Crime Analysis: San Francisco, Summer 2014

Prasad Bandaru

24 January 2016

The criminal incident data from San Francisco for the summer of 2014 has been made public. I would like to make an visual analysis on these incidents and try to answer 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 thefts most common?
- How do incidents vary by neighborhood? Which incidents are most common in the city center? In what areas or neighborhoods are robberies or thefts most common?
- How do incidents vary month to month in the Summer 2014 dataset?
- Which incident types tend to correlate with each other on a day-by-day basis?

Dataset overview

Load the required libraries and the san francisco crime incidents data.

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("lubridate")
library("ggplot2")
library("caret")
```

```
## Loading required package: lattice
```

```
san <- read.csv("sanfrancisco_incidents_summer_2014.csv")
```

Initial look at the data reveals that it contains 13 fields. Of these category, date, time and area details are important for my current analysis. There are 34 categories of crime as listed below.

```
str(san)
```

```
## 'data.frame':    28993 obs. of  13 variables:
##  $ IncidntNum: int  140734311 140736317 146177923 146177531 140734220 140734349 140734349 140734349
```

```
##  $ Category  : Factor w/ 34 levels "ARSON","ASSAULT",..: 1 20 16 16 20 7 7 6 21 30 ...
##  $ Descript  : Factor w/ 368 levels "ABANDONMENT OF CHILD",..: 15 179 143 143 132 247 239 93 107 347
##  $ DayOfWeek : Factor w/ 7 levels "Friday","Monday",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Date      : Factor w/ 92 levels "06/01/2014","06/02/2014",..: 92 92 92 92 92 92 92 92 92 92 ...
##  $ Time      : Factor w/ 1379 levels "00:01","00:02",..: 1370 1365 1351 1351 1344 1334 1334 1334 132
##  $ PdDistrict: Factor w/ 10 levels "BAYVIEW","CENTRAL",..: 1 4 8 7 7 8 8 8 3 2 ...
##  $ Resolution: Factor w/ 16 levels "ARREST, BOOKED",..: 12 12 12 12 12 1 1 1 12 2 ...
##  $ Address   : Factor w/ 8055 levels "0 Block of 10TH ST",..: 6843 4022 1098 6111 5096 1263 1263 1263
##  $ X         : num  -122 -122 -122 -122 -123 ...
##  $ Y         : num  37.7 37.8 37.8 37.8 37.8 ...
##  $ Location  : Factor w/ 8732 levels "(37.7080829769301, -122.419241455854)",..: 1970 3730 5834 4802
##  $ PdId      : num  1.41e+13 1.41e+13 1.46e+13 1.46e+13 1.41e+13 ...
```

```
levels(san$Category)
```

```
##  [1] "ARSON"                       "ASSAULT"
##  [3] "BRIBERY"                     "BURGLARY"
##  [5] "DISORDERLY CONDUCT"          "DRIVING UNDER THE INFLUENCE"
##  [7] "DRUG/NARCOTIC"               "DRUNKENNESS"
##  [9] "EMBEZZLEMENT"                "EXTORTION"
## [11] "FAMILY OFFENSES"             "FORGERY/COUNTERFEITING"
## [13] "FRAUD"                       "GAMBLING"
## [15] "KIDNAPPING"                  "LARCENY/THEFT"
## [17] "LIQUOR LAWS"                 "LOITERING"
## [19] "MISSING PERSON"              "NON-CRIMINAL"
## [21] "OTHER OFFENSES"              "PORNOGRAPHY/OBSCENE MAT"
## [23] "PROSTITUTION"                "ROBBERY"
## [25] "RUNAWAY"                     "SECONDARY CODES"
## [27] "STOLEN PROPERTY"             "SUICIDE"
## [29] "SUSPICIOUS OCC"              "TRESPASS"
## [31] "VANDALISM"                   "VEHICLE THEFT"
## [33] "WARRANTS"                    "WEAPON LAWS"
```

Enriching data

I enriched the data by converting the Date field to proper date format. Additionally I deduct Hour, Month and YDay (day of the year) and store them in separate fields. In this way it is easier to use the data in the analysis. Also, I add an additional count field and set it to 1. This field is for easy aggregation.

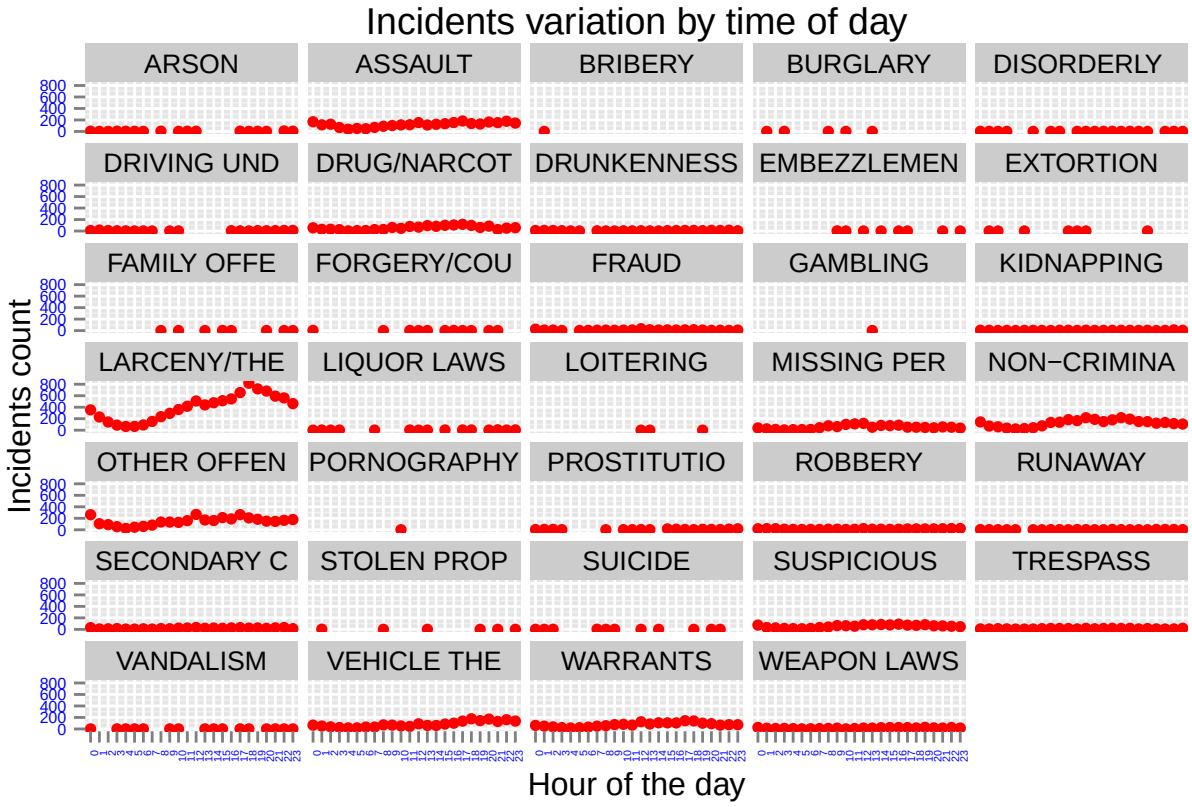
```
san$Date <- as.POSIXct(strptime(san$Date, format = "%m/%d/%Y"))
san$Hour <- as.factor(hour(as.POSIXlt(strptime(san$Time, format = "%H:%M"))))
san$Month <- as.factor(month(san$Date, label = TRUE))
san$YDay <- yday(san$Date)
san$Count <- c(1)
head(san$Date)
```

```
## [1] "2014-08-31 CEST" "2014-08-31 CEST" "2014-08-31 CEST" "2014-08-31 CEST"
## [5] "2014-08-31 CEST" "2014-08-31 CEST"
```

How incidents vary by time of day

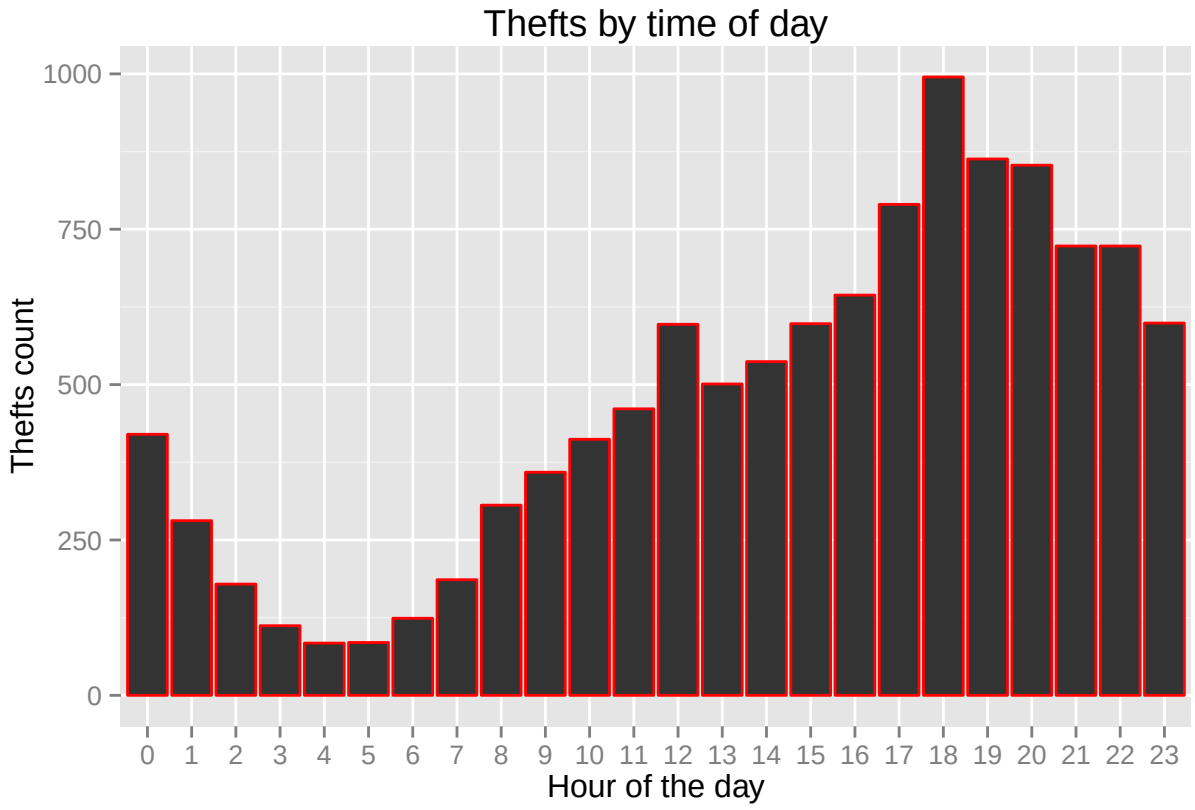
I consider only the Hour part of the time and plot number of incidents for each hour spread out for each category. The plot reveals that theft has a higher variation over the day. Also, the number of thefts are significantly higher in the evening.

```
counts <- aggregate(Count ~ Category + Hour, data = san, sum)
counts$Category <- substr(counts$Category, 1, 11)
ggplot(counts, aes(Hour, Count)) +
  facet_wrap(~ Category, ncol = 5) +
  labs(title = "Incidents variation by time of day", x = "Hour of the day", y = "Incidents count") +
  theme(axis.text.y = element_text(size = 6, color = "blue")) +
  theme(strip.text = element_text(size = 10)) +
  theme(axis.text.x = element_text(size = 4, angle = 90, hjust = 1, color = "blue")) +
  geom_point(color = c("red"))
```



The number of thefts increases slowly over the day and peaks around 6pm. Then it gradually comes down and is at minimum between 4 am and 5 am.

```
robbery <- san[san$Category == "VEHICLE THEFT" | san$Category == "LARCENY/THEFT",]
ggplot(robbery, aes(Hour)) +
  labs(title = "Thefts by time of day", x = "Hour of the day", y = "Thefts count") +
  geom_histogram(color = c("red"))
```



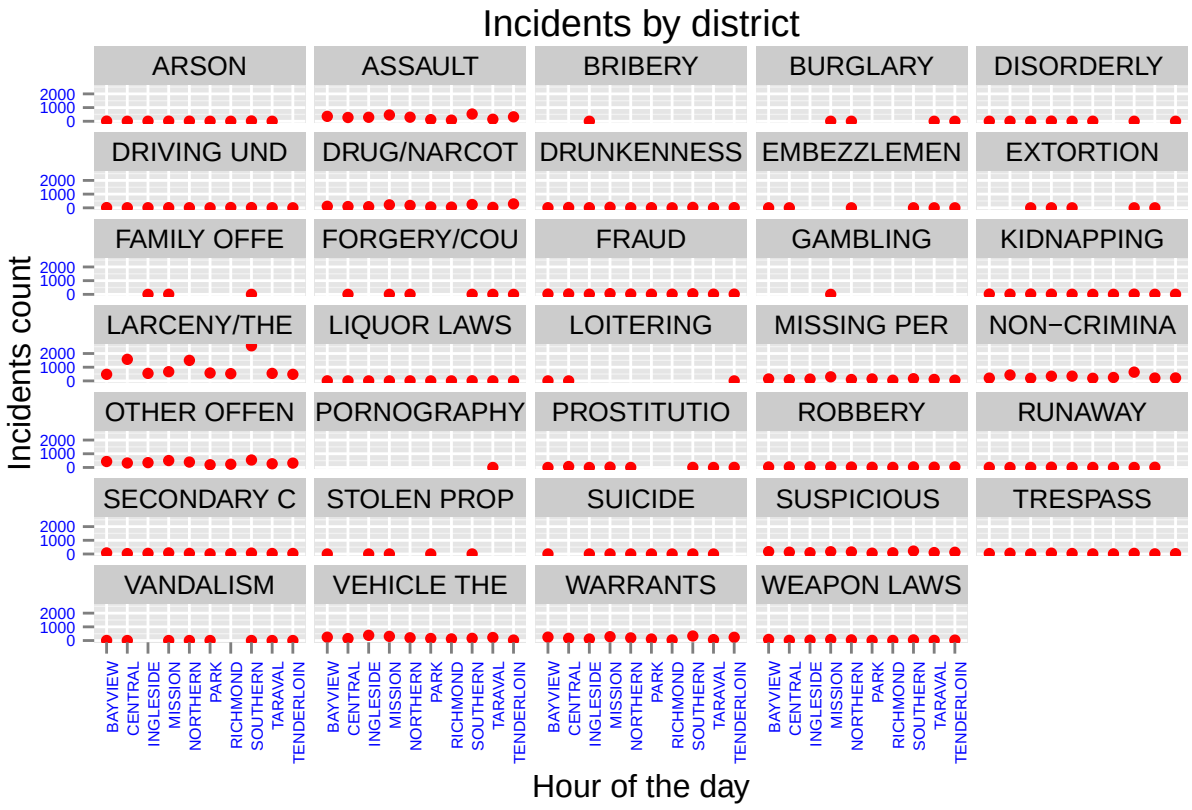
How incidents vary by neighborhood

Further I split the incidents by district. It reveals that thefts are more in central, northern and southern districts compared to other districts.

```
areawise <- aggregate(Count ~ Category + PdDistrict, data = san, sum)
levels(areawise$PdDistrict)
```

```
##  [1] "BAYVIEW"    "CENTRAL"    "INGLESIDE"  "MISSION"    "NORTHERN"
##  [6] "PARK"       "RICHMOND"   "SOUTHERN"   "TARAVAL"    "TENDERLOIN"
```

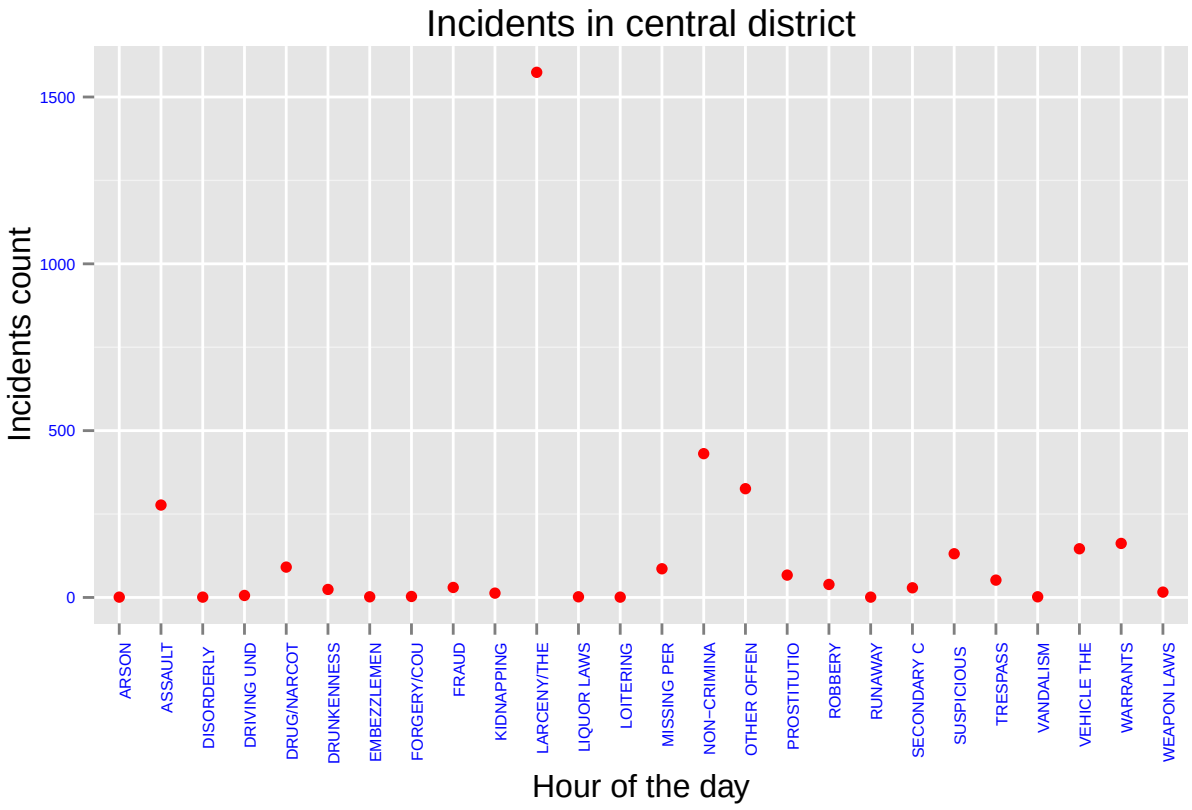
```
areawise$Category <- substr(areawise$Category, 1, 11)
ggplot(areawise, aes(PdDistrict, Count)) +
  facet_wrap(~ Category, ncol = 5) +
  labs(title = "Incidents by district", x = "Hour of the day", y = "Incidents count") +
  theme(axis.text = element_text(size = 6, color = "blue")) +
  theme(strip.text = element_text(size = 10)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_point(color = c("red"))
```



How incidents vary in the central district

The following chart gives a better view of the incidents in the central district.

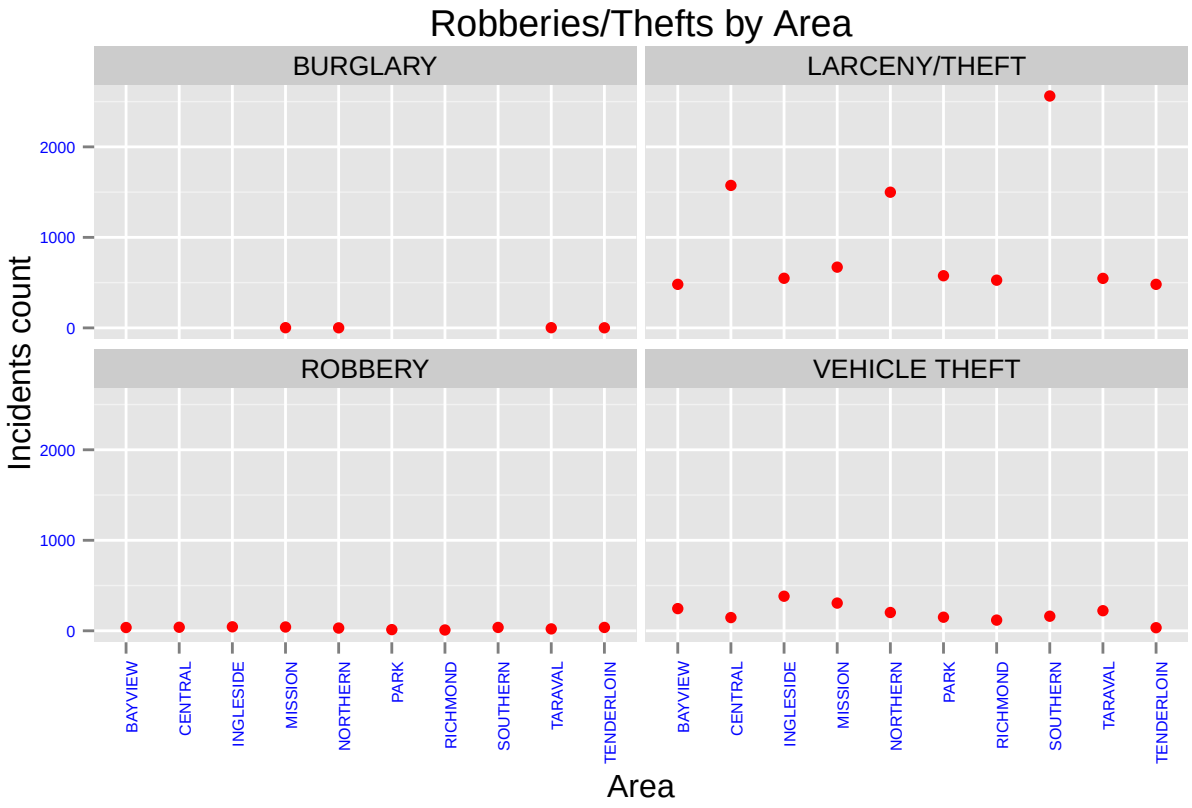
```
central <- areawise[areawise$PdDistrict == "CENTRAL", ]
ggplot(central, aes(Category, Count)) +
  labs(title = "Incidents in central district", x = "Hour of the day", y = "Incidents count") +
  theme(axis.text = element_text(size = 6, color = "blue")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_point(color = c("red"))
```



Areas with most common robberies or thefts

I filtered the categories: ROBBERY, VEHICLE THEFT, LARCENCY/THEFT and BURGLARY, to find the areas where these incidents are common. Robbery is common in all districts, while Burglary happened only in the districts: MISSION, NORTHERN, TARAVAL and TENDERLOIN. LARCENCY/THEFT is higher among these categories and is particular higher in districts: CENTRAL, NORTHERN and SOUTHERN.

```
thefts <- san[san$Category == "ROBBERY" | san$Category == "VEHICLE THEFT" |
                san$Category == "LARCENY/THEFT" | san$Category == "BURGLARY", ]
common <- aggregate(Count ~ Category + PdDistrict, data = thefts, sum)
ggplot(common, aes(PdDistrict, Count)) +
  labs(title = "Robberies/Thefts by Area", x = "Area", y = "Incidents count") +
  facet_wrap(~ Category) +
  theme(axis.text = element_text(size = 6, color = "blue")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_point(color = c("red"))
```



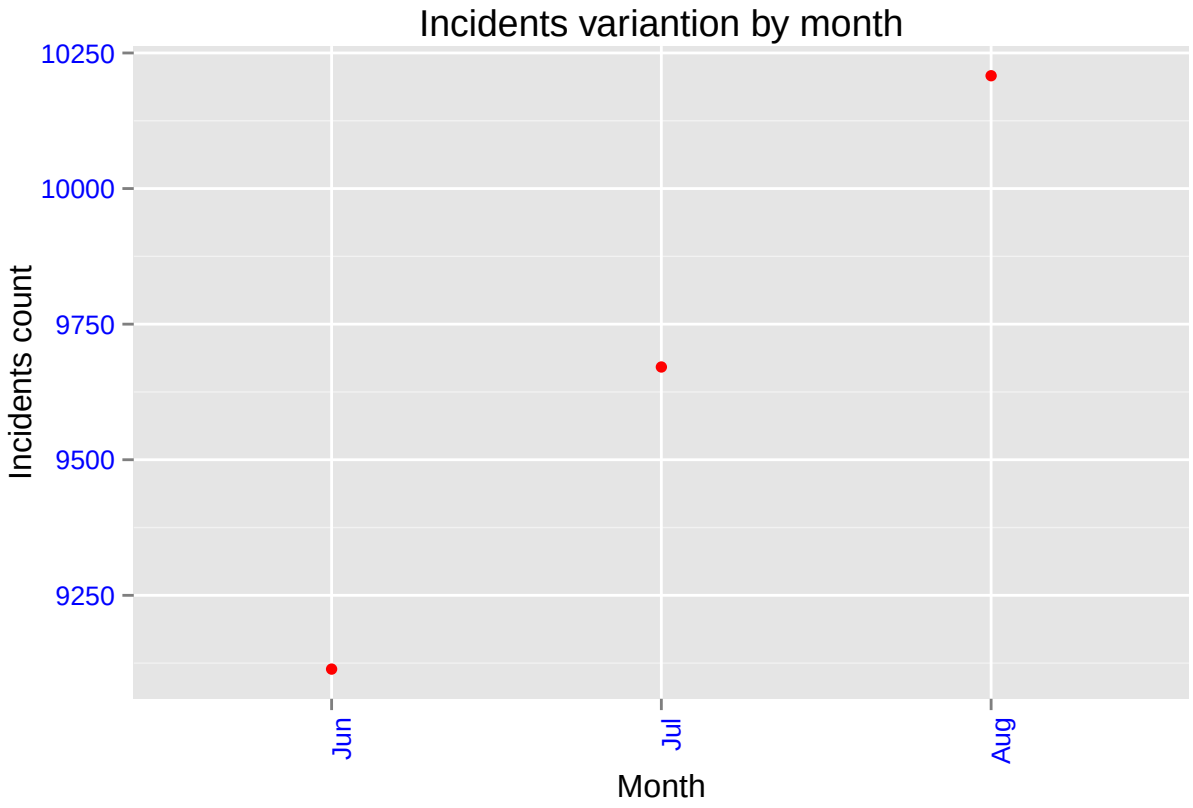
Incidents variation for the given Summer months

The dataset consists only of the summer months: June, July and August. Plotting the total incidents per month reveals that the crimes are linearly increasing. This means that there is a need for increased police officials in July compared to June and some more increase for police officials in August compared to July.

```
levels(san$Month)
```

```
##  [1] "Jan" "Feb" "Mar" "Apr" "May" "Jun" "Jul" "Aug" "Sep" "Oct" "Nov"
## [12] "Dec"
```

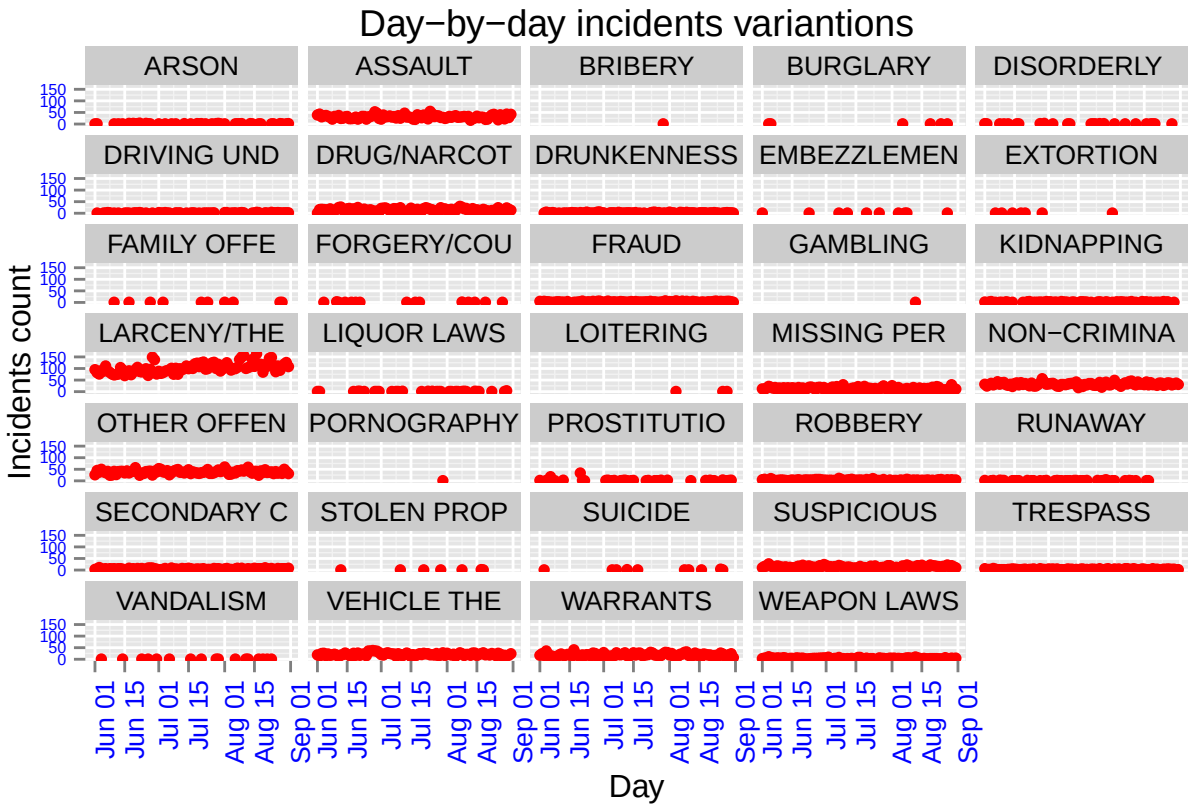
```
months <- aggregate(Count ~ Month, data = san, sum)
ggplot(months, aes(Month, Count)) +
  labs(title = "Incidents variantion by month", x = "Month", y = "Incidents count") +
  theme(axis.text = element_text(size = 10, color = "blue")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_point(color = c("red"))
```



Incident types correlation on a day-by-day basis

Plotting the various incidents on a day-by-day basis tends to show many correlations, say, between ARSON, KIDNAPPING, FRAUD, etc. But, it's difficult so say this which ones have better correlation from such a graph.

```
daily <- aggregate(Count ~ Date + Category, data = san, sum)
daily$Category <- substr(daily$Category, 1, 11)
ggplot(daily, aes(Date, Count)) +
  facet_wrap(~ Category, ncol = 5) +
  labs(title = "Day-by-day incidents variantions", x = "Day", y = "Incidents count") +
  theme(axis.text.y = element_text(size = 6, color = "blue")) +
  theme(axis.text.x = element_text(size = 10, color = "blue", angle = 90, hjust = 1)) +
  geom_point(color = c("red"))
```



Finding top 10 correlations

With the help of functions from the caret library and using the correlation function (cor), an attempt is made to find the top 10 correlations: the top 5 positive and the top 5 negative correlations.

```
sdaily <- subset(san, select = c(Category, YDay))
dmy <- dummyVars("~ .", data = sdaily)
ndaily <- data.frame(predict(dmy, newdata = sdaily))
cdaily <- cor(ndaily)
fdaily <- as.data.frame(as.table(cdaily))
sset <- subset(fdaily, Var1 != Var2)
```

The top 5 positive correlations are:

```
head(sset[order(sset$Freq, decreasing=TRUE),], 5)
```

```
##                                       Var1
## 560                                   YDay
## 1206                Category.LARCENY.THEFT
## 210                                   YDay
## 1196 Category.DRIVING.UNDER.THE.INFLUENCE
## 630                                   YDay
##                                       Var2       Freq
## 560                 Category.LARCENY.THEFT 0.05418746
## 1206                                  YDay 0.05418746
## 210  Category.DRIVING.UNDER.THE.INFLUENCE 0.01356544
```

```
## 1196                                YDay 0.01356544
## 630                  Category.LOITERING 0.01264506
```

The top 5 negative correlations are:

```
head(sset[order(sset$Freq),], 5)
```

```
##                         Var1                      Var2       Freq
## 546 Category.OTHER.OFFENSES  Category.LARCENY.THEFT -0.2607822
## 716  Category.LARCENY.THEFT Category.OTHER.OFFENSES -0.2607822
## 545   Category.NON.CRIMINAL  Category.LARCENY.THEFT -0.2375465
## 681  Category.LARCENY.THEFT   Category.NON.CRIMINAL -0.2375465
## 51   Category.LARCENY.THEFT        Category.ASSAULT -0.2313134
```