Seattle Incidents Report for last 10 years

Fatma Cemile Serce

February 29, 2016

In this study, I created a Crime Dashboard for the city of Seattle. I used real Full Seattle incident dataset for last 5 years.

Loaded Libraries

```
library(dplyr)
library(ggplot2)
```

Research Questions

In this study, I tried to find answers to the following questions:

- How do incidents vary by time of day?
- Which incidents are most common in the evening?
- During what periods of the day are robberies most common?
- In what areas or neighborhoods are robberies or thefts most common?
- How do incidents vary month to month in the dataset?

A Closer Look at Data

The data contains 610.590 incidents(observations) recorded in the city of Seattle since 1969. There are 19 attributes/fields. Regarding the research questions, I figured out five important fields to work with: *Offense Type*,*Occurred Date or Date Range Start*, *Location*, *Month*, and *Year*.

```
dataset <- read.csv(file="Seattle_Police_Department_Police_Report_Incident.csv",header=TRUE)
glimpse(dataset)
```

```
## Observations: 610,590
## Variables: 19
## $ RMS.CDW.ID                      (int) 644054, 644055, 644048, 6440...
## $ General.Offense.Number          (dbl) 201661434, 201661445, 201661...
## $ Offense.Code                    (fctr) 4199, 5015, 2404, 2404, 240...
## $ Offense.Code.Extension          (int) 0, 1, 1, 1, 8, 2, 2, 0, 0, 1...
## $ Offense.Type                    (fctr) LIQUOR LAW VIOLATION, WARRA...
## $ Summary.Offense.Code            (fctr) 4100, 5000, 2400, 2400, 240...
## $ Summarized.Offense.Description  (fctr) LIQUOR VIOLATION, WARRANT A...
## $ Date.Reported                   (fctr) 02/20/2016 03:24:00 PM, 02/...
## $ Occurred.Date.or.Date.Range.Start (fctr) 02/20/2016 03:24:00 PM, 02/...
## $ Occurred.Date.Range.End         (fctr) , , 02/20/2016 12:30:00 PM,...
## $ Hundred.Block.Location          (fctr) NICKERSON ST / QUEEN ANNE A...
## $ District.Sector                    (fctr) Q, E, Q, J, Q, S, E, E, G, ...
## $ Zone.Beat                          (fctr) Q2, E1, Q3, J1, Q3, S1, E1,...
## $ Census.Tract.2000                  (dbl) 6000.4008, 7500.3014, 7000.5...
## $ Longitude                          (dbl) -122.3577, -122.3209, -122.3...
## $ Latitude                           (dbl) 47.64977, 47.62153, 47.62571...
## $ Location                           (fctr) (47.649772644, -122.3576812...
## $ Month                              (int) 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ Year                               (int) 2016, 2016, 2016, 2016, 2016...
```

There are 147 Offense Types observed in the data.

```
levels(dataset$Offense.Type)[1:10]
```

```
##  [1] "[INC - CASE DC USE ONLY]"   "ANIMAL-BITE"
##  [3] "ANIMAL-CRUELTY"             "ANIMAL-OTH"
##  [5] "ASSLT-AGG-BODYFORCE"        "ASSLT-AGG-GUN"
##  [7] "ASSLT-AGG-POLICE-BODYFORCE" "ASSLT-AGG-POLICE-GUN"
##  [9] "ASSLT-AGG-POLICE-WEAPON"    "ASSLT-AGG-WEAPON"
```

Data Munging

Date Conversion

Here is the sample date format of a given incident.

```
## [1] 02/20/2016 03:24:00 PM
## Levels: 02/20/2016 03:24:00 PM
```

In order to answer the questions above, I need to retrieve the day of the week and the time of the day information from the given date format. I used *POSIXct* and *POSIXlt* function to convert between character representations and objects of classes representing calendar dates and times. I created two new fields, Hour and DayOfWeek, from Date field given in the data.

```
dataset$Occurred.Date.or.Date.Range.Start <- as.POSIXct(strptime(dataset$Occurred.Date.or.Date.Range.Sta
    "%m/%d/%Y %I:%M:%S %p"))
dataset$DayOfWeek <- as.factor(weekdays(dataset$Occurred.Date.or.Date.Range.Start))
dataset$Hour <- as.factor(as.POSIXlt(dataset$Occurred.Date.or.Date.Range.Start)$hour)
```

Subsetting Data

The data contains incidents records since 1969. However, there are few incidents recorded until 2010. Therefore, I prefered to use only last 5 years data, and I created a subset as in the following.

```
last5year_data <- subset(dataset, dataset$Year>2010 & dataset$Year<2016)
```

Calculating the Counts

I am interested in the number of incidents (frequencies) by time of day, day of week, month of a year and a year. In order to calculate the frequencies, I created a new field, Count, and assign 1 as an initial value. Then I calculated the counts by *Offense Type* and *Hour*.

```
last5year_data$Count<-c(1)
count_values <- aggregate(Count~Offense.Type+Hour,data=last5year_data,sum)
```

Top 10 Crimes

There are 202 different Offense Types, and it is not possible to visualize all of them in one plot. Actually, you can do that but no one can understand anything. Therefore, I prefered to work with Top 10 crimes, and subset the data to keep data only for top 10 crimes, which are given below. As you can see, THEFT-CARPOWL is the most common crime.

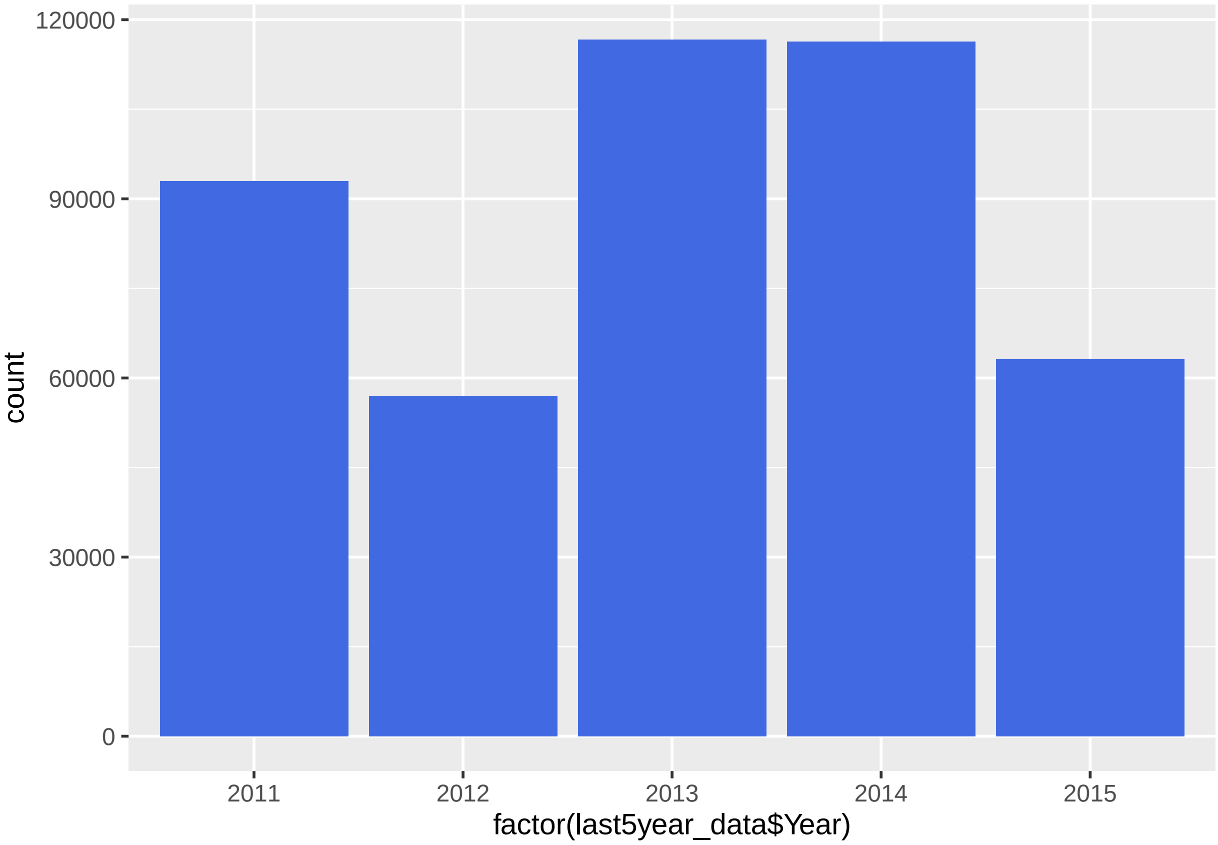
```
common_crimes <- aggregate(Count~Offense.Type,data=last5year_data,sum)
sorted_common_crimes <- common_crimes$Offense.Type[order(common_crimes$Count,decreasing = TRUE)]
top10_crimes <- sorted_common_crimes[1:10]
top10_dataset <- subset(count_values, Offense.Type %in% top10_crimes[1:10])
tail(top10_crimes,n=10)
```

```
##  [1] THEFT-CARPROWL                 THEFT-OTH
##  [3] VEH-THEFT-AUTO                 PROPERTY DAMAGE-NON RESIDENTIA
##  [5] BURGLARY-FORCE-RES             ASSLT-NONAGG
##  [7] DISTURBANCE-OTH                THEFT-SHOPLIFT
##  [9] PROPERTY FOUND                 THEFT-BUILDING
## 202 Levels: [INC - CASE DC USE ONLY] ANIMAL-BITE ... WEAPON-UNLAWFUL USE
```

Incidents Variation by Year (last 10 years)

What happened in 2013 and 2014 in Seattle? As you can see, the crime rate was very very high in 2013 and 2014.

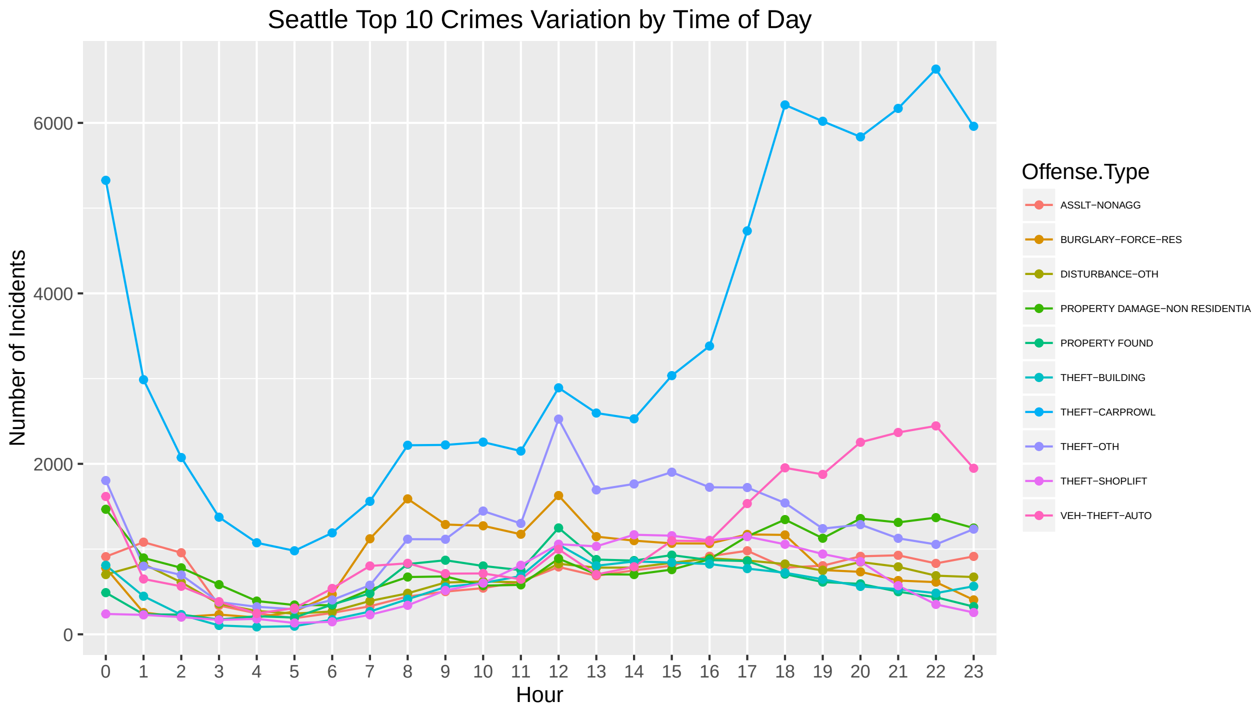
```
ggplot(data = last5year_data, aes(x = factor(last5year_data$Year))) +
    geom_bar(stat = "count", fill = "royalblue")
```



Top 10 Crimes Variation by Time of Day

Be careful between 18:00 and 22:00, these are the hot times for THEFT-CARPOWL.

```
ggplot(data = top10_dataset, aes(Hour, Count)) +
    labs(title = "Seattle Top 10 Crimes Variation by Time of Day",
        y = "Number of Incidents") + theme(legend.text = element_text(size = 5)) +
    geom_line(aes(colour = Offense.Type,
        group = Offense.Type)) + geom_point(aes(colour = Offense.Type))
```

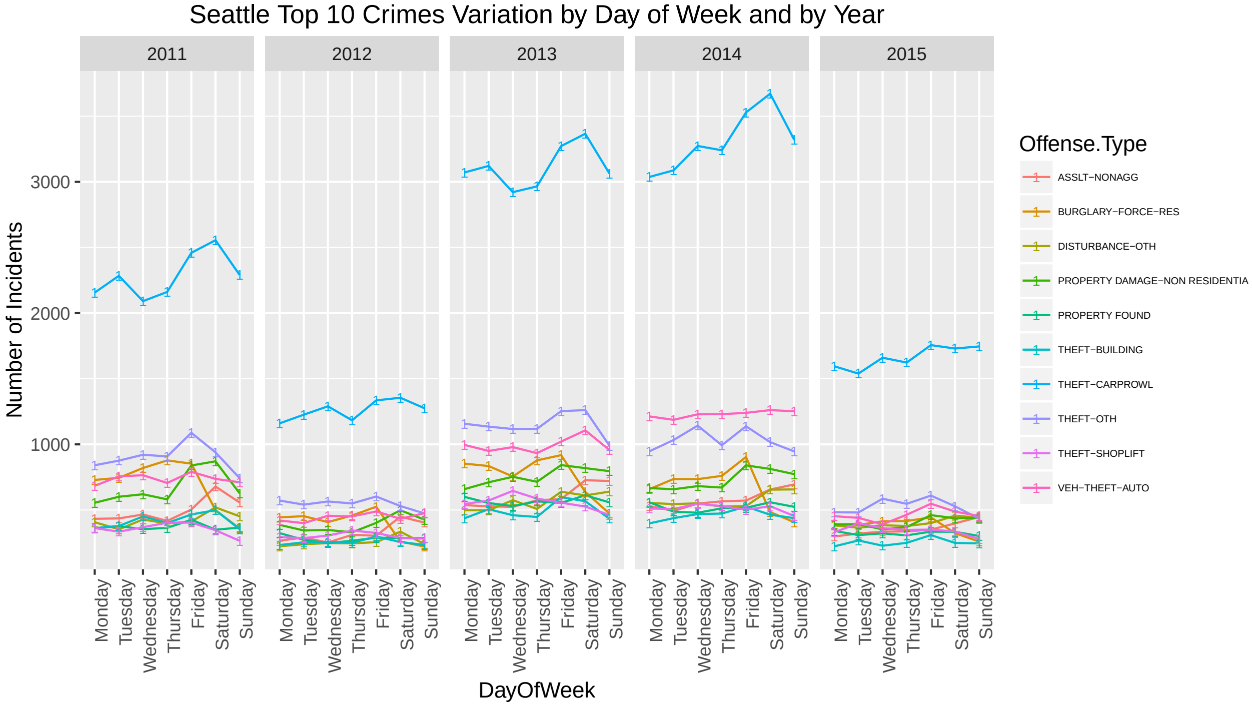


Top 10 Crimes Variation by Day of Week and by Year

As you can see, THEFT-CARPOWL tends to increase in weekend, especially in Saturday. This is consistently true for almost all last 5 years. To draw the plot, I calculated the count values based on Offense.Type, Year and DayOfWeek.

```
count_values_day <- aggregate(Count ~ Offense.Type +
    DayOfWeek + Year, data = last5year_data,
    sum)
top10_dataset <- subset(count_values_day,
    Offense.Type %in% top10_crimes[1:10])
top10_dataset$DayOfWeek <- factor(top10_dataset$DayOfWeek,
    levels = c("Monday", "Tuesday", "Wednesday",
        "Thursday", "Friday", "Saturday",
        "Sunday"))

ggplot(data = top10_dataset, aes(DayOfWeek,
    Count)) + labs(title = "Seattle Top 10 Crimes Variation by Day of Week and by Year",
    y = "Number of Incidents") + facet_grid(~Year) +
    theme(legend.text = element_text(size = 5)) +
    geom_line(aes(colour = Offense.Type,
        group = Offense.Type)) + geom_point(shape = "18",
    aes(colour = Offense.Type), size = 2) +
    theme(axis.text.x = element_text(angle = 90,
        hjust = 1))
```



Extra. . .

I wanted to try different setup for data analysis of the same data. I installed and used *elasticsearch. kibana* and *logstash* for storing, retrieving and visualizing the data (last 10 years data). I found it quite effective. Once you import your data into elasticsearch, you can create any visualization you want in kibana, and by just configuring filters, all visualization panels are updated automatically. Here is the link for the URL that you can access the analysis results.

http://datatistics.blogspot.com/