Analysis of Crimes in Boston

Natalya Shelchkova

May 4, 2018

Introduction

In order to analyze the trend of crime in Boston data was gathered from Analyze Boston which contains the date, location, and crime that was committed between June 15th, 2015 and May 2nd, 2018 amongst other things.

Description of Variables

The original data file contained a number of variables, some of which were changed or excluded for the analysis. A breakdown of the variables can be seen below:

- **INCIDENT_NUMBER:** Internal BPD report number
- **OFFENSE_CODE:** Numerical code of offense description
- **OFFENSE_CODE_GROUP:** Internal categorization of offense description
- **OFFENSE_DESCRIPTION:** Primary descriptor of incident
- **DISTRICT:** What district the crime was reported in
- **REPORTING_AREA:** Reporting area number associated with where the crime was reported from
- **SHOOTING:** Indicating a shooting took place
- **OCCURRED_ON_DATE:** Earliest date and time the incident could have taken place
  - This variable was further broken down to:
    - **YEAR**
    - **MONTH**
    - **DAY_OF_WEEK**
    - **HOUR**
- **UCR_PART:** Universal Crime Reporting number
- **STREET:** Street name the incident took place
  - Which was given in more detail using:
    - **LAT**
    - **LONG**
    - **LOCATION:** (Lat, Long)

Certain variables, such as the incident number, UCR part, and offense description were filtered out, and the OCCURRED_ON_DATE variable was split into the month, day, year, and time, in order to extract the day of the month values.

```r
# Read data
crime_data <- read.csv("crime.csv")

# Filter data
crime_data.v1 <- crime_data %>%
  select(-INCIDENT_NUMBER, -OFFENSE_DESCRIPTION, -UCR_PART)

# Seperate OCCURRED_ON_DATE to include the day of the month

crime_data.v2 <- crime_data.v1 %>%
  separate(OCCURRED_ON_DATE, c("Date","Time"), sep = " ") %>%
  separate(Date, c("MM","DAY","YY"), sep = "/") %>%
  select(-YY, -MM, -Time)

saveRDS(crime_data.v2, "filtered_crime.RDS")
```

Analysis by Date

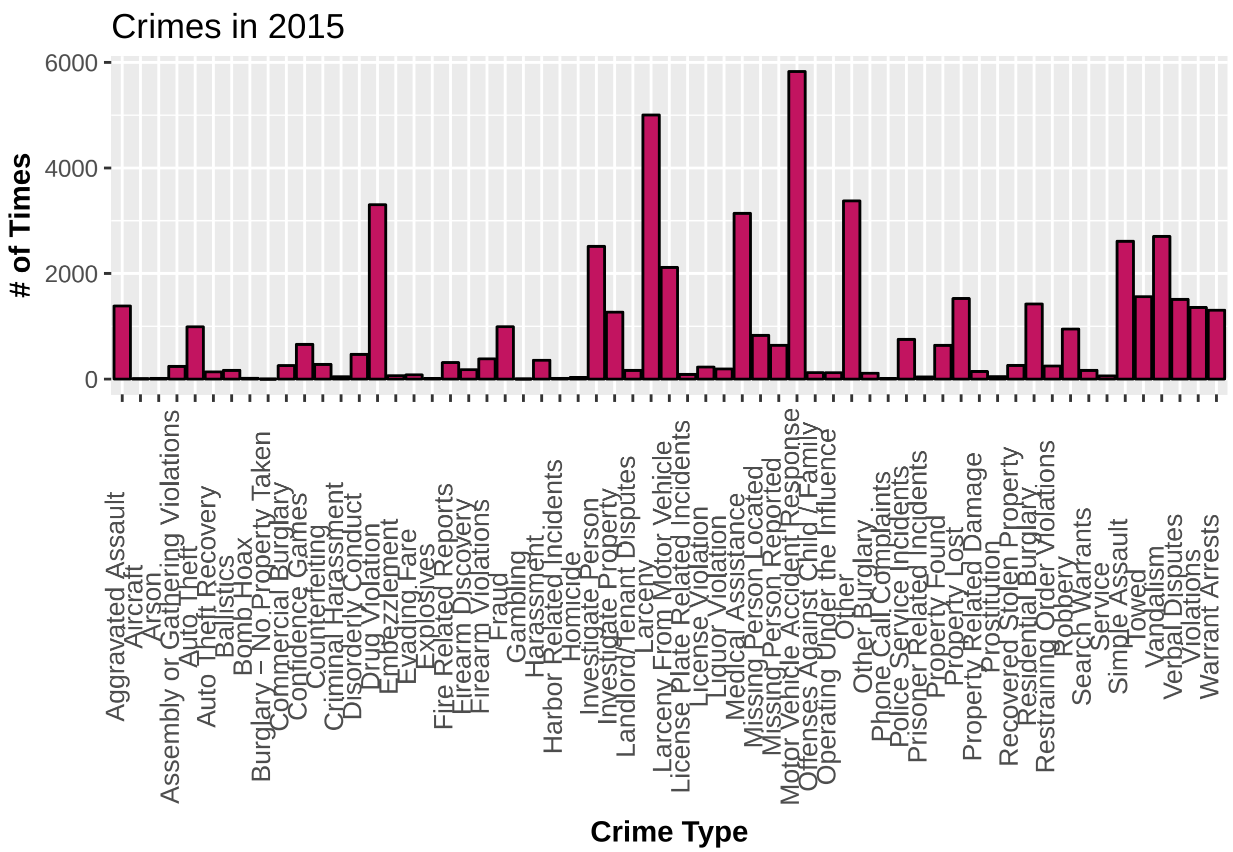
Given how crime tends to follow certain trends, such as more crimes occuring during holidays or late at night, it is important to look at the distribution of crimes at various time scales.

Analysis by Year

By grouping crime data by year, we can look to see whether overall crime values have increased in the years following 2015. However, encounter a problem since the data file only contains a partial list of crimes from 2015, and given that the current year is not yet over, also only a partial list of crimes for 2018. This means that the only two full years of crime data come from 2016 and 2017. However, we can still look at the distribution of the different types of crime that occurred each year, along with the overall pattern of how crime rates change throughout the course of the year.

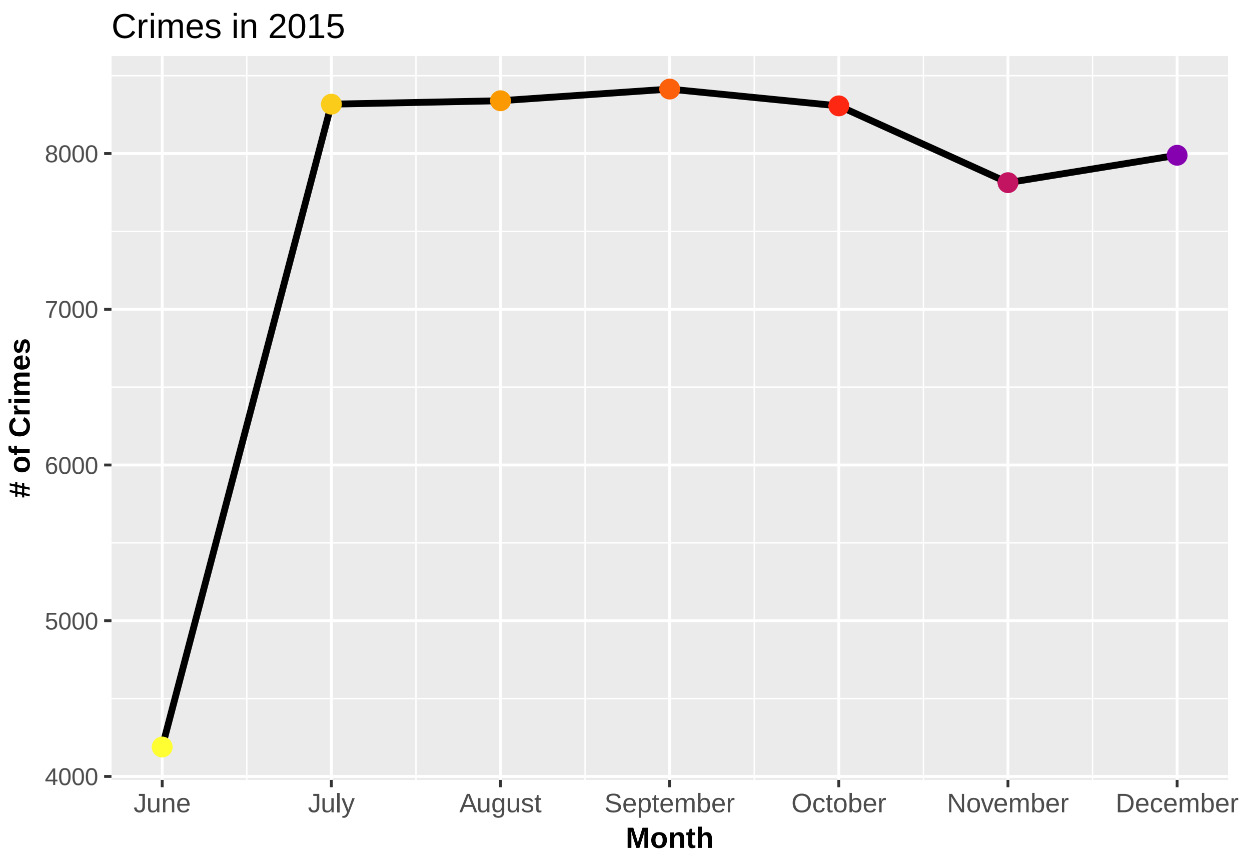
```
# Extract the data for a specific year
crime_2015 <- crime %>%
    filter(YEAR == 2015)

# Crime Type
ggplot(crime_2015, aes(x=OFFENSE_CODE_GROUP)) +
  geom_bar( stat="count", color = "black", fill = "#C21460") +
  theme(axis.text.x = element_text(size = 10, angle = 90, vjust = 0.1),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"),
        legend.text = element_text(size = 10),
        legend.title = element_text(face = "bold", hjust = -1)) +
  labs(title="Crimes in 2015",
       x="Crime Type",
       y = "# of Times")
```



```
# Trend of crime rates over the year
crime_per_month_2015 <- crime_2015 %>%
    select(MONTH) %>%
    group_by(MONTH) %>%
    summarize(`CRIME TOTALS` = n()) %>%
    mutate(Proportion = `CRIME TOTALS` / sum(`CRIME TOTALS`))

ggplot(crime_per_month_2015, aes(x = 6:12, y = `CRIME TOTALS`)) +
  geom_line(size = 1.2)+
  geom_point(size = 3, colour = month_colors[6:12]) +
  theme(axis.text.x = element_text(size = 10),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"),
        legend.text = element_text(size = 10)) +
  labs(title="Crimes in 2015",
       x= "Month",
       y = "# of Crimes") +
  scale_x_continuous(breaks = 6:12,
                     labels= month.name[6:12])
```



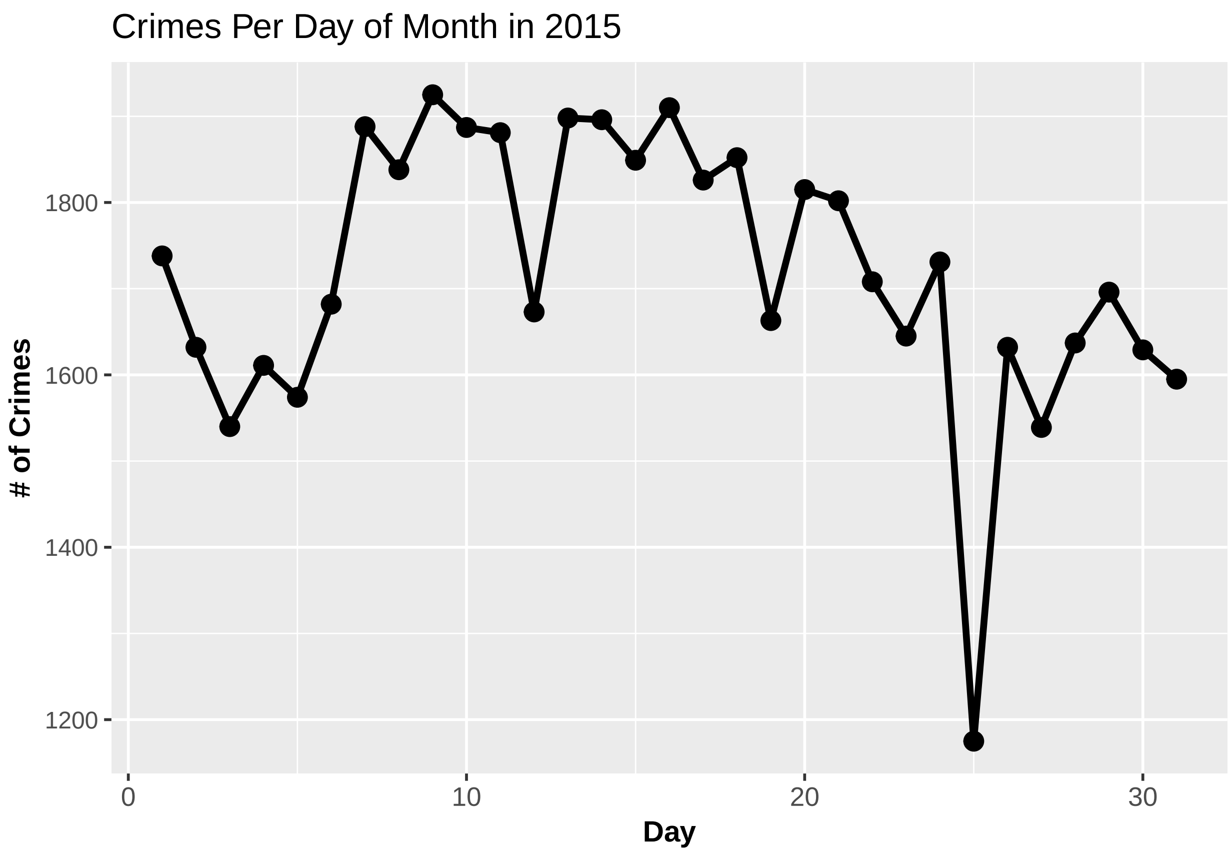
While the example code above is for the year 2015 specifically, it can be used for the subsequent years by adjusting the filtering parameters.

Analysis by Month

There are many urban legends saying that police are more likely to pull people over for minor traffic violations at the end of the month to meet quotas, or how more crimes occur on full moons. Thus, we can also look to see how the rate of crime changes based on the day of the month.

```
# Crime rates over the course of the month
crime_per_day_2015 <- crime_2015 %>%
    select(DAY) %>%
    group_by(DAY) %>%
    summarize(`CRIME TOTALS` = n()) %>%
    mutate(Proportion = `CRIME TOTALS` / sum(`CRIME TOTALS`))

ggplot(crime_per_day_2015, aes(x = 1:31, y = `CRIME TOTALS`)) +
  geom_line(size = 1.2)+
  geom_point(size = 3) +
  theme(axis.text.x = element_text(size = 10),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"),
        legend.text = element_text(size = 10)) +
  labs(title="Crimes Per Day of Month in 2015",
       x= "Day",
       y = "# of Crimes")
```

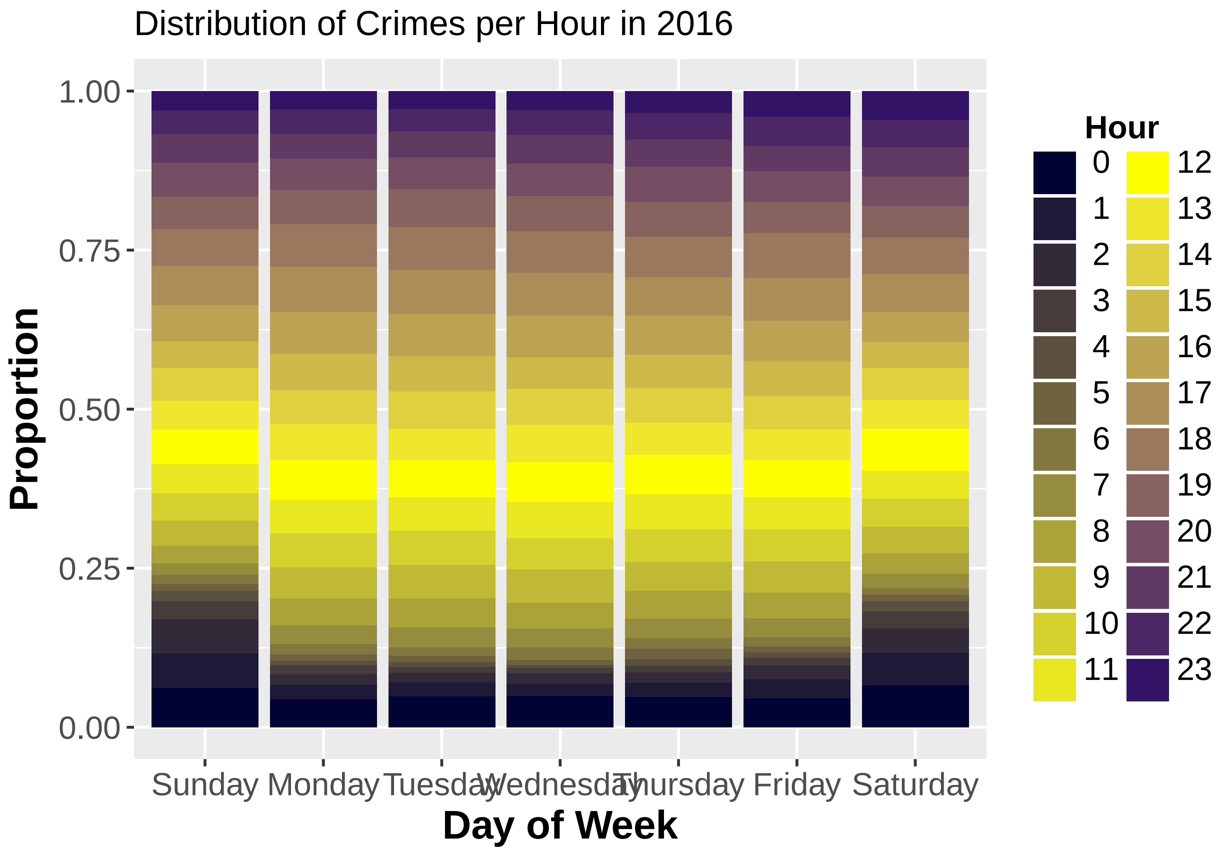


Analysis by Week

We can look to see if day of the week has an influence on crime rates. It would make sense that more crimes, escpecially at night, would occur during the weekends due to the 9-5 work week schedule that is prevelant in todays society.

```
# Crime rates based on the day of the week and hour of the day
crime_per_week_2015 <- crime_2015 %>%
    select(DAY_OF_WEEK, HOUR) %>%
    group_by(DAY_OF_WEEK, HOUR) %>%
    summarize(`CRIME TOTALS` = n()) %>%
    mutate(Proportion = `CRIME TOTALS` / sum(`CRIME TOTALS`))

ggplot(crime_per_week_2015, aes(x = DAY_OF_WEEK, y = Proportion, fill = HOUR)) +
  geom_bar(stat = "identity") +
  scale_fill_gradient2(low = "#000333", mid = "yellow", midpoint = 12, high = "#000066",
                       labels = 0:23,
                       breaks = 0:23,
                       guide = guide_legend(
                         title.position = "top",
                         title = "Hour",
                         label.position = "right",
                         label.hjust = 0.5,
                         label.vjust = 1,
                         title.vjust = 1,
                         title.hjust = .5,
                         title.theme = element_text(face = "bold", angle = 0),
                         label.theme = element_text(angle = 0)
                       )
  ) +
  labs(title="Distribution of Crimes per Hour in 2016",
       x= "Day of Week",
       y = "Proportion") +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_text(face = "bold", size = 15),
        axis.title.y = element_text(face = "bold", size = 15)) +
  scale_x_discrete(limits=c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday","Saturday"))
```

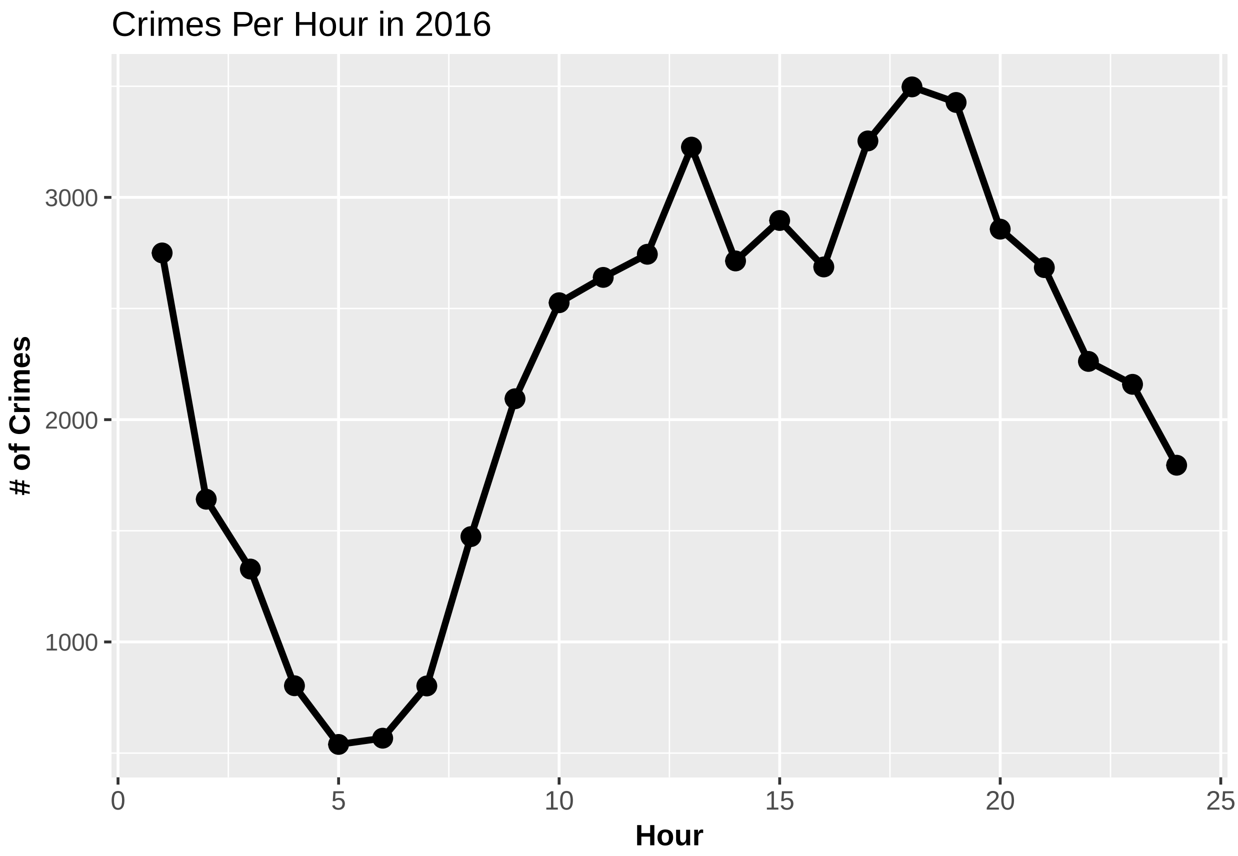


Analysis by Hour

As can be seen above, there are dips and troughs in the crime rate throughout the course of the day, and these can be further seen by isolating crime rates to the hour alone.

```
# Crime trends over the course of a day
 crime_per_hour_2015 <- crime_2015 %>%
    select(HOUR) %>%
    group_by(HOUR) %>%
    summarize(`CRIME TOTALS` = n()) %>%
    mutate(Proportion = `CRIME TOTALS` / sum(`CRIME TOTALS`))
```

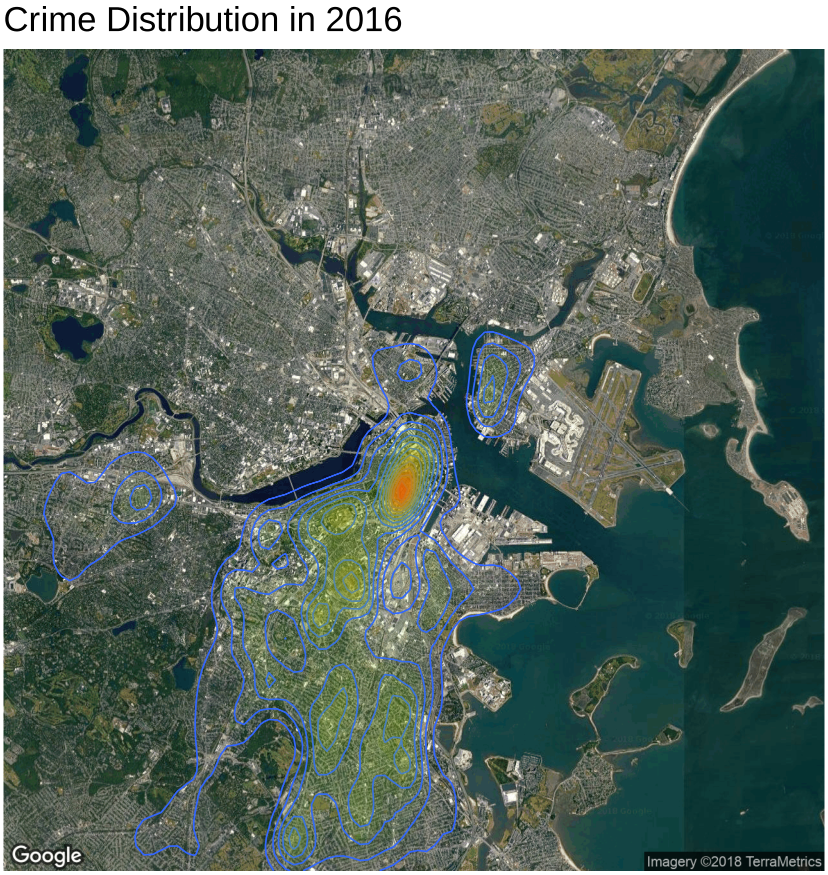
```
ggplot(crime_per_hour_2015, aes(x = 1:24, y = `CRIME TOTALS`)) +
  geom_line(size = 1.2)+
  geom_point(size = 3) +
  theme(axis.text.x = element_text(size = 10),
        axis.title.x = element_text(face = "bold"),
        axis.title.y = element_text(face = "bold"),
        legend.text = element_text(size = 10)) +
  labs(title="Crimes Per Hour in 2016",
       x= "Hour",
       y = "# of Crimes")
```



Analysis by Location

Crimes tend to occur in clusters; areas with high levels of crime will continue to have high levels of crime, and areas which experience very little crime will continue seeing this trend. Thus, looking at location can provide some insights on the distribution of crime within a city.

```
ggmap(boston_map, extent = "device") +
  geom_density2d(data = crime_2015,
                 aes(x = Long, y = Lat), size = 0.3) +
  stat_density2d(data = crime_2015, aes(x = Long, y = Lat,
                                       fill = ..level.., alpha = ..level..),
                 size = 0.01, bins = 16, geom = "polygon") +
  scale_fill_gradient(low = "green", high = "red",
                      guide = FALSE) +
  scale_alpha(range = c(0, 0.3), guide = FALSE) +
  labs(title="Crime Distribution in 2016")
```



Statistics

To look at whether the crime trends are significant, a chi-squared *goodness-of-fit* test was used. This will allow us to look at whether the difference in crime rates is expected or not by assuming that the crime rates should be the same regardless of the time period. An alpha of 0.05 was used to determine significance. For all the chi-squared tests performed the p-values were well below 0.05 which means that time period and crime rate are not independent.

```
# Run chi-squared test
crime_2015 <- crime %>% filter(YEAR == 2015)
week_2015 <- crime_2015 %>% select(DAY_OF_WEEK) %>% group_by(DAY_OF_WEEK) %>% summarize(TOTAL = n())
Number_of_Crimes_Committed_Each_Day_of_the_Week_in_2015 <- week_2015["TOTAL"]
chisq.test(Number_of_Crimes_Committed_Each_Day_of_the_Week_in_2015)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  Number_of_Crimes_Committed_Each_Day_of_the_Week_in_2015
## X-squared = 197.61, df = 6, p-value < 2.2e-16
```

```
week_chisq <- chisq.test(week_2015["TOTAL"])

# make table with expected, observed, and residual values
```

```
week_table <- cbind(week_chisq$observed, week_chisq$expected, week_chisq$residuals)
colnames(week_table) <- c("Obs.","Exp.","Res.")
rownames(week_table) <- c("Friday","Monday","Saturday","Sunday","Thursday","Tuesday","Wednesday")

week_table
```

```
##           Obs.     Exp.       Res.
## Friday    8024 7623.857   4.582766
## Monday    7795 7623.857   1.960069
## Saturday  7385 7623.857  -2.735589
## Sunday    6582 7623.857 -11.932208
## Thursday  7797 7623.857   1.982975
## Tuesday   7913 7623.857   3.311503
## Wednesday 7871 7623.857   2.830484
```