A brief criminal incident analysis in San Francisco of Summer 2014

Prasad Bandaru

13 February 2016

The criminal incident data from San Francisco for the summer of 2014 has been made public. I made a visual analysis on these incidents to find out the top 5 types of crime based on the time of the day and make following analysis. * figure out the districts where these crime are most * pickup the top crime and the corresponding district and analyse it for given period in the input data and also by day-of-week * finally figure out an exact location where the top crime is most concentrated

This report is created using Rmarkdown and the source code is available linked github

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

Dataset overview

Load the san francisco crime incidents data. Initial look at the data reveals that it contains 13 fields. Of these category, date, time, dayofweek and location details are important for my current analysis. There are 34 categories of crime.

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

```
## 'data.frame':    28993 obs. of  13 variables:
##  $ IncidntNum: int  140734311 140736317 146177923 146177531 140734220 140734349 140734349 140734349 1
##  $ 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 126
##  $ 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 ...
```

```r
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

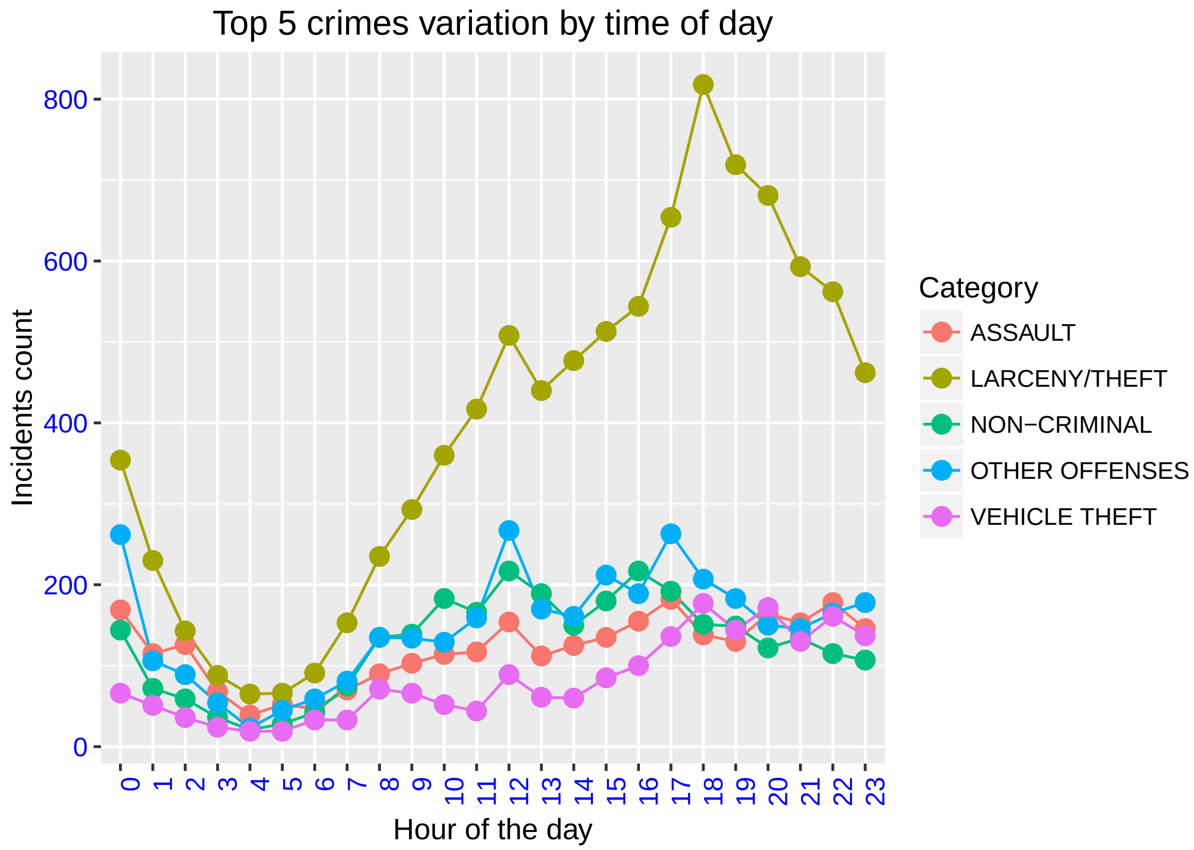
I enriched the data by converting the Date field to proper date format. Additionally I deduct Hour and store them in a separate field. I also re-arranged the DayOfWeek from Monday to Sunday. I included an additional count field to make it easier for aggregation.

```r
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$DayOfWeek <- factor(san$DayOfWeek,
                 levels = c("Monday", "Tuesday",  "Wednesday", "Thursday",
                            "Friday", "Saturday", "Sunday"))
san$Count <- c(1)
```

Top crimes variation by time of day

I plot the hourly incidents count for the top 5 crime category. The plot reveals that theft has a very high variation over the day and also it is most committed crime. The number of thefts increases slowly over the day and peaks at 6pm. Then it gradually comes down and is at minimum between 4 am and 5 am.

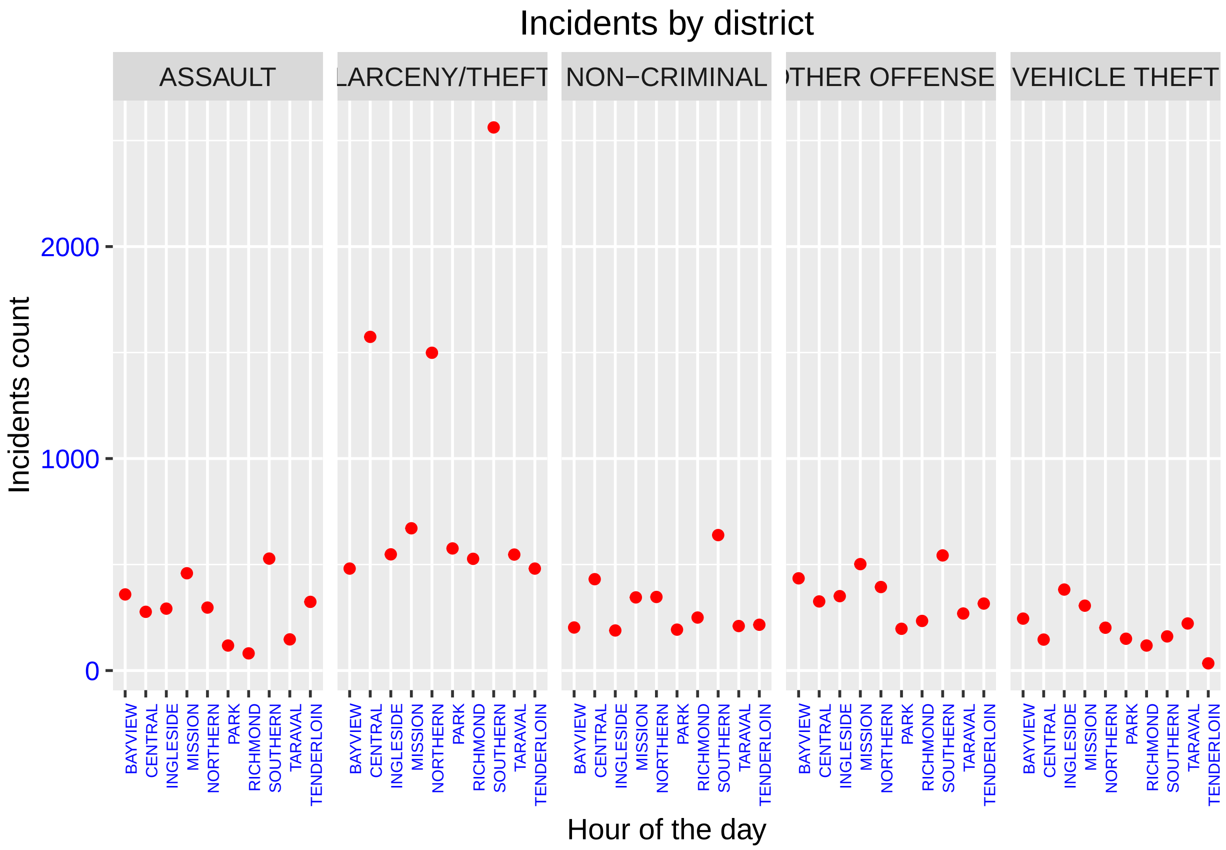
```r
counts <- aggregate(Count ~ Category + Hour, data = san, sum)
topCrimes <- aggregate(Count ~ Category, data = san, sum)
topX <- topCrimes$Category [order (topCrimes$Count, decreasing = TRUE)]
#counts$Category <- substr(counts$Category, 1, 11)
ggplot(data = subset(counts, Category %in% topX [1 : 5]), aes(Hour, Count)) +
  labs(title = "Top 5 crimes variation by time of day", x = "Hour of the day", y = "Incidents count") +
  theme(axis.text.y = element_text(size = 10, color = "blue")) +
  theme(strip.text = element_text(size = 10)) +
  theme(axis.text.x = element_text(size = 10, angle = 90, hjust = 1, color = "blue")) +
  geom_line(aes(colour=Category, group=Category)) +
  geom_point(aes(colour=Category), size=3)
```



Top crimes variation by district

Further splitting the top 5 incidents by district reveals that the number of thefts are more in central, northern and southern districts compared to other districts. In the southern district the thefts are 5 times more than other lower theft districts. So, we will dig more into the theft in the southern district.

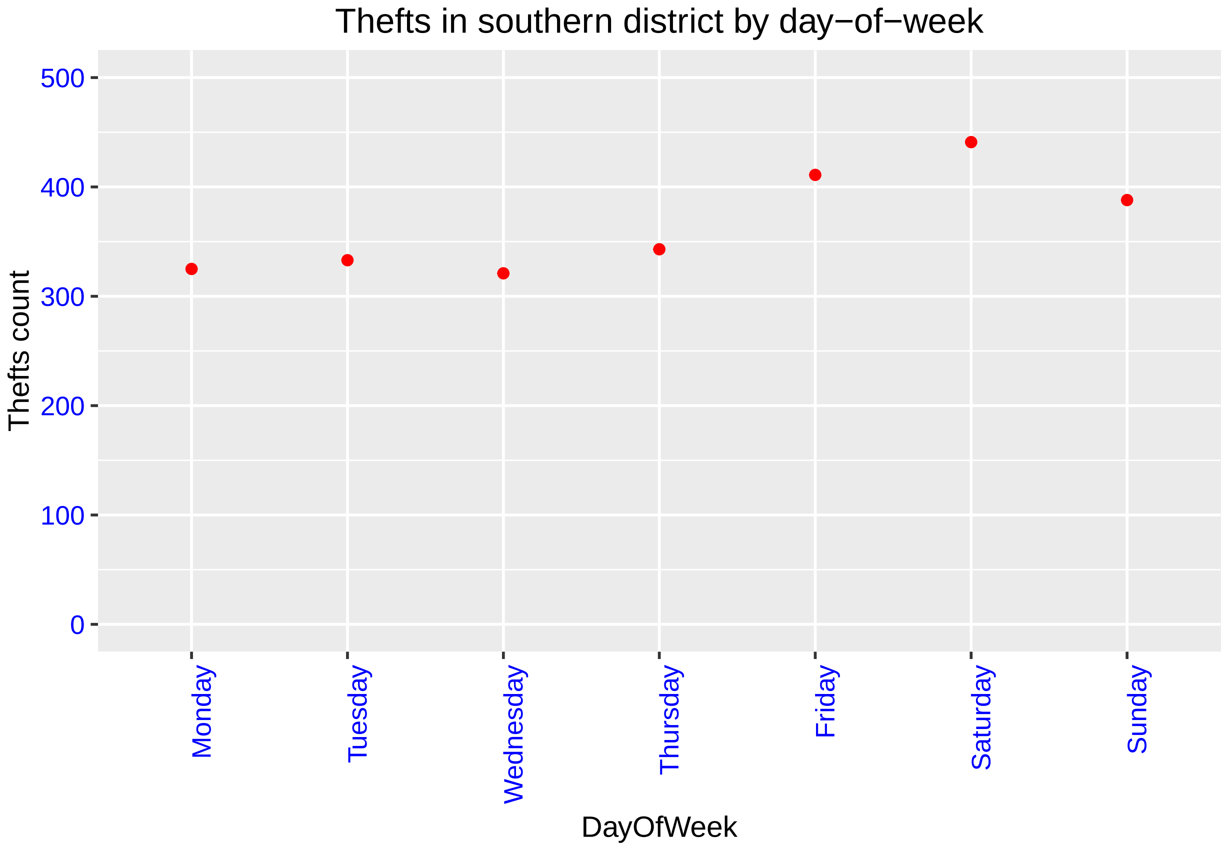
```
areawise <- aggregate(Count ~ Category + PdDistrict, data = san, sum)
ggplot(data = subset(areawise, Category %in% topX [1 : 5]), aes(PdDistrict, Count)) +
  facet_grid(~ Category) +
  labs(title = "Incidents by district", x = "Hour of the day", y = "Incidents count") +
  theme(axis.text = element_text(size = 10, color = "blue")) +
  theme(strip.text = element_text(size = 10)) +
  theme(axis.text.x = element_text(size = 6, angle = 90, hjust = 1)) +
  geom_point(color = c("red"))
```


Incidents by district
ASSAULT
LARCENY/THEFT
NON−CRIMINAL
OTHER OFFENSE
VEHICLE THEFT
Incidents count
2000
1000
0
BAYVIEW
CENTRAL
INGLESIDE
MISSION
NORTHERN
PARK
RICHMOND
SOUTHERN
TARAVAL
TENDERLOIN
Hour of the day

Thefts variation in the southern district by day-of-week

Plotting the thefts count by day-of-week reveals that there are only few variations over the week days. And somewhat higher value during the weekend.

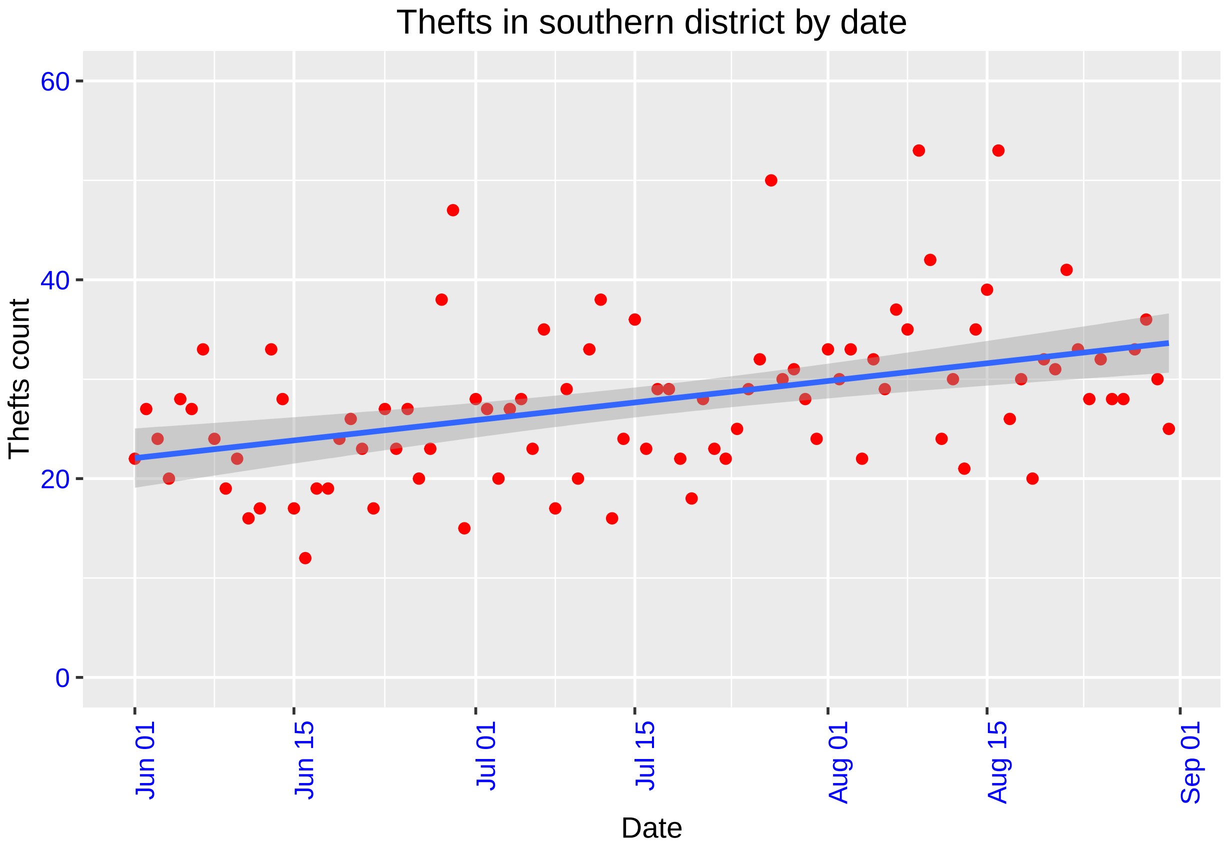
```r
southern <- san[san$PdDistrict == "SOUTHERN" & san$Category == "LARCENY/THEFT", ]
counts <- aggregate(Count ~ DayOfWeek, data = southern, sum)
ggplot(counts, aes(DayOfWeek, Count)) +
  labs(title = "Thefts in southern district by day-of-week", x = "DayOfWeek", y = "Thefts count") +
  theme(axis.text.y = element_text(size = 10, color = "blue")) + ylim(0,500) +
  theme(axis.text.x = element_text(size = 10, color = "blue", angle = 90, hjust = 1)) +
  geom_point(color = c("red"))
```



Thefts variation in the southern district by date

Plotting the thefts count by date reveals that the thefts tend to increase over the summer months in the southern district. It assume that it might be due to the weather factor, where increasing temperatures could be leading to more thefts. However, I don't make any conclusions on this correlation!

```
counts <- aggregate(Count ~ Date, data = southern, sum)
ggplot(counts, aes(Date, Count)) +
  labs(title = "Thefts in southern district by date", x = "Date", y = "Thefts count") +
  theme(axis.text.y = element_text(size = 10, color = "blue")) + ylim(0,60) +
  theme(axis.text.x = element_text(size = 10, color = "blue", angle = 90, hjust = 1)) +
  geom_point(color = c("red")) +
  geom_smooth(method = "lm", se = TRUE)
```

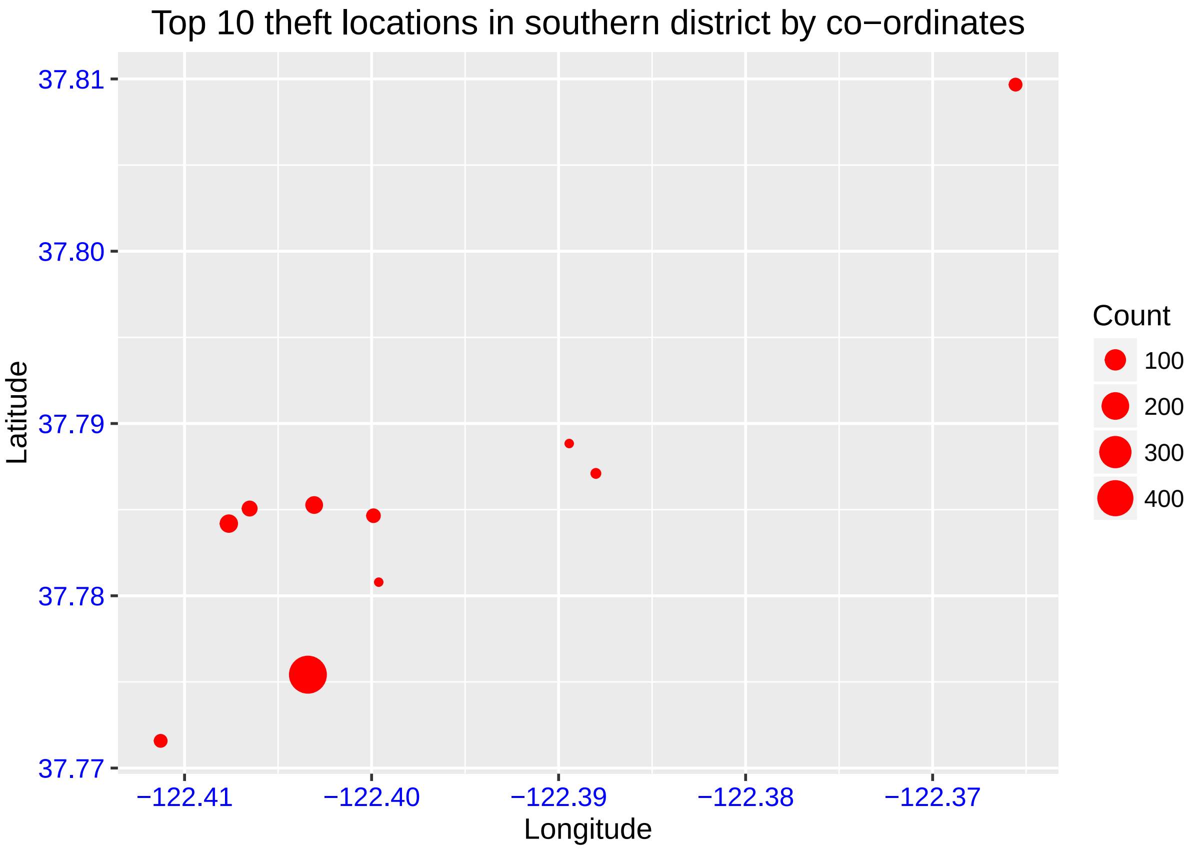


Location of highest theft

Plotting the concentration of the theft crimes and further digging into the data reveals that 800 Block of BRYANT ST in Southern district San Francisco has the highest concetration of thefts. The figure is close to 5% of all thefts in San Francisco and is close to 18% of all thefts in that district. This is really a very very high figure and there is something pecuiar about this location in San Francisco that would be interesting to make further analysis.

```
allThefts = aggregate(Count ~ X + Y + PdDistrict, data = subset(san, san$Category == "LARCENY/THEFT"), s
southThefts <- aggregate(Count ~ X + Y, data = subset(allThefts, allThefts$PdDistrict == "SOUTHERN"), su
topThefts <- head(southThefts[order(southThefts$Count, decreasing=TRUE),], 10)

ggplot(topThefts, aes(X,Y)) +
  labs(title = "Top 10 theft locations in southern district by co-ordinates", x = "Longitude", y = "Lati
  theme(axis.text = element_text(size = 10, color = "blue")) +
  geom_point(color = c("red"), aes(size = Count))
```



```
highestTheft <- head(southThefts[order(southThefts$Count, decreasing=TRUE),], 1)
highestTheftAddress <- subset(san, san$X == highestTheft$X & san$Y == highestTheft$Y,
                       select = c(Address, PdDistrict))

cat(paste(round(highestTheft$Count/sum(southThefts$Count)*100, digits = 2), "% of all thefts in Souther
```

```
## 17.72 % of all thefts in Southern district, San Francisco occur at the Address:
##  800 Block of BRYANT ST
```

```
cat(paste(round(highestTheft$Count/sum(allThefts$Count)*100, digits = 2), "% of all thefts in San Franci
```

```
## 4.8 % of all thefts in San Francisco occur at the Address:
##  800 Block of BRYANT ST , SOUTHERN District
```