MA 615 Midterm Project Report
Chicago Crimes in 2016

Lu LI U67923571

Note: This is a collaboration project, with Ningxi Wei
Github: https://github.com/elenaluuu/Chicago-Crime-Project

Overview

The dataset used for this project focus on Crimes in Chicago from 2001 to 2016.[1] This project focuses on data wrangling using R software, including packages *tidyr*, *dplyr* and *ggplot.* The original dataset includes 22 variables and 1048575 observations. This large dataset is reduced into a table with 3 variables: *"Primary Type", "Location Description", "Community Area"* that we are interested. We further tidied this table up into 77 independent tables respect to Community areas. Descriptive statistics are shown with tables and plots subsequently, and some suggestions for further analysis is argued at the end.

Tidy data

The original dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to Oct. 17th 2016, minus the most recent seven days. It contains 22 variables, some of them are date of crimes, time of crimes, FBI code, satellite positions, etc. I omitted these variables with no interests and left 3 variables with most interests (could be used for further analysis): "data.Primary.Type", "data.Location.Description", and "data.Community.Area".

```
df <-subset(data, select =
c("data.Year","data.Primary.Type","data.Location.Description","data.Community.Area"))
write.csv(df, file = "CrimesSelected.csv")
```

I further reduced the selected dataset by focusing on data in 2016, from Jan 1st to Oct.17th.

```
d1a <-subset(d1, data.data.Year>=2016)
```

Now I rename the other three variables into "Primary Type", "Location Description", "Community Area". Next, I rearrange the table with community Area code in increasing numerical order, sort crime type and location description in alphabetical increasing order.

[1] The data is obtained from https://catalog.data.gov/dataset/crimes-2001-to-present-398a4

```
colnames(d1a) <- c("Primary Type", "Location Description", "Community Area")
d1b <- d1a[order(d1a$`Community Area`,d1a$`Primary Type`,d1a$`Location Description`),]
```

	Primary Type	Location Description	Community Area
3664995	ARSON	APARTMENT	1
101106	ARSON	OTHER	1
21778	ARSON	RESIDENCE-GARAGE	1
20734	ASSAULT	ALLEY	1
20920	ASSAULT	ALLEY	1

Table 1: First five rows of table *d1b*, functioned as the fundamental dataset for further analysis.

At this stage, the table is tidy. But to be aware, since each desired variable have lots of possible values (more than 20), it is still hard to do further analysis using this table, due to large number of observation (approximately 100 thousands) and resulting low speed of R programming running. Therefore, we further split this table by 77 community areas and counting occurrences for each type of crimes. Take community area 1 as example:

```
df <- data.frame(data=read.csv("Crimes2016.csv",header = TRUE, stringsAsFactors = TRUE))
dfCA1 <- filter(df,df$data.Community.Area==1)
crimeTypeCount1<-as.data.frame(table(dfCA1$data.Primary.Type))
colnames(crimeTypeCount1)<-c("Crime Type","Occurrence")
#rank by Occurrence
crimeTypeCount1<-arrange(crimeTypeCount1,desc(Occurrence))
write.csv(crimeTypeCount1, file = "CrimesTypeCount01.csv")
```

	Crime Type	Occurrence
1	THEFT	708
2	BATTERY	587
3	CRIMINAL DAMAGE	327
4	ASSAULT	183
5	OTHER OFFENSE	178

Table 2: First five rows of *crimeTypreCount1*, same table structure for the remaining 76 areas.

The advantage for separating community areas is that observations in each table are largely reduced to approximately 5,000 to 10,000, which increases programming running speed rapidly.

Moreover, it is easy to combine any tables for specific analysis. For example, if you want to analysis crimes in the central area of Chicago, you can combine tables of area 8,32 and 33.[2]

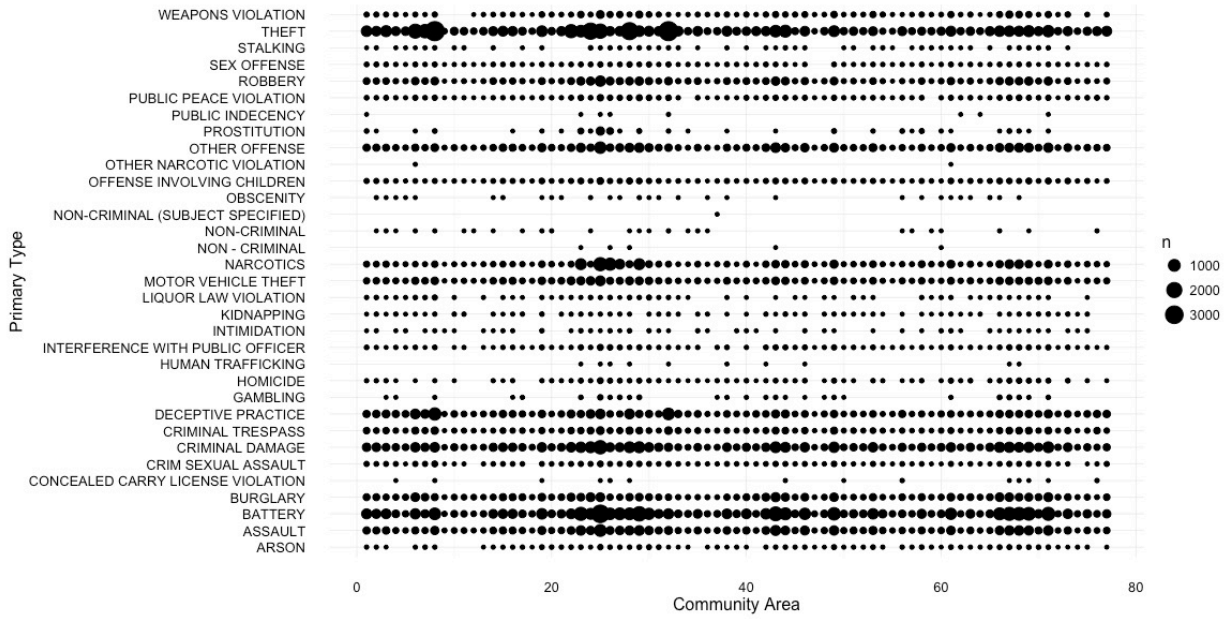
Note that the new collection of tidy tables only contains information about crime type and its occurrence in each community area, while leaving no information for location type. It is because we only want to focus on crime type and community areas for descriptive statistics and plots later in this report. If analysis about location description is further considered, then another collection of tables containing locations should be produced in a similar approach.

Descriptive Statistics and Plots

In this report, I focus on two perspectives: crime occurrence in 77 community areas and top 5 crime type in community area 25.

First, I look at how crime type varies in 77 community areas in Chicago.

```
ggplot(data=df,aes(x=df$data.Community.Area,y=df$data.Primary.Type)) +
  geom_count()+
  xlab("Community Area") +
  ylab("Primary Type") +
  theme_minimal()
```

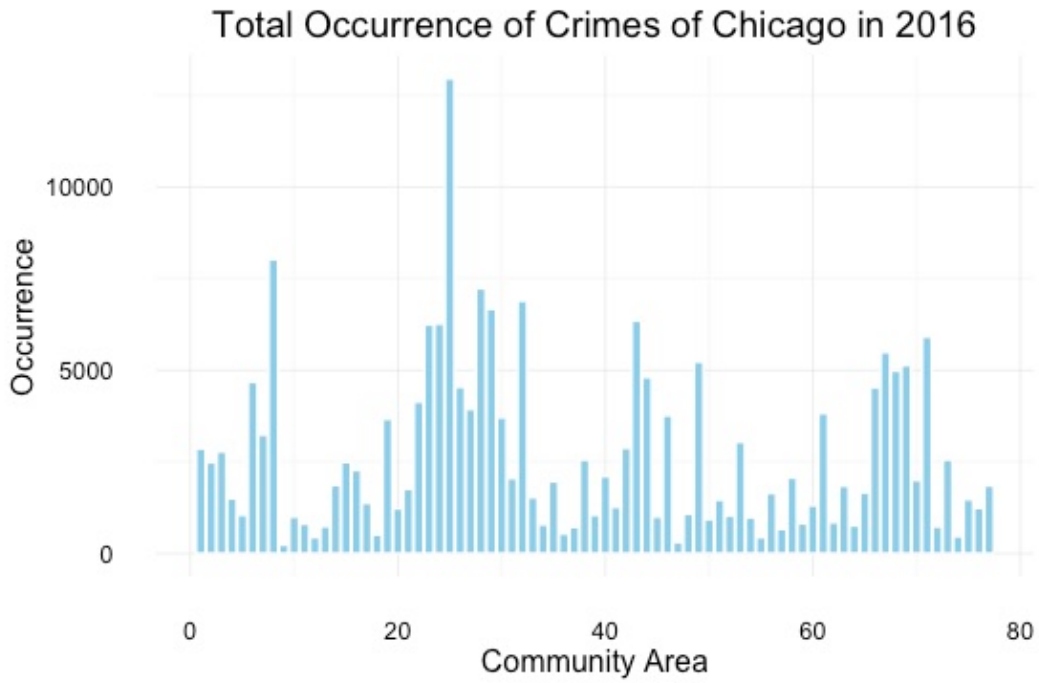


[2] Community areas by side: https://en.wikipedia.org/wiki/Community_areas_in_Chicago

It can be seen that some crime types happen equally among community areas, such as sex offense, criminal trespass, crime sexual assault; while some of other types varies according to areas ---- for example, theft crime is popular in area 8,9,10 and from 23 to 30; battery is popular in area 18 -27 and 65-70.

Next, consider the total crime occurrence in 77 community areas.

```
ggplot(data=df,aes(x=df$data.Community.Area)) +
  geom_bar(stat="count",color="White",fill="Sky Blue") +
  ggtitle("Total Occurance of Crimes of Chicago in 2016") +
  xlab("Community Area") + ylab("Occurance") + theme_minimal()
```



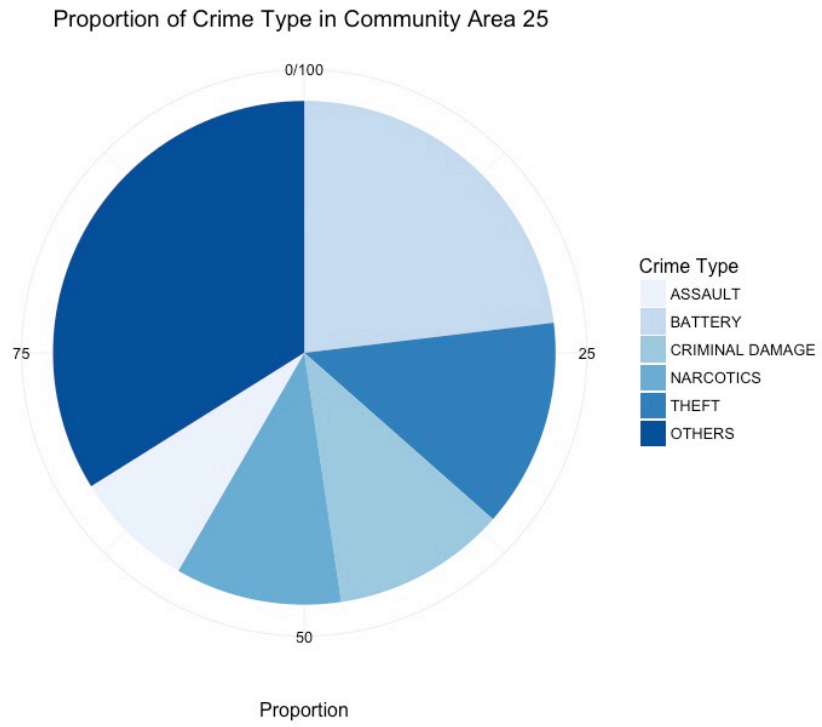
From the histogram, we can see that Community Area 25 has the largest number of crime occurrence, following by Area 8 and Area 28. Most of areas have crime occurrence below 5000, and half of them are below 2500. Table 3 shows that three top-ranking community areas contributes approximately 13.61% of total crimes in Chicago area.

	Community.Area	Occurrence	Prop
1	25	12954	6.25
2	8	8024	3.87
3	28	7231	3.49

Table 3: Top3 community areas in total numbers of crimes happened in Chicago

Secondly, I precisely focus on Community Area 25 and its top 5 crime types.

```
bp <- ggplot(Perc25, aes(x="", y=Perc25$Percentage, fill=Perc25$Crime.Type))+
  geom_bar(width = 1, stat = "identity") +
  ggtitle("Proportion of Crime Type in Community Area 25 ")
pie <- bp + coord_polar(theta = "y") + xlab("")+ylab("Proportion") + labs(fill='Crime Type')
pie + scale_fill_brewer(palette="Blues") + theme_minimal()
```



The most common crime types in Area 25 are assault, battery, criminal damage, narcotics and theft, approximately 67% of total crimes. Battery has the largest proportion (23.09%) among the five.

Possible Analysis in the further

Based on the approach discussed in this report, we can further analysis how different crime types spread among community areas in Chicago. Specifically, is there any similarity and differentiations among 4 community area by side? What's more, we can analysis the relationship between crime type and locations in Chicago. What is the most common place for theft? Which part of Chicago should people be aware of in terms of battery?