Crime in Philadelphia: Data Analysis Summary using R

Bobby Arrington

barringtontx

10 June, 2019

Introduction

```
In this is "Capstone Project: "Choose Your Own", We will summarise a crime dataset for the
city of Philadelphia to answer the following research questions:

  1. How has crime evolved over time in the city of Philadelphia?
  2. What time of day do most crime occur?
  3. Which districts are more potentially dangerous?
  4. What Day(S) have highest crime rate and type?
```

Data and Method/Analysis

```
The analysis is done with a sample of the crime dataset from [Philadelphia Crime Data]
("https://kaggle.com/mchirico/philadelphiacrimedata/version/19") which contains all
the crime incidents that occurred in city of Philadelphia from 2006 to 2017.

We first prepare the data, creating, transforming and cleaning the variables we are
interested in. Then, we perform statistical analysis through line graphs, bar graphs and
heat-maps which will answer each of our research questions.

The report has been done with **R markdown** and it is code reproducible.
```

**Results*

```r
#Activate Libraries
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```r
library(zoo)
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(knitr)
library(plotly)
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------------

## v tibble  2.1.2     v purrr   0.3.2
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts -----------------------------------------------------------------------------------------
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x plotly::filter()         masks dplyr::filter(), stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

Second, we load the data into the R environment.

```
# Read csv in R
pdx = read.csv("https://cyo.arringtonadventures.com/crime/crime.csv",header = T)
```

Third, we create three new variables **Count**, **Year** and **Day**.
Additionally, we will rename District from Dc_Dist.

```
# Create a variable count with value 1
pdx$Count <- 1

# Extract year from Date
pdx$Year <- substring(pdx$Dispatch_Date,1,4)

# Extract Day from Dispatch_Date
pdx$Day <- wday(ymd(pdx$Dispatch_Date))
```

```r
# Rename District from Dc_Dist
colnames(pdx)[1] <- "District"
```

Fourth, we group Category from existent in the variables Text_General_Code.

```r
pdx$Category[pdx$Text_General_Code == "Thefts" |
             pdx$Text_General_Code == "Motor vehicle theft" |
             pdx$Text_General_Code == "Theft from vehicle" |
             pdx$Text_General_Code == "Recovered stolen motor vehicle" |
             pdx$Text_General_Code == "Embezzlement" |
             pdx$Text_General_Code == "Forgery and counterfeiting" |
             pdx$Text_General_Code == "Receiving stolen property"] <- "Thefts"
pdx$Category[pdx$Text_General_Code == "Vandalism/criminal mischief" |
             pdx$Text_General_Code == "Fraud"] <- "Criminal Damage"
pdx$Category[pdx$Text_General_Code == "Narcotic / drug law violations"] <- "Narcotics"
pdx$Category[pdx$Text_General_Code == "Other Assaults" |
             pdx$Text_General_Code == "Aggravated assault firearm"|
             pdx$Text_General_Code == "Aggravated assault no firearm"] <- "Assault"
pdx$Category[pdx$Text_General_Code == "Burglary non-residential" |
             pdx$Text_General_Code == "Burglary residential"] <- "Burglary"
pdx$Category[pdx$Text_General_Code == "Robbery no firearm" |
             pdx$Text_General_Code == "Robbery Firearm"]  <- "Robbery"
pdx$Category[pdx$Text_General_Code == "Arson" |
              pdx$Text_General_Code == "Gambling violations" |
              pdx$Text_General_Code == "Liquor law violations" |
              pdx$Text_General_Code == "Offenses against family and children" |
              pdx$Text_General_Code == "Public drunkenness" |
              pdx$Text_General_Code == "Vagrancy/loitering" |
              pdx$Text_General_Code == "Disorderly conduct" |
              pdx$Text_General_Code == "Prostitution and commercialized vice" |
              pdx$Text_General_Code == "Other sex offenses (not commercialized)" |
              pdx$Text_General_Code == "Rape" |
              pdx$Text_General_Code == "Weapon violations"|
              pdx$Text_General_Code == "Homicide - criminal" |
              pdx$Text_General_Code == "All other offenses" |
              pdx$Text_General_Code == "Homicide - gross negligence" |
              pdx$Text_General_Code == "Homicide - justifiable" |
              pdx$Text_General_Code == "Disorderly conduct" |
              pdx$Text_General_Code == "Offenses against family and children" |
              pdx$Text_General_Code == "Other assaults" |
              pdx$Text_General_Code == "Driving under the influence" |
              pdx$Text_General_Code == "Public Drunkenness" |
              pdx$Text_General_Code == "Homicide - Criminal" |
              pdx$Text_General_Code == "Burglary non-residential"] <- "Others"
```

The next step is to drop all those columns we do not need to answer our research questions.

```r
# Drop all variables we are not interested in
pdx <- pdx[, -c(2,3,5,7,8,9,11,12,13,14)]
```

Then, we clean the dataset of missing values and remove all values from 2017 - this last year is not complete.

```r
# Remove NAs
pdx <- pdx[complete.cases(pdx),]
```

```r
# Remove 2017 rows
pdx <- pdx[!pdx$Year > 2016,]
```

Finally, we show the the dataset ready for exploration.

```r
# Show first 6 records
head(pdx)
```

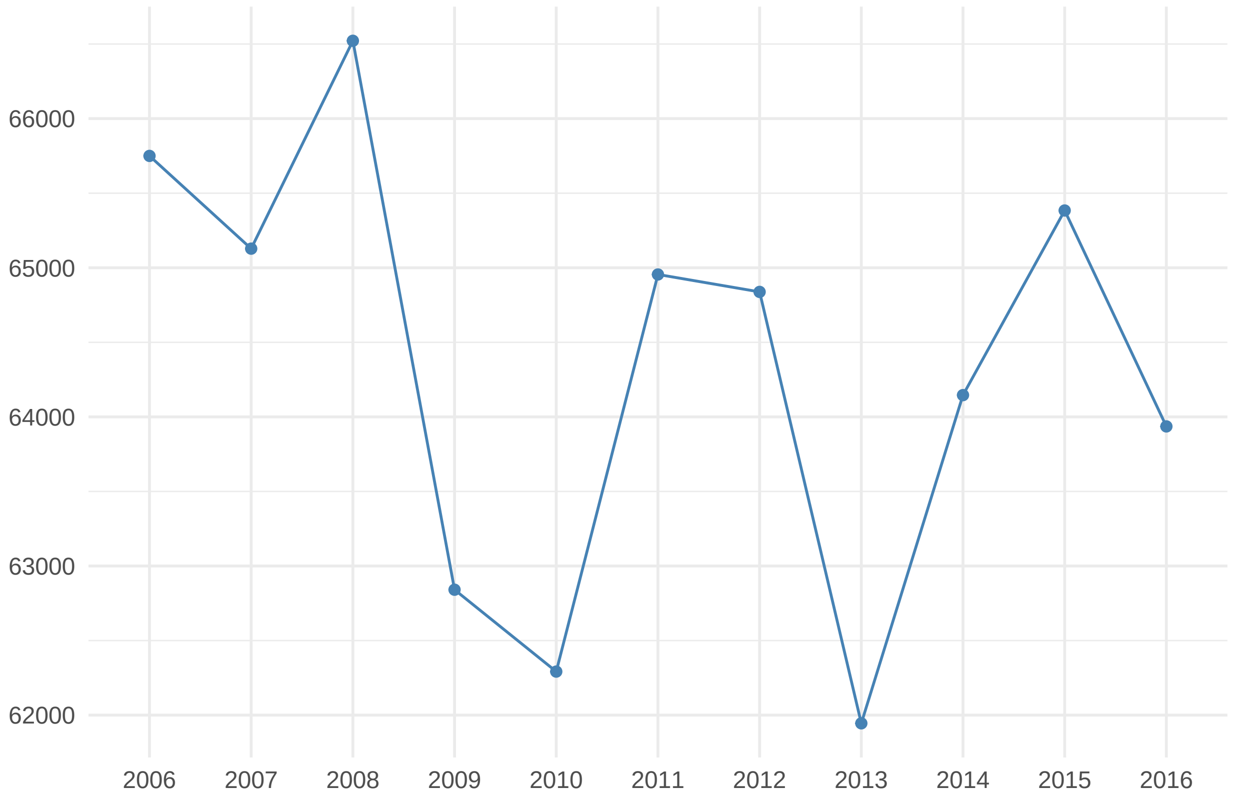
```
##    District Dispatch_Date Hour Text_General_Code Count Year Day Category
## 1         18    2009-10-02   14    Other Assaults     1 2009   6  Assault
## 3         25    2009-08-07   15    Other Assaults     1 2009   6  Assault
## 6         17    2015-04-25   12            Thefts     1 2015   7   Thefts
## 7         23    2009-02-10   14    Other Assaults     1 2009   3  Assault
## 12         1    2009-02-09   22    Other Assaults     1 2009   2  Assault
## 13        22    2015-10-06   18            Thefts     1 2015   3   Thefts
```

Data Exploration

How has crime evolved over time in the city of Philadelphia?

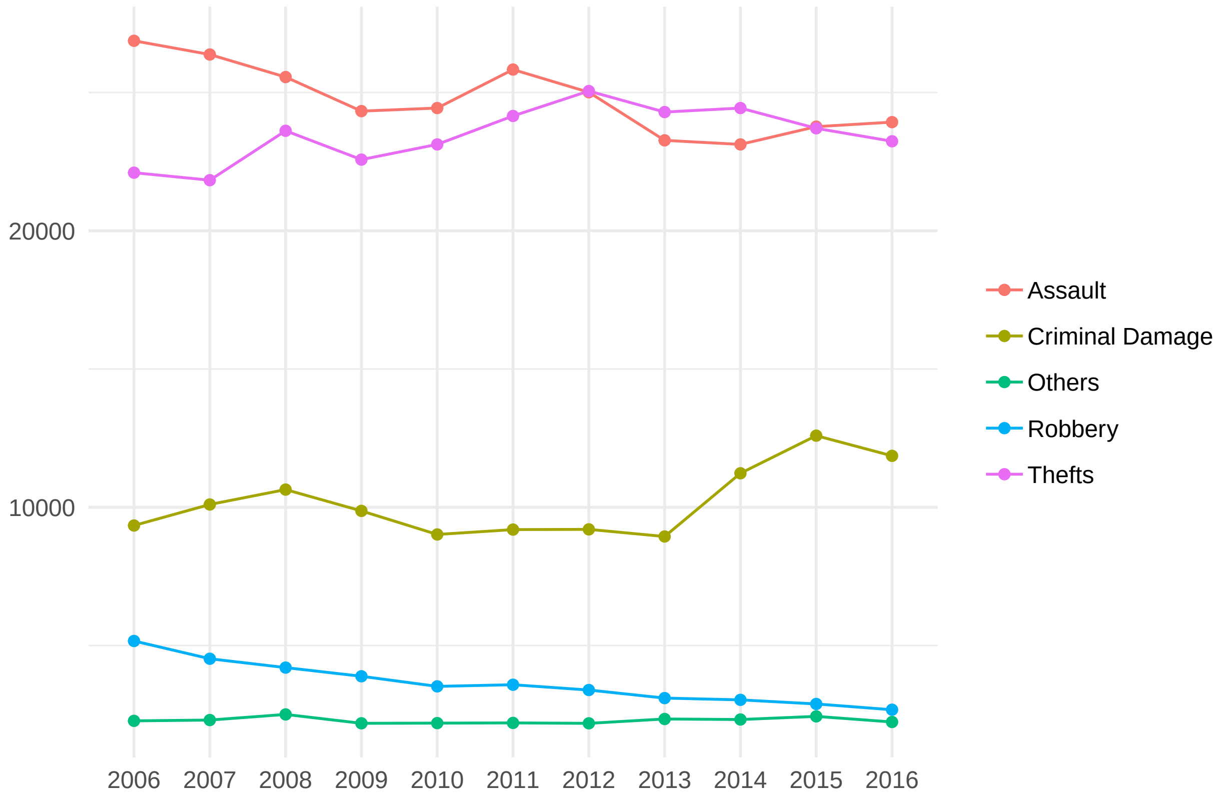
```r
#To answer this question we plot the number of crimes per year from 2006 to 2016.
#The graph shows that crime in the city of Philadelphia have a drop in 2007 followed
#by a spike in 2008 to highest level in 2008. In 2010 and 2013 it drop to its lowest
#levels. It spiked again in 2015, but it never reached the high of 2008.


# Create aggregated object

dd_aggr <- aggregate(Count ~ Year, data = pdx, FUN = sum)

# Plot the graph
ggplot(dd_aggr, aes(x=Year, y= Count, group = 1)) +
  geom_line(colour = "steelblue") +
  geom_point(colour = "steelblue") +
  theme_minimal() +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  ggtitle("Figures 1. Crimes evolution 2006-2016")
```

Figures 1. Crimes evolution 2006−2016



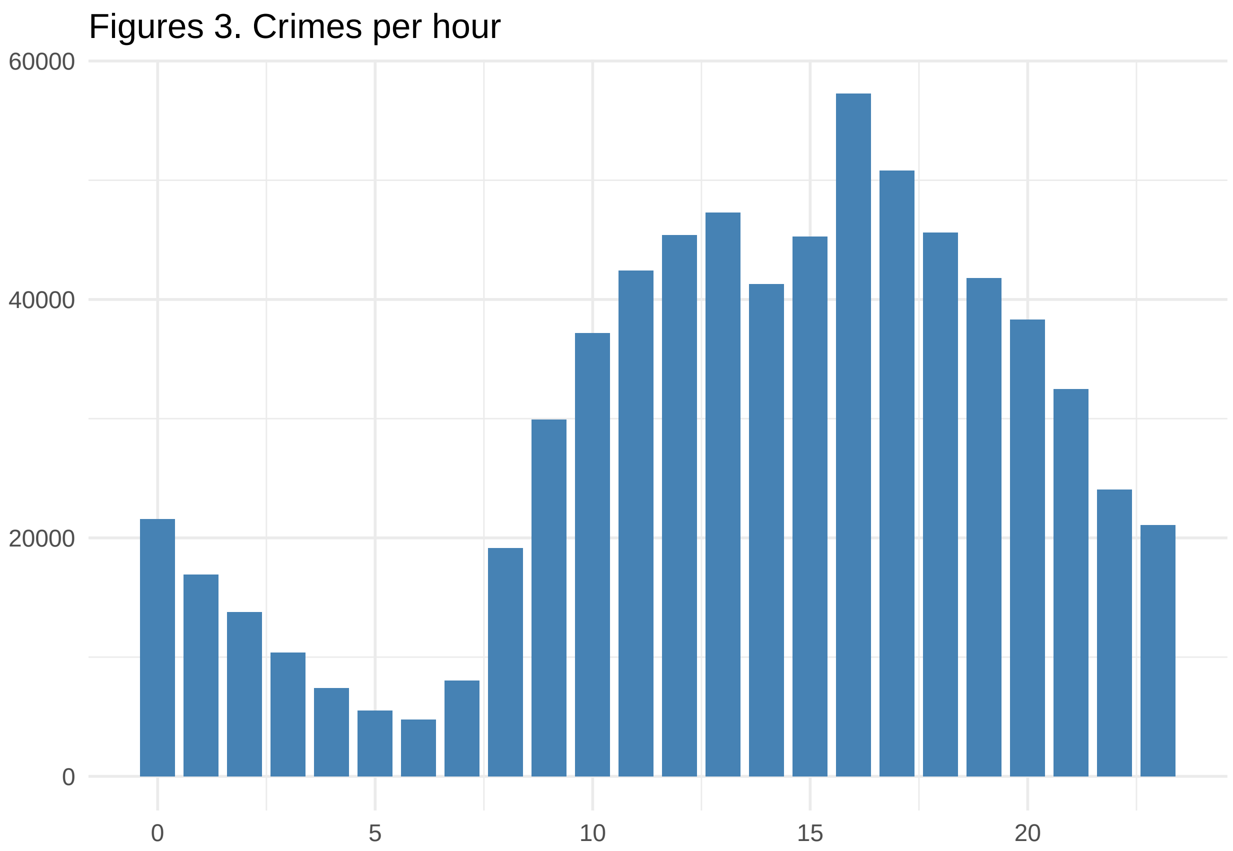
```
## Figure 2 depicts the annual frequency of crimes per type and their trend.
## The most common types of crime are Theft and Assaults seen to run in parallel
## as the top crimes between 2006-2016. Whereas, Robbery and Other are at the
## chart.

# Create aggregated object

dd_aggr2 <- aggregate(Count ~ Category + Year, data = pdx, FUN = sum)

# Plot the graph

ggplot(data=dd_aggr2, aes(x=Year, y=Count, group = Category, colour = Category)) +
  geom_line() +
  geom_point() +
  theme_minimal() +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  theme(legend.title=element_blank()) +
  ggtitle("Figures 2. Crimes evolution per type of crime 2006-2016")
```

Figures 2. Crimes evolution per type of crime 2006−2016

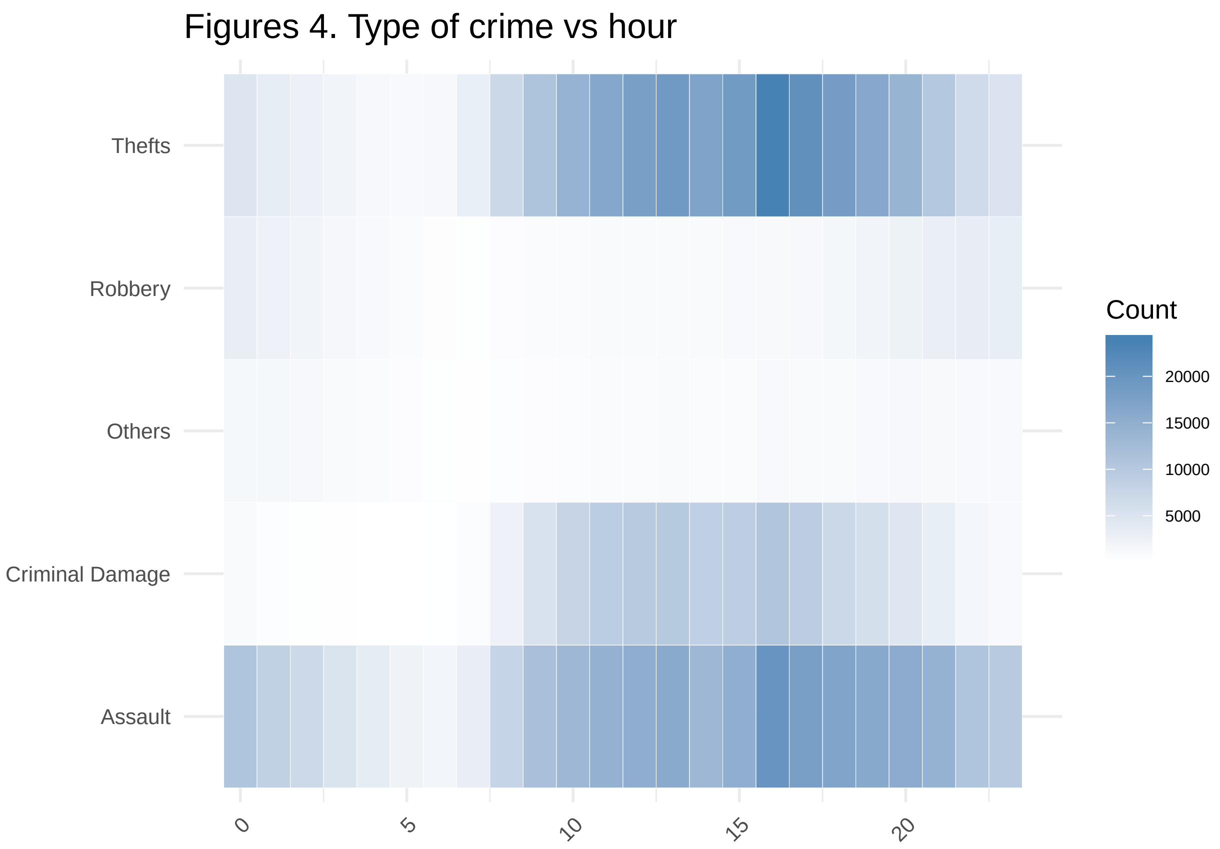


What time of day do most crime occur?

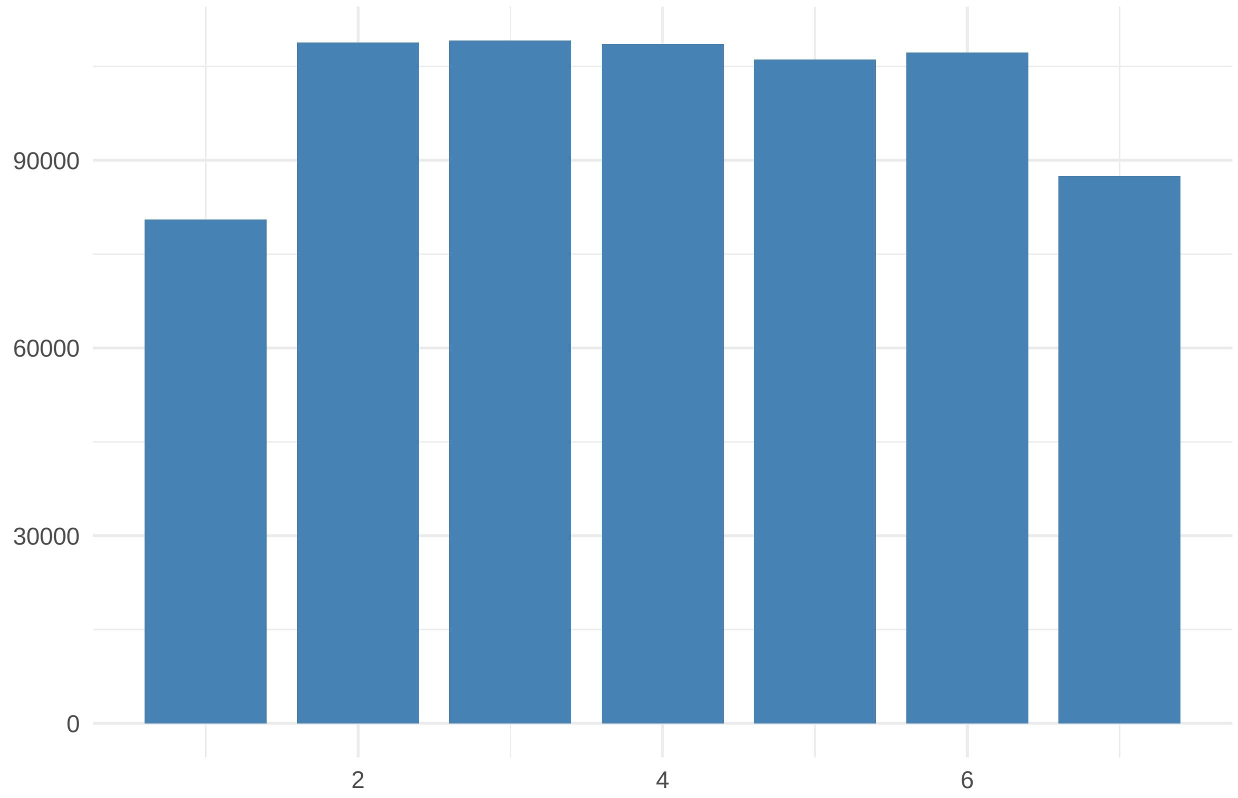
```
## The following bar graph (Figure 3) shows the number of crimes increases gradually
## from 06:00 in the morning (the hour with less crimes) until 16:00 in the afternoon
## (the hour with the most crimes). The hours of 18:00 and 05:00 shows a stead decrease
## in crime.


ggplot(pdx, aes(x=Hour)) +
  geom_bar(stat="Count", width=0.8, fill = "steelblue") +
  theme(axis.text.x = element_text(angle = 0, hjust = 1)) +
  labs(x = "Hour", y = "Number of crimes") +
  theme_minimal() +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  ggtitle("Figures 3. Crimes per hour")
```



```
## The heat-map in (Figure 4) shows the distribution of number of crimes per hour and type.
## For example, we can see that the peak hours of Theft and Others are at 16:00, 17:00 and 12:00.
## Assaults concentrate between 08:00 to 14:00 and 16:00 to 22:00. Other types are more evenly
## distributed throughout the day.

dd_aggr3 <- aggregate(Count ~ Category + Hour, data = pdx, FUN = sum)

# Plot graph

p1 <- ggplot(data = dd_aggr3, aes(x = Hour, y = Category)) +
  geom_tile(aes(fill = Count), color = "white") +
  scale_fill_gradient(low = "white", high = "steelblue")

p1 + theme_minimal() +
  theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 6),
        axis.text.y = element_text(size= 8),
        axis.text.x = element_text(size = 8, angle = 45, hjust = 1)) +
  ggtitle("Figures 4. Type of crime vs hour")
```

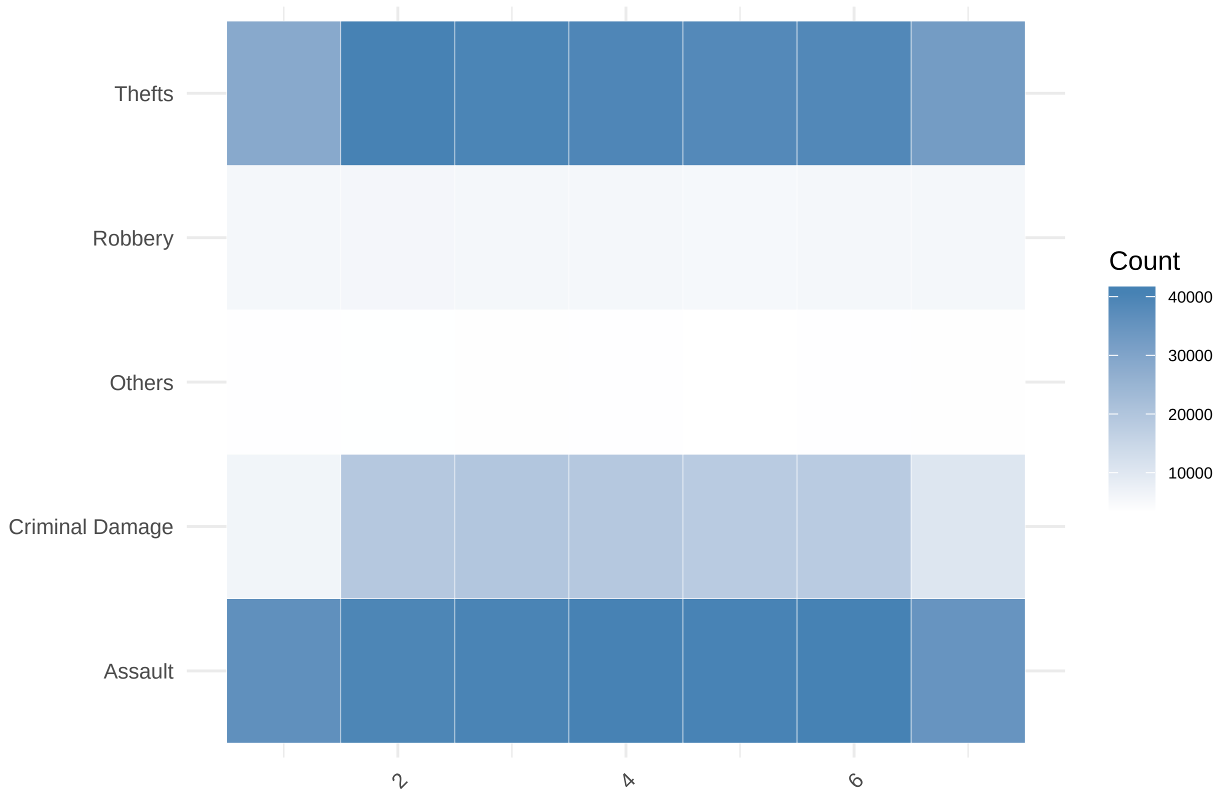


```
## what day of week do most crime occur?
## The bar-chart in (Figure 5) shows the distribution of number of crimes per day.
## For example, we can see that Monday, Tuesday and Wednesday are about even.
## 1 = Sunday, 2 = Monday, 3 = Tuesday, 4 = Wednesday, 5 = Thusday, 6 = Friday, 7 = Saturday

ggplot(pdx, aes(x=Day)) +
  geom_bar(stat="Count", width=0.8, fill = "steelblue") +
  theme(axis.text.x = element_text(angle = 0, hjust = 1)) +
  labs(x = "Day", y = "Number of crimes") +
  theme_minimal() +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
ggtitle("Figures 5. Crimes per day")
```


Figures 5. Crimes per day

```r
##  The heat-map show distrbution of crimes per day and type
## The heat-map in (Figure 5) shows the distribution of number of crimes vs day.
## For example, we can see that Monday, Tuesday and Wednesday are about even
## for Thefts and Assualts.
## 1 = Sunday, 2 = Monday, 3 = Tuesday, 4 = Wednesday, 5 = Thusday, 6 = Friday, 7 = Saturday

dd_aggr2a <- aggregate(Count ~ Category + Day, data = pdx, FUN = sum)

# Plot graph

p1 <- ggplot(data = dd_aggr2a, aes(x = Day, y = Category)) +
  geom_tile(aes(fill = Count), color = "white") +
  scale_fill_gradient(low = "white", high = "steelblue")

p1 + theme_minimal() +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 6),
        axis.text.y = element_text(size= 8),
        axis.text.x = element_text(size = 8, angle = 45, hjust = 1)) +
  ggtitle("Figures 6. Type of crime vs day")
```

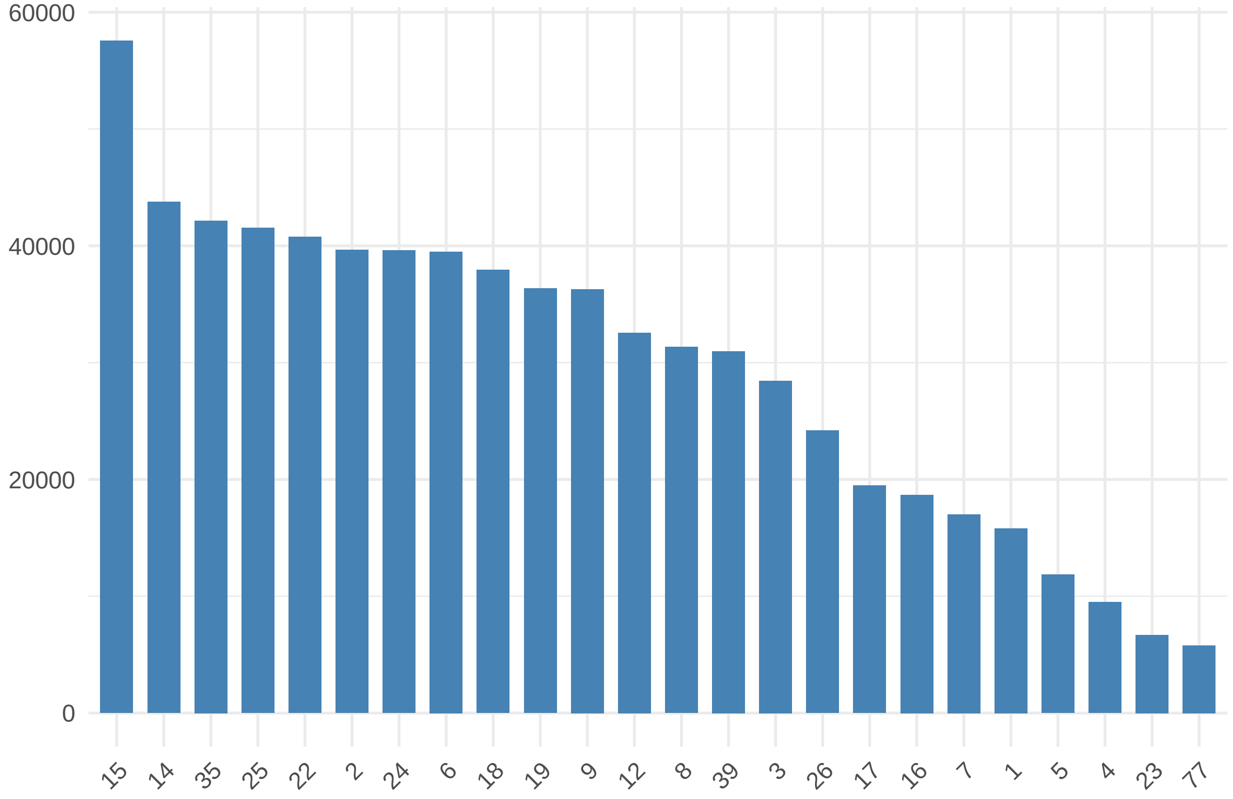
Figures 6. Type of crime vs day



Which districts are more potentially dangerous?

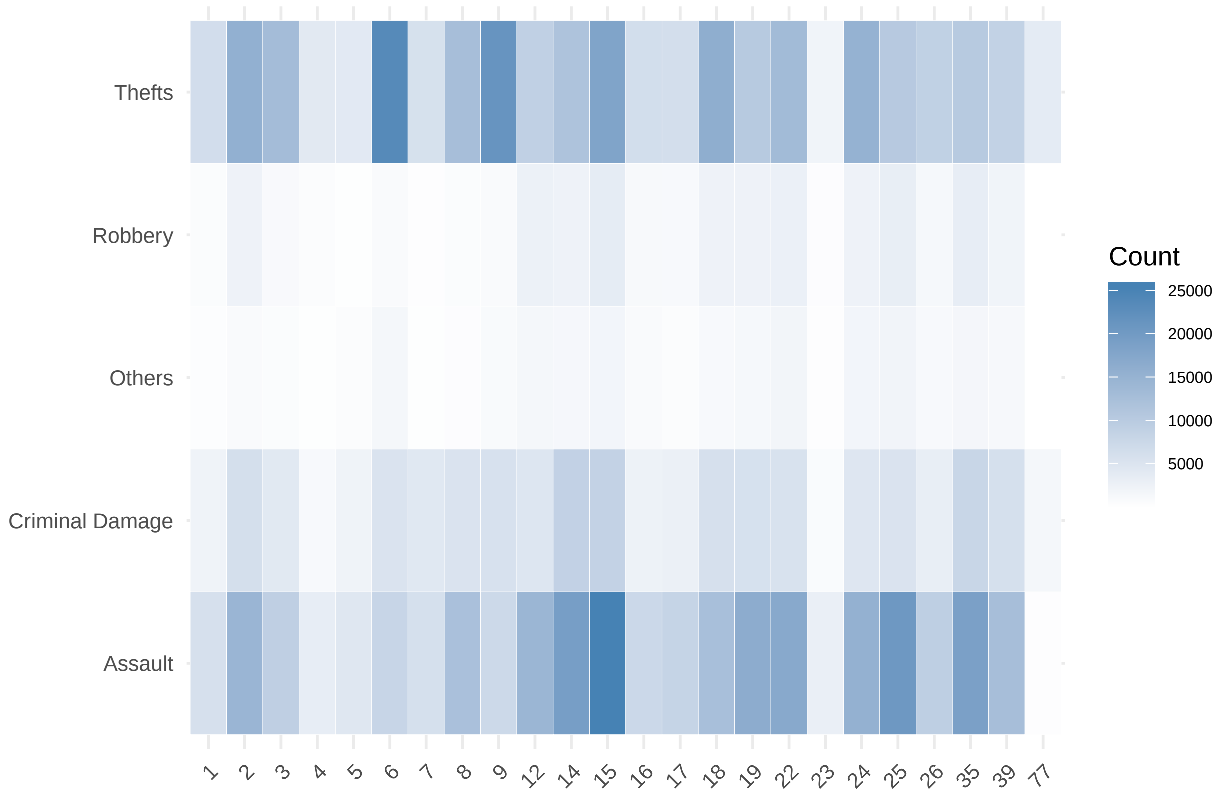
```
# We visualise the number of crimes per districts. The most dangerous district seems
# number 15, with more than 150000 records in the 10 years, while district 77 with
# 5000 seems the safest.

pdx$District<- as.factor(pdx$District)
dd_sub <- subset(pdx, District!="92")

# Create aggregated object
dd_aggr6 <- aggregate(Count ~ District, data = dd_sub, FUN = sum)

# Order values
dd_aggr6$District <- factor(dd_aggr6$District, levels = dd_aggr6$District[order(-dd_aggr6$Count)])

# Plot the graph
ggplot(dd_aggr6, aes(x=District, y = Count)) +
  theme_minimal() +
  geom_bar(stat="identity", width=0.7, fill = "steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "District", y = "Number of crimes") +
  theme(axis.title.x=element_blank()) +
  theme(axis.title.y=element_blank()) +
  ggtitle("Figures 7. Crimes per district")
```

Figures 7. Crimes per district



```r
# There are some interesting findings in the relations between the type of crime and districts
# where they occurred. For example we can see that districts 6,9 and 15 are particularly,
# dangerous in terms of Theft that 14 and 15 stand out in terms of Assualt,

dd_aggr7 <- aggregate(Count ~ Category + District, data = dd_sub, FUN = sum)
# Plot the graph
p3<-ggplot(data = dd_aggr7, aes(x = District, y = Category)) +
  geom_tile(aes(fill = Count), color = "white") +
  scale_fill_gradient(low = "white", high = "steelblue")
p3+ theme_minimal()+ theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 6),
        axis.text.y = element_text(size= 8),
        axis.text.x = element_text(size = 8, angle = 45, hjust = 1)) +
  ggtitle("Figures 8. Type of crimes vs district")
```

Figures 8. Type of crimes vs district



Conclusion ‘” The research questions that were asked in the Introducton were answered in the charts and graphs in the results section above.

Summary: 1. Crimes per distict were the highest in district 15 and lowest in district 77. 2. Types of crimes vs day of the week were under **Assualt** followed by **Thefts*. 3. Crimes per day were slightly higher beginning Monday, Tuesday and Wednesday 4. Types of crimes vs hour under the **Thefts** and **Assualt** Category were highest 16:00 and 18:00 hours and between 22:00 and 01:00 hours. 5. Crimes evolution per type of crime beteeen 2000-2016 has shown as higher incident under the **Other** and **Assault** more than the other Cateories. While **Narcotics**, **Burglary**, **Criminal Damage**, **Robbery** are more or less linear. 6. Crimes evolution in general between 2006-2016 has shown as a marked decrease 2007 and 2013, but reaching it highest level in 2008.

Based on the above analysis, district **77** appears to be the safest and district **15** as potentially the most dangerous.

`**The End**`