DC area criminal incident analysis

Alex, Bo, Jingyi

2020/4/26

1. Introduction

Our project is criminal cases analyze in the DC area, in the project we hope we can explore the relationship between criminal case numbers and other possible predictor variables. By this topic, we hope to raise people's awareness of public safety. Using all the data available from the Metropolitan Police Department, we can begin to understand how crime is evolving and understand whether local law enforcement is prioritizing the effective resources to address it.

In this project, the main focus is exploring the question of what is the trend in the number of crimes changes from 2018 to 2019 in the DC area. Then we propose four hypotheses as the research framework. In detail, the hypotheses are (1) Some types of crime's number increased in 2019 than in 2018. (2) Due to Christmas Day and New Year, we think the total number of crimes may occur the most at the end of the year until the beginning of the next year(from November to January of the following year). (3) The month can affect the number of crimes. (4) Compared with the evening and day, there are more crimes at midnight.

Our team members are Yueyang Liu, Bo Wang, Jingyi Ge. Each of us did different jobs and completed the project together. Yueyang is s responsible for collecting and tidying original data and producing the shiny app. Bo and Jingyi are responsible for report analysis. And study hypotheses together by all team members.

2. About this data

This data is published by the Metropolitan Police Department, which includes information on criminal cases that happened in the DC area. Data are presented for 2018 to Feb 2020. There are 29 variables in the data set.

```
##  [1] "NEIGHBORHOOD_CLUSTER" "CENSUS_TRACT"         "offensegroup"
##  [4] "LONGITUDE"            "END_DATE"             "offense.text"
##  [7] "SHIFT"                "YBLOCK"               "DISTRICT"
## [10] "WARD"                 "YEAR"                 "offensekey"
## [13] "BID"                  "sector"               "PSA"
## [16] "ucr.rank"             "BLOCK_GROUP"          "VOTING_PRECINCT"
## [19] "XBLOCK"               "BLOCK"                "START_DATE"
## [22] "CCN"                  "OFFENSE"              "OCTO_RECORD_ID"
## [25] "ANC"                  "REPORT_DAT"           "METHOD"
## [28] "location"             "LATITUDE"
```

Here is our data link which is convenient for everyone to download and use. https://drive.google.com/file/d/1gzomTXvi0qrqx7_jUOz7uurleJ_2KIZ0/view?usp=sharing

3. Tidy data

First of all, we get the original data from the official website. Then we use 'lubridate' and 'tidyverse' packages to tidy our data. We also noticed that our data contains missing values, and therefore we dropped those "NA" rows so that we can make a clearer analysis. We generate data and time columns so that we can analyze the relationship between crime numbers and time/date. And to the crime time and location analysis, we use the Leaflet map.
The variables we used in the project are list below.

```
##  [1] "offense.text" "SHIFT"        "DISTRICT"     "YEAR"         "sector"
##  [6] "REPORT_DAT"   "Month"        "Year"         "Day"          "Hour"
## [11] "Time"
```

We also designed the shiny app to show our data, the link is listed below: https://alex-nkg.shinyapps.io/DC_Crime/

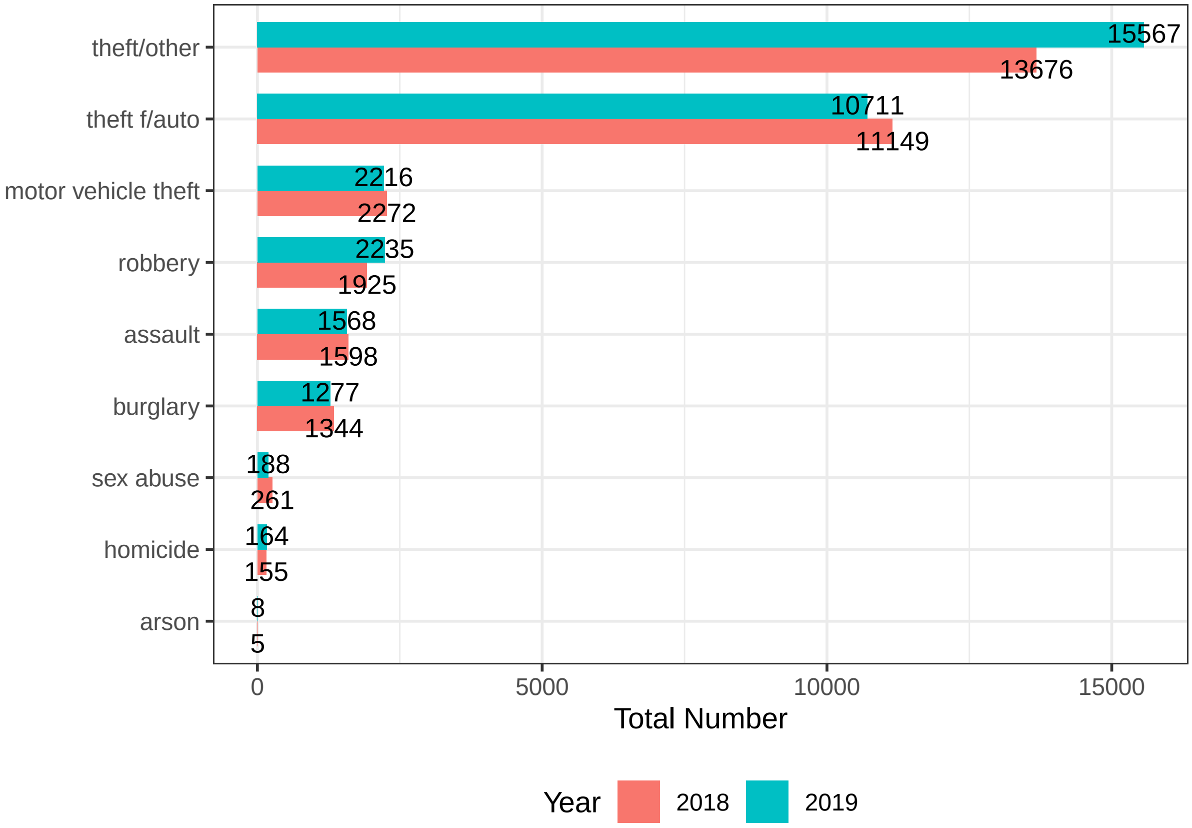
4. Analysis, visualization models

4.1 Crime trends from 2018 to 2019.

Hypothesis: Some types of crime number increased in 2019 than in 2018

We would like to compare crime data from 2018 to 2019. We grouped the data frame by offense type and year. Then use a ggplot to draw the bar plot respectively.

This plot shows the comparison of the crime number between 2018 and 2019 by type:

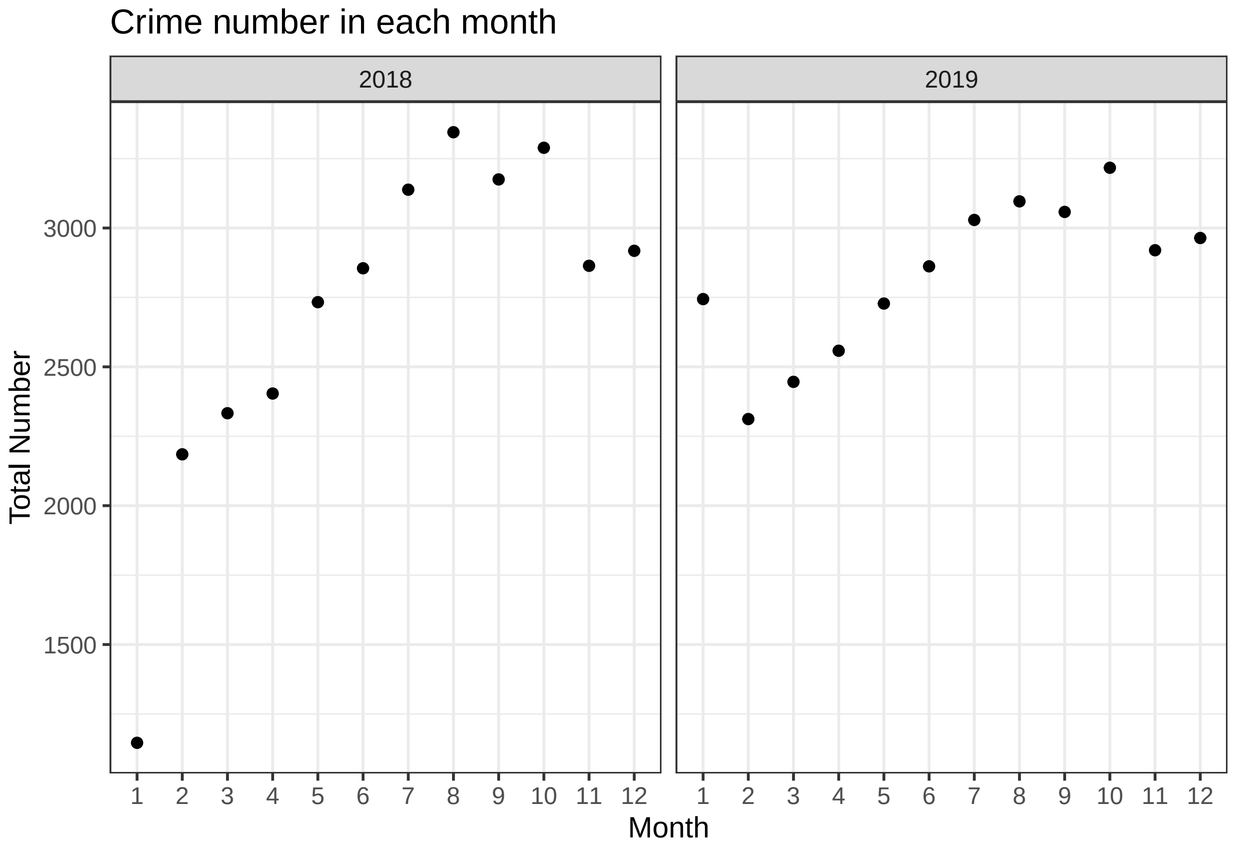


From the plot, we can confirm our first hypothesis as some specific type of crime cases increased slightly from 2018 to 2019, including theft and robbery. Meanwhile, auto theft and sex abuse cases decreased from 2018 to 2019, while other kinds of criminals remain steady in these 2 years. According to this estimation, we may say the public safety situation is improving.

4.2 The relation between crimes and the month

Hypothesis: due to Christmas Day and New Year, we think the total number of crimes may occur the most at the end of the year until the beginning of the next year(from November to January of the following year).

We grouped the data frame by year and month and drew a scatter plot to exam the relationship between crime numbers and months.



According to this plot, we reject hypothesis 2, the crime number gradually increased from Feb and reached its peak point in August. Crimes occur the most between August and October.

From the result of hypothesis 2, we did a deeper study since we would like to figure out the relationship between the month and the number of crimes.

Going deeper, we developed a regression model for these two variables.

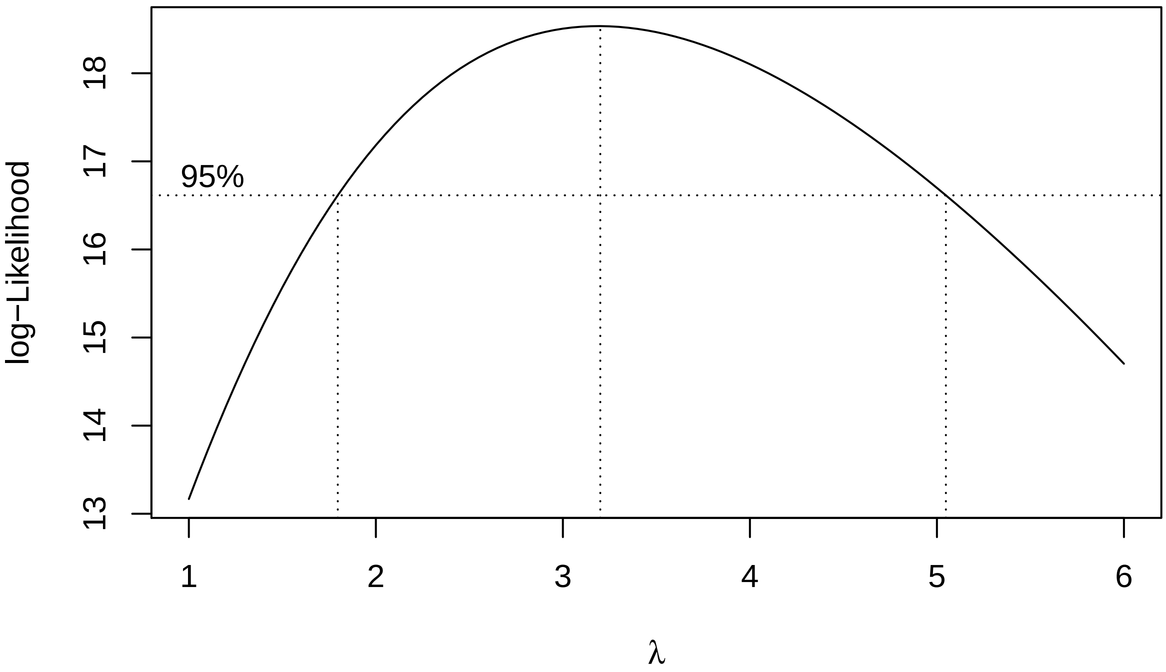
$Y_i = \beta_0 + \beta_1 X_1$
β_0 is the intercept on Y.
β_1 is the effect on Y(Total crime cases) for changes in X(month) given the other variables in the model.

```
## [1] "R^2:"

## [1] 0.5217264

## [1] "P value:"

## [1] 6.732482e-05
```

We see the p-value is 6.732e-05, but the R-squared is only about 0.5217 which does not fit the model very well, we still need to optimize this model. From the plot the Jan's data is far from other points, to figure it is an outlier or not we did the outlier test here.

```
##    rstudent unadjusted p-value Bonferroni p
## 1 -5.203306         3.7144e-05   0.00089145
```

According to the outlier test, we can confirm that Jan's data is an outlier data, but we still plan to keep this outlier since we don't know how it will affect the data. We see a caved shape between the month and crime number in the plot, indicate that a transformation on the responses y is necessary.
To find the best λ, we apply the Box-Cox transformation to the responses Y:

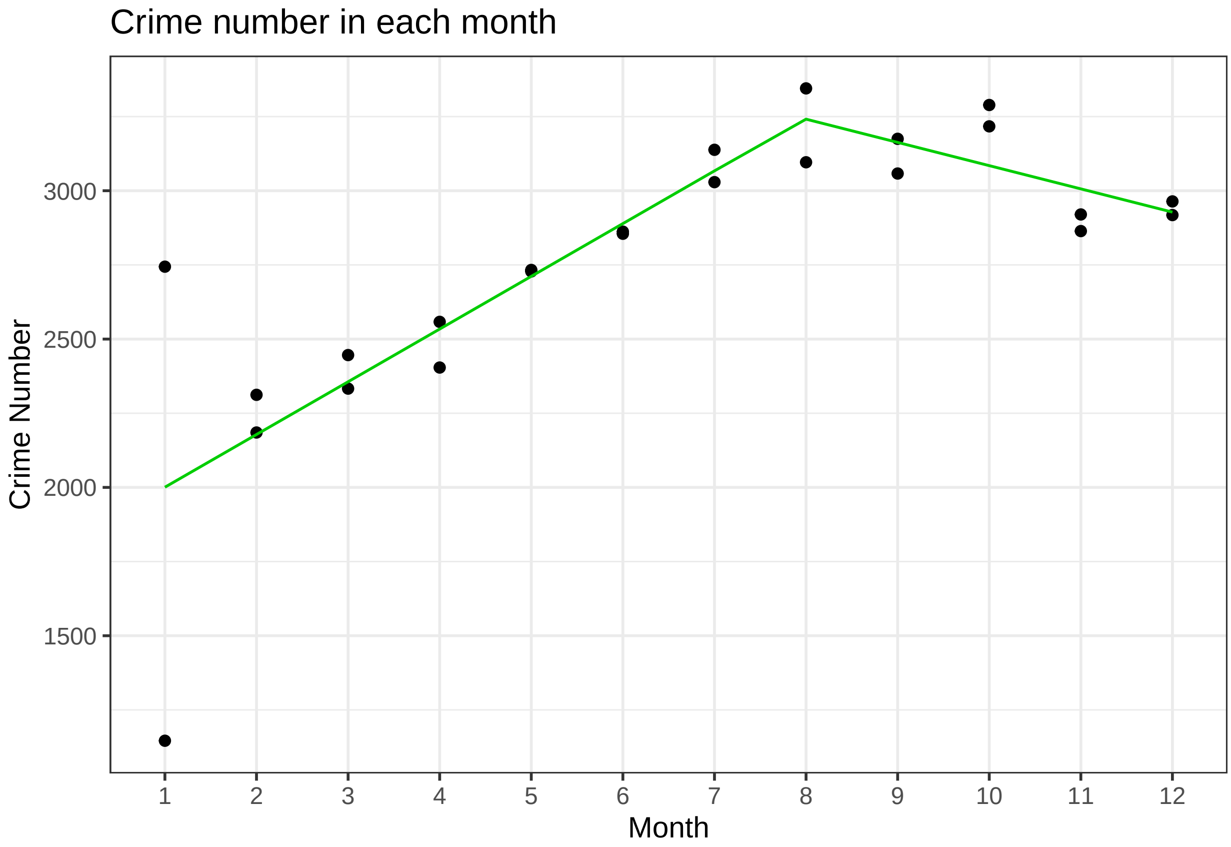


according to the plot, we set $\lambda = 3.2$

```
## [1] "R^2"

## [1] 0.5698974
```

The R-squared is 0.5699, which is better than the previous one. Now the model is:$Y^{3.1} = 2.258e^{10}$+$4.283e^{09}X$
We also tired a piecewise function to fit the model.

```
## [1] "R^2"

## [1] 0.7138857
```

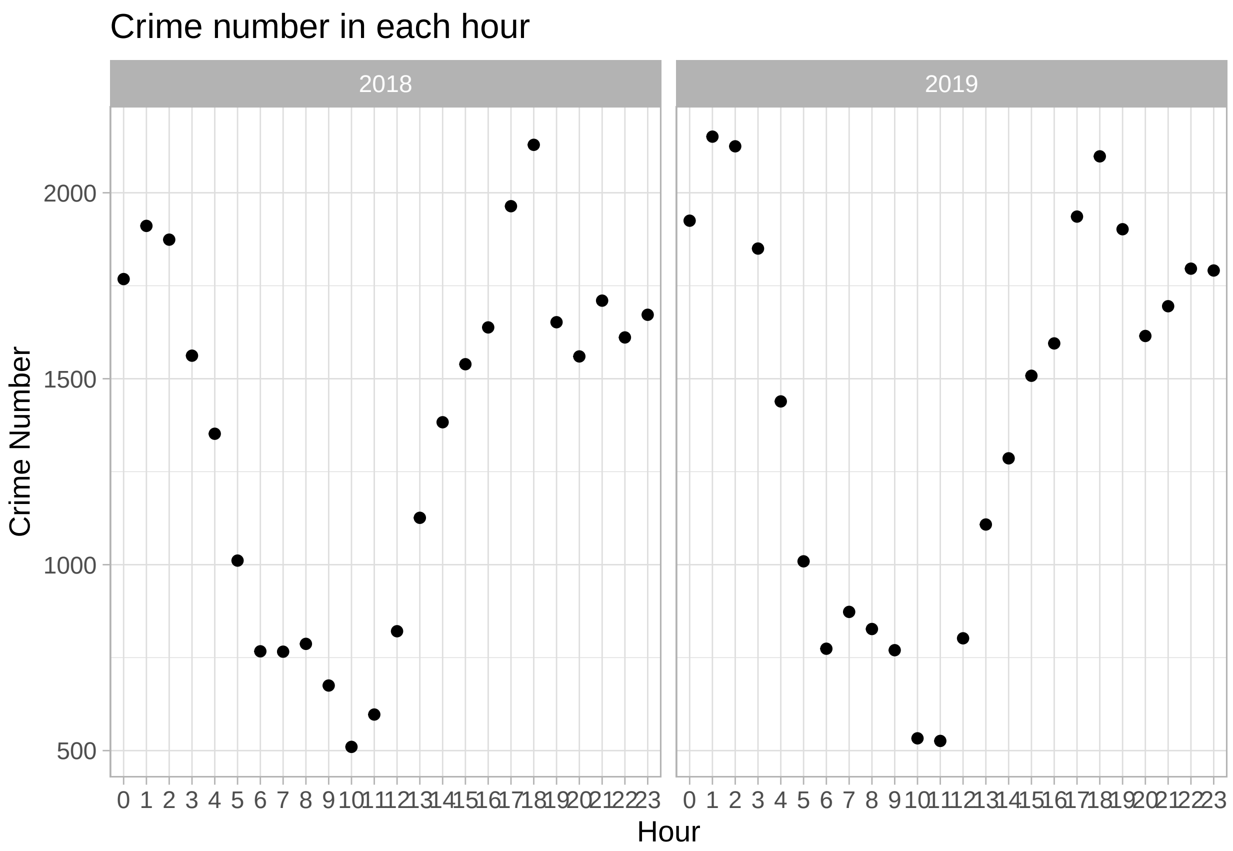


In the piecewise function model, the R-squared value is 0.7139, and the p-value is small, we can accept that there is a piecewise linear relationship between the month and the crime number.

4.3 Crime numbers in different periods of time

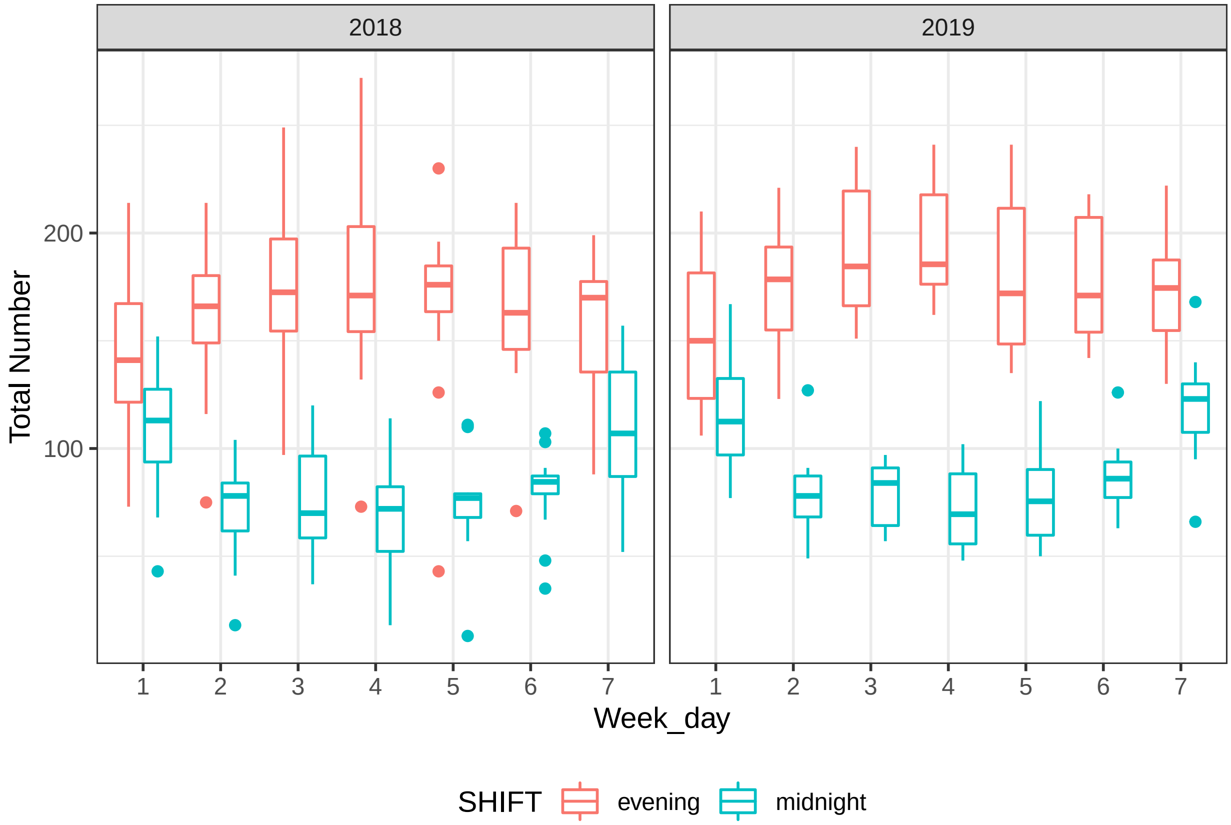
Hypothesis: Compared with the day and evening, there are more crimes at midnight.

First, we grouped the data frame by the year and hour, and calculated the total number of the crimes, then use a scatter plot to show their distribution.



According to the point plot, we find that: compared with day, the crimes are more active at night. Thus, Later we will plot our data with only night cases.

In the original data frame, the `SHIFT` variable was set as day(11am - 7pm), evening(8pm - 3am), midnight(4am – 10am). Then we use `filter` function to get rid of the day time value and use boxplot to compare the change in the total number of crimes at night and midnight every weekday.



According to the plot we find there are more crimes in the evening than at midnight. Thus, we will reject our hypothesis 4.
Residents can be reminded that there are more thieves during that time, and law enforcement agencies need to maintain sufficient workforces at night to cope with the criminal acts.

Conclusion

After further study, we can draw a conclusion of the what tendency of the crime change from 2018 to 2019 in the DC area is.
(1)The first hypothesis is some types of criminal cases increased in 2019 than in 2018, and we have proven it is true. The theft and robbery increased slightly from 2018 to 2019.
(2) When mentioning the month and the crimes. Our second hypothesis is due to Christmas Day and New Year, we think the total number of crimes may occur the most at the end of the year until the beginning of the next year(from November to January of the following year). However, we have to reject the hypothesis. The period of the crimes occurs the most is from August to October. After analysis, we think the month can affect the number of crimes. We take January as outliers, then use a piecewise function to fit the linear model. Finally, we accept the hypothesis that there is a piecewise linear relationship between the month and the crime number.
(3) In common idea, criminal cases likely happen during the period from midnight time to dawn, when people are usually sleepy and tired. Our hypothesis is compared with the evening, and there are more crimes at midnight. However, according to the plot, we reject our last hypothesis. The crimes are more active in the evening rather than midnight.

In summary, through this project, we want to show people that Building on DC's public safety improvements will take more than policing alone, we need a holistic approach that involves social services, the police, nonprofits, and residents working together, only through collaboration can the city build safer communities—creating the social supports, stability, and opportunities that will bring down even the most persistent pockets of violent crime.

Limitation

We have only found the data for 2018 and 2019, and if we can get more data, we may able to do the time series data mining. The original data doesn't include the population distribution in each area; we cannot study the relationship between the number of crimes and the number of people, and cannot determine the crime rate. There is still has biases in the original data frame, the division of day and night and midnight is different from the common sense which may affect our hypothesis judgment.

Reference

https://mpdc.dc.gov/page/district-crime-data-glance

Moore,H.M., & Trojanowicz,C.R.(1988). Policing and the Fear of Crime. U.S. Department of Justice. No.3 https://www.ncjrs.gov/pdffiles1/nij/111459.pdf

Session Info

```r
sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.4
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] MASS_7.3-51.5   car_3.0-7       carData_3.0-3   broom_0.5.6    
##  [5] gganimate_1.0.5 plotly_4.9.2.1  lubridate_1.7.8 forcats_0.5.0  
##  [9] stringr_1.4.0   dplyr_0.8.5     purrr_0.3.4     readr_1.3.1    
## [13] tidyr_1.0.2     tibble_3.0.1    ggplot2_3.3.0   tidyverse_1.3.0
## 
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.4.6      lattice_0.20-38   prettyunits_1.1.1 utf8_1.1.4       
##  [5] assertthat_0.2.1  digest_0.6.25     R6_2.4.1          cellranger_1.1.0 
##  [9] backports_1.1.6   reprex_0.3.0      evaluate_0.14     httr_1.4.1       
## [13] pillar_1.4.3      rlang_0.4.5       progress_1.2.2    curl_4.3         
## [17] lazyeval_0.2.2    readxl_1.3.1      rstudioapi_0.11   data.table_1.12.8
## [21] rmarkdown_2.1     labeling_0.3      foreign_0.8-75    htmlwidgets_1.5.1
## [25] munsell_0.5.0     compiler_3.6.3    modelr_0.1.6      xfun_0.13        
## [29] pkgconfig_2.0.3   htmltools_0.4.0   tidyselect_1.0.0  rio_0.5.16       
## [33] fansi_0.4.1       viridisLite_0.3.0 crayon_1.3.4      dbplyr_1.4.3     
## [37] withr_2.2.0       grid_3.6.3        nlme_3.1-144      jsonlite_1.6.1   
## [41] gtable_0.3.0      lifecycle_0.2.0   DBI_1.1.0         magrittr_1.5     
## [45] scales_1.1.0      zip_2.0.4         cli_2.0.2         stringi_1.4.6    
## [49] farver_2.0.3      fs_1.4.1          xml2_1.3.2        ellipsis_0.3.0   
## [53] generics_0.0.2    vctrs_0.2.4       openxlsx_4.1.4    tools_3.6.3      
## [57] glue_1.4.0        tweenr_1.0.1      hms_0.5.3         abind_1.4-5      
## [61] yaml_2.2.1        colorspace_1.4-1  rvest_0.3.5       knitr_1.28       
## [65] haven_2.2.0
```