NYPD-Shooting-Incidents

Overview

This paper provides an analysis of publicly available data on shootings in New York City from 2006 to 2020 provided by the New York City Police Department. The analysis was developed in conjunction with the course work for the 3rd week of the Data Science as a Field course delivered online via the University of Colorado Boulder on Coursera.

Coursework and associated materials can be accessed via https://www.coursera.org/learn/data-science-as-a-field/lecture/gBSD6/intro-to-r-markdown.

This document aims to provide the steps neccesary to repeat the findings through your own work.

We will cover

- The packages leveraged
- An overview of the data we have gathered to conduct analysis
- The process for cleaning and preparing the data
- An analysis on if incident rates are improving or worsening over time
- Analysis of murder rates
- An analysis on demographic patterns in the data
- Predicability of the shooting rates as an indicator of the murder rate
- Key conclusions and

External Libraries Leveraged

- library(tidyverse)
- library(lubridate)
- library(magrittr)
- library(readxl)
- library(ggthemes)
- library(sf)
- library(tmap)
- library(tinytex)
- library(zoo)

```
library(tidyverse)
library(lubridate)
library(magrittr)
library(readxl)
library(ggthemes)
library(sf)
library(tmap)
library(tinytex)
library(zoo)
```

Data load

NYPD Shooting Incident data collated by the NYPD can be found via Data.gov

https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

```r
# Sets the NYC Shooting Data CSV
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

shooting_Data <- read_csv(url_in)
```

Initial analysis for data cleansing

```r
summary(shooting_Data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23568       Length:23568      Length:23568
##  1st Qu.: 55317014   Class :character   Class1:hms        Class :character
##  Median : 83365370   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :102218616                      Mode  :numeric
##  3rd Qu.:150772442
##  Max.   :222473262
##
##    PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000   Length:23568       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000   Class :character   FALSE:19080
##  Median : 69.00   Median :0.0000   Mode  :character   TRUE :4488
##  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       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999900   1st Qu.:182565
##  Mode  :character   Mode  :character   Median :1007645   Median :193482
##                                        Mean   :1009363   Mean   :207312
##                                        3rd Qu.:1016807   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##
##     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
##
```

Based on our planned analysis and summary of the loaded data we will

- Remove un-needed columns
  - Incident_Key - we do not need the unique ids for the individual incidents as we are analyzing aggregate data
  - Jurisdicition_Code - not needed for spatial, demographic, and murder rate analysis
  - Occur_Time - we are not conducting a time based analysis on incidents
  - Precinct - Data not leveraged
  - Location_Desc - data not leveraged
  - X_COORD_CD - data not leveraged
  - Y_COORD_CD - data not leveraged
  - Latitude - data not leveraged
  - Longitude - data not leveraged
  - Long_Lat - data not leveraged
- Fix column data types
  - OCCUR_DATE - from char to date
- Filter out NA data as we are interested in data with perps and victims
  - PERP_AGE_GROUP
  - PERP_SEX
  - PERP_RACE
- Rename multiple columns for easier coding

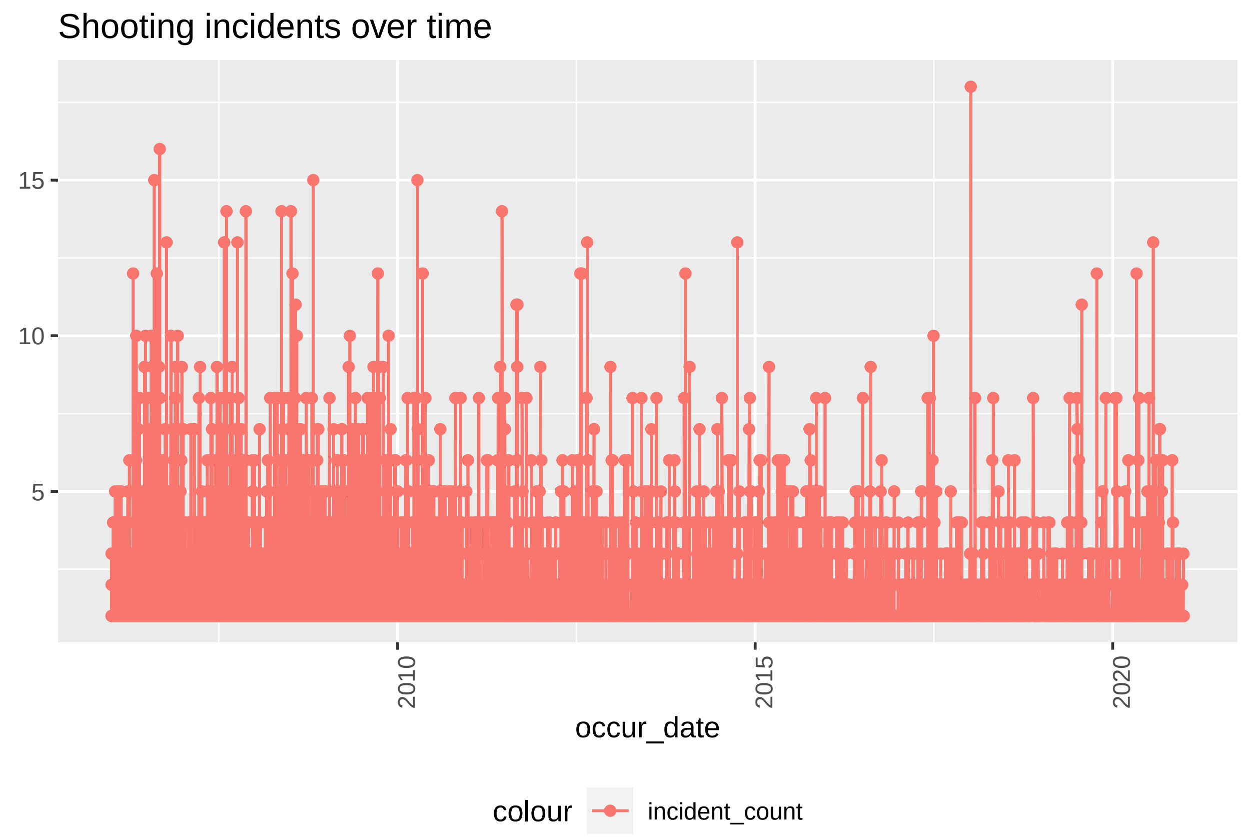
```
cleanWorkingData <- shooting_Data %>%
  select(-c(INCIDENT_KEY, OCCUR_TIME, PRECINCT, JURISDICTION_CODE, LOCATION_DESC, X_COORD_CD, Y_COORD_C
  mutate(occur_date = mdy(OCCUR_DATE)) %>%
  filter(PERP_AGE_GROUP!='NA') %>%
  filter(PERP_SEX!='NA') %>%
  filter(PERP_RACE!='NA') %>%
  mutate(perp_age_group = factor(PERP_AGE_GROUP)) %>%
  mutate(perp_sex = factor(PERP_SEX)) %>%
  mutate(perp_race = factor(PERP_RACE)) %>%
  mutate(vic_age_group = factor(VIC_AGE_GROUP)) %>%
  mutate(vic_sex = factor(VIC_SEX)) %>%
  mutate(vic_race = factor(VIC_RACE)) %>%
  mutate(vic_age_group = factor(VIC_AGE_GROUP)) %>%
  mutate(boro = factor(BORO))  %>%
  rename(murder_flag = 'STATISTICAL_MURDER_FLAG')

cleanWorkingData <- cleanWorkingData %>%
  select(-c(OCCUR_DATE, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE, BORO))

summary(cleanWorkingData)
```

```
##  murder_flag       occur_date          perp_age_group perp_sex
##  Mode :logical   Min.   :2006-01-01   18-24  :5448   F:  334
##  FALSE:12233     1st Qu.:2008-04-02   25-44  :4613   M:13305
##  TRUE :2876      Median :2010-07-10   UNKNOWN:3156   U: 1470
##                  Mean   :2011-09-26   <18    :1354
##                  3rd Qu.:2015-01-04   45-64  : 481
##                  Max.   :2020-12-29   65+    :  54
##                                       (Other):   3
##                            perp_race    vic_age_group  vic_sex
##  AMERICAN INDIAN/ALASKAN NATIVE:   2   <18    :1788   F: 1576
##  ASIAN / PACIFIC ISLANDER      : 120   18-24  :5714   M:13521
##  BLACK                         :9855   25-44  :6400   U:   12
##  BLACK HISPANIC                :1081   45-64  :1033
##  UNKNOWN                       :1835   65+    : 117
##  WHITE                         : 255   UNKNOWN:  57
##  WHITE HISPANIC                :1961
##                             vic_race                boro
##  AMERICAN INDIAN/ALASKAN NATIVE:    7   BRONX        :4497
##  ASIAN / PACIFIC ISLANDER      :  235   BROOKLYN     :5744
##  BLACK                         :10325   MANHATTAN    :1994
##  BLACK HISPANIC                : 1490   QUEENS       :2308
##  UNKNOWN                       :   68   STATEN ISLAND: 566
##  WHITE                         :  477
##  WHITE HISPANIC                : 2507
```

Incident Rate Analysis

For this section we want to look at the incident rate over time both in aggregate and by individual burb. To do this we will begin with adding incident counts and grouping them by date.

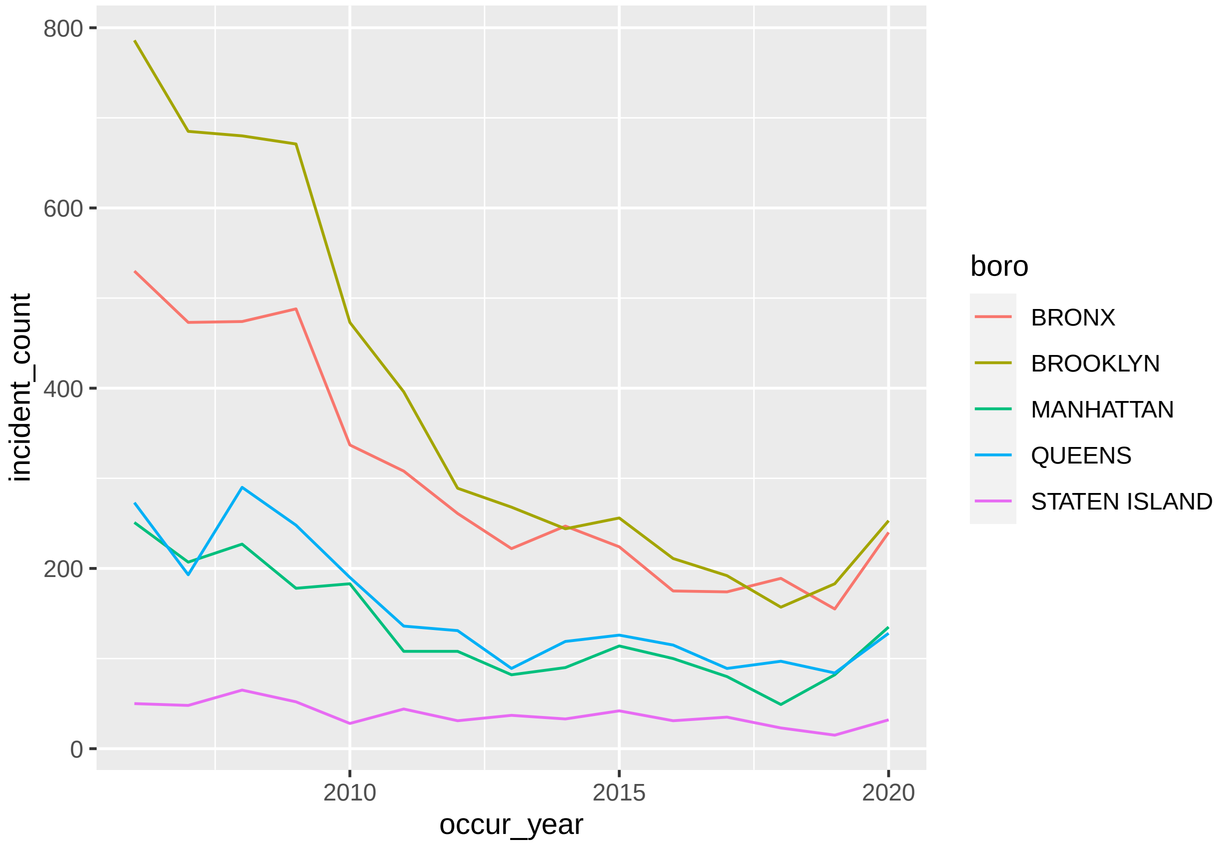
```
incident_rate_over_time <- cleanWorkingData

graphable_incident_data <- incident_rate_over_time %>%
  group_by(boro, occur_date) %>%
  tally(name="incident_count")

new_graph <- graphable_incident_data %>%
  filter(incident_count > 0) %>%
  ggplot(aes(x = occur_date, y = incident_count)) +
  geom_line(aes(color = "incident_count")) +
  geom_point(aes(color = "incident_count")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90))+
  labs(title = "Shooting incidents over time", y= NULL)

new_graph
```



```
incident_by_boro <- incident_rate_over_time  %>%
  mutate(occur_year = year(occur_date)) %>%
  group_by(occur_year, boro) %>%
  summarise(incident_count = n()) %>%
  select(boro, occur_year, incident_count)
  ungroup
```

```
## function (x, ...)
## {
##     UseMethod("ungroup")
## }
## <bytecode: 0x0000000012c7d040>
## <environment: namespace:dplyr>
```

```
ggplot(incident_by_boro, aes(occur_year, incident_count, colour = boro)) +
  geom_line()
```



Incident Rate Conclusion

Incident rates from 2006 to 2020 are declining both in aggregate as well as individually across boros. There is a strong correlation in terms of trend across boros over time. Each boro is showing a similar decline in violent crime with a perp and victim in general incident rates.

Brooklyn has seen the sharpest decline in incident rates moving from the highest rate nearing 800 per year down to approximately 250 in 2020 briefly coming in under the Bronx in 2016.

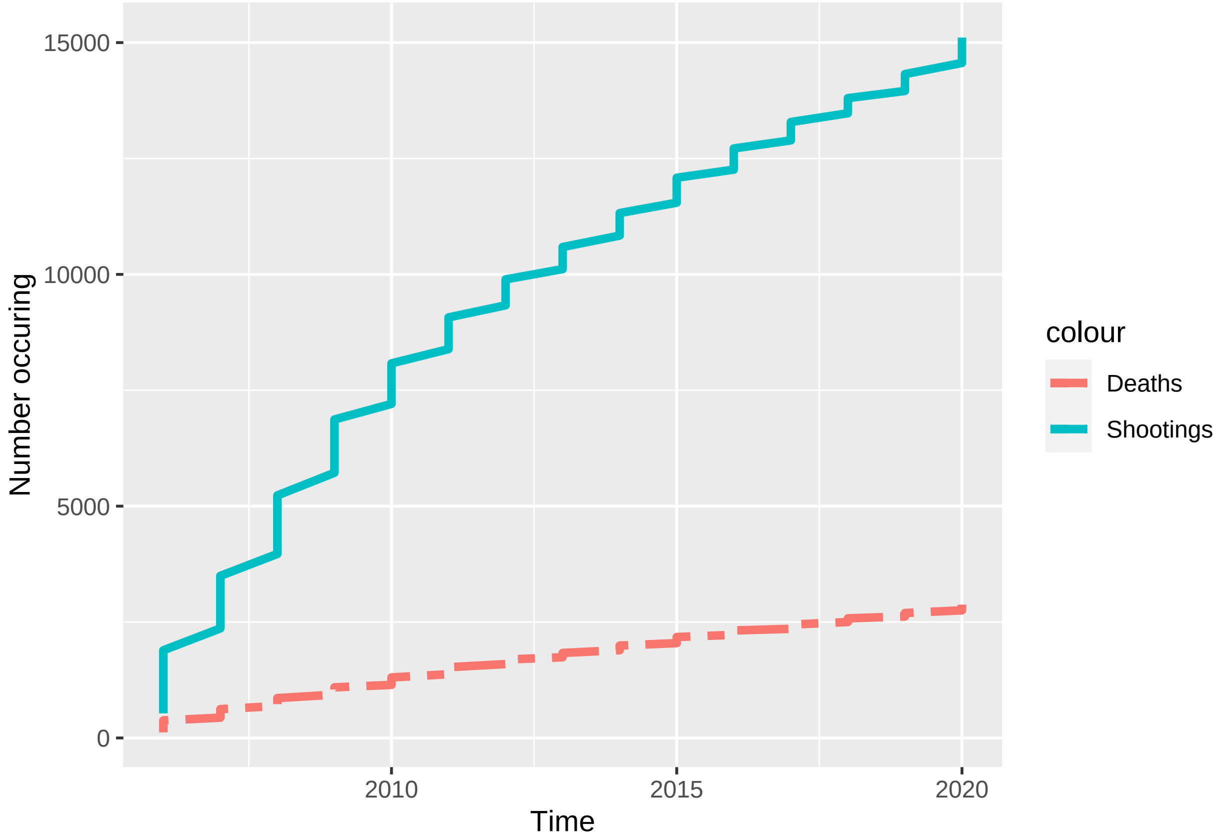
Each Boro shows a sharp increase in the number of incidents in 2020. Further analysis at this stage is not feasible.

Total analysis of murders as compared to incidents

```
total_incidents_rollup <- incident_rate_over_time %>%
  mutate(occur_year = year(occur_date)) %>%
  group_by(occur_year, boro) %>%
  summarise(incident_count = n(), murders = sum(murder_flag == TRUE)) %>%
  ungroup()
```

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

```
total_incidents_rollup$incident_to_date <-  cumsum(total_incidents_rollup$incident_count)
total_incidents_rollup$murders_to_date <-  cumsum(total_incidents_rollup$murders)

colors <-c("Deaths" = "red", "Shootings" = "orange")

total_incident_graph <- total_incidents_rollup %>%
  filter(incident_count > 0) %>%
  ggplot(aes(x = occur_year)) +
  geom_line(aes(y = incident_to_date, color="Shootings"), size = 1.5) +
  geom_line(aes(y = murders_to_date, color="Deaths"), linetype="twodash",size = 1.5) +
  xlab("Time") +
  ylab("Number occuring")

total_incident_graph
```



total_incident_graph

Incidents as a predicator of murder rates

```
mod <- lm(murders_to_date ~ incident_to_date, data = total_incidents_rollup)
summary(mod)
```

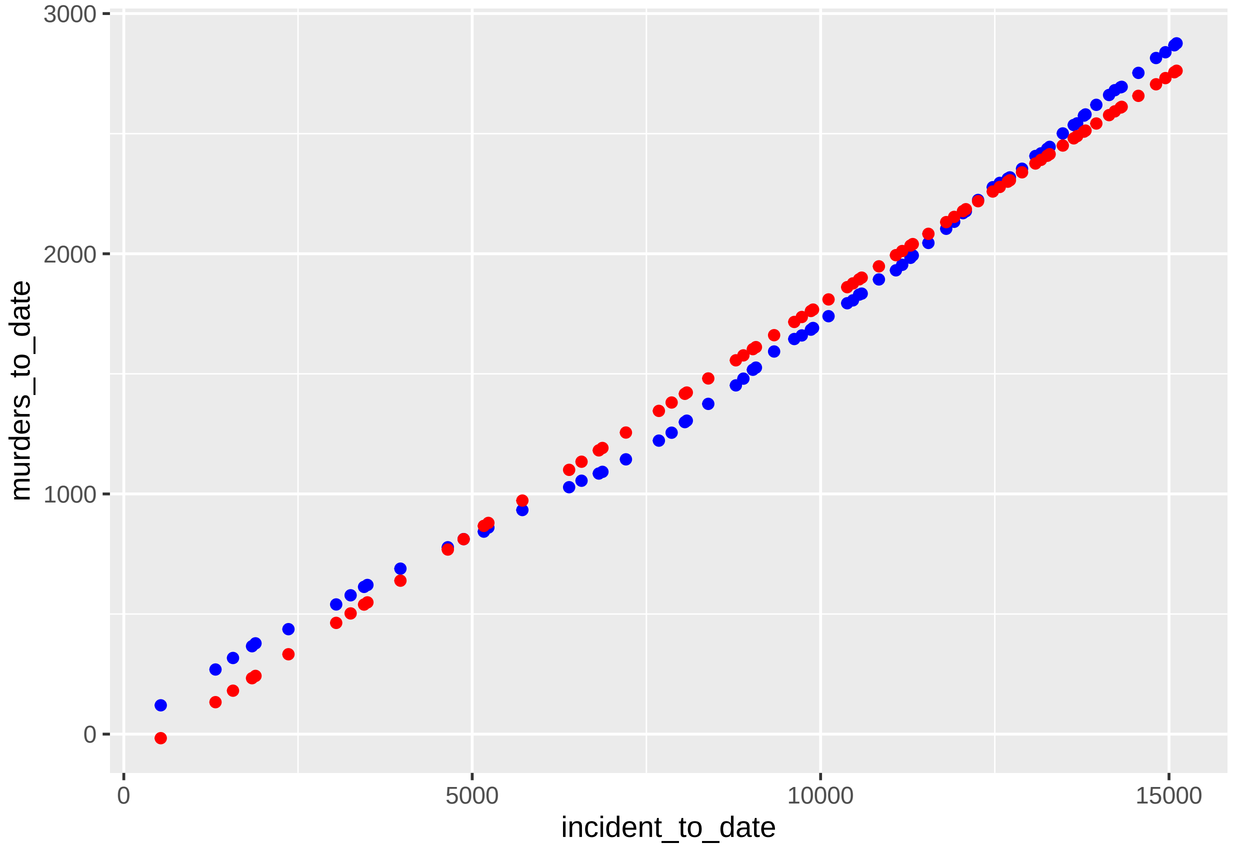
```
##
```

```
## Call:
## lm(formula = murders_to_date ~ incident_to_date, data = total_incidents_rollup)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -125.619  -70.186    0.297   73.079  136.800
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.178e+02  2.378e+01  -4.954 4.55e-06 ***
## incident_to_date  1.906e-01  2.285e-03  83.394  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.51 on 73 degrees of freedom
## Multiple R-squared:  0.9896, Adjusted R-squared:  0.9895
## F-statistic:  6955 on 1 and 73 DF,  p-value: < 2.2e-16
```

```
predictions <- total_incidents_rollup %>%
  mutate(prediction = predict(mod))

predictions
```

```
## # A tibble: 75 x 7
##    occur_year boro       incident_count murders incident_to_date murders_to_date
##         <dbl> <fct>               <int>   <int>            <int>           <int>
##  1       2006 BRONX                 530     120              530             120
##  2       2006 BROOKLYN              786     149             1316             269
##  3       2006 MANHATTAN             251      48             1567             317
##  4       2006 QUEENS                273      49             1840             366
##  5       2006 STATEN IS~             50      12             1890             378
##  6       2007 BRONX                 473      59             2363             437
##  7       2007 BROOKLYN              685     103             3048             540
##  8       2007 MANHATTAN             207      38             3255             578
##  9       2007 QUEENS                193      35             3448             613
## 10       2007 STATEN IS~             48       8             3496             621
## # ... with 65 more rows, and 1 more variable: prediction <dbl>
```

```
predictions %>% ggplot() +
  geom_point(aes(x = incident_to_date, y = murders_to_date ), color = "blue") +
  geom_point(aes(x = incident_to_date, y = prediction ), color = "red")
```



Conclusion

While not a sizable cognative leap…shootings are indeed strong predictor of murders.

Demographic analysis

```
summary(cleanWorkingData)
```

```
##  murder_flag       occur_date          perp_age_group perp_sex
##  Mode :logical   Min.   :2006-01-01   18-24  :5448   F:  334
##  FALSE:12233     1st Qu.:2008-04-02   25-44  :4613   M:13305
##  TRUE :2876      Median :2010-07-10   UNKNOWN:3156   U: 1470
##                  Mean   :2011-09-26   <18    :1354
##                  3rd Qu.:2015-01-04   45-64  : 481
##                  Max.   :2020-12-29   65+    :  54
##                                       (Other):   3
##                           perp_race     vic_age_group  vic_sex
##  AMERICAN INDIAN/ALASKAN NATIVE:   2   <18    :1788   F: 1576
##  ASIAN / PACIFIC ISLANDER      : 120   18-24  :5714   M:13521
##  BLACK                         :9855   25-44  :6400   U:   12
##  BLACK HISPANIC                :1081   45-64  :1033
##  UNKNOWN                       :1835   65+    : 117
##  WHITE                         : 255   UNKNOWN:  57
##  WHITE HISPANIC                :1961
##                            vic_race                  boro
##  AMERICAN INDIAN/ALASKAN NATIVE:    7   BRONX        :4497
##  ASIAN / PACIFIC ISLANDER      :  235   BROOKLYN     :5744
##  BLACK                         :10325   MANHATTAN    :1994
##  BLACK HISPANIC                : 1490   QUEENS       :2308
##  UNKNOWN                       :   68   STATEN ISLAND: 566
##  WHITE                         :  477
##  WHITE HISPANIC                : 2507
```

Demographic conclusion

The perps are dramatically african american males over time with the majority of the offender ages being 18-24. Incidents occuring are largelyt African American male ages 18-24 against african american males ages 18-24.

Males represent the materially significant count in terms of both perps and victims.

Shooting rates continue to decline slowly in aggregate over time.

Number of shootings is a very positive and intuitive correlary to the number of deaths that will occur.

Bias analysis

There might be bias in the data in terms of the reporting of certain crimes based on demographics. Additionally it is important to consider the predominance of a particular ethnic or age group based on total population within a given boro. As an example if the general population of NYC is largely african american males ages 18-24 and24-44 then the statistical significance of the finding isnt relative. If on the other hand the population is blended or of an alternative background this information would be mark stark… in which case why is the number disproportionate?