
table(shootings_by_perpetrator_sex$Perpetrator_Sex)
```

```
##
##         Male       Female Unidentified
##         4274          274          845
```

Create data table for shootings by the Victim's Sex

```
shootings_by_victim_sex<- NYPD%>% group_by(Date,Victim_Sex) %>%
  summarize(Shootings= sum(Shootings))
```

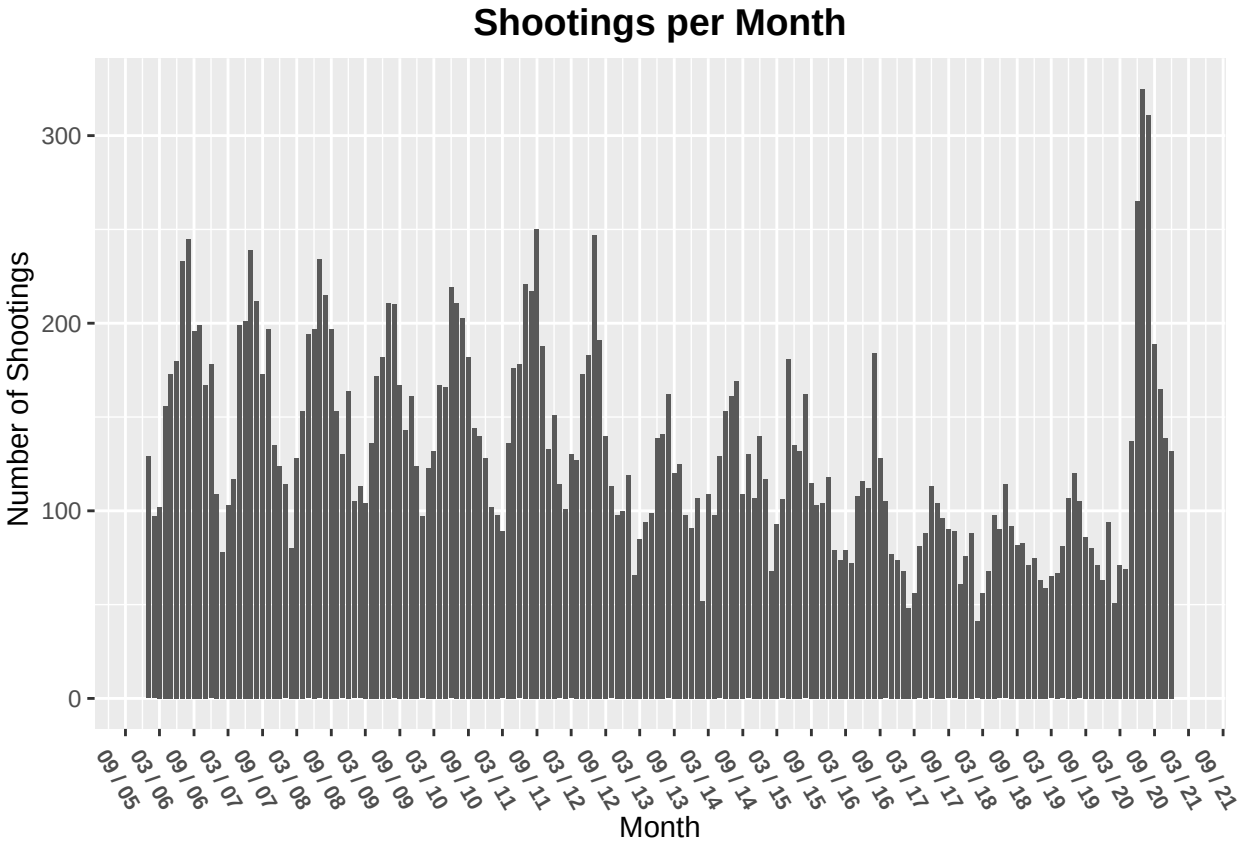
```
## `summarise()` has grouped output by 'Date'. You can override using the `.groups` argument.
```

Create table for shootings by the Victim's Sex

```
table(shootings_by_victim_sex$Victim_Sex)
```

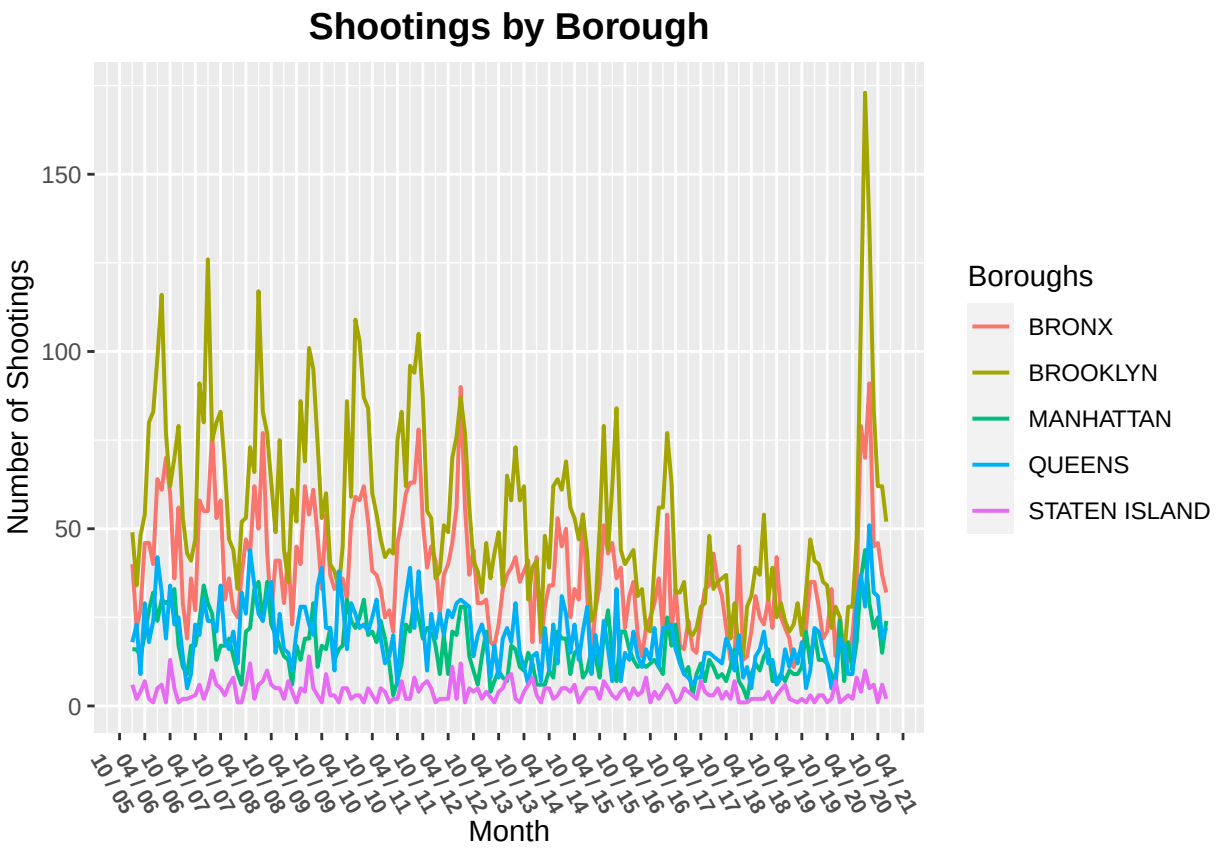
```
##
##         Male       Female Unidentified
##         4995         1429            9
```

```
shootings_by_day$Month <- as.Date(cut(shootings_by_day$Date, breaks = "1 month"))

ggplot(shootings_by_day,aes(Month,Shootings))+ stat_summary(fun = sum, geom = "bar")+
  ggtitle("Shootings per Month") + scale_x_date(date_labels = "%m / %y",date_breaks = "6 month") +
  theme(axis.text.x = element_text(face = "bold", angle = 300, size = 8),
  plot.title = element_text(face = "bold",size = 14, hjust = 0.5))+
  labs( x= "Month", y = " Number of Shootings")
```



There are a significant drop off in the number of shootings in the winter months and a significant increase during the summer. Shootings spiked in the summer of 2020 and appeared to be above monthly averages in the months following.

```
shootings_by_Borough$Month <- as.Date(cut(shootings_by_Borough$Date, breaks = "1 month"))

ggplot(shootings_by_Borough,aes(Month,Shootings, group = Borough))+
stat_summary(fun = sum, geom = "line", size = 0.7, mapping = aes(color = factor(Borough)))+
ggtitle("Shootings by Borough") + scale_x_date(date_labels = "%m / %y",date_breaks = "6 month")+
theme(axis.text.x = element_text(face = "bold", angle = 300, size = 8),
plot.title = element_text(face = "bold",size = 14, hjust = 0.5)) +
guides(color = guide_legend(title = "Boroughs"))+labs( x= "Month", y = "Number of Shootings")
```



In the line graph above, labeled "Shootings by Borough", there are the largest number of shootings in Brooklyn and the least number of shootings in Staten Island.

Modeling

I indexed every row so a linear model could be run.

```
shootings_by_month_num <- shootings_by_month %>% mutate(month_num = 1:nrow(shootings_by_month))
head(shootings_by_month_num)
```

```
##   date_month Shootings month_num
## 1 2006-01-01       129         1
## 2 2006-02-01        97         2
## 3 2006-03-01       102         3
## 4 2006-04-01       156         4
## 5 2006-05-01       173         5
## 6 2006-06-01       180         6
```

```
tail(shootings_by_month_num)
```

```
##     date_month Shootings month_num
## 175 2020-07-01       325       175
## 176 2020-08-01       311       176
## 177 2020-09-01       189       177
## 178 2020-10-01       165       178
## 179 2020-11-01       139       179
## 180 2020-12-01       132       180
```

Here I am creating the linear model and storing into *mod*

```
mod<- lm(Shootings ~ shootings_by_month_num$month_num , data = shootings_by_month_num)
```

Here I took a summary of the linear model. The summary shows that the model is a very poor fit for the data with an adjusted R-squared of .1737.

```
summary(mod)
```

```
##
## Call:
## lm(formula = Shootings ~ shootings_by_month_num$month_num, data = shootings_by_month_num)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.861 -32.958  -8.648  27.679 230.239
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      169.86921    7.20717  23.569  < 2e-16 ***
## shootings_by_month_num$month_num  -0.42919    0.06906  -6.214 3.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.15 on 178 degrees of freedom
## Multiple R-squared:  0.1783, Adjusted R-squared:  0.1737
## F-statistic: 38.62 on 1 and 178 DF,  p-value: 3.552e-09
```

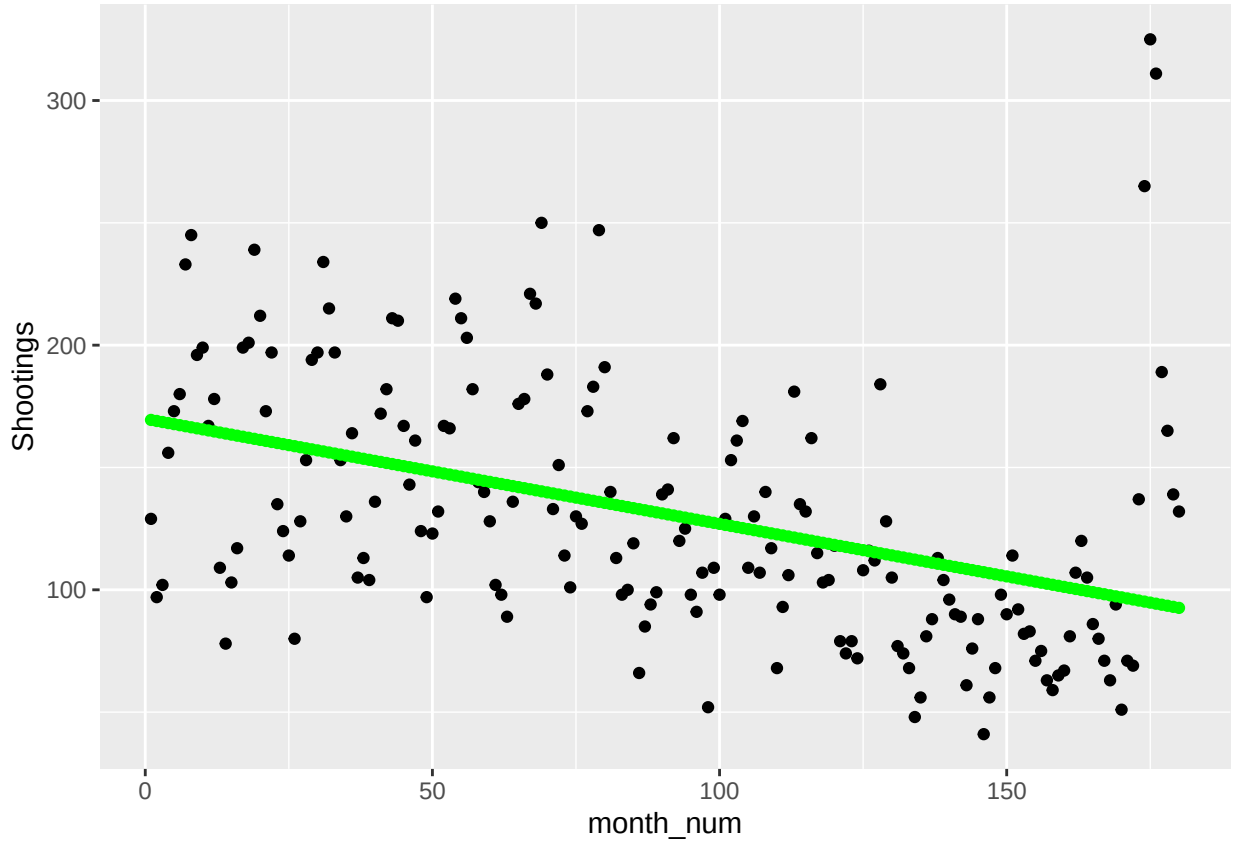
To check that the model is a poor fit, let's graph the model with the original values.

```
mod_pred <- shootings_by_month_num %>% mutate(pred = predict(mod))
head(mod_pred,10)
```

```
##    date_month Shootings month_num     pred
## 1  2006-01-01       129         1 169.4400
## 2  2006-02-01        97         2 169.0108
## 3  2006-03-01       102         3 168.5817
## 4  2006-04-01       156         4 168.1525
## 5  2006-05-01       173         5 167.7233
## 6  2006-06-01       180         6 167.2941
## 7  2006-07-01       233         7 166.8649
## 8  2006-08-01       245         8 166.4357
## 9  2006-09-01       196         9 166.0065
## 10 2006-10-01       199        10 165.5773
```

This graph confirms that our model is a very poor fit.

```
mod_pred %>% ggplot() + geom_point(aes(x = month_num, y = Shootings)) +
  geom_point(aes(x = month_num, y = pred), color = "green")
```



More Analysis

Based on the first line graph, labeled "Shootings Per Month", we can confirm whether or not there is a significant difference in shootings in warmer and colder months.

Here is a tibble of the total shooting by month.

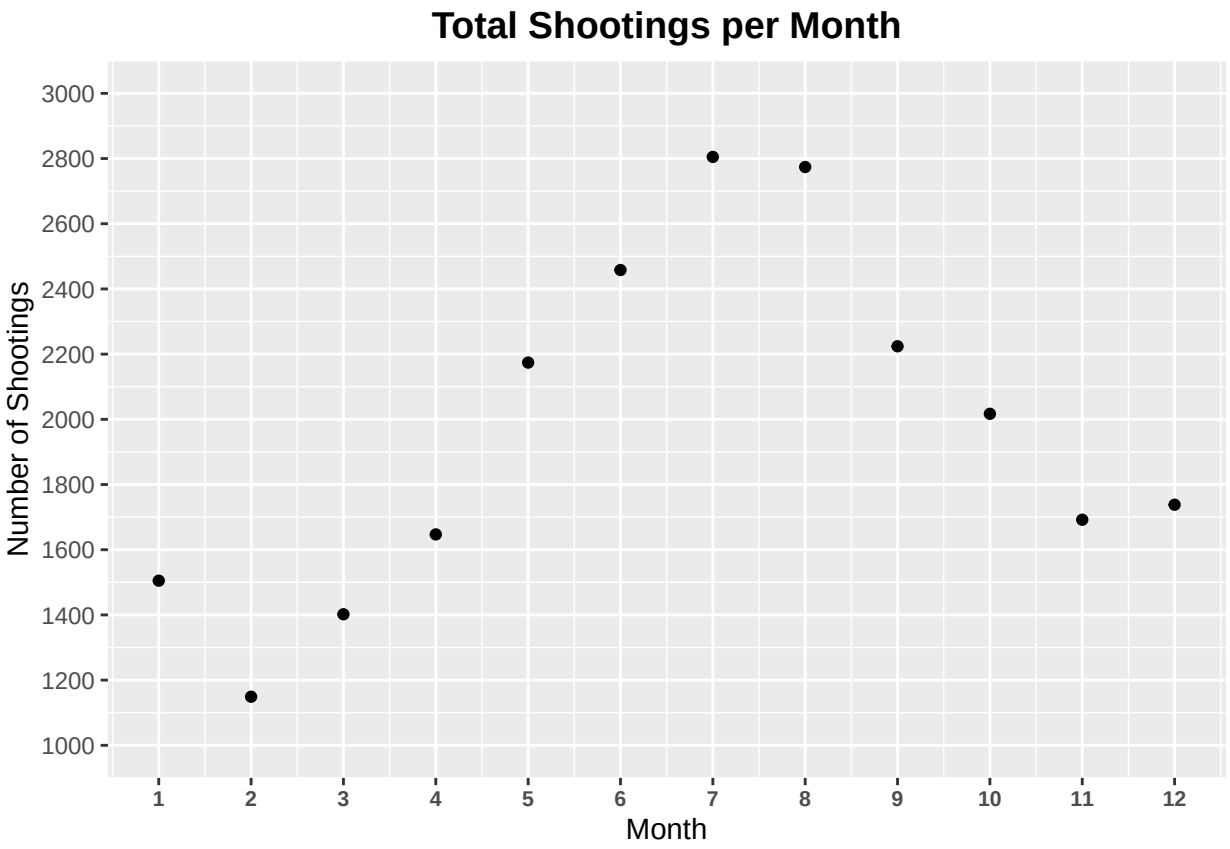
```
new_num =0
newmod <- shootings_by_month_num %>% mutate(date_month = month(date_month)) %>%
select(date_month, Shootings) %>% group_by(date_month) %>% summarize(Shootings= sum(Shootings))

head(newmod,12)
```

```
## # A tibble: 12 x 2
##    date_month Shootings
##         <dbl>     <dbl>
##  1          1      1505
##  2          2      1149
##  3          3      1402
##  4          4      1647
##  5          5      2174
##  6          6      2458
##  7          7      2805
##  8          8      2774
##  9          9      2224
## 10         10      2017
## 11         11      1692
```

```
## 12          12       1738
```

Here is a graph of the number of shootings per month. There is clearly a greater number of shootings in warmer months than colder months.

```
ggplot(newmod,aes(date_month,Shootings))+geom_point()+scale_x_continuous(n.breaks= 12)+
  scale_y_continuous(limits = c(1000,3000),n.breaks= 15)+
  ggtitle("Total Shootings per Month")+
  theme(axis.text.x = element_text(face = "bold", size = 8),
  plot.title = element_text(face = "bold",size = 14, hjust = 0.5)) +
  labs( x= "Month", y = "Number of Shootings")
```



Potential Bias

This is only the second Data Science project I have ever done so there may have been some bias when not accounting for NA values and choosing variables that may not be as effective. I have not finished the statistics pathway courses yet either so the model may not be nearly as good of a predictor as I would have hoped. I have limited experience with R but the way I chose to manipulate the dates could have affected the accuracy of the model. I was also learning different ways to cleanse data, apply analysis, and create visualizations as I was coding. As result, there may be some bias in the simplicity of the code that could skew my conclusion. I could have added a weather API and run more advanced statistical techniques to analyze the correlation between the shootings per month and the climate. Also my graphs could be biased from the standpoint that I chose certain bin widths and scales that could make the results appear different than they should seem due to my lack of experience with scaling graphs in R.

Conclusion

We can conclude that there many more male shooters and shooting victims than female shooters and shooting victims. There will need to be more research done to determine whether this gap is consistent around the U.S or whether the extreme sex gap disparity in shooting in cities is specific to New York City. We can also conclude that the Bronx has the highest number of shootings while Staten Island has the least. This is likely due to a variety of factors such as population and income. Months with warmer weather consistently had more shootings than months with cooler weather. A possible cause for this is people tend to be inside more when the temperature is cooler. There was a large spike in shootings in Summer 2020. A potential cause could be the decreased restrictions on activity from Covid-19 in Summer 2020. Addressing issues with shootings will require more thorough research of the variety of factors that cause the shootings as well as ways to possibly decrease the number of shootings in the future.