Week 3 Project

Bradley Cumming

2022-04-10

Week 3 Project

This is an R Markdown document for the Week 3 Project Assignment in the course "Data Science as a Field". This document is a reproducible work flow to download, tidy, summarize, visualize, and analyze Historical NYPD Shooting Incident data.

Import Libraries

First import the following libraries to gain access to relevant functions:

```
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- 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
```

Import Data

To import the data, use the link to obtain the csv file and read the data into R using the "read.csv()" function.

```
url_in <- 'https://data.cityofnewyork.us/api/views/5ucz-vwe8/rows.csv?accessType=DOWNLOAD'

data <- read.csv(url_in)
```

Tidy Data

The objective is to discover the date and time statistics that the shooting incidents occurred.

To tidy the data:

1. The "OCCUR_DATE" variable column is not a time object so a new column must be mutated in and calle

2. I will group the incidents by BORO.

3. In order to obtain meaningful summary statistics, the "OCCUR_TIME" column must be converted from a

4.The irrelevant columns must be removed and the data must be sorted (arranged) by a column that

```
data <- data %>%
    mutate(DATE = mdy(OCCUR_DATE)) %>%
    mutate(TIME = str_remove(OCCUR_TIME, "1970-01-01 ")) %>%
    mutate(TIME = sapply(strsplit(TIME,":"),
              function(x) {x <- as.numeric(x)
                           x[1]+x[2]/60}))%>%
    group_by(BORO) %>%
    select(-c(INCIDENT_KEY, PRECINCT, OCCUR_TIME, OCCUR_DATE, JURISDICTION_CODE, LOCATION_DESC,
            PERP_SEX, PERP_RACE, VIC_SEX, VIC_RACE, PERP_AGE_GROUP, VIC_AGE_GROUP,
            X_COORD_CD, Y_COORD_CD, New.Georeferenced.Column, Latitude, Longitude,
    arrange(DATE)%>%
    select(DATE, everything())
```

Summary of Data

The summary indicates that 2011 total shooting incidents occurred within the New York City jurisdiction. The summary statistics for the TIME variable indicates that an equal amount of incidents occurred between the time interval of 12:00 AM to 4:20 PM and time interval of 4:20 PM to 11:59 PM. In other words, the number of shooting incididents is more concentrated at night than in the morning, which provides evidence that the night time is more dangerous than the morning.

```
summary(data)
```

```
##       DATE                BORO           LOC_OF_OCCUR_DESC  LOC_CLASSFCTN_DESC
##  Min.   :2022-01-01   Length:844         Length:844         Length:844
##  1st Qu.:2022-02-28   Class :character   Class :character   Class :character
##  Median :2022-04-12   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2022-04-08
##  3rd Qu.:2022-05-21
##  Max.   :2022-06-30
##       TIME
##  Min.   : 0.01667
##  1st Qu.: 4.79583
```

```
##  Median :15.70000
##  Mean   :13.55136
##  3rd Qu.:20.25000
##  Max.   :23.96667
```

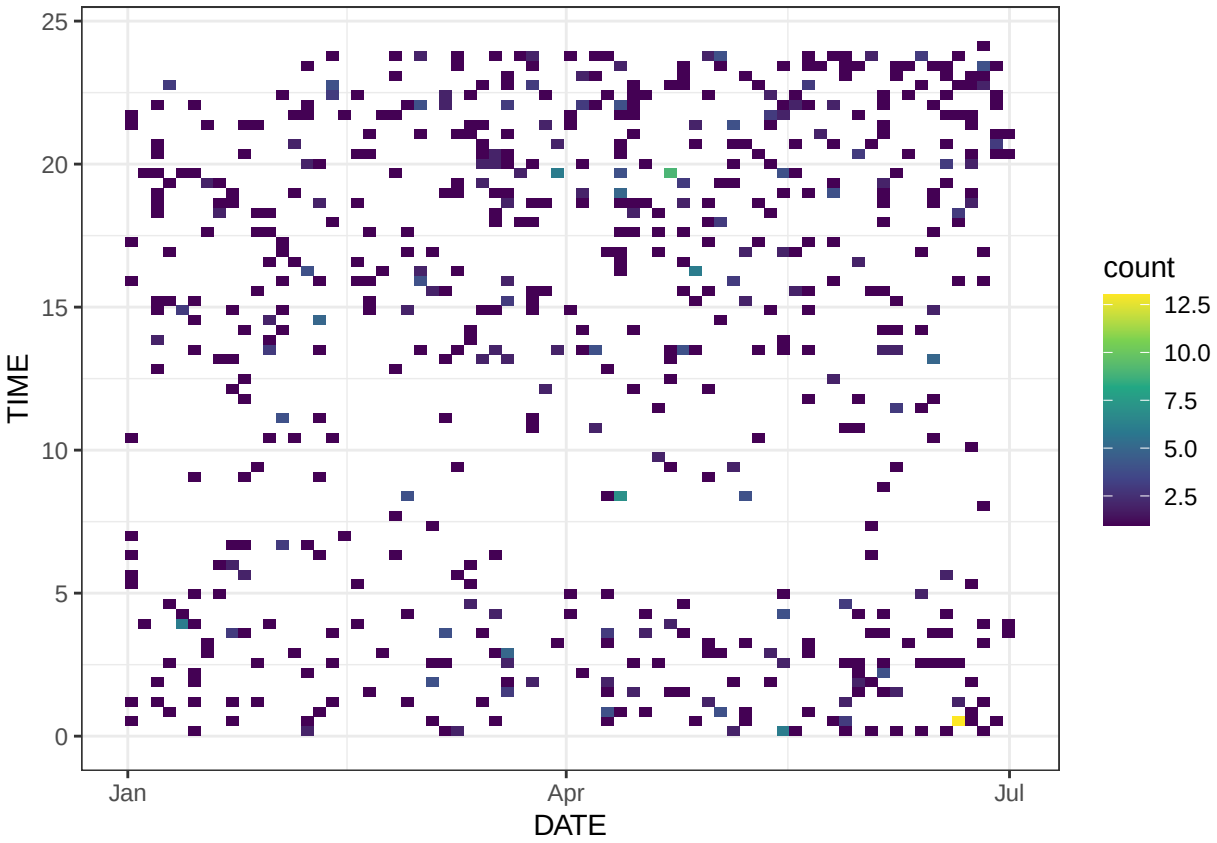
Visualization of Data

A possible benefit of data analysis for this kind of data is to enable better policing through better informed staffing and resource allocation decisions. Perhaps, the NYPD would like to know when they should dispatch more police officers for certain times of the day and year when extra backup may be necessary. To visualize the most dangerous (most incidents) and safest (least incidents) time periods throughout the year, the best tool is a variation of a scatterplot that graphs the incidents by date and time. The most dangerous dates and times will have a cluster of incidents, while the safest dates and times will have a void of incidents. The scatter-plot can be visualized with (1) color gradients and (2) contour maps.

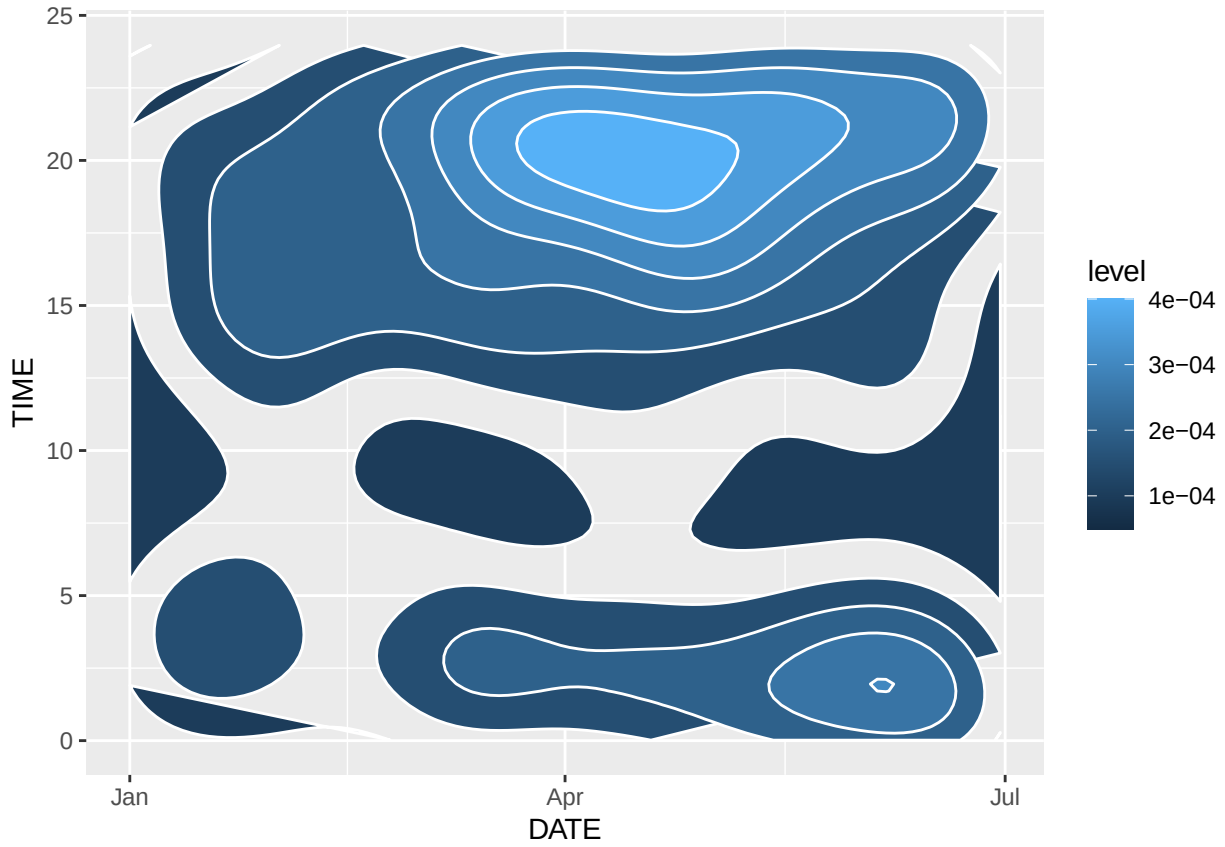
```
data_visualization <- data %>%
      ggplot(aes(x=DATE, y=TIME) ) +
      geom_bin2d(bins = 70) +
      scale_fill_continuous(type = "viridis") +
      theme_bw()


data_visualization2 <- data %>%
      ggplot(aes(x=DATE, y=TIME) ) +
      stat_density_2d(aes(fill = ..level..), geom = "polygon", colour="white")

plot(data_visualization)
```




```
plot(data_visualization2)
```



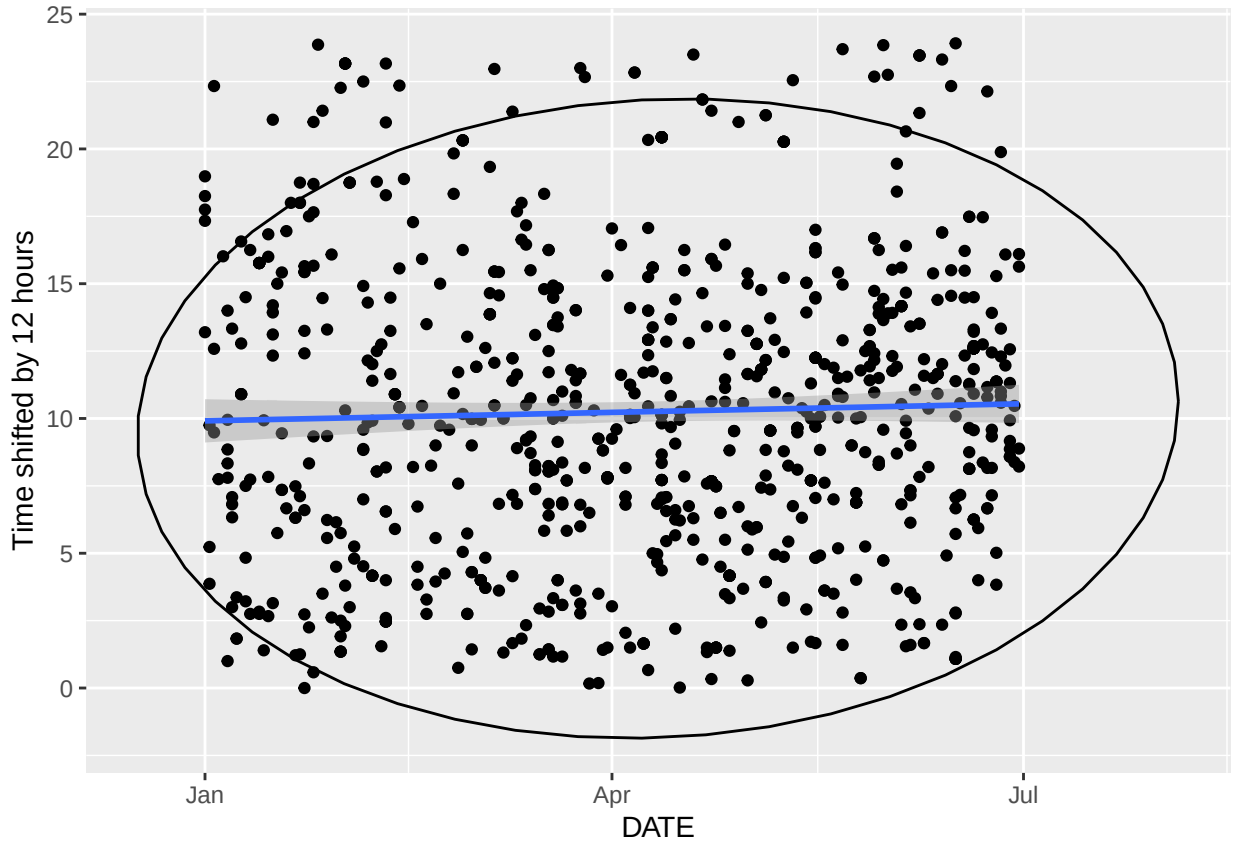
Model Data

To model the data, the TIME column will need to be adjusted by 12 hours to account for most incidences occurring during the night which will represent the data in a truer form. A linear model will show which hours are most likely to have shooting incidents occur.

```
data <- data %>%
    mutate(ADJUSTEDTIME = sapply(TIME, function(x) ifelse(x >= 12, x-12, x + 12)))

linear_model <- ggplot(data, aes(x=DATE, y=ADJUSTEDTIME)) +
                  geom_point()+
                  geom_smooth(method=lm)+
                  labs(y = "Time shifted by 12 hours") +
                  stat_ellipse()
linear_model
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Analysis

A visual analysis of the data suggests that extra policing power should be considered during the late night of the summer months as this is where the highest concentration of shooting incidents occur.

Color Gradient Scatter Plot

The color gradient scatter plot indicates that between the hours of 8pm and 3am have the highest number of incidents while the hours between 7am and 11am have relatively few incidents. The late hours of June 2021 were particularly active with shooting incidents, and a date in November 2021 at approximately 5pm was the most active time of the year.

Contour Map and Linear Model

The contour map indicates that mid-morning of late August/early September is the safest time of the year in New York City in regards to shooting incidents while mid-June to mid-July around approximately 10pm to 3am is the most dangerous time of the year.

Conclusion

According to the analysis, the NYPD should dedicate extra resources and attention to their police force for the summer months, especially during the night. The NYPD should consider allocating more of their budget and manpower to these summer months and less in the winter months. However, a potential bias in the

analysis is that a presupposition is made thats says preventing and preparing for shooting incidents is the highest priority for the NYPD. Other data sources (i.e. traffic accidents, burglaries, stabbings, other violent crimes, etc) may indicate that more priority should be considered elsewhere. For example, if the number of traffic accidents and violations increases during the winter, then an argument could be made that the NYPD should not necessarily reduce the police force but rather they should simply equip police officer to better police traffic instead of policing violence.