NYPD Shooting Incident

Mike Prodo

10/28/2021

The Data

The data being used for this project is a list of every shooting incident that occurred in New York City from the years 2006 through 2020. This data and information about it can be found through the link https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic.

Libraries used in this project are "chron."

Reading in the data:

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

Initial look at the data:

```
summary(NYPD_data)
```

```
##   INCIDENT_KEY            OCCUR_DATE         OCCUR_TIME                BORO
##  Min.   :  9953245   07/05/2020:   47   23:30:00:  159   BRONX        :6700
##  1st Qu.: 55317014   09/04/2011:   31   21:00:00:  128   BROOKLYN     :9722
##  Median : 83365370   07/26/2020:   29   22:30:00:  126   MANHATTAN    :2921
##  Mean   :102218616   08/11/2007:   26   1:30:00 :  125   QUEENS       :3527
##  3rd Qu.:150772442   08/15/2020:   25   23:00:00:  121   STATEN ISLAND: 698
##  Max.   :222473262   09/04/2006:   25   0:30:00 :  118
##                      (Other)   :23385   (Other) :22791
##     PRECINCT      JURISDICTION_CODE                   LOCATION_DESC
##  Min.   :  1.00   Min.   :0.0000                               :13581
##  1st Qu.: 44.00   1st Qu.:0.0000    MULTI DWELL - PUBLIC HOUS: 4230
##  Median : 69.00   Median :0.0000    MULTI DWELL - APT BUILD  : 2551
##  Mean   : 66.21   Mean   :0.3323    PVT HOUSE                :  858
##  3rd Qu.: 81.00   3rd Qu.:0.0000    GROCERY/BODEGA           :  572
##  Max.   :123.00   Max.   :2.0000    BAR/NIGHT CLUB           :  558
##                   NA's   :2         (Other)                  : 1218
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX          PERP_RACE
##  false:19080                    :8459    : 8425   BLACK         :9855
##  true : 4488             18-24  :5448   F:  334                 :8425
##                          25-44  :4613   M:13305   WHITE HISPANIC:1961
##                          UNKNOWN:3156   U: 1504   UNKNOWN       :1869
##                          <18    :1354             BLACK HISPANIC:1081
##                          45-64  : 481             WHITE         : 255
##                          (Other):  57             (Other)       : 122
##  VIC_AGE_GROUP   VIC_SEX                             VIC_RACE
##  <18     : 2525   F: 2195   AMERICAN INDIAN/ALASKAN NATIVE:    9
##  18-24   : 9000   M:21353   ASIAN / PACIFIC ISLANDER      :  320
##  25-44   :10287   U:   20   BLACK                         :16846
##  45-64   : 1536             BLACK HISPANIC                : 2244
##  65+     :  155             UNKNOWN                       :  102
##  UNKNOWN:   65             WHITE                         :  615
##                             WHITE HISPANIC                : 3432
##    X_COORD_CD      Y_COORD_CD       Latitude       Longitude
##  1017119:   80   183909 :   80   Min.   :40.51   Min.   :-74.25
##  1008276:   52   183623 :   58   1st Qu.:40.67   1st Qu.:-73.94
##  1026387:   52   262634 :   52   Median :40.70   Median :-73.92
##  1008427:   51   183518 :   51   Mean   :40.74   Mean   :-73.91
##  1046405:   49   183798 :   49   3rd Qu.:40.82   3rd Qu.:-73.88
##  1017141:   48   187113 :   49   Max.   :40.91   Max.   :-73.70
##  (Other):23236   (Other):23229
##                                           Lon_Lat
##  POINT (-73.88151172399995 40.67141166300007) :   80
##  POINT (-73.84760778699997 40.88745131300004) :   52
##  POINT (-73.91339091999998 40.670655072000045):   52
##  POINT (-73.91284696199995 40.670366460000025):   51
##  POINT (-73.77590919399995 40.680048726000045):   49
##  POINT (-73.88143295699997 40.67110691100004) :   48
##  (Other)                                      :23236
```

For this project we will be looking at the number of shootings on each day of the week, as well as what time these shootings occur. The two variables we need are OCCUR_DATE and OCCUR_TIME.

Cleaning data to obtain the information we need:

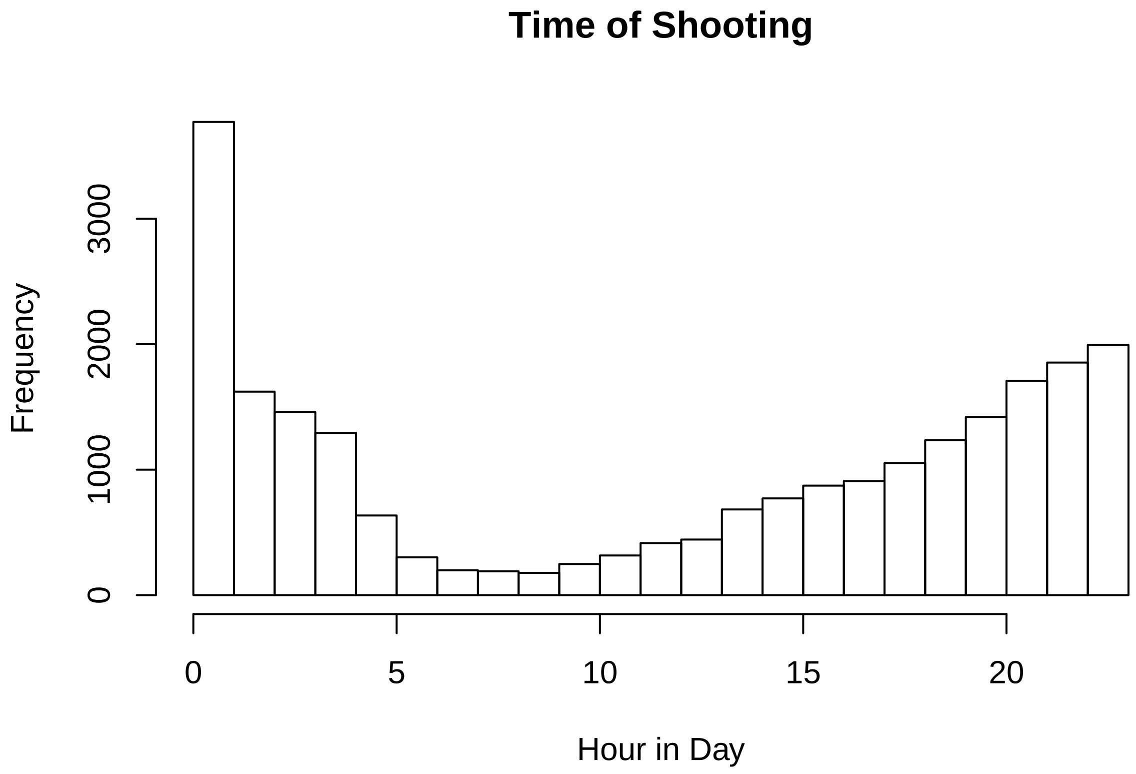
```
# creating new data frame with OCCUR_TIME and OCCUR_DATE variables
NYPD = NYPD_data[,2:3]

# Convert variables to appropriate types
NYPD$OCCUR_DATE = as.Date(NYPD$OCCUR_DATE, format = "%m/%d/%Y")
NYPD$OCCUR_TIME = chron(times. = NYPD$OCCUR_TIME, format = "h:m:s")
```

Time of Shooting Analysis

The first visual to look at is a histogram of the time of each shooting occurrence.

```
hist(x = as.numeric(substr(NYPD$OCCUR_TIME,1,2)), breaks = 0:23,
     xlab = "Hour in Day", main = "Time of Shooting")
```



As we can see by the histogram, most shootings occur at night. The shape of the histogram is a parabola, so we will look at fitting a quadratic model to the data.

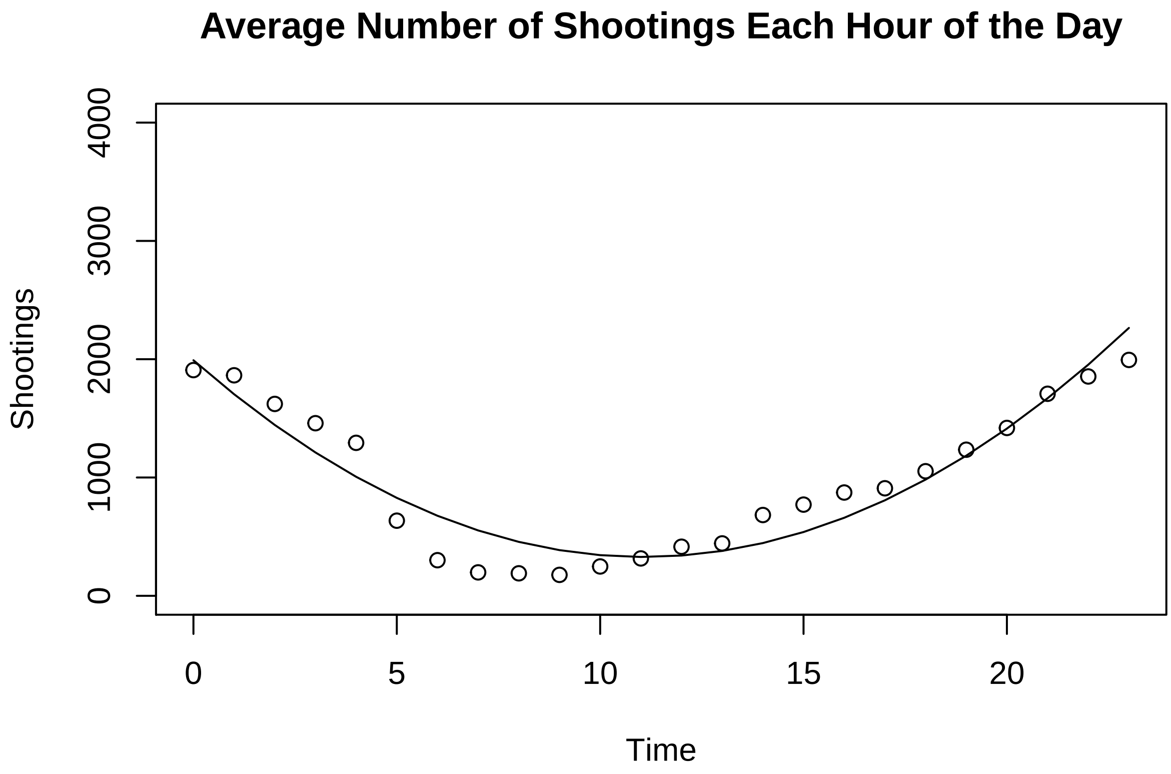
```
# Create time data frame
TimeData = data.frame(matrix(ncol = 3, nrow = 24))
colnames(TimeData) = c("Time","TimeSquared","Shootings")
TimeData$Time = 0:23
TimeData$TimeSquared = TimeData$Time^2
for (i in 1:24){
  TimeData$Shootings[i] = sum(as.numeric(substr(NYPD$OCCUR_TIME,1,2)) == TimeData$Time[i])
}
model = lm(Shootings~Time + TimeSquared, data = TimeData)
summary(model)
```

```
##
## Call:
## lm(formula = Shootings ~ Time + TimeSquared, data = TimeData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -375.62 -122.93   44.31  164.34  286.90
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1991.1385   117.1003   17.00 9.37e-14 ***
## Time        -300.6052    23.5835  -12.75 2.37e-11 ***
## TimeSquared   13.5864     0.9903   13.72 5.93e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 207.4 on 21 degrees of freedom
## Multiple R-squared:  0.9014, Adjusted R-squared:  0.892
## F-statistic:    96 on 2 and 21 DF,  p-value: 2.725e-11
```

The quadratic model for predicting the number of shootings in each hour of the day is $1991.14 - 300.61x + 13.59x^2$ where x is the hour of the day 0 to 23. The Adjusted R-squared value is 0.892 which signifies that this quadratic model is a good fit for the data.

Fitting the quadratic model on the plot.

```
plot(x = TimeData$Time, y = TimeData$Shootings, ylim = range(0:4000),
     main = "Average Number of Shootings Each Hour of the Day", xlab = "Time",
     ylab = "Shootings")
pred = predict(model, x = TimeData$Time)
lines(x = 0:23, y = pred, type="l")
```



Looking at the plot, the quadratic model looks like a good fit for the data.

Weekday Analysis

Now let's look at what day of the week shootings occur by adding a weekday variable.

```r
# Get weekday of date
NYPD$Weekday = weekdays(NYPD$OCCUR_DATE)
```

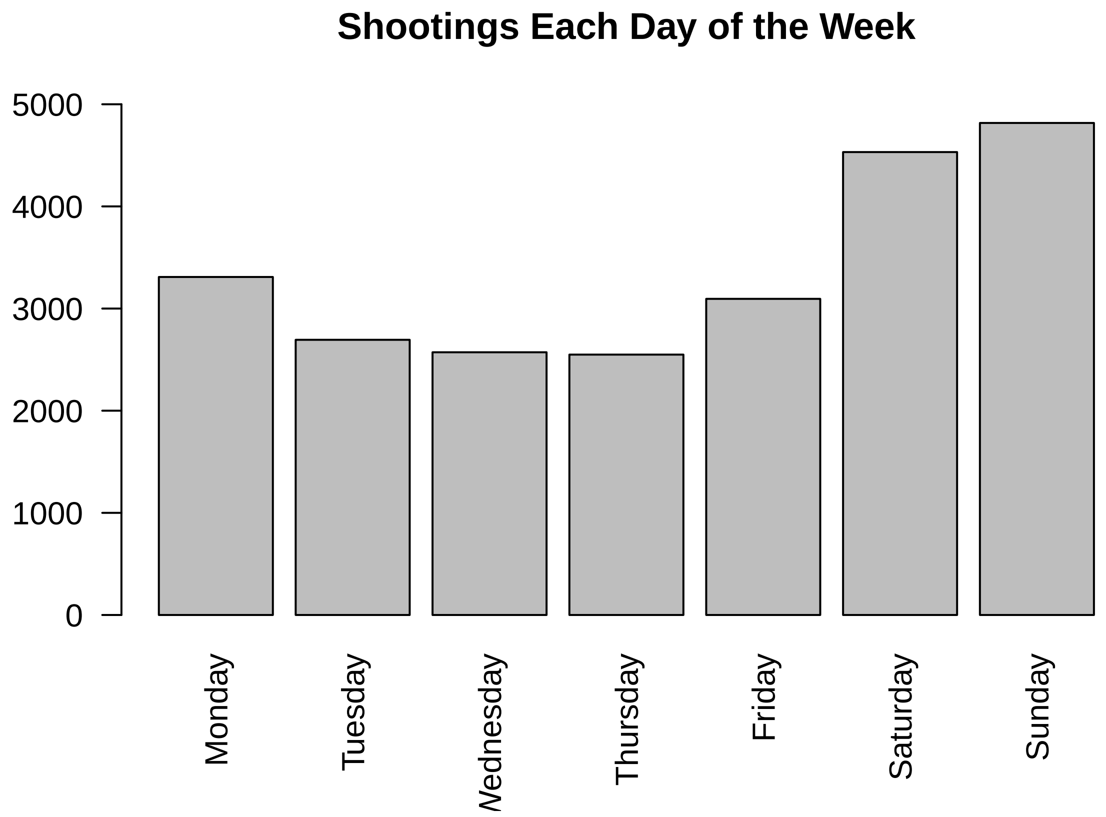
Making a table for the number of shootings each day of the week.

```r
# Create table
day = data.frame(matrix(nrow=7,ncol=2))
colnames(day) = c("Day","Shootings")

# Fill in table
day$Day = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday")
for (i in 1:7){
  day$Shootings[i] = sum(NYPD$Weekday==day$Day[i])
}
```

Creating a bar plot for number of shootings each day of the week.

```r
barplot(height = day$Shootings, names.arg = day$Day, ylim = c(0,5000),
        main = "Shootings Each Day of the Week", las = 2)
```



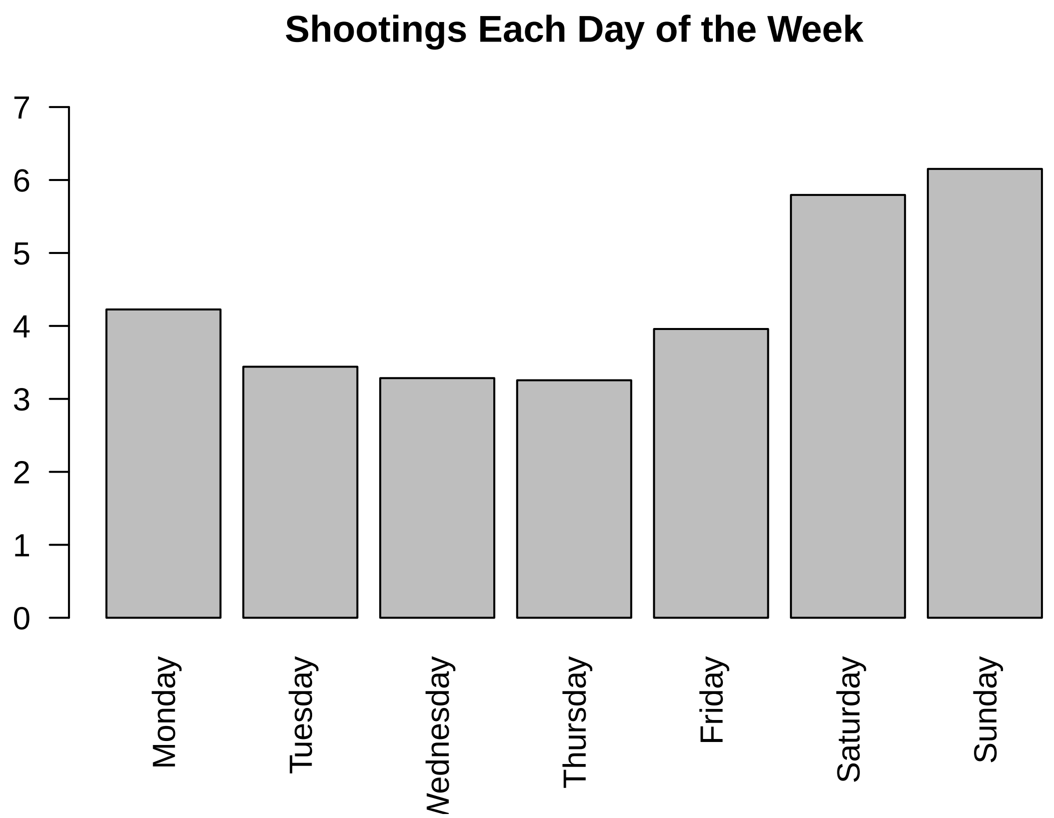
Adding average number of shootings each day variable.

```r
for (i in 1:7){
  day$Count[i] = sum(format(seq(from = min(NYPD$OCCUR_DATE), to = max(NYPD$OCCUR_DATE),
                                by = "day"), "%w") == i)
  # The numerical value in R for Sunday is 0 instead of 7
  if (i == 7){
    day$Count[i] = sum(format(seq(from = min(NYPD$OCCUR_DATE), to = max(NYPD$OCCUR_DATE),
                                  by = "day"), "%w") == 0)
  }

  day$Average[i] = day$Shootings[i]/day$Count[i]
}
```

Looking at bar plot for average number of shootings each day.

```
barplot(height = day$Average, names.arg = day$Day, ylim = c(0,7),
        main = "Shootings Each Day of the Week", las = 2)
```



After looking at the bar plot, it appears that the number of shootings each day Monday through Sunday is quadratic, so we will fit a quadratic model to the data. Monday through Sunday will be indicated by their day values, Monday being 1 through Sunday being 7.

Creating a quadratic model for predicting the number of shootings on a given weekday.

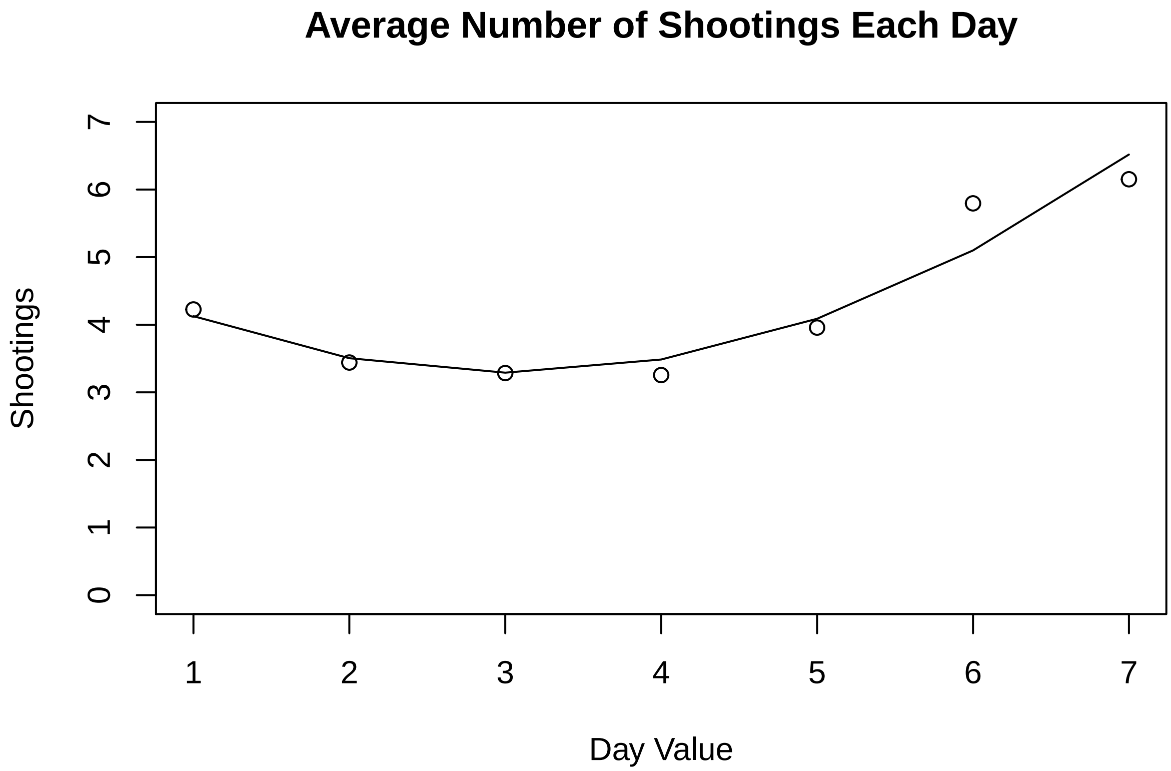
```
dayValues = 1:7
dayValuesSquared = dayValues^2
model2 = lm(day$Average~(dayValues + dayValuesSquared))
summary(model2)
```

```
##
## Call:
## lm(formula = day$Average ~ (dayValues + dayValuesSquared))
##
## Residuals:
##         1         2         3         4         5         6         7
##  0.099803 -0.063944 -0.006173 -0.230076 -0.130342  0.696504 -0.365773
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.15605    0.65328   7.893  0.00139 **
## dayValues        -1.23386    0.37439  -3.296  0.03005 *
## dayValuesSquared  0.20406    0.04574   4.461  0.01115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4192 on 4 degrees of freedom
## Multiple R-squared:  0.9187, Adjusted R-squared:  0.8781
## F-statistic: 22.61 on 2 and 4 DF,  p-value: 0.006605
```

The quadratic model for predicting the number of shootings on a given day is $5.156 - 1.234x + 0.204x^2$ where x is the day value. The Adjusted R-squared value is 0.8781 which signifies that this quadratic model is a good fit for the data.

Let's plot the data and prediction and see how it fits.

```
plot(x = 1:7, y = day$Average, ylim = range(0:7), main = "Average Number of Shootings Each Day",
     xlab = "Day Value", ylab = "Shootings")
pred = predict(model2, x = 1:7)
lines(pred, type="l")
```



As we can see by the plot, the quadratic model is in fact a good fit for the data.

Conclusion

In conclusion, most of the shootings in New York City happen during the night hours. The number of shootings that will happen can be pretty well predicted by the quadratic model $1991.14 - 300.61x + 13.59x^2$ where x is the hour in military time 0:23 with an Adjusted R-Squared value of 0.892. If you wanted to predict the number of shootings each hour in a given day, the model would be $0.363 - 0.055x + 0.002x^2$, where again the x is the hour in military time 0:23.

The average number of shootings in New York City each day during the week, Monday through Friday, is between 3 and 4, whereas the average number of shootings on the weekend in New York City is around 6. Looking at a given week Monday through Sunday, the relationship between what day it is and the average number of shootings is quadratic. The quadratic model $5.156 - 1.234x + 0.204x^2$ fits the data well with an Adjusted R-Squared value of 0.8781.

Possible bias in this data set is that most shootings occur at night before people go to bed, and this leads to the recording of shootings technically being the next day. This bias has been mitigated by looking at both the day shootings occur and what time they occur. These findings can be used to possibly inform the New York police department about which days and times might need more or less police presence around the city.