Shooting Data Analysis Project

S

2/24/2023

Introduction

In this project, I analyzed shooting data provided publicly by the New York City government. The analysis indicated clear concentrations of gun violence with respect to the time of day and location. I then built a model that fits a curve around the time of prior shooting events. I tested that model on a different 8 years of data that the model had not seen before. If future shootings occur in part along the same pattern of prior events, the model suggests a possible optimization of future police staffing to either 1) prevent shootings through deterrence, or 2) be present to more easily catch the perpetrators after the crime.

Import, Clean, and Tidy the Data

The standard R libraries I use are tidyverse, stringr, ggplot2, dplyr, and lubridate.

First, I import the data as as csv file.

Some of the variables I discard becuase I do not believe they are relevant to this analysis. The discarded variables are INCIDENT_KEY, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, and Lon_Lat.

Some of the other variables required factoring (for categorical types) or accounting for missing data. For the variables with missing data or data that I deemed to be incorrect (such as age values of more than 1,000), I usually omitted the NA values. The exception of this was the LOCATION_DESC variable, since I hypothesized that the lack of a descriptor for a shooting event indicated it was a different type of location (and hence inserted 'OTHER') and not unknown (like the age of the perpetrator might be in a lot of cases). This solidifies some of the sampling bias innate within the data set, but allows for simpler statistical analysis.

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

#creating separate date and time variables as well as a combined date_time variable
dates = as.POSIXct(df$OCCUR_DATE,format = "%m/%d/%y", tz = "America/New_York")
times = as.POSIXct(df$OCCUR_TIME,format = "%H:%M:%S", tz = "America/New_York")
f = "%m/%d/%Y %H:%M:%S"
dt = as.POSIXct(paste(df$OCCUR_DATE,df$OCCUR_TIME), format = f, tz = "America/New_York")

boroughs = factor(df$BORO)

precinct = df$PRECINCT

locations = factor(df$LOCATION_DESC)
levels(locations)[match("",levels(locations))] <- "OTHER"

murder_flag = factor(df$STATISTICAL_MURDER_FLAG)
```

```r
age_factors = c("<18", "18-24", "25-44", "45-64", "65+") #removing cases where the perp age range is un

perp_age = factor(df$PERP_AGE_GROUP,age_factors)
perp_age = na.omit(perp_age)

sex_levels = c("M", "F") #The data set has some rows where the the value is U, which I'm assuming is un
perp_sex = factor(df$PERP_SEX, sex_levels)
perp_sex = na.omit(perp_sex)

race_levels = c("ASIAN / PACIFIC ISLANDER","BLACK", "BLACK HISPANIC", "WHITE HISPANIC", "WHITE", "AMERIC

perp_race = factor(df$PERP_RACE, race_levels)
perp_race = na.omit(perp_race)

vic_age = factor(df$VIC_AGE_GROUP,age_factors)
vic_age = na.omit(vic_age)

vic_sex = factor(df$VIC_SEX, sex_levels)
vic_sex = na.omit(vic_sex)

vic_race = factor(df$PERP_RACE, race_levels)
vic_race = na.omit(perp_race)

lat = df$Latitude
long = df$Longitude

#print summaries
print("Summary of Event Dates/Times")
```

```
## [1] "Summary of Event Dates/Times"
```

```r
summary(dt)
```

```
##                       Min.                      1st Qu.
## "2006-01-01 02:00:00.0000" "2009-05-10 04:05:00.0000"
##                     Median                         Mean
## "2012-08-26 01:05:00.0000" "2013-06-14 04:24:56.1064"
##                    3rd Qu.                         Max.
## "2017-07-01 00:20:15.0000" "2021-12-31 19:23:00.0000"
```

```r
print("Summary of Boroughs")
```

```
## [1] "Summary of Boroughs"
```

```r
summary(boroughs)
```

```
##         BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND
##          7402         10365          3265          3828           736
```

```r
print("Summary of Precicnt Number")
```

```
## [1] "Summary of Precicnt Number"
```

```r
summary(precinct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   44.00   69.00   65.87   81.00  123.00 
```

```r
print("Summary of Location Description")
```

```
## [1] "Summary of Location Description"
```

```r
summary(locations)
```

```
##                      OTHER                        ATM                       BANK 
##                      14977                          1                          3 
##             BAR/NIGHT CLUB         BEAUTY/NAIL SALON                CANDY STORE 
##                        588                        105                          6 
##                CHAIN STORE                 CHECK CASH          CLOTHING BOUTIQUE 
##                          5                          1                         14 
##            COMMERCIAL BLDG                 DEPT STORE             DOCTOR/DENTIST 
##                        265                          9                          1 
##                 DRUG STORE       DRY CLEANER/LAUNDRY          FACTORY/WAREHOUSE 
##                         11                         31                          6 
##                  FAST FOOD                GAS STATION             GROCERY/BODEGA 
##                         99                         61                        622 
##       GYM/FITNESS FACILITY                   HOSPITAL                HOTEL/MOTEL 
##                          3                         47                         32 
##              JEWELRY STORE               LIQUOR STORE               LOAN COMPANY 
##                         12                         36                          1 
##    MULTI DWELL - APT BUILD MULTI DWELL - PUBLIC HOUS                       NONE 
##                       2664                       4559                        175 
##           PHOTO/COPY STORE                  PVT HOUSE          RESTAURANT/DINER 
##                          1                        893                        194 
##                     SCHOOL                 SHOE STORE             SMALL MERCHANT 
##                          1                          9                         25 
## SOCIAL CLUB/POLICY LOCATI          STORAGE FACILITY         STORE UNCLASSIFIED 
##                         66                          1                         35 
##                SUPERMARKET            TELECOMM. STORE              VARIETY STORE 
##                         19                          5                         11 
##                VIDEO STORE 
##                          2 
```

```r
print("Summary of Perpetrator Age")
```

```
## [1] "Summary of Perpetrator Age"
```

```r
summary(perp_age)
```

```
##   <18 18-24 25-44 45-64   65+
##  1463  5844  5202   535    57
```

```r
print("Summary of Perpretrator Race")
```

```
## [1] "Summary of Perpretrator Race"
```

```r
summary(perp_race)
```

```
##       ASIAN / PACIFIC ISLANDER                          BLACK
##                            141                          10668
##                 BLACK HISPANIC                 WHITE HISPANIC
##                           1203                           2164
##                          WHITE AMERICAN INDIAN/ALASKAN NATIVE
##                            272                              2
```

```r
print("Summary of Perpetrator Sex")
```

```
## [1] "Summary of Perpetrator Sex"
```

```r
summary(perp_sex)
```

```
##     M     F
## 14416   371
```

```r
print("Summary of Victim Age")
```

```
## [1] "Summary of Victim Age"
```

```r
summary(vic_age)
```

```
##   <18 18-24 25-44 45-64   65+
##  2681  9604 11386  1698   167
```

```r
print("Summary of Victim Race")
```

```
## [1] "Summary of Victim Race"
```

```r
summary(vic_race)
```

```
##       ASIAN / PACIFIC ISLANDER                          BLACK
##                            141                          10668
##                 BLACK HISPANIC                 WHITE HISPANIC
##                           1203                           2164
##                          WHITE AMERICAN INDIAN/ALASKAN NATIVE
##                            272                              2
```

```r
print("Summary of Victim Sex")
```

```
## [1] "Summary of Victim Sex"
```

```r
summary(vic_sex)
```

```
##     M     F
## 23182  2403
```

```r
print("Summary of Latitude")
```

```
## [1] "Summary of Latitude"
```

```r
summary(lat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   40.51   40.67   40.70   40.74   40.82   40.91
```

```r
print("Summary of Longitude")
```

```
## [1] "Summary of Longitude"
```

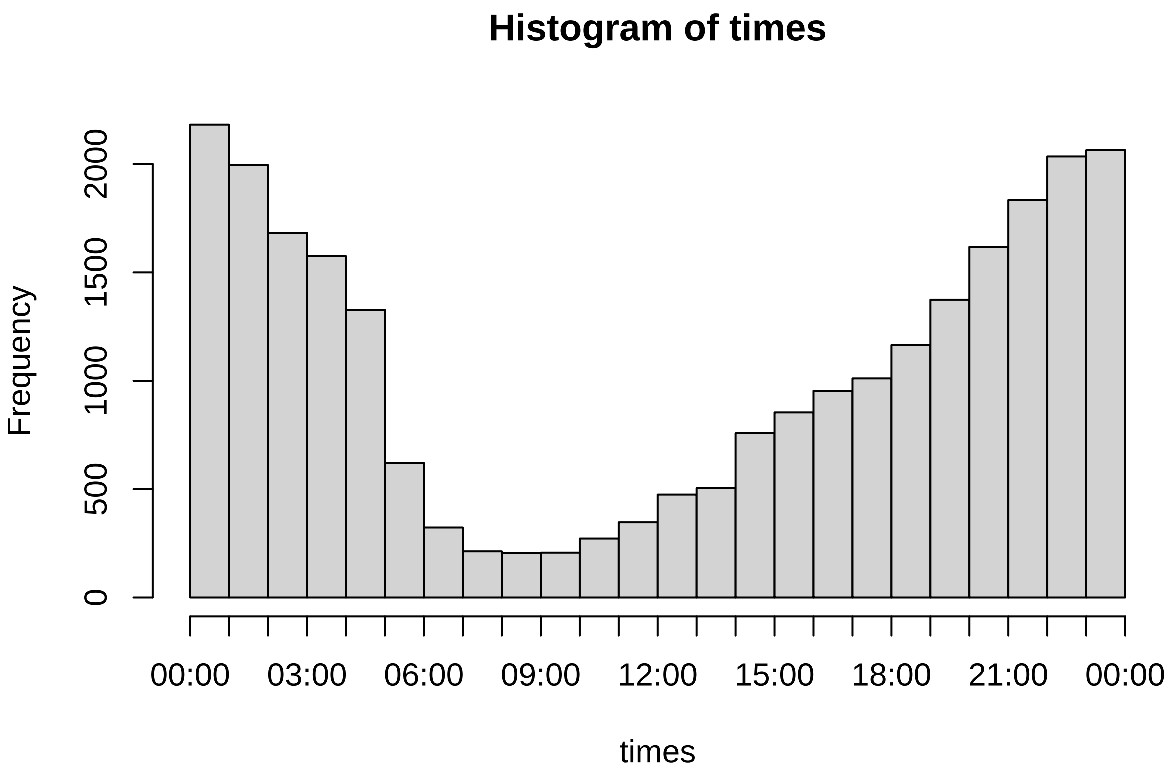
```r
summary(long)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -74.25  -73.94  -73.92  -73.91  -73.88  -73.70
```

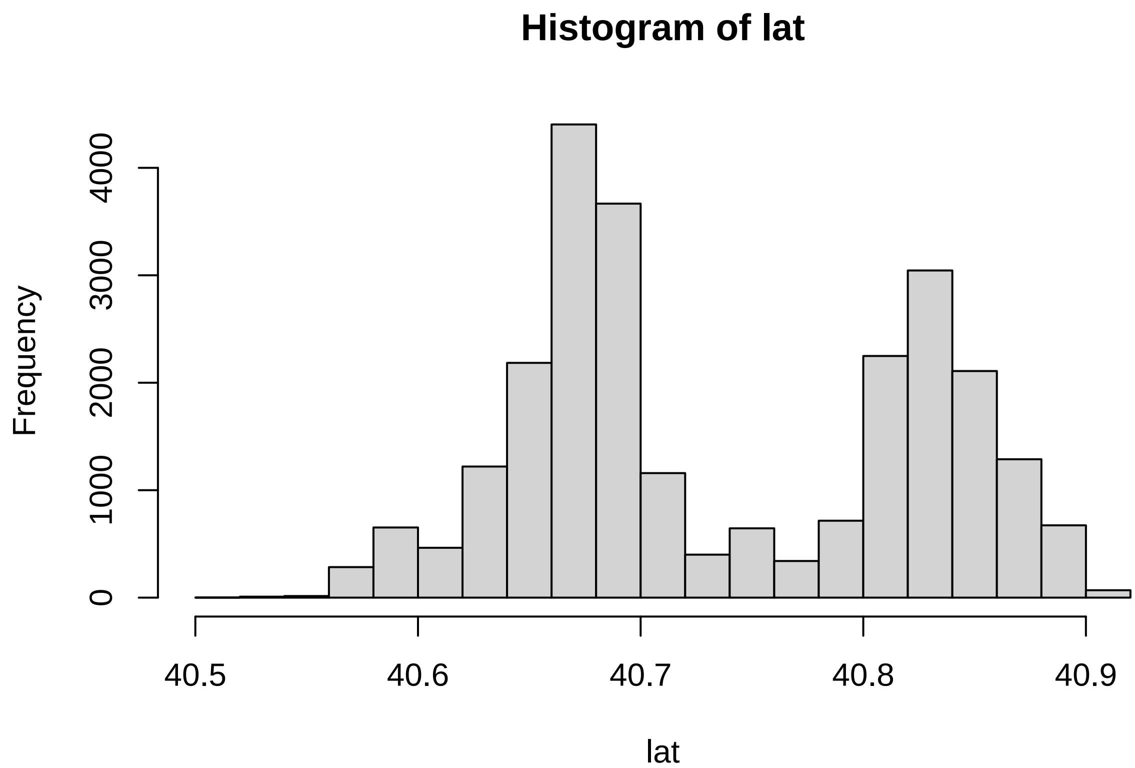
Generate Visualizations of the Data

The three plots below show histograms of time, latitude, and longitude with respect to shooting frequency. Hour of the day and longitude are clearly clustered around single points, whereas latitude is bi-modally distributed. Note that while the time of day looks bi-modally distributed, remember that time is a continuous variable that loops back, so 11pm is adjacent to midnight, which is adjacent to 1am…

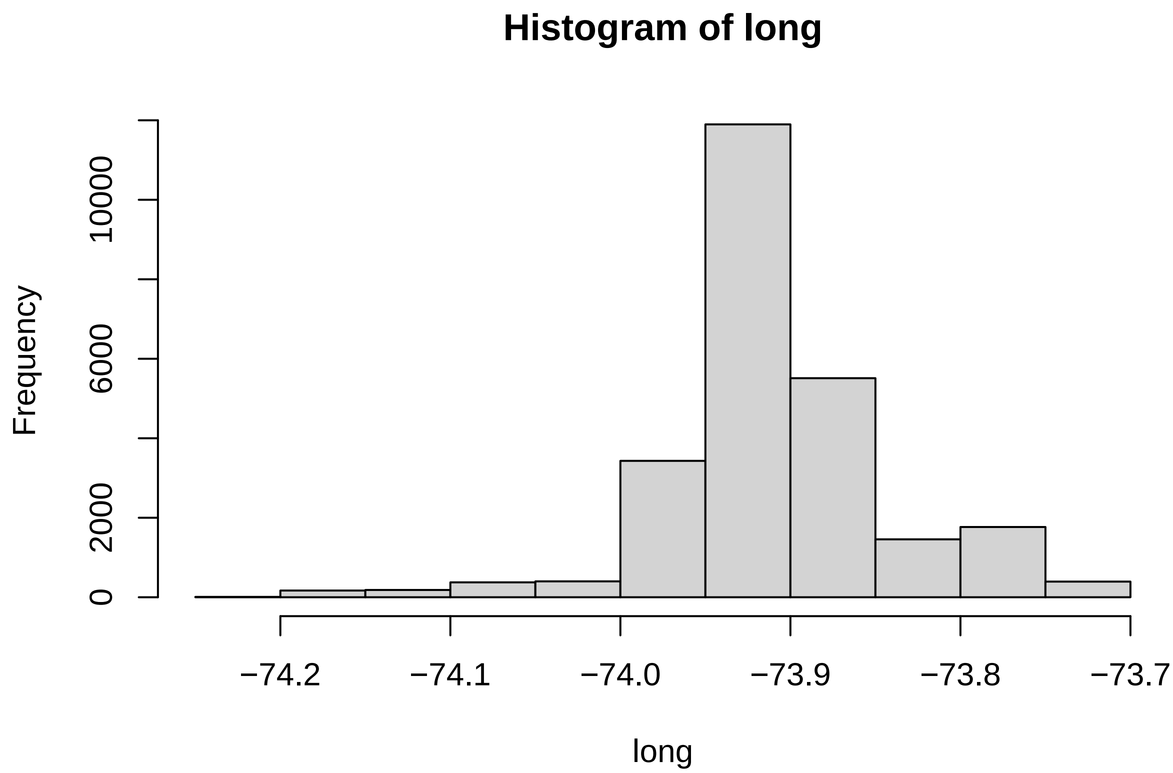
```r
hist(times,breaks="hours",freq = TRUE)
```


Histogram of times
Frequency
0
500
1000
1500
2000
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
00:00
times

```
hist(lat)
```


Histogram of lat
Frequency
0
1000
2000
3000
4000
40.5
40.6
40.7
40.8
40.9
lat

```
hist(long)
```



Generate an Analysis Model on the Data

Focusing on the three variables I generated visualizations of, I built a model that seeks to predict the proportion of shootings over each hour in a day. I did they by training a 5-degree polynomial model on the shooting occurrences from 2006 through 2013. I then applied that model to 2014-2018 shooting data to test it on a data set separate from the training.

```
df$OCCUR_DATE = as.POSIXct(df$OCCUR_DATE, format = "%m/%d/%Y")
d1 = df[df$OCCUR_DATE < "2014-01-01",]
d2 = df[df$OCCUR_DATE >= "2014-01-01",]

d1t = as.POSIXct(d1$OCCUR_TIME,format = "%H:%M:%S", tz = "America/New_York")
d2t = as.POSIXct(d2$OCCUR_TIME,format = "%H:%M:%S", tz = "America/New_York")

c1 = cut(d1t, breaks = "1 hour")
c1_summary = summary(c1)
prop_of_shootings1 = numeric(24)
time_bins = numeric(24)
total1 = sum(c1_summary)

for (i in 1:24){
  time_bins[i] = i-1
  prop_of_shootings1[i] = c1_summary[[i]]/total1
}
```

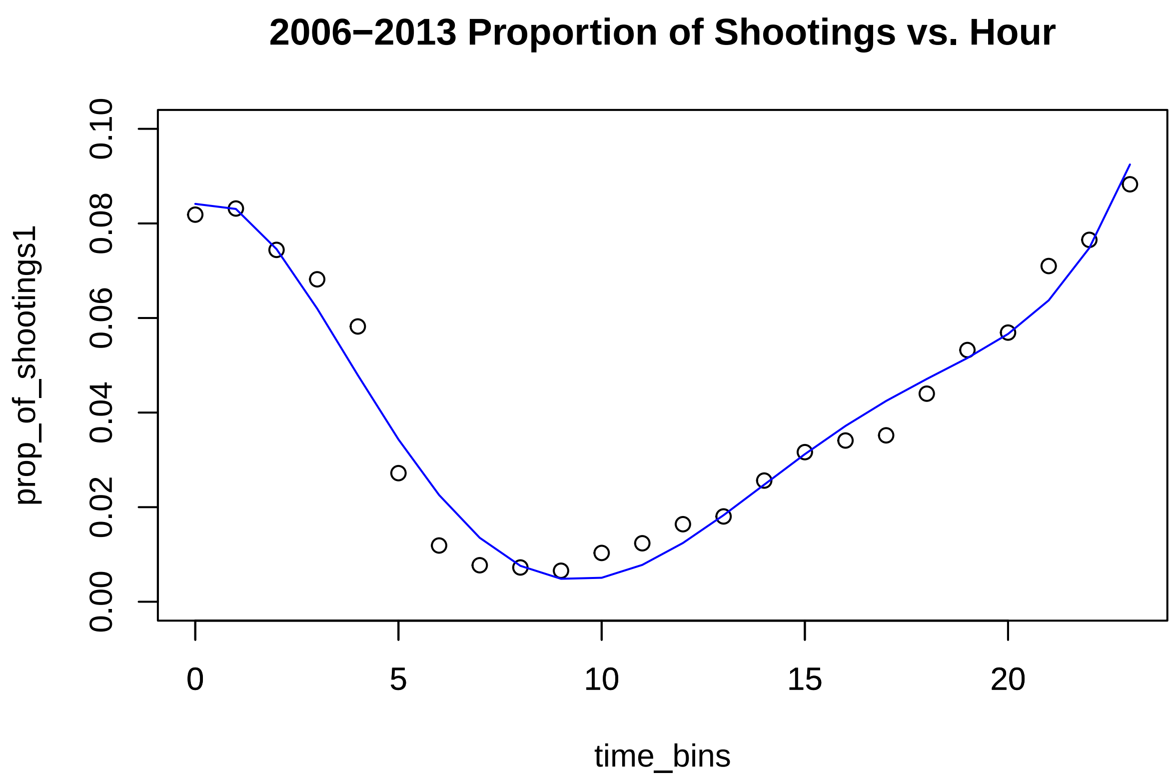
```r
model = lm(prop_of_shootings1 ~ poly(time_bins,5))

summary(model)
```

```
##
## Call:
## lm(formula = prop_of_shootings1 ~ poly(time_bins, 5))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0106832 -0.0030899  0.0001881  0.0022792  0.0102914
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.041667   0.001143  36.441  < 2e-16 ***
## poly(time_bins, 5)1  0.010064   0.005601   1.797 0.089193 .
## poly(time_bins, 5)2  0.127228   0.005601  22.713 1.06e-14 ***
## poly(time_bins, 5)3 -0.021326   0.005601  -3.807 0.001290 **
## poly(time_bins, 5)4 -0.011427   0.005601  -2.040 0.056305 .
## poly(time_bins, 5)5  0.026963   0.005601   4.814 0.000139 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005601 on 18 degrees of freedom
## Multiple R-squared:  0.9689, Adjusted R-squared:  0.9603
## F-statistic: 112.2 on 5 and 18 DF,  p-value: 6.417e-13
```

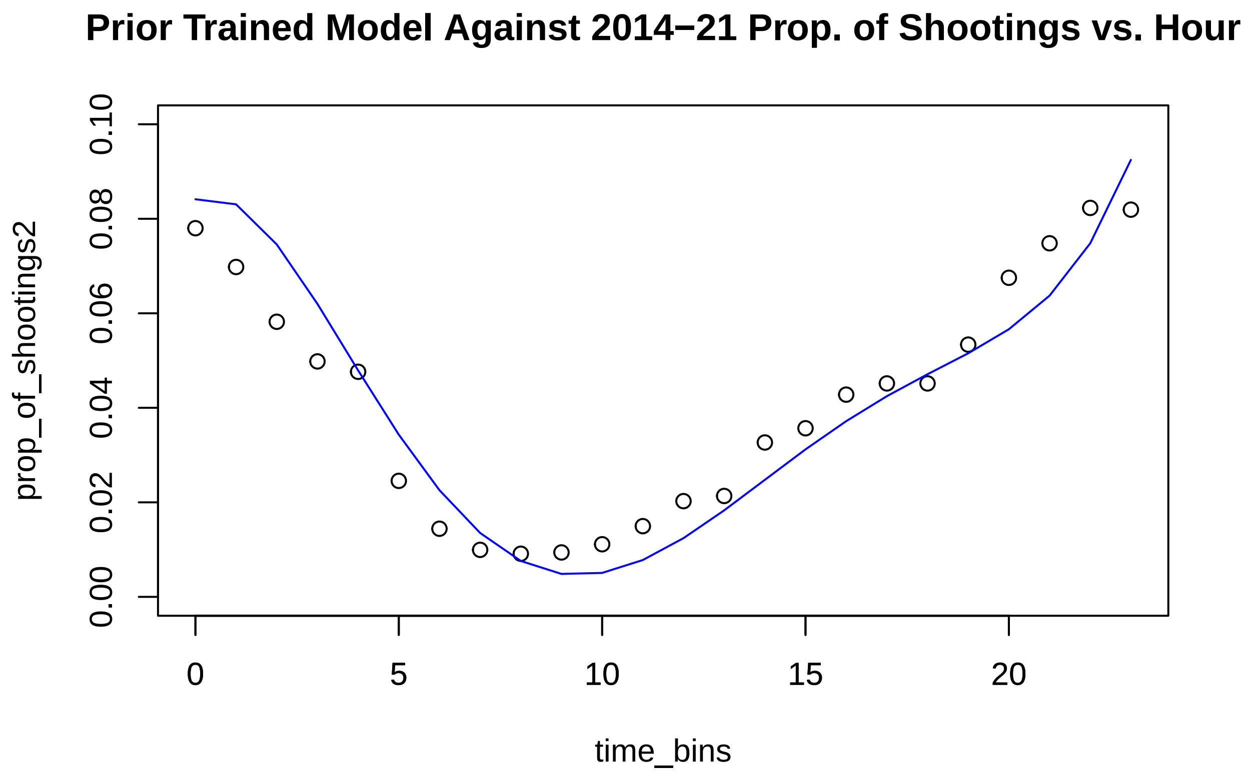
```r
predicted1 = predict(model,data.frame(x=time_bins))
rmse1 = sqrt(mean(prop_of_shootings1 - predicted1)^2)
print(rmse1)
```

```
## [1] 2.486439e-17
```

```r
frame()
plot(time_bins,prop_of_shootings1,ylim = c(0,.1))
lines(time_bins,predicted1,col='blue')
title("2006-2013 Proportion of Shootings vs. Hour")
axis(side=1,at=seq(0,23,5))
```




```
c2 = cut(d2t, breaks = "1 hour")
c2_summary = summary(c2)
prop_of_shootings2 = numeric(24)
total2 = sum(c2_summary)

for (i in 1:24){
  prop_of_shootings2[i] = c2_summary[[i]]/total2
}
frame()
plot(time_bins,prop_of_shootings2,ylim = c(0,.1))
lines(time_bins,predicted1,col='blue')
title("Prior Trained Model Against 2014-21 Prop. of Shootings vs. Hour")
axis(side=1,at=seq(0,23,5))
```




```
rmse2 = sqrt(mean(prop_of_shootings2 - predicted1)^2)
print(rmse2)
```

```
## [1] 2.334645e-17
```

Conclusion

In this project, I imported, cleaned, and visualized data on gun violence in New York City. As an analysis, I built a 5 degree polynomial model trained on the first 8 years of shooting data that predicts overall proportion of shootings per each hour of the day. I then applied that model's prediction the second 8 years of shooting data as a test. The root mean square error remained low. This suggests that if the NYPD seeks to better police the city, it would be better served to increase staffing during night hours rather than during the day - and exactly per the polynomial regression model distribution would be even better.

A potential bias in the data is that it only accounts for shooting events that were recorded. It is quite possible that not all shootings are reported, particularly in neighborhoods where bystanders may fear retatliation should they report shootings to the police. This should be considered when applying any statistical conclusions from this project to real-world policing policies.

A personal bias I might have in this analysis is that I actually grew up for several years in New York City and remember my parents telling me not to go out exploring at night. To mitigate this error, my analysis focuses on what the data shows - not what lessons my parents taught me as a child. Furthermore, I separated my training and testing data when building the polynomial regression model. That way, I could conduct a rigorous test of my model (which shows that there are more shootings at night) without it being biased by data it had already seen.