ShootingIncident

Student Name (Removed for assessment)

2024-01-06

Step 1: Import data

This bellow code import data from https://catalog.data.gov/dataset

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

data = read_csv(url, show_col_types = FALSE)
```

Step 2: Tidy and Transform Data

Print a summary of the data

```
summary(data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class  :character  Class1:hms        Class  :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class  :character  1st Qu.: 44.00   1st Qu.:0.0000    Class  :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA’s   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class  :character  FALSE:22046             Class  :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA’s   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA’s   :10
```

Select interested features

Select interested features only

```
data <- data %>%
  select(c(OCCUR_DATE, OCCUR_TIME, BORO, LOCATION_DESC, STATISTICAL_MURDER_FLAG,
           PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE,
           Latitude, Longitude))
```

Transform data

Convert date and time to date types

```
datat <- data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

There are missing data in some columns, such as PERP_AGE_GROUP or PERP_SEX, PERP_RACE.

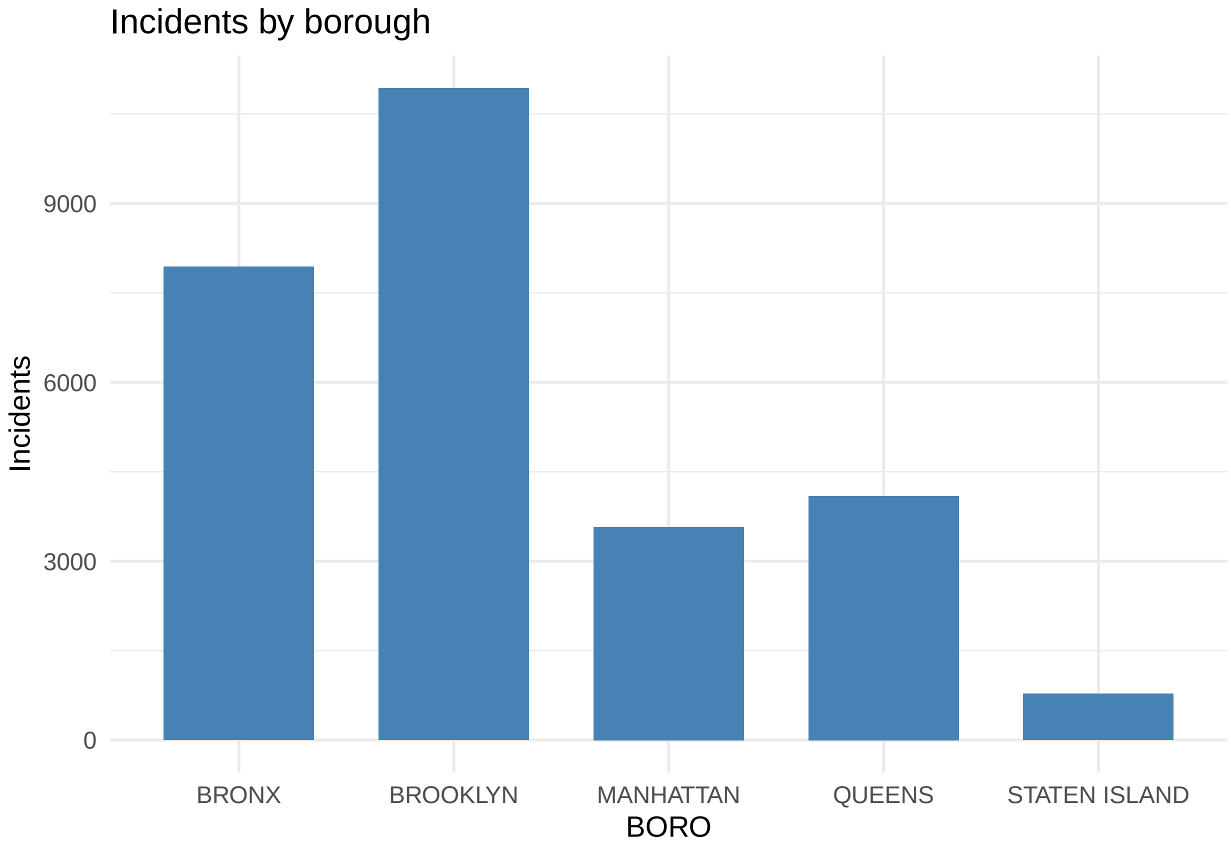
There are some way to handle it:

- Replace NA with a median of the total value (e.g. age median for PERP_AGE_GROUP)
- Adding a new type for NA value, such as “UNKNOWN” for missing value of PERP_SEX

Step 3: Add Visualizations and Analysis

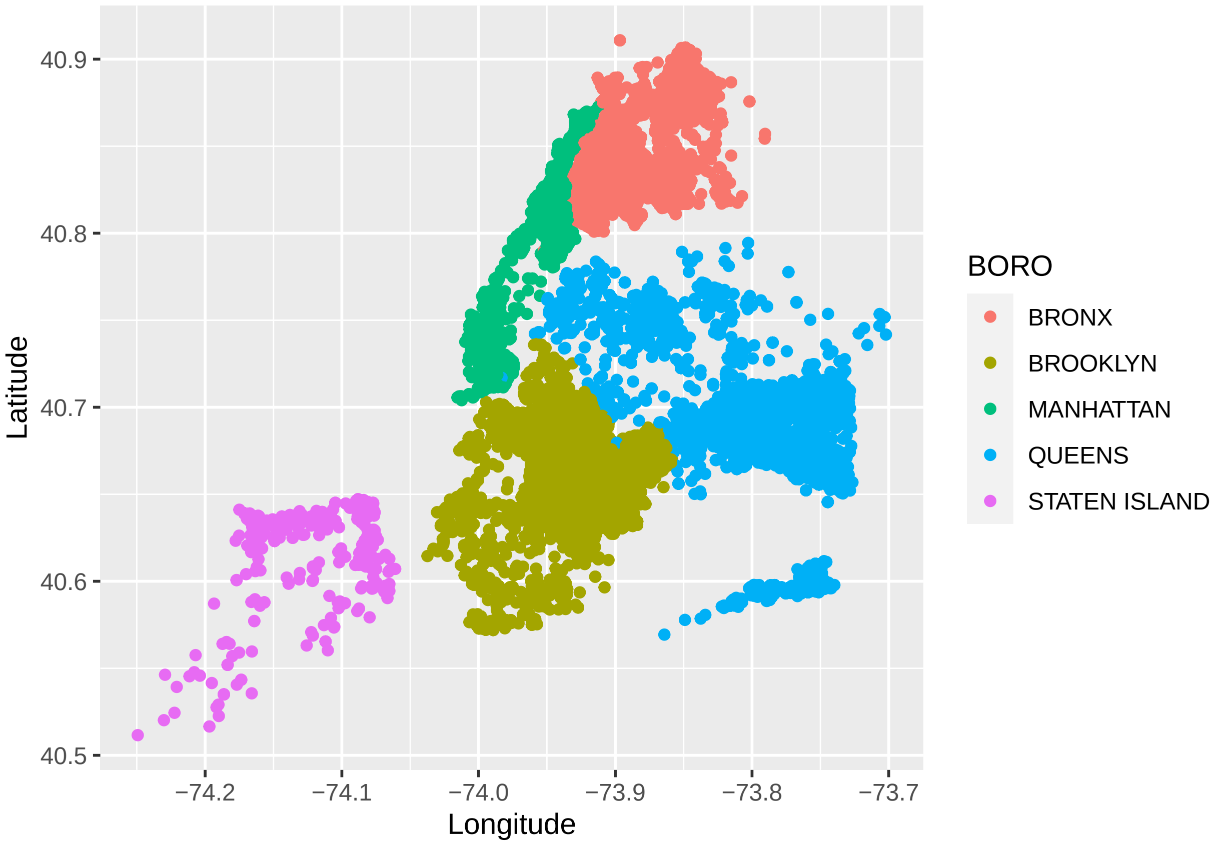
Showing number of incidents by borough

```r
data %>%
  ggplot(aes(x=BORO))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  labs(title = "Incidents by borough", y = "Incidents") +
  theme_minimal()
```



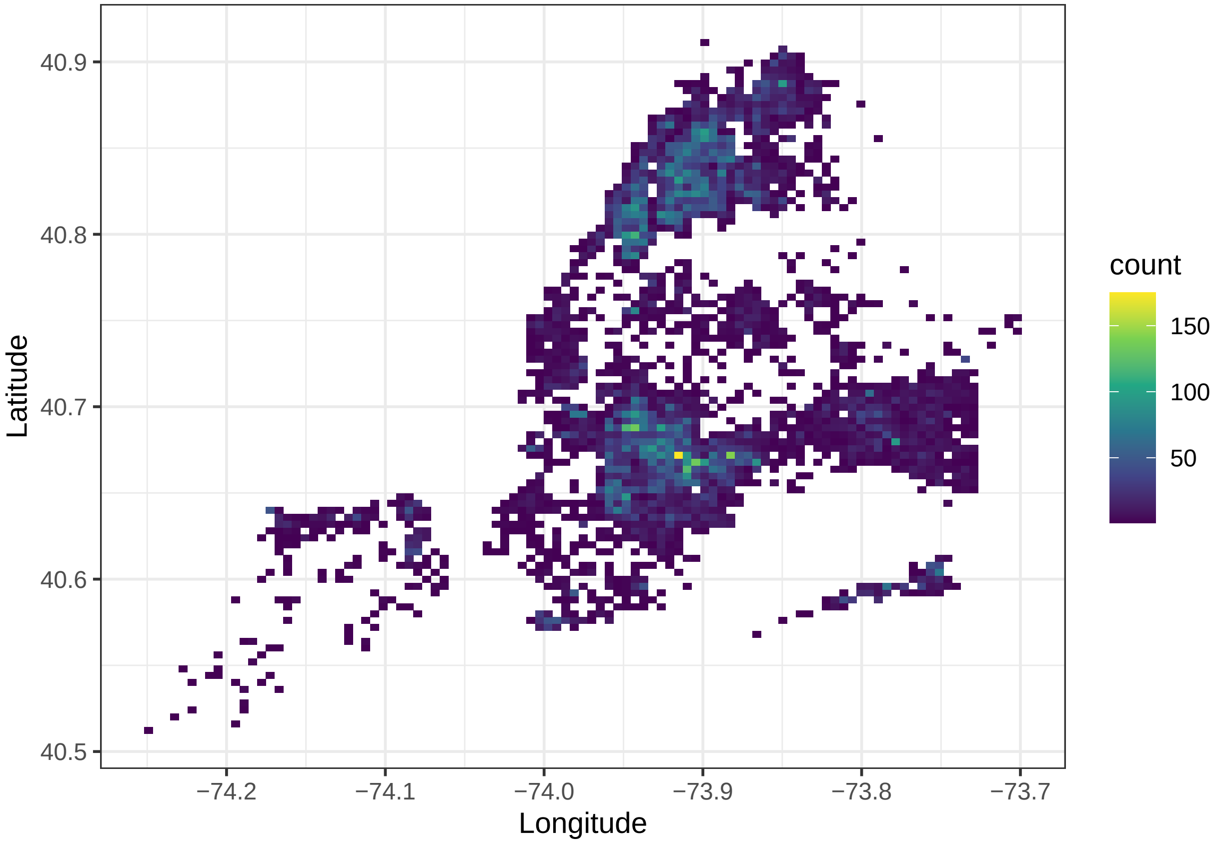
Since there's lat/long data, let's plot it in 2D map by borou to see the spacial distribution

```r
data %>%
  ggplot(aes(x=Longitude, y=Latitude)) +
  geom_point(aes(color=BORO))
```



Plot the data with density

```
data %>%
  ggplot(aes(x=Longitude, y=Latitude)) +
  geom_bin2d(bins = 100) +
  scale_fill_continuous(type = "viridis") +
  theme_bw()
```



Observation:

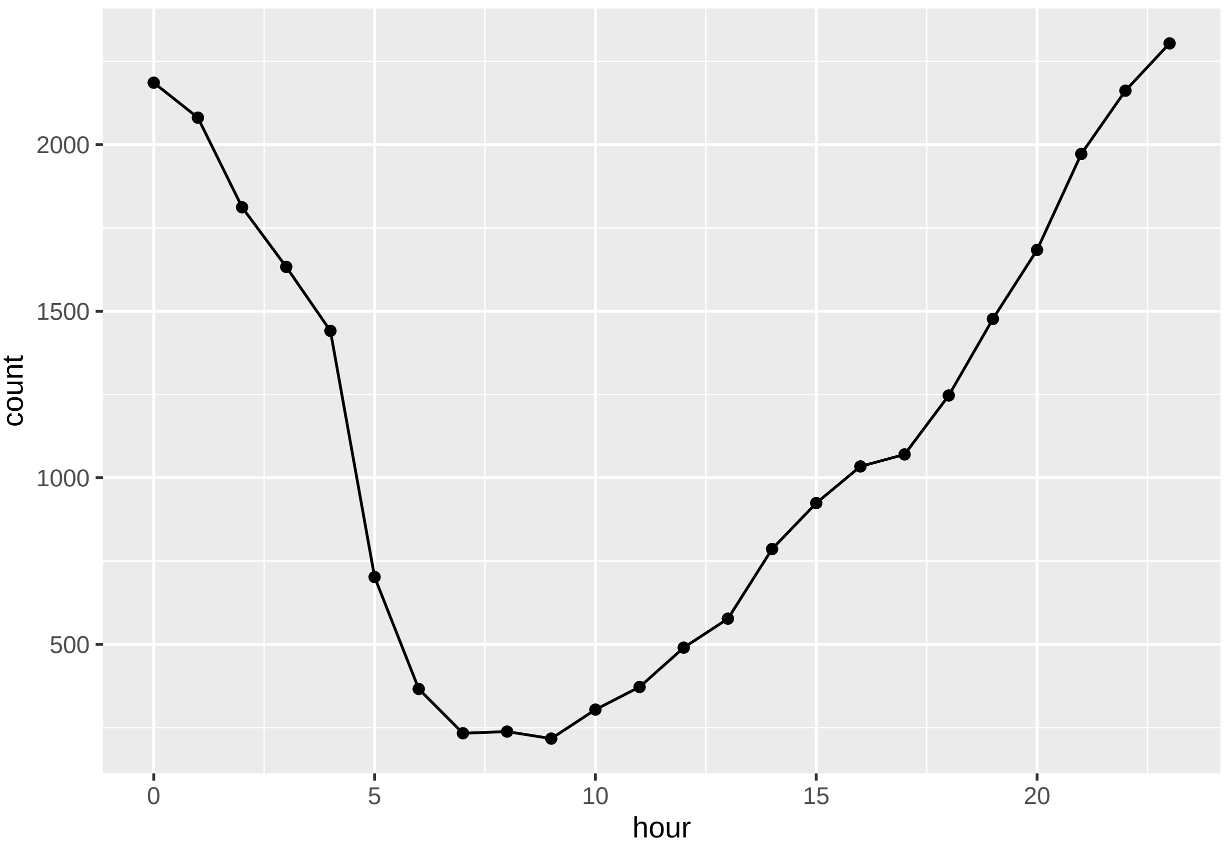
- There are high number of incidents in center of BROOKLYN, and between MANHATAN & BRONX

Create a new variable for hours

```
data <- data %>%
  mutate(hour = hour(OCCUR_TIME))
```

Plot the incident by hours

```
data %>%
  ggplot(aes(x=hour))+
  geom_line(stat="count") +
  geom_point(stat="count")
```

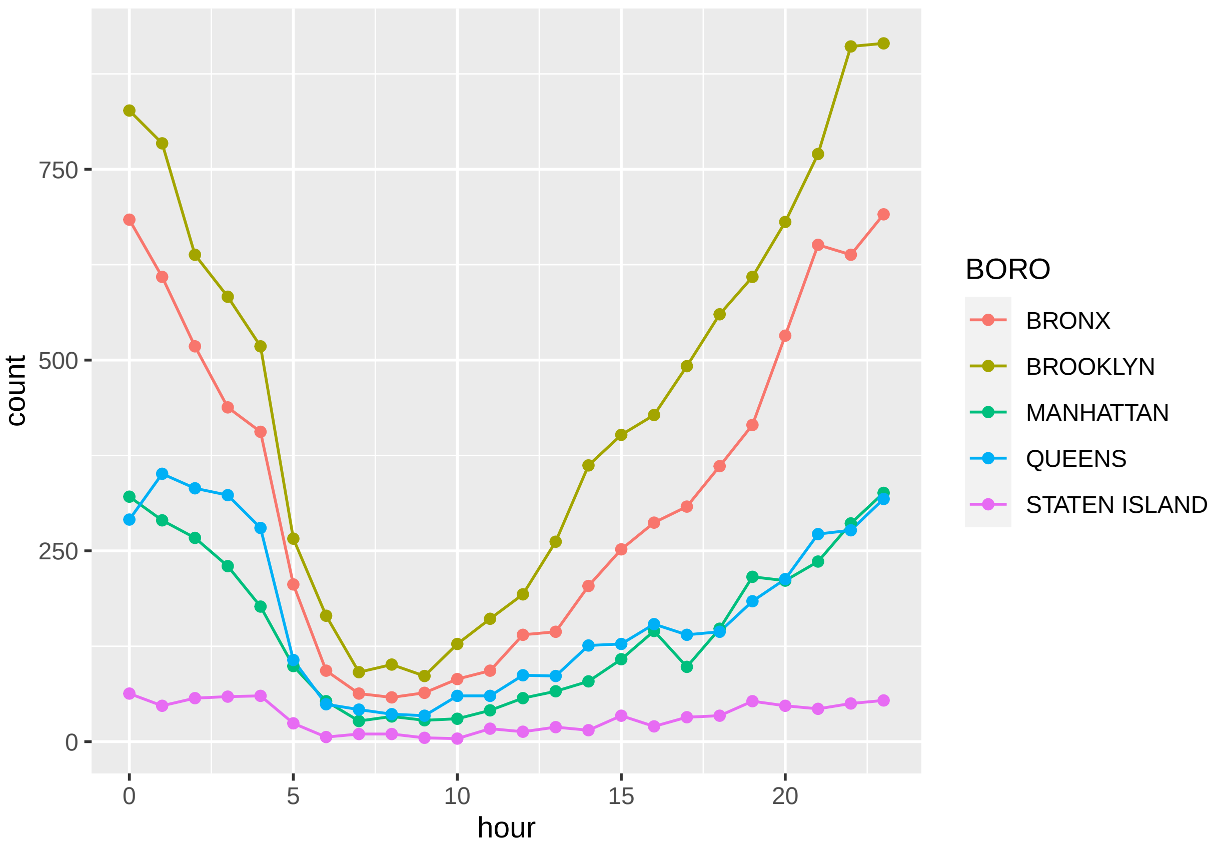


Observation:

- The number of incident increase significantly on evening and mid-night

Plot the incidents in hours, counting by borough

```
data %>%
  ggplot(aes(x=hour, col=BORO))+
  geom_line(stat="count") +
  geom_point(stat="count")
```



Modeling data

```
data_totals_by_hour <- data %>%
  count(hour)

summary(data_totals_by_hour)
```

```
##       hour             n         
##  Min.   : 0.00   Min.   : 217.0  
##  1st Qu.: 5.75   1st Qu.: 460.5  
##  Median :11.50   Median :1052.0  
##  Mean   :11.50   Mean   :1138.0  
##  3rd Qu.:17.25   3rd Qu.:1716.0  
##  Max.   :23.00   Max.   :2304.0  
```

From the above visualization, let try a quadratic model between the number of incident and hour.

Firstly, create a new variable hour2:

```
data_totals_by_hour <- data_totals_by_hour %>%
  mutate(hour2=hour^2)
```

Then create a model

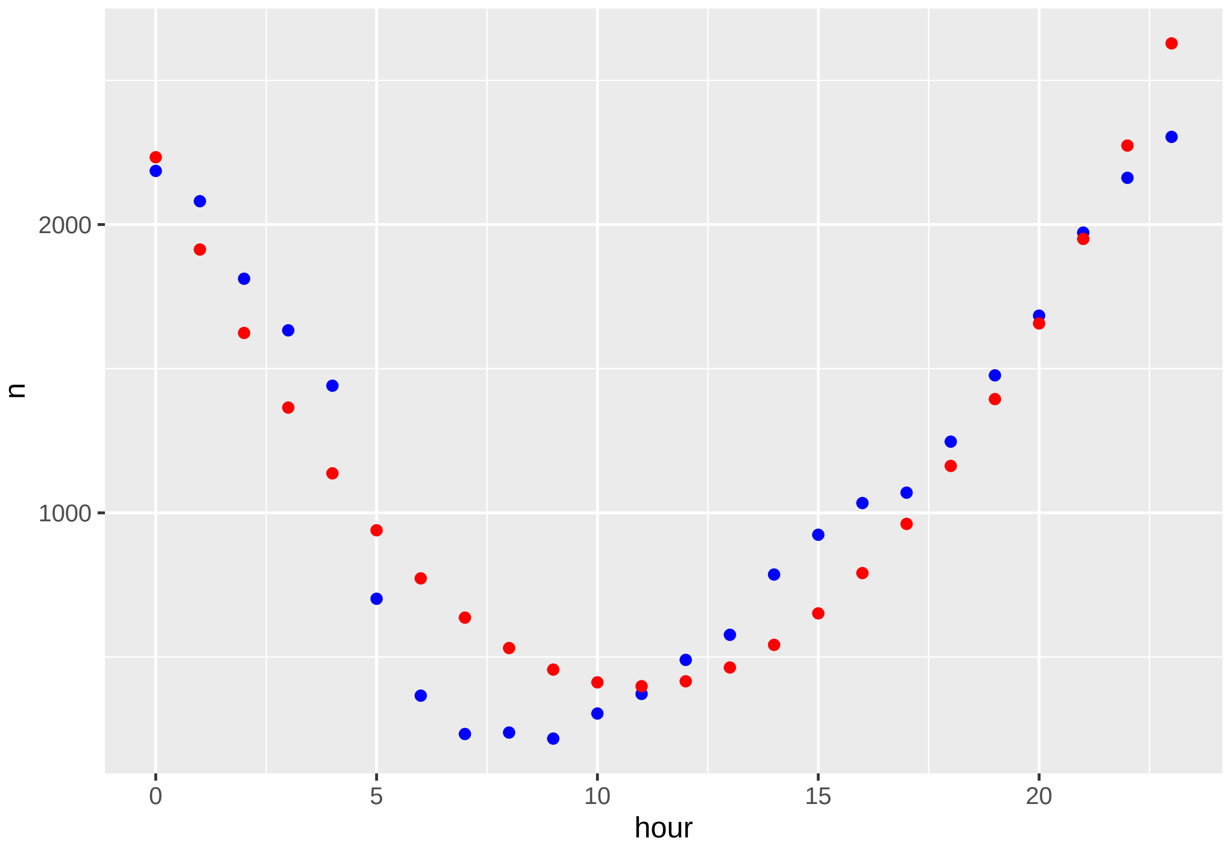
```r
quadraticModel <- lm(n ~ hour + hour2, data=data_totals_by_hour)
summary(quadraticModel)
```

```
##
## Call:
## lm(formula = n ~ hour + hour2, data = data_totals_by_hour)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -406.73 -143.32   50.61  172.71  303.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2233.526    130.753   17.08 8.56e-14 ***
## hour        -335.455     26.333  -12.74 2.40e-11 ***
## hour2         15.331      1.106   13.87 4.86e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 231.6 on 21 degrees of freedom
## Multiple R-squared:  0.9044, Adjusted R-squared:  0.8952
## F-statistic: 99.28 on 2 and 21 DF,  p-value: 1.981e-11
```

Let plot the model prediction

```r
data_totals_by_hour_pred <- data_totals_by_hour %>%
  mutate(pred = predict(quadraticModel))

data_totals_by_hour_pred %>%
  ggplot() +
  geom_point(aes(x = hour, y = n), color = "blue") +
  geom_point(aes(x = hour, y = pred), color = "red")
```



Conclusion

- There is a relationship between the time of the day (hour), and the chance that an shooting incident happens.
- The relation ship can be represented by a quadratic model between the hour of the day and the number of the incidents

Bias:

- People tend to think day light is safer than evening or night
- Personally, I think dense area with high population might likely to have more incidents. The future improvement could be include the population of the areas into the data set.
- I didn’t check gender or race into the report. One way to improve is to consider theses factor as well.

Session info

```
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31)
## Platform: x86_64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.1
##
```

```
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRblas.0.dylib 
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRlapack.dylib;  LAPACK
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## time zone: America/Los_Angeles
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4    
##  [5] purrr_1.0.2     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1   
##  [9] ggplot2_3.4.4   tidyverse_2.0.0
## 
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4        generics_0.1.3    stringi_1.8.3     hms_1.1.3        
##  [5] digest_0.6.33     magrittr_2.0.3    evaluate_0.23     grid_4.3.2       
##  [9] timechange_0.2.0  fastmap_1.1.1     fansi_1.0.6       viridisLite_0.4.2
## [13] scales_1.3.0      cli_3.6.2         rlang_1.1.2       crayon_1.5.2     
## [17] bit64_4.0.5       munsell_0.5.0     withr_2.5.2       yaml_2.3.8       
## [21] tools_4.3.2       parallel_4.3.2    tzdb_0.4.0        colorspace_2.1-0 
## [25] curl_5.2.0        vctrs_0.6.5       R6_2.5.1          lifecycle_1.0.4  
## [29] bit_4.0.5         vroom_1.6.5       pkgconfig_2.0.3   pillar_1.9.0     
## [33] gtable_0.3.4      glue_1.6.2        xfun_0.41         tidyselect_1.2.0 
## [37] highr_0.10        rstudioapi_0.15.0 knitr_1.45        farver_2.1.1     
## [41] htmltools_0.5.7   rmarkdown_2.25    labeling_0.4.3    compiler_4.3.2
```