Stop Data Analysis of Minneapolis Police Department 2017

Final Project Report - Spring 2020

Name : M. M. NABI

Email : mn918@msstate.edu

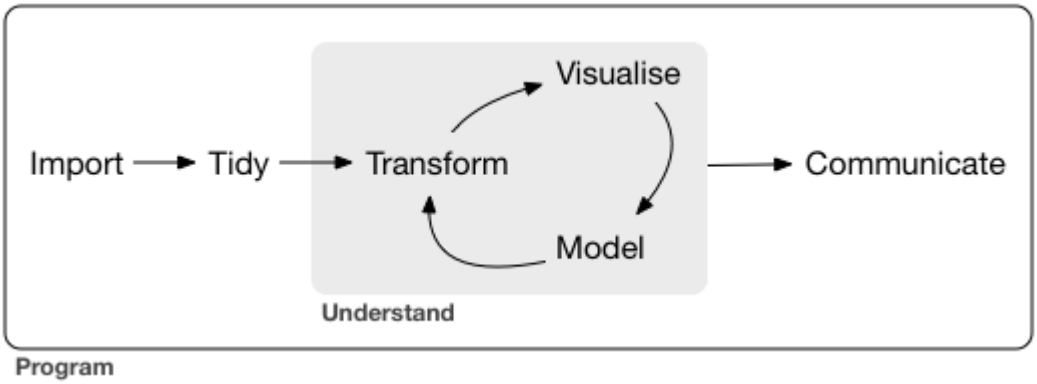
Date: 04-24-2020

Contents

Introduction

The purpose of this project is to analyze a dataset in order to get a general idea and a visualization of the aesthetics of a certain activity. This project will help us to find out some hidden information by using several graphs or chats. As we know that several methods are employed to display a dataset. The key viewpoint of this project would thus help to explain the fundamental features of data analysis using R Language.

Overall Data analysis process is as below



In the R language, there are stages of data science for data analysis. First, the data needs to be imported into the system. After that, the biggest challenge is to make it tidy. A tidy dataset helps us to perceive the overall structure for analysis. After that, it is easy to transform, model and visualize for better decision making. In the end, it gives a better view of communication.

Data Source

For this project, the police stop data in 2017 of the Minneapolis Police Department is used here.

Data source link :

http://opendata.minneapolismn.gov/datasets/police-stop-data

This dataset is in CSV format from the source location. In this dataset, it has 51857 observations (rows) for 14 variables (columns).

Data Analysis

Let's look at the dataset. What types of data is there?

```
str(Police_record)
```

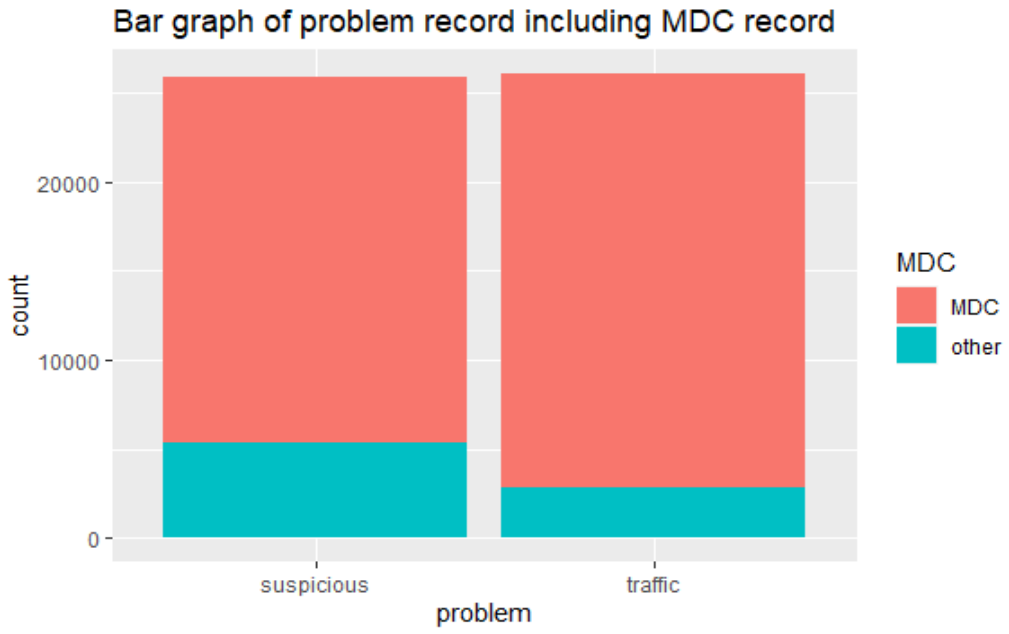
```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 51920 obs. of  15
variables:
##  $ X1            : num  6823 6824 6825 6826 6827 ...
##  $ idNum         : chr  "17-000003" "17-000007" "17-000073" "17-000092" ..
.
##  $ date          : POSIXct, format: "2017-01-01 00:00:42" "2017-01-01 00:0
3:07" ...
##  $ problem       : chr  "suspicious" "suspicious" "traffic" "suspicious" .
..
##  $ MDC           : chr  "MDC" "MDC" "MDC" "MDC" ...
##  $ citationIssued: logi  NA NA NA NA NA NA ...
##  $ personSearch  : chr  "NO" "NO" "NO" "NO" ...
##  $ vehicleSearch : chr  "NO" "NO" "NO" "NO" ...
##  $ preRace       : chr  "Unknown" "Unknown" "Unknown" "Unknown" ...
##  $ race          : chr  "Unknown" "Unknown" "White" "East African" ...
##  $ gender        : chr  "Unknown" "Male" "Female" "Male" ...
##  $ lat           : num  45 45 44.9 44.9 45 ...
##  $ long          : num  -93.2 -93.3 -93.3 -93.3 -93.3 ...
##  $ policePrecinct: num  1 1 5 5 1 1 1 2 2 4 ...
##  $ neighborhood  : chr  "Cedar Riverside" "Downtown West" "Whittier" "Whit
tier" ...
```

According to the dataset, it is observed that it has incident records with GPS (LAT LONG) and timestamp. It includes a record of problems, issued citation, person and vehicle searched parameters. It has also race and gender of the identifiers as well as which neighborhood they are from. This factor with 84 levels gives the name of the neighborhood of the incident. Data is also collected via an in-vehicle computer called MDC. There is other data source (foot/bicycle/horseback) without MDC which are marked as other sources in the dataset. Most of the data are character and number data. There are some logical data and time formated data(POSIXct).

Data Analysis (Visualization)

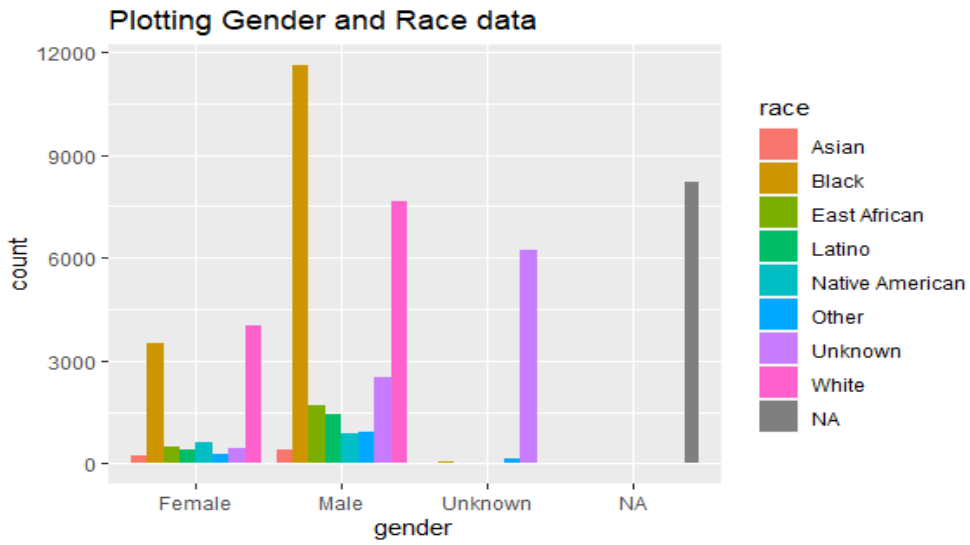
First, we will look at some simple data visualization based on the problem and MDC record parameter. From the plot, it is observed that both suspicious and traffic problem is similar in number. But it is noticeable that the MDC record is much higher compared to the other records.

```
ggplot(data = Police_record) +
 geom_bar(mapping = aes(x = problem, y = ..count.., group = MDC, fill = MDC))
+ggtitle("Bar graph of problem record including MDC record")
```



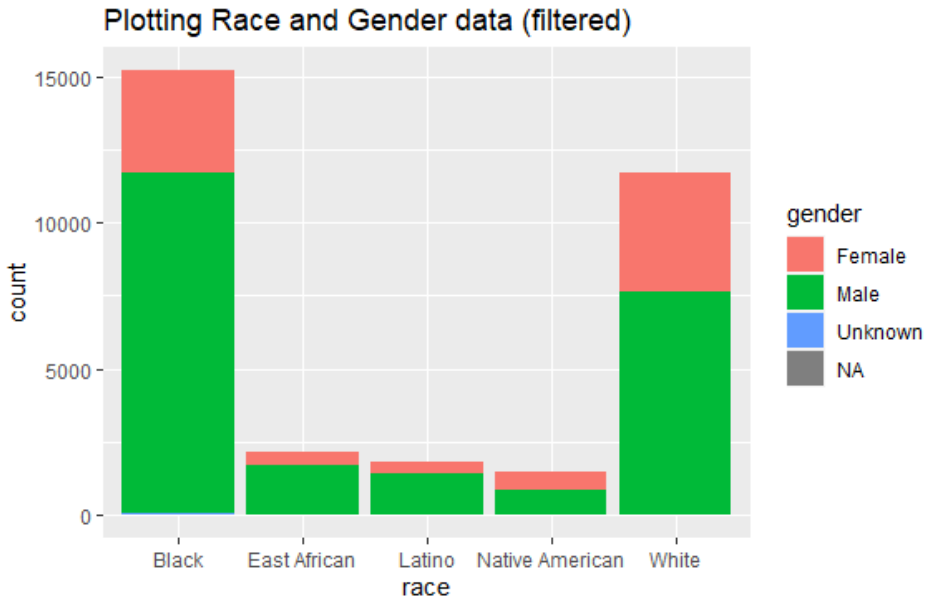
Next, we will see how gender records and race involve in these activities. If we categorize the gender in the x-axis and divide the race using color aesthetic, we can observe that males are contributing a large number in this dataset. There are some other records seems like taking a large number of data which are Unknow gender and NA gender. From this analysis, the number of Black males is taking the highest number of records here. After that White male is taking the second-highest number position.

```
ggplot(data = Police_record) +
  geom_bar(mapping = aes(x = gender, fill = race), position = "dodge")+
        ggtitle("Plotting Gender and Race data")
```



If we try to filter some data in this analysis and rearrange it in a different way then, it becomes much more understandable for us to decide which people are taking place in this dataset. There are comparatively few people like East African, Latino and Native American people taking place in this dataset. But it is obvious that the number of male people is grater in all races.

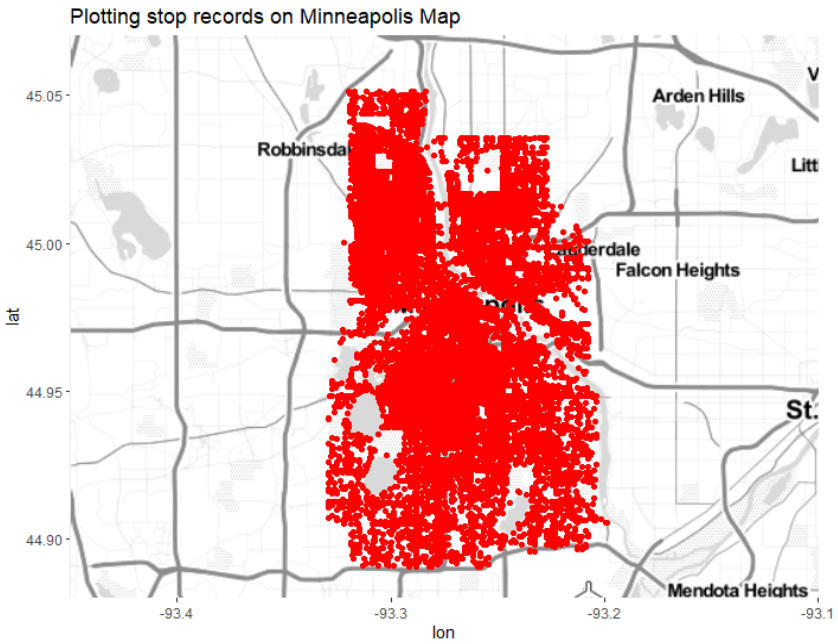
```
u <- Police_record %>%
  select(race, gender) %>%
  filter(race %in% c("Black","White","Native American","East African", "Latin
o"))
ggplot(data = u) +
 geom_bar(mapping = aes(x = race, y = ..count.., group = gender, fill = gende
r))+
        ggtitle("Plotting Race and Gender data (filtered)")
```



Data Analysis (Visualization on Map)

As mentioned earlier, it has GPS data, so we can visualize it by plotting it in the Minneapolis map. To make it more clear on the map, we set the Base map zoom level at 11. As it has more than 50 thousand data points, so it seems all regions will cover the whole plot.

```
Map_Minneapolis  <- c(left = -93.45, bottom = 44.88, right = -93.10, top = 45
.07)
Base_map <- get_stamenmap(Map_Minneapolis, zoom = 11, maptype = "toner-lite")
%>% ggmap()

Base_map + geom_point(data=Police_record[sample(nrow(Police_record), 50000),
], aes(x=long, y=lat),color = "red")+
        ggtitle("Plotting stop records on Minneapolis Map")
```



If we try to see which neighborhood area has more stop record, we can filter out which has counts less then 1500 records. So, we find only 7 neighborhood area grater than 1500 records.

```
w <- Police_record %>%
      count(neighborhood) %>%
      filter(n > 1500)
w
```

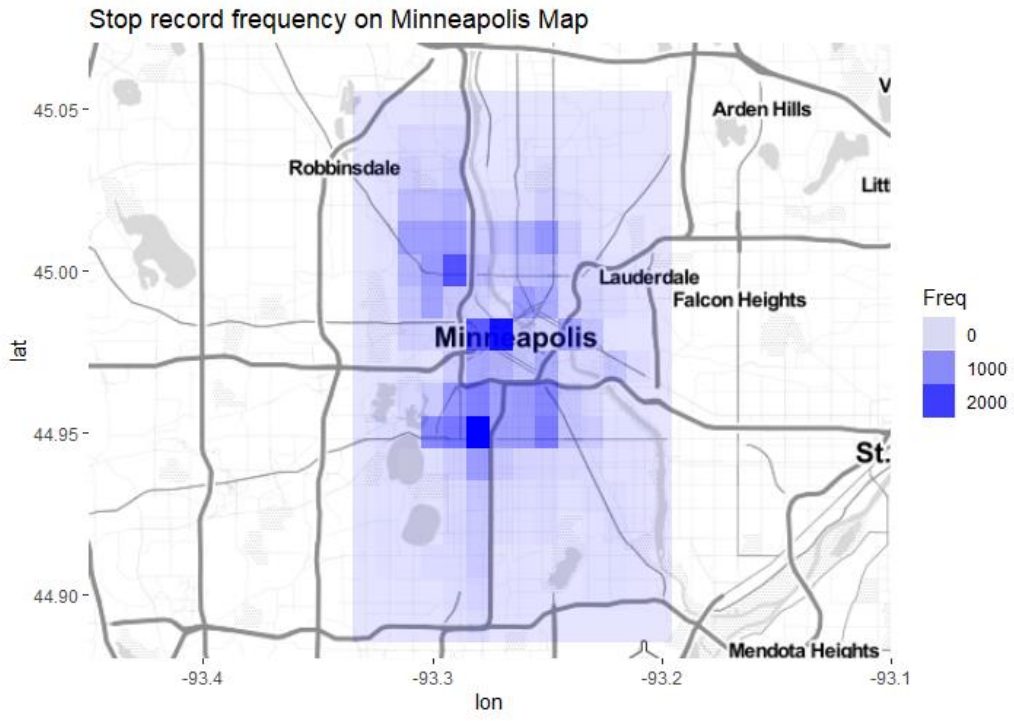
```
## # A tibble: 7 x 2
##   neighborhood      n
##   <chr>         <int>
## 1 Downtown West  4409
## 2 Hawthorne      2031
## 3 Jordan         2075
## # … with 4 more rows
```

Map Visualization based on the record frequency

Now, we will see the stop record is distributed on the map based on the frequency. This plot will help us to visualize in the map that which area having a more stop record.

```
LatLonCounts <- as.data.frame(table(round(Police_record$long, 2), round(Polic
e_record$lat, 2)))

LatLonCounts$Long <- as.numeric(as.character(LatLonCounts$Var1))
LatLonCounts$Lat <- as.numeric(as.character(LatLonCounts$Var2))

Base_map + geom_tile(data = LatLonCounts, aes(x = Long, y = Lat, alpha = Freq
```

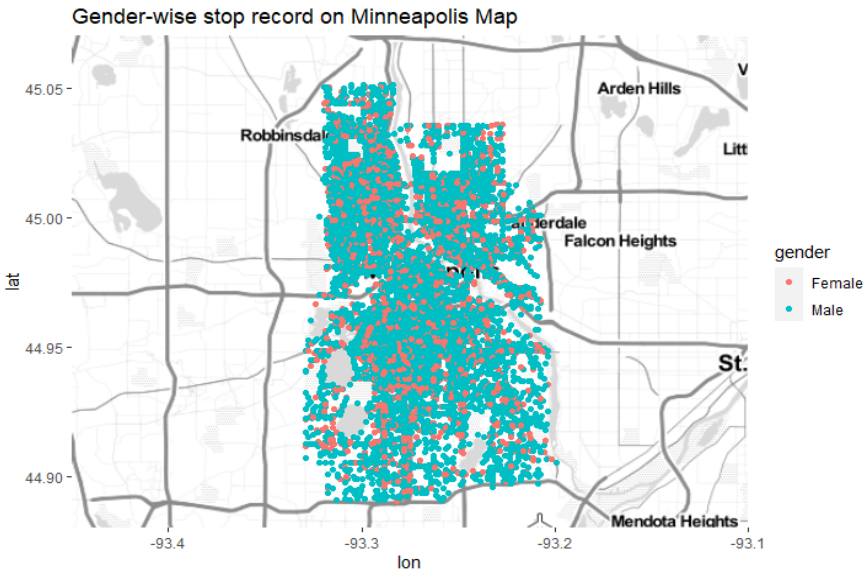
```
), fill = "blue")+
        ggtitle("Stop record frequency on Minneapolis Map")
```



Map Visualization based on gender (MALE & Female)

Now, we will observe how gender is distributed in the map. We will only filter the male and female gender for more visualize the plot. As mentioned earlier, it is obvious that the male gender is taking more areas to stop records.

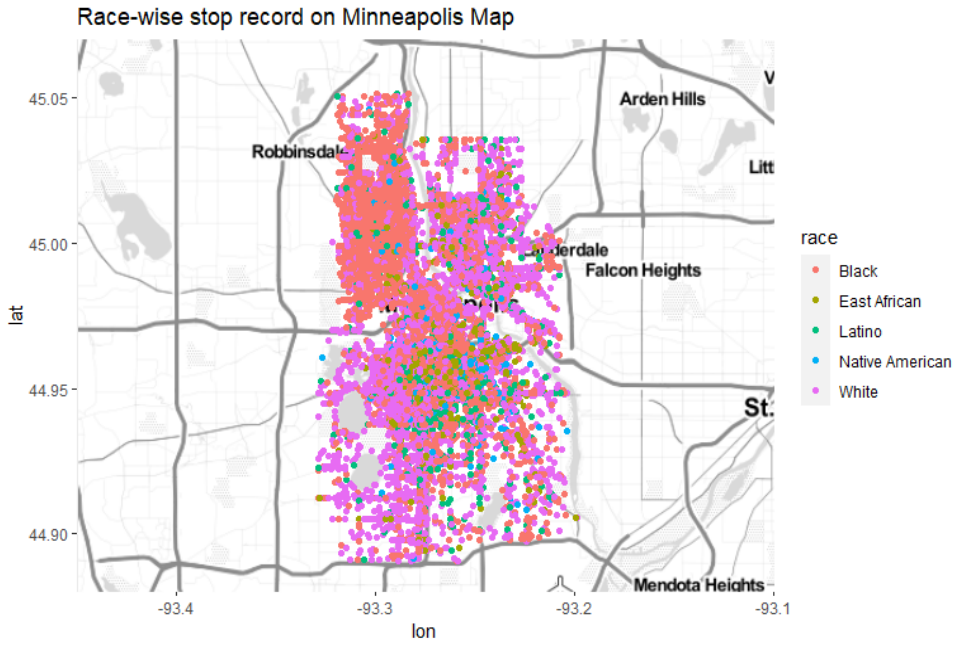
```
q <- Police_record %>%
  select(gender, lat, long) %>%
  filter(gender %in% c("Male", "Female"))
Base_map + geom_point(data = q, aes(x = long, y = lat, color = gender)) +
        ggtitle("Gender-wise stop record on Minneapolis Map")
```



Map Visualization based on Race

Same as previous, we can see that Black people are covering more areas on the map especially the Northeast region. In the center of an area covering more white people and few Latino peoples are there in the southern part.

```
r <- Police_record %>%
  select(race, lat, long) %>%
  filter(race %in% c("Black","White","Native American","East African", "Latin
o"))
Base_map + geom_point(data = r, aes(x = long, y = lat, color = race)) +
        ggtitle("Race-wise stop record on Minneapolis Map")
```



Data Analysis (Based on Date & Time)

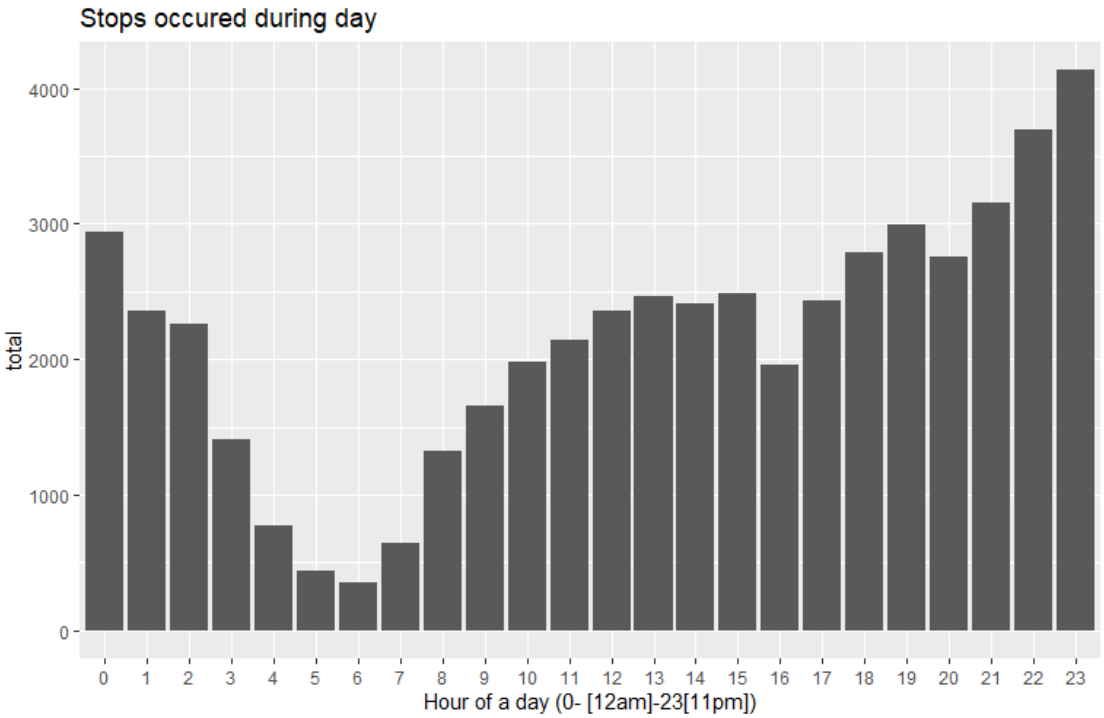
In this dataset, we have the time date column which is in POSIXct format and it will be extracted to find the Hour, Day, Month, Week, etc. Then, that specific date/time data will be grouped and plot for the desired output.

```
comb <- Police_record
comb$date <- as.POSIXct(comb$date, format = "%m/%d/%Y %H:%M:%S")

comb$Time <- format(as.POSIXct(comb$date, format = "%m/%d/%Y %H:%M:%S"), form
at="%H:%M:%S")
comb$date <- ymd_hms(comb$date)
comb$day <- factor(day(comb$date))
comb$month <- factor(month(comb$date, label = TRUE))
comb$year <- factor(year(comb$date))
comb$dayofweek <- factor(wday(comb$date, label = TRUE))


comb$hour <- factor(hour(hms(comb$Time)))
comb$minute <- factor(minute(hms(comb$Time)))
comb$second <- factor(second(hms(comb$Time)))
```

Data Analysis (Based on Hour)

Now, total stops data will be examined based on the whole day/ hourly basis. From the plot, we can see that during the early monring the rate of stop record is minimum. It goes an increasing order during the period, especially during the mid-night. There is a slight decrease during the afternoon (4 pm).

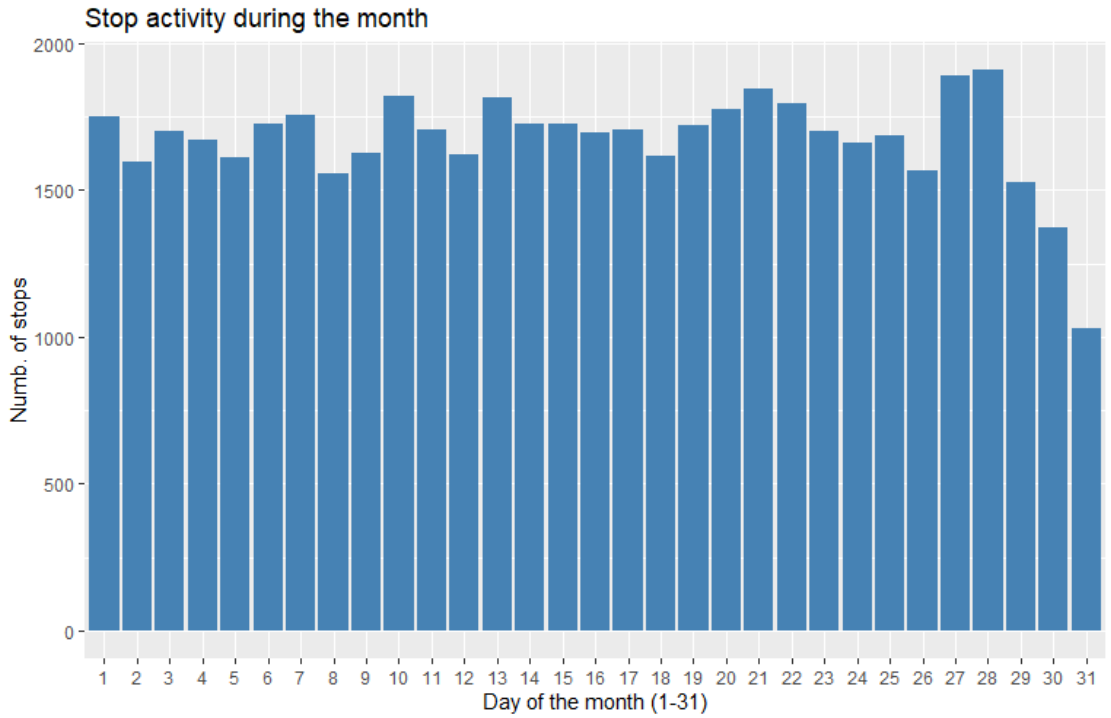
```
hour_data <- comb %>%
           group_by(hour) %>%
               dplyr::summarize(Total = n())

ggplot(hour_data, aes(hour, Total)) +
        geom_bar( stat = "identity") +
           ggtitle("Stops occured during day") +
          labs(x="Hour of a day (0- [12am]-23[11pm])", y="total")
```



Data Analysis (Based on days in month)

Based on the days in the month analysis, it doesn't have any exact pattern. It's near about the same every day in the month. But we see, there is a sudden decrease in the 31st day. It is because every month doesn't have 31st days. So, it seems reasonable.

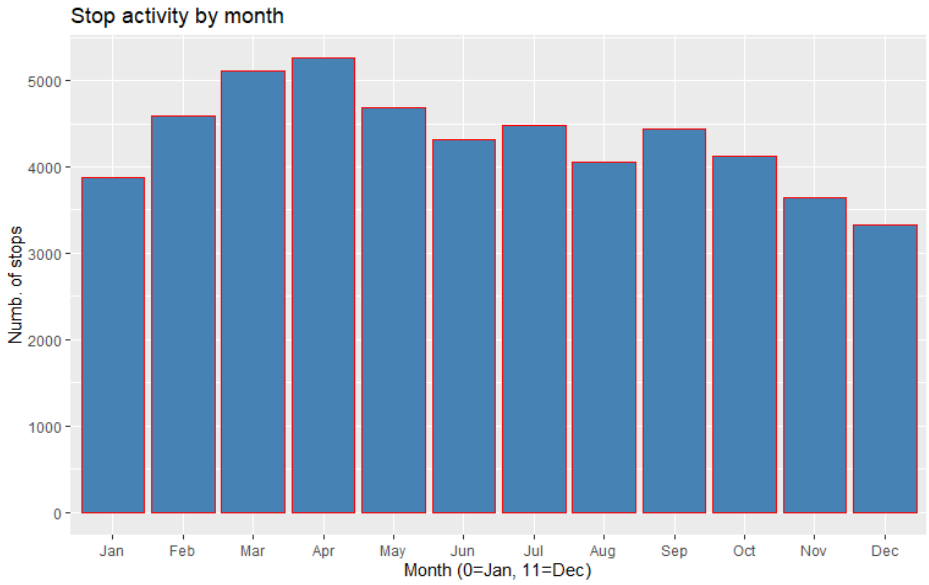
```
day_data <- comb %>%
  group_by(day) %>%
    dplyr::summarize(Total = n())

ggplot(day_data, aes(day, Total)) +
        geom_bar( stat = "identity", fill = "steelblue") +
          labs(x="Day of the month (1-31)", y="Numb. of stops", title="Stop a
ctivity during the month")
```



Data Analysis (Based on Month)

Here, we will see the pattern of the plot based on the months. From the plot, we can see that during the beginning of the year the number of records is in increasing trend. It gets the highest record in April. It becomes decreased during the summertime and goes down to December. Most probably during the summertime, everyone goes out of the city and the same as Christmas eve.

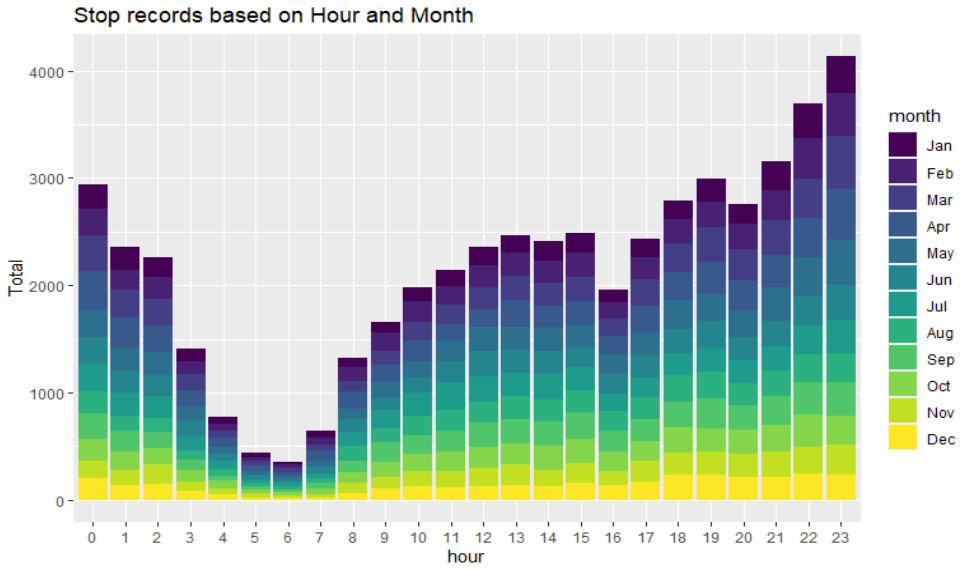
```
month_data <- comb %>%
  group_by(month) %>%
    dplyr::summarize(Total = n())

ggplot(month_data, aes(month, Total)) +
        geom_bar( stat = "identity", fill = "steelblue", color = "red") +
          labs(x="Month (0=Jan, 11=Dec)", y="Numb. of stops", title="Stop act
ivity by month")
```



Data Analysis (Based on Hour-Month)

If we try to look at the pattern based on the month and hour basis in the same plot, we can observe that early morning (6 am) every month has the lowest record of stop activity by the police department. The plot also tells us that the monthly data on an hourly basis in December has the highest number of stop records. Throughout the day it is in increasing order and becomes decreased after midnight.

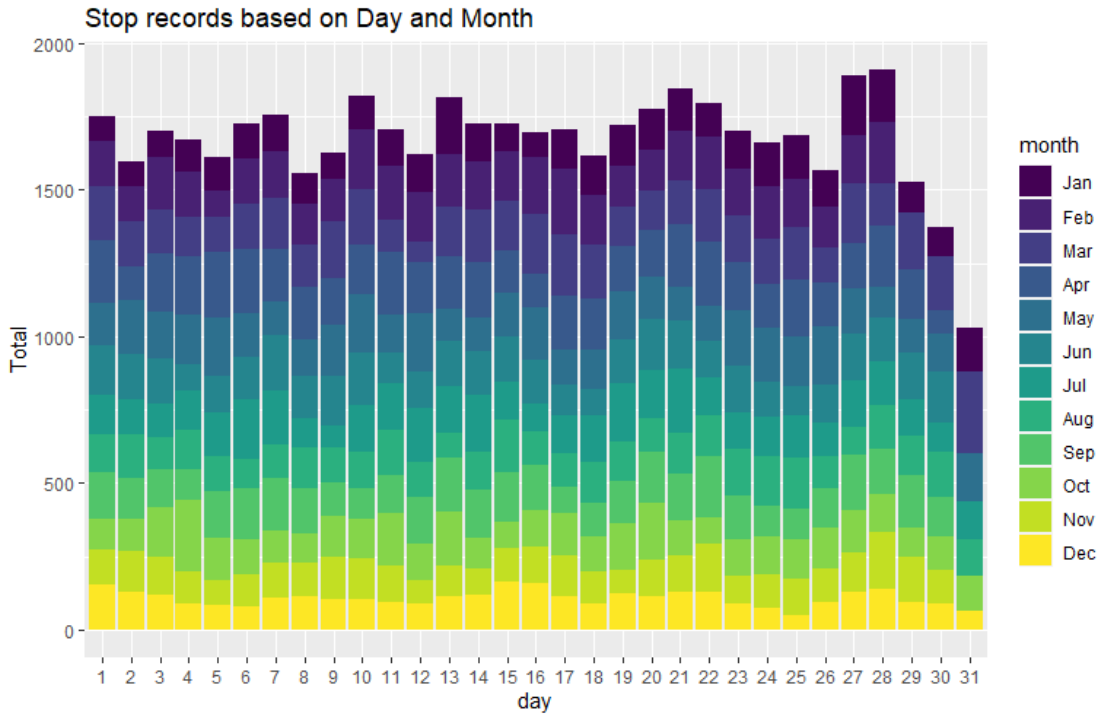
```
month_hour <- comb %>%
          group_by(month, hour) %>%
             dplyr::summarize(Total = n())
ggplot(month_hour, aes(hour, Total, fill = month)) +
       geom_bar( stat = "identity") +
          ggtitle("Stop records based on Hour and Month")
```



Data Analysis (Based on Days and Month)

Now, we will try to see how day-wise data evolve on a monthly basis. In the plot, we see that it has almost a similar trend throughout the days but the main variation is coming in the months' data. The changes basically made by the monthly data. As before, we get a decreased value on the 31st day of the month.

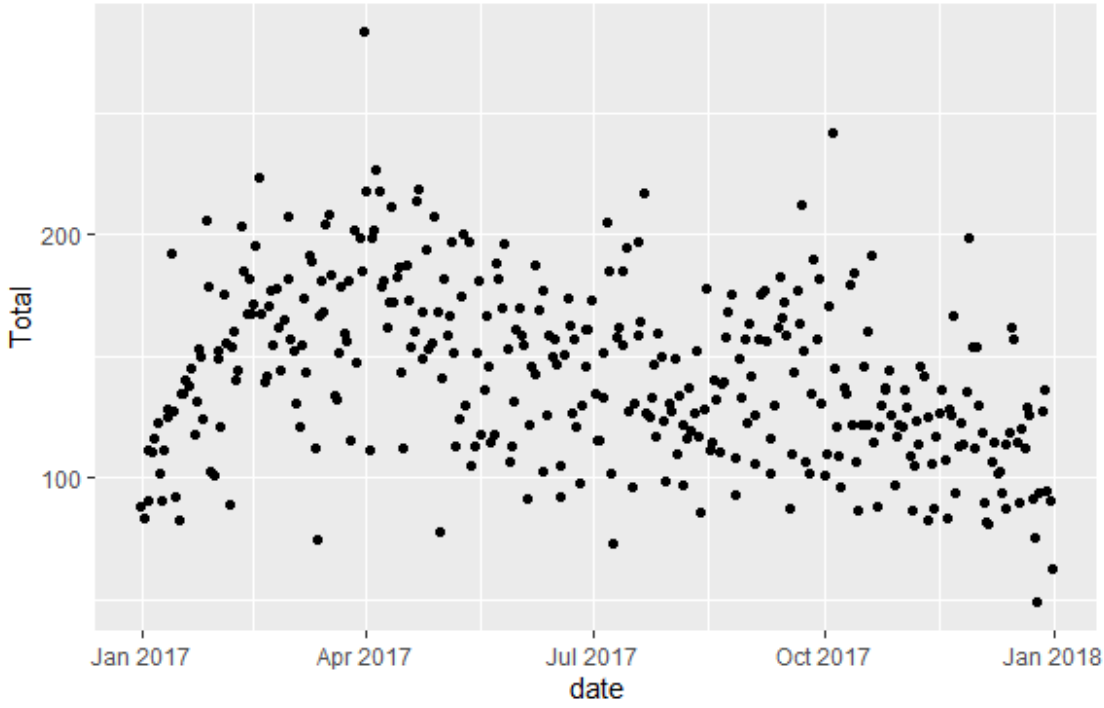
```
day_month_group <- comb %>%
         group_by(month, day) %>%
             dplyr::summarize(Total = n())
ggplot(day_month_group, aes(day, Total, fill = month)) +
        geom_bar( stat = "identity") +
           ggtitle("Stop records based on Day and Month")
```



Modeling

We have already seen different data analyses of the Minneapolis Police Department's stop data. As the dataset has a record of each date. So, we can count the number of stops record date-wise. The below plot shows us how the data is distributed throughout the year based on each day.

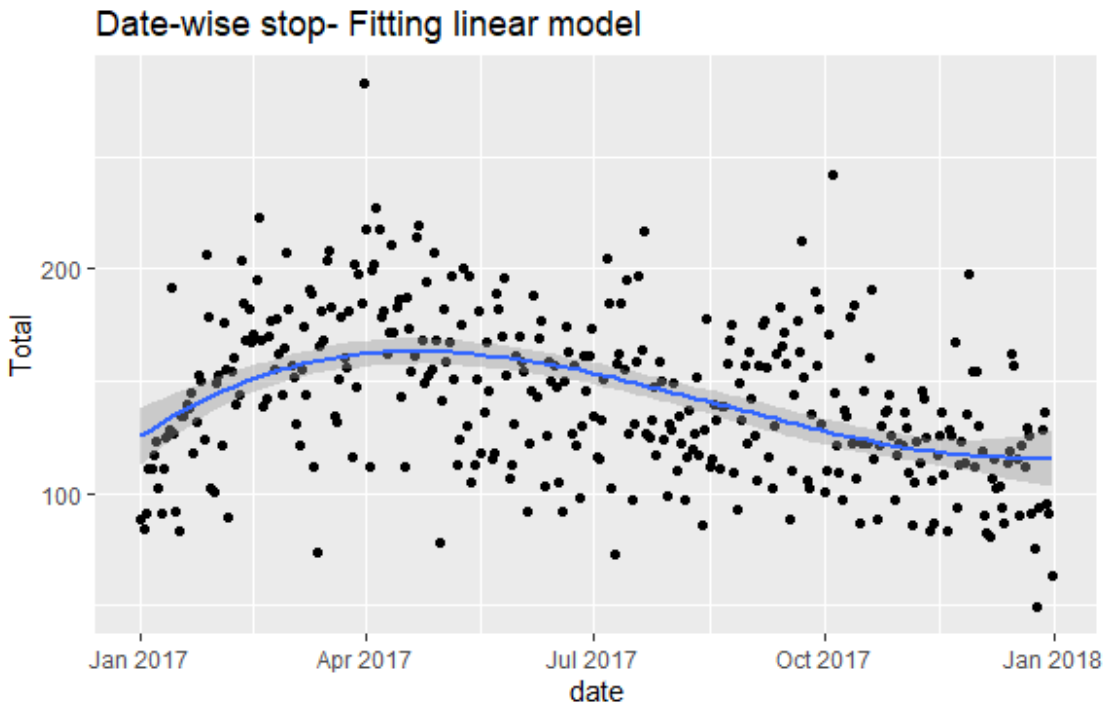
```
comb <- Police_record
comb$date <- as.Date(comb$date,"%m/%d/%Y %H:%M:%S")

## Warning in as.POSIXlt.POSIXct(x, tz = tz): unknown timezone '%m/%d/%Y %H:%
M:%S'

Date_data <- comb %>%
            group_by(date) %>%
                dplyr::summarize(Total = n())

ggplot(Date_data, mapping=aes(x=date, y= Total)) +
        geom_jitter()
```



Fitting a linear Model

Let's try to fit a model based on the record in each day. This model seems reasonable in some points. We have calculated the coefficient and rood squared value. But we can try any other model which can fit in a better way.

```
ggplot(Date_data, aes(date, Total)) +
  geom_point()+
  geom_smooth(method = lm, formula = y ~ splines::bs(x, 3))+
        ggtitle("Date-wise stop- Fitting linear model")
```



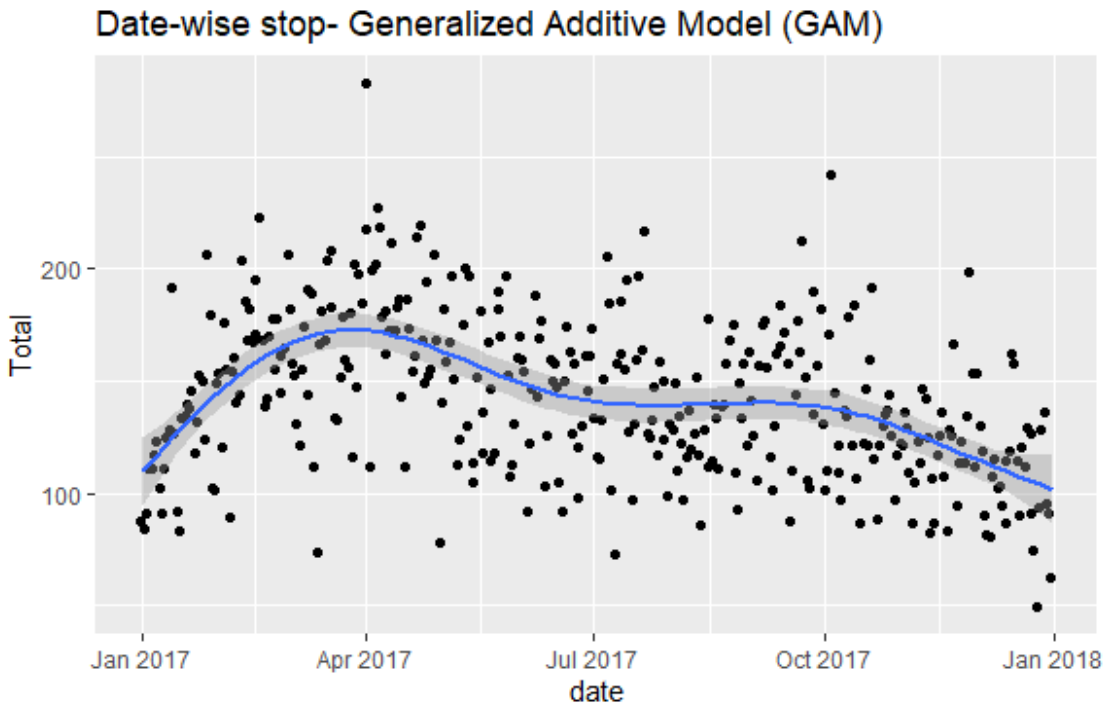

```
model_1 <- lm(Total ~ splines::bs(date, 3), data = Date_data)
coef(model_1) %>% str()

##  Named num [1:4] 125.29 93.18 -15.68 -9.42
##  - attr(*, "names")= chr [1:4] "(Intercept)" "splines::bs(date, 3)1" "spli
nes::bs(date, 3)2" "splines::bs(date, 3)3"

summary(model_1)$r.squared * 100

## [1] 21.75583
```

Fitting a Generalized Additive Model (GAM)

Now, we will try another type of linear model called generalized additive model (GAM) where the linear predictor depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions. This plot looks more fitted to our dataset. We also tried to find the co-efficient and root squared value. After observing the model, it looks like the stop record goes down linearly. It might have two reasons. As we have already seen that at the end of the year, the count data become decreased as people may go out for vacation. The second reason, the police dept. may have already taken some steps so people are aware of the rules. So the number comes to decrease.

```
ggplot(Date_data, mapping=aes(x=date, y= Total)) +
        geom_jitter()+
  geom_smooth(method = "gam")+
        ggtitle("Date-wise stop- Generalized Additive Model (GAM)")
```

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```




```
model_2 <- gam(Total ~ date , data = Date_data)
coef(model_2) %>% str()
```

```
##  Named num [1:2] 2090.55 -0.112
##  - attr(*, "names")= chr [1:2] "(Intercept)" "date"
```

```
summary(model_2)$r.squared * 100
```

```
## numeric(0)
```

Conclusion:

From the analysis of Stop Data of Minneapolis Police Department 2017, we can generally tell that male people are more like to stop by the police and black and white race people are in the top number in the dataset. From the date and time analysis, we have seen major changes in early morning data which become increased during midnight. From the monthly data, April got the most record. If the police department checks this data analysis they can easily make any decision where and when to take necessary action.

References:

[1] Online: "http://opendata.minneapolismn.gov/datasets/police-stop-data" Minneapolis Police Department 2017 Stop Data.

[2] Online: "https://ggplot2.tidyverse.org/reference/geom_smooth".

[3] Online: "https://rstudio-pubs-static.s3.amazonaws.com/407566_47415e9536964db2939600c448809bcd.html".