NYPD Shooting

D. Ikoma

2022-09-11

Objective of analysis

New York is a quite attractive city for foreign travelers. But, gun shooting incidents are of critical concern for foreign travelers in New York. Therefore we would like to investigate the trend of gun shooting incidents in New York. Especially the trend of areas and time slots is helpful for travelers avoiding shooting incidents.

Clear memory

At first, we clear memories in advance.

```
rm(list=ls())
gc();gc()
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 460281 24.6     992988 53.1   644200 34.5
## Vcells 828769  6.4    8388608 64.0  1635000 12.5
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 460703 24.7     992988 53.1   644200 34.5
## Vcells 829743  6.4    8388608 64.0  1635000 12.5
```

Import libraries

We import libraries to use this analysis.

```
# ```{r import_libraries, echo = TRUE, message = FALSE}
Sys.setenv(LANGUAGE="en")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- 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
```

Read data

We read data from web site.

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/"
file_names <- c("rows.csv")
urls <- str_c(url_in, file_names)
data_NYPD <- read_csv(urls[1])
```

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
spec(data_NYPD)
```

```
## 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()
## )
```

Tidy data

We investigate basic data and statistics. Next we remove some data and missing values for our analysis.

```
data_NYPD
```

```
## # A tibble: 25,596 x 19
##    INCID~1 OCCUR~2 OCCUR~3 BORO  PRECI~4 JURIS~5 LOCAT~6 STATI~7 PERP_~8 PERP_~9
##      <dbl> <chr>   <time>  <chr>   <dbl>   <dbl> <chr>   <lgl>   <chr>   <chr>
##  1  2.36e8 11/11/~ 15:04   BROO~      79       0 <NA>    FALSE   <NA>    <NA>
##  2  2.31e8 07/16/~ 22:05   BROO~      72       0 <NA>    FALSE   45-64   M
##  3  2.31e8 07/11/~ 01:09   BROO~      79       0 <NA>    FALSE   <18     M
##  4  2.38e8 12/11/~ 13:42   BROO~      81       0 <NA>    FALSE   <NA>    <NA>
##  5  2.24e8 02/16/~ 20:00   QUEE~     113       0 <NA>    FALSE   <NA>    <NA>
##  6  2.28e8 05/15/~ 04:13   QUEE~     113       0 <NA>    TRUE    <NA>    <NA>
##  7  2.27e8 04/14/~ 21:08   BRONX      42       0 COMMER~ TRUE    <NA>    <NA>
##  8  2.38e8 12/10/~ 19:30   BRONX      52       0 <NA>    FALSE   <NA>    <NA>
##  9  2.25e8 02/22/~ 00:18   MANH~      34       0 <NA>    FALSE   <NA>    <NA>
## 10  2.25e8 03/07/~ 06:15   BROO~      75       0 <NA>    TRUE    25-44   M
## # ... with 25,586 more rows, 9 more variables: PERP_RACE <chr>,
## #   VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>, X_COORD_CD <dbl>,
## #   Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>, Lon_Lat <chr>, and
## #   abbreviated variable names 1: INCIDENT_KEY, 2: OCCUR_DATE, 3: OCCUR_TIME,
## #   4: PRECINCT, 5: JURISDICTION_CODE, 6: LOCATION_DESC,
## #   7: STATISTICAL_MURDER_FLAG, 8: PERP_AGE_GROUP, 9: PERP_SEX
```

```
summary(data_NYPD)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms         Class :character
##  Median : 86437258   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  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:25596       Length:25596       Length:25596       Length:25596
##  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:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  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
##
```

```
data_NYPD <- data_NYPD %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  select(-c(LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_RACE)) %>%
  drop_na(JURISDICTION_CODE)
data_NYPD
```

```
## # A tibble: 25,594 x 15
##    INCIDENT_KEY OCCUR_DATE OCCUR~1 BORO  PRECI~2 JURIS~3 STATI~4 VIC_A~5 VIC_SEX
##           <dbl> <date>     <time>  <chr>   <dbl>   <dbl> <lgl>   <chr>   <chr>
##  1    236168668 2021-11-11 15:04   BROO~      79       0 FALSE   18-24   M
##  2    231008085 2021-07-16 22:05   BROO~      72       0 FALSE   25-44   M
##  3    230717903 2021-07-11 01:09   BROO~      79       0 FALSE   25-44   M
##  4    237712309 2021-12-11 13:42   BROO~      81       0 FALSE   25-44   M
##  5    224465521 2021-02-16 20:00   QUEE~     113       0 FALSE   25-44   M
##  6    228252164 2021-05-15 04:13   QUEE~     113       0 TRUE    25-44   M
##  7    226950018 2021-04-14 21:08   BRONX      42       0 TRUE    18-24   M
##  8    237710987 2021-12-10 19:30   BRONX      52       0 FALSE   25-44   M
##  9    224701998 2021-02-22 00:18   MANH~      34       0 FALSE   25-44   M
## 10    225295736 2021-03-07 06:15   BROO~      75       0 TRUE    25-44   M
## # ... with 25,584 more rows, 6 more variables: VIC_RACE <chr>,
## #   X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>, Longitude <dbl>,
## #   Lon_Lat <chr>, and abbreviated variable names 1: OCCUR_TIME, 2: PRECINCT,
## #   3: JURISDICTION_CODE, 4: STATISTICAL_MURDER_FLAG, 5: VIC_AGE_GROUP
```

```
summary(data_NYPD)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME          BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:25594      Length:25594
##  1st Qu.: 61593633   1st Qu.:2009-05-10   Class1:hms        Class :character
##  Median : 86437258   Median :2012-08-26   Class2:difftime   Mode  :character
##  Mean   :112382536   Mean   :2013-06-13   Mode  :numeric
##  3rd Qu.:166660833   3rd Qu.:2017-06-30
##  Max.   :238490103   Max.   :2021-12-31
##     PRECINCT      JURISDICTION_CODE STATISTICAL_MURDER_FLAG VIC_AGE_GROUP
##  Min.   :  1.00   Min.   :0.0000    Mode :logical           Length:25594
##  1st Qu.: 44.00   1st Qu.:0.0000    FALSE:20666             Class :character
##  Median : 69.00   Median :0.0000    TRUE :4928              Mode  :character
```

```
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25594       Length:25594       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000010   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194030
##                                        Mean   :1009455   Mean   :207893
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25594
##  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
```

We group shooting criminals counts by borough to analyze area trends.

```
data_by_BORO <- data_NYPD %>%
  group_by(BORO) %>%
  count(BORO) %>%
  summarize(cases = sum(n)) %>%
  ungroup()
data_by_BORO
```

```
## # A tibble: 5 x 2
##   BORO          cases
##   <chr>         <int>
## 1 BRONX          7402
## 2 BROOKLYN      10365
## 3 MANHATTAN      3264
## 4 QUEENS         3827
## 5 STATEN ISLAND   736
```

Also, we group them by hour to know the dangerous time slots.

```
tibble_opt <- list(
  "tibble.print_max" = 100,
  "tibble.print_min" = 20
)
options(tibble_opt)

data_by_hour <- data_NYPD %>%
  mutate(hour = hour(OCCUR_TIME)) %>%
  group_by(hour) %>%
  count(hour) %>%
  summarize(hours = sum(n)) %>%
  ungroup()
data_by_hour
```

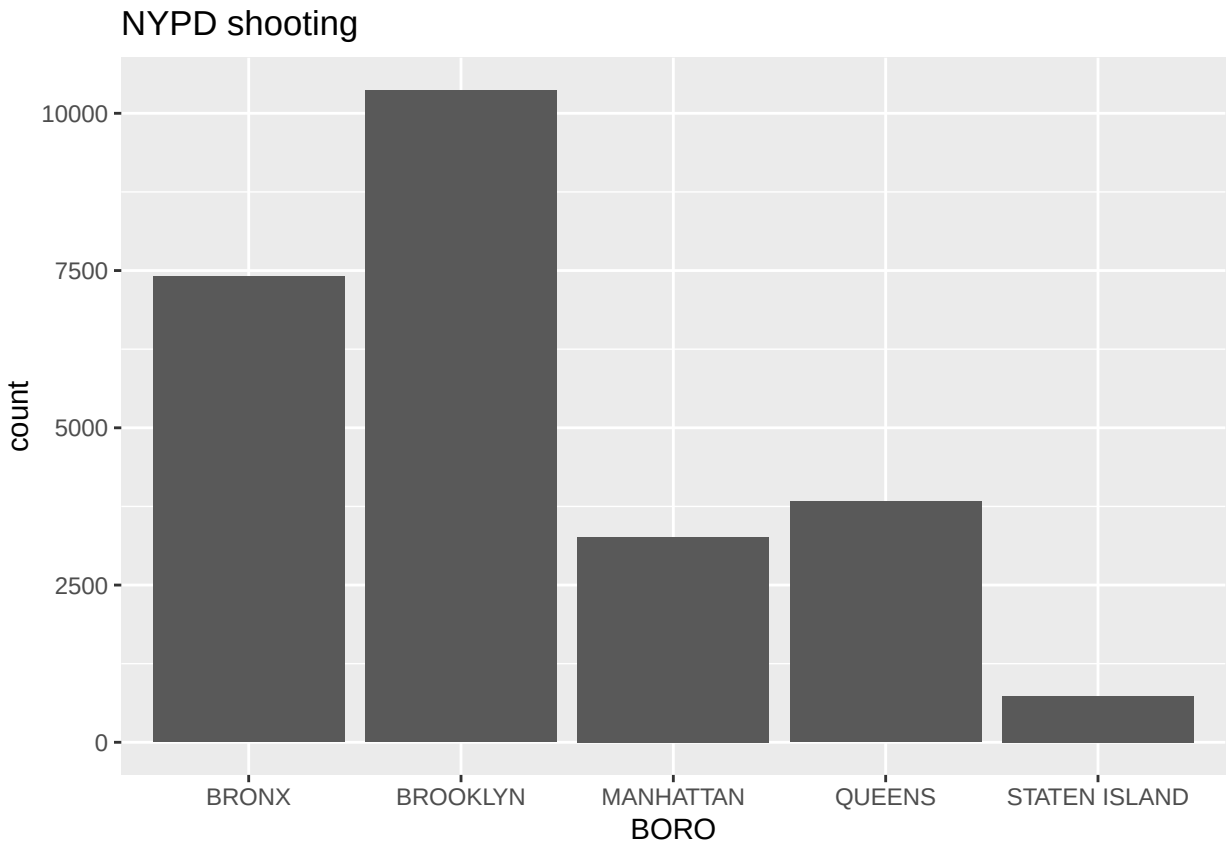
```
## # A tibble: 24 x 2
```

```
##     hour hours
##    <int> <int>
##  1     0  2053
##  2     1  1981
##  3     2  1726
##  4     3  1544
##  5     4  1374
##  6     5   667
##  7     6   332
##  8     7   222
##  9     8   206
## 10     9   199
## 11    10   273
## 12    11   345
## 13    12   462
## 14    13   498
## 15    14   733
## 16    15   854
## 17    16   968
## 18    17  1010
## 19    18  1139
## 20    19  1364
## 21    20  1573
## 22    21  1859
## 23    22  2022
## 24    23  2190
```

Visualize data

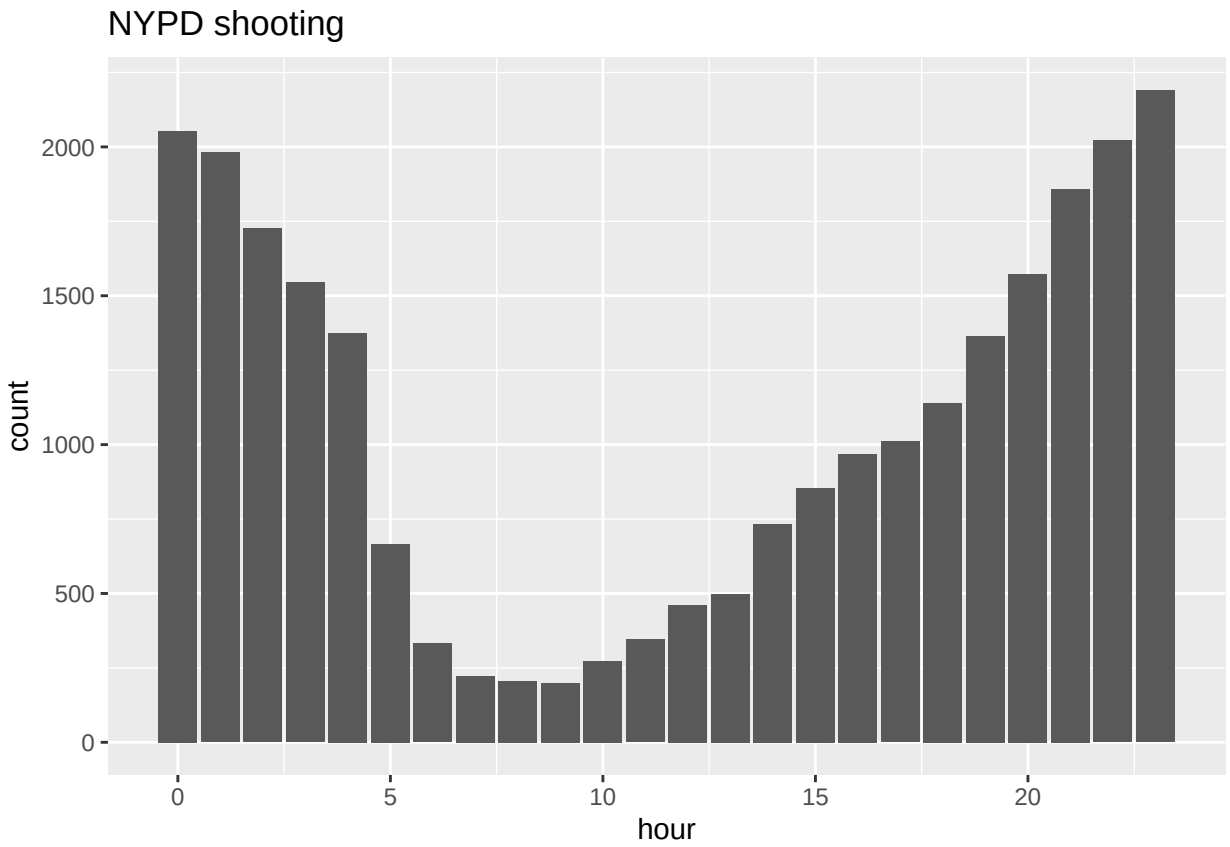
Here is a graphic representation of shooting counts by borough.

```
data_by_BORO %>%
  ggplot(aes(x = BORO, y = cases)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0)) +
  labs(title = "NYPD shooting", y = "count")
```



We also visualize the shooting incidents counts by hour.

```
data_by_hour %>%
  ggplot(aes(x = hour, y = hours)) +
  geom_bar(stat = "identity") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0)) +
  labs(title = "NYPD shooting", y = "count")
```



Modeling

We model the number of criminal count by hour. We use a polynomial regression model.

```
mod <- lm(hours ~ hour + I(hour^2) + I(hour^3) + I(hour^4), data = data_by_hour)

summary(mod)
```

```
##
## Call:
## lm(formula = hours ~ hour + I(hour^2) + I(hour^3) + I(hour^4),
##     data = data_by_hour)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -276.03 -117.85   16.07   90.75  345.23
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2283.24145  139.52761  16.364 1.18e-12 ***
## hour        -365.90498   87.90206  -4.163 0.000528 ***
## I(hour^2)      9.52703   16.03870   0.594 0.559517
## I(hour^3)      1.01719    1.05863   0.961 0.348697
## I(hour^4)     -0.03276    0.02282  -1.436 0.167374
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 172.1 on 19 degrees of freedom
## Multiple R-squared:  0.9471, Adjusted R-squared:  0.936
## F-statistic: 85.05 on 4 and 19 DF,  p-value: 7.451e-12
```

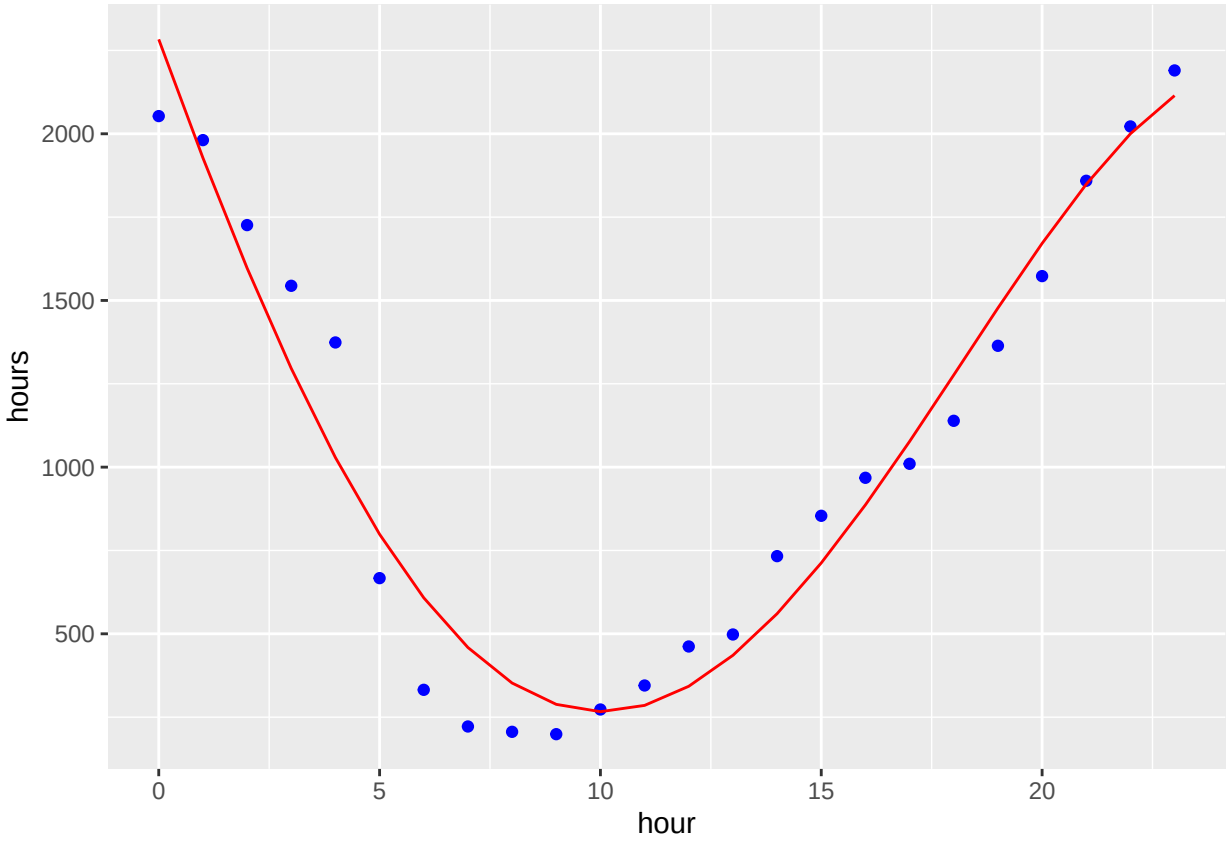
```
data_by_hour %>% slice_min(hours)
```

```
## # A tibble: 1 x 2
##    hour hours
##   <int> <int>
## 1     9   199
```

```
data_by_hour %>% slice_max(hours)
```

```
## # A tibble: 1 x 2
##    hour hours
##   <int> <int>
## 1    23  2190
```

```
x_grid <- seq(0, 23)
new_df <- tibble(hours = x_grid)
data_by_hour_pred <- data_by_hour %>% mutate(pred = predict(mod))
data_by_hour_pred %>% ggplot() +
  geom_point(aes(x = hour, y = hours), color = "blue") +
  geom_line(aes(x = hour, y = pred), color = "red")
```



Conclusions

Brooklyn has the highest number of gun crimes in New York City by borough. Brooklyn has the most population, so next we need to look at the number of crimes per capita.

In addition, an analysis of the time period in which crimes occur shows that the number of crimes is low in the morning, with a minimum at 9:00 a.m., increasing in the evening and peaking at 11:00 p.m.

Regarding bias, it is necessary to investigate the effects of races, residents income, educational level, etc. with objective data and conduct an analysis that eliminates the bias.

```
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22621)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=Japanese_Japan.utf8  LC_CTYPE=Japanese_Japan.utf8
## [3] LC_MONETARY=Japanese_Japan.utf8 LC_NUMERIC=C
## [5] LC_TIME=Japanese_Japan.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
##  [1] lubridate_1.8.0 forcats_0.5.2   stringr_1.4.1   dplyr_1.0.9
##  [5] purrr_0.3.4     readr_2.1.2     tidyr_1.2.0     tibble_3.1.8
##  [9] ggplot2_3.3.6   tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] assertthat_0.2.1    digest_0.6.29       utf8_1.2.2
##  [4] R6_2.5.1            cellranger_1.1.0    backports_1.4.1
##  [7] reprex_2.0.2        evaluate_0.16       highr_0.9
## [10] httr_1.4.4          pillar_1.8.1        rlang_1.0.4
## [13] googlesheets4_1.0.1 curl_4.3.2          readxl_1.4.1
## [16] rstudioapi_0.14     rmarkdown_2.16      labeling_0.4.2
## [19] googledrive_2.0.0   bit_4.0.4           munsell_0.5.0
## [22] broom_1.0.1         compiler_4.2.1      modelr_0.1.9
## [25] xfun_0.32           pkgconfig_2.0.3     htmltools_0.5.3
## [28] tidyselect_1.1.2    fansi_1.0.3         crayon_1.5.1
## [31] tzdb_0.3.0          dbplyr_2.2.1        withr_2.5.0
## [34] grid_4.2.1          jsonlite_1.8.0      gtable_0.3.0
## [37] lifecycle_1.0.1     DBI_1.1.3           magrittr_2.0.3
## [40] scales_1.2.1        cli_3.3.0           stringi_1.7.8
## [43] vroom_1.5.7         farver_2.1.1        fs_1.5.2
## [46] xml2_1.3.3          ellipsis_0.3.2      generics_0.1.3
## [49] vctrs_0.4.1         tools_4.2.1         bit64_4.0.5
## [52] glue_1.6.2          hms_1.1.2           parallel_4.2.1
## [55] fastmap_1.1.0       yaml_2.3.5          colorspace_2.0-3
## [58] gargle_1.2.0        rvest_1.0.3         knitr_1.40
## [61] haven_2.5.1
```