Chicago Crime: Data Analysis and Visualisations using R

Student ID: 201081646

November 3, 2018

1 Introduction

This report is the first assessment for the **MATH5741M Statistical Theory and Methods** module. Its objective is to summarise statistically a dataset sample of crimes in the city of Chicago and due to limit has restic to answer the following research questions:

- How has crime evolved over time in the city of Chicago?
- What time of day do most crime occur?
- In which locations is crime more likely to happen?

2 Data and methods

The dataset analysed is a sample of the original data of crimes extracted from the Chicago Police Department which content the crimes that occurred in the city of Chicago from 2001 to present.

For the analysis, first, we prepare the data creating, transforming and simplifying variables, as well as cleaning the dataset keeping the variables we are interested in. Secondly, to answer our research questions we perform the an analysis based on line graphs and heat-maps. Finally, we summarise the findings.

The report has been done with `Rmarkdown` but unfortunately does not include all the `R` code cells written for its performance[1]. However, it is available for consultation in this link https://github.com/eugenividal/Chicago-Crime-Data-Analysis.

3 Results

3.1 Data preparation

First, we load the libraries we will need for the project and get the data into the `R` environment.

```
# load libraries
library(ggplot2)
library(ggmap)
library(lubridate)
library(scales)
library(zoo)
library(dplyr)
library(knitr)
```

```
# Read csv in R
dd=read.csv("http://www1.maths.leeds.ac.uk/~charles/math5741/crime.csv",header=T)
```

[1] In this report it is not included the code used to group categories in variables `Primary.Type` and `Location.Description`, neither the code used to generate the visualisations

Second, we create the new variables (`Count`, `Month_year`, `Hour`) based on the existing ones, and give them the right format for later exploration.

```
# Create a variable count with value 1
dd$Count <- 1
# Convert Date from factor to date
dd$Date <- mdy_hms(dd$Date)
# Extract hour from Date
dd$Hour <- substring(dd$Date, 12,13)
# Drop time from Date
dd$Date <- as.Date(dd$Date, format="%m/%d/%Y")
# Drop days from Date
dd$Month_year <- as.Date(as.yearmon(dd$Date, "%Y-%m"))
```

?as.Date

Third, we group the categories of the variables `Primary.Type` and `Location.Description` to simplify the analysis and call them `Type_grouped` and `Location_grouped` respectively.

Next, we drop all variables we do not need to answer our research questions.

```
# Drop all variables we are not interested in
dd <- dd[, -c(1:2, 4:11, 13:18)]
```

Then, we clean the dataset of missing values.

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

Finally, we show the the dataset ready for exploration.

```
# Show first 5 records
head(dd)
```

```
##         Date District Count Hour Month_year Type_grouped Location_grouped
## 1 2013-07-20       19     1   00 2013-07-01       Batery           Street
## 2 2013-07-20       19     1   01 2013-07-01       Others           Street
## 3 2013-07-19        2     1   21 2013-07-01      Assault        Apartment
## 4 2013-07-20        9     1   02 2013-07-01    Narcotics           Street
## 5 2013-07-12        3     1   17 2013-07-01        Theft           Street
## 6 2013-07-20        9     1   01 2013-07-01       Batery        Apartment
```

3.2 Data exploration

3.2.1 How has crime evolved over time in the city of Chicago?

To answer the first question we plot in Figure 1 the number of crimes per month from 2001 to the present. The graph shows that crime in the city of Chicago has been decreasing consistently over the whole period. The wave-shape of the graph also shows that there is a clear periodic pattern per months of the year.

```
dd_aggr <- aggregate(Count ~ Date, data = dd, FUN = sum)
# add cumulative column to the data frame
dd_aggr[,"cum_Count"] <- cumsum(dd_aggr$Count)
# cumulative sum over rolling date range
```

In Figure 2 we plot the same graph per type of crime ordered per importance, being Theft the most significant and deceptive practice the less. Except the deceptive practice crime, which at present keeps in similar values as in 2001, the rest of type have been falling to a greater or lesser extent.

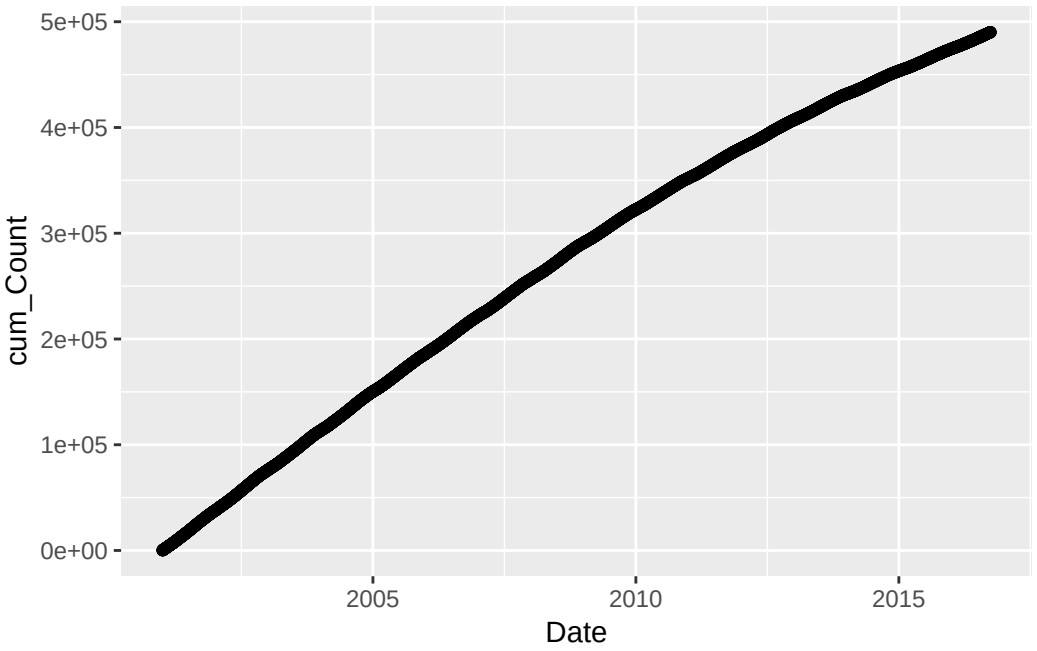

Figure 1: Crimes evolution

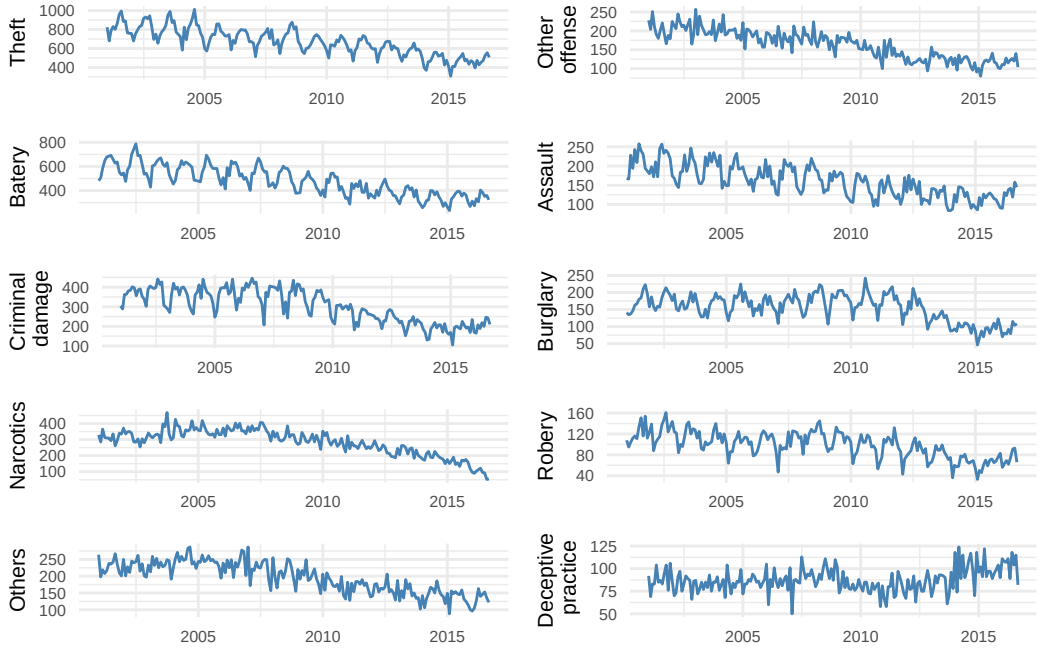

Figure 2: Evolution per type of crime

3.2.2 What time of day do most crime occur?

The crimes are concentrated in hours

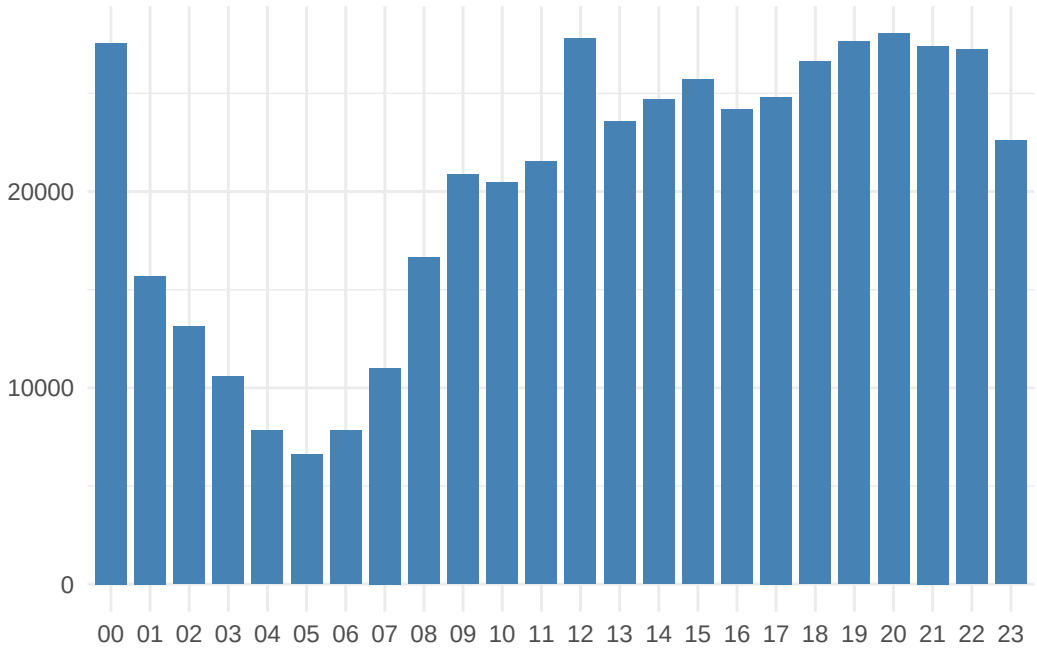

Figure 3: Crimes per hour

Figure 3 a head map between type of crime and hour. The peak hour of crimes morning, other crimes peak during the and the final group of crimes peak at night.

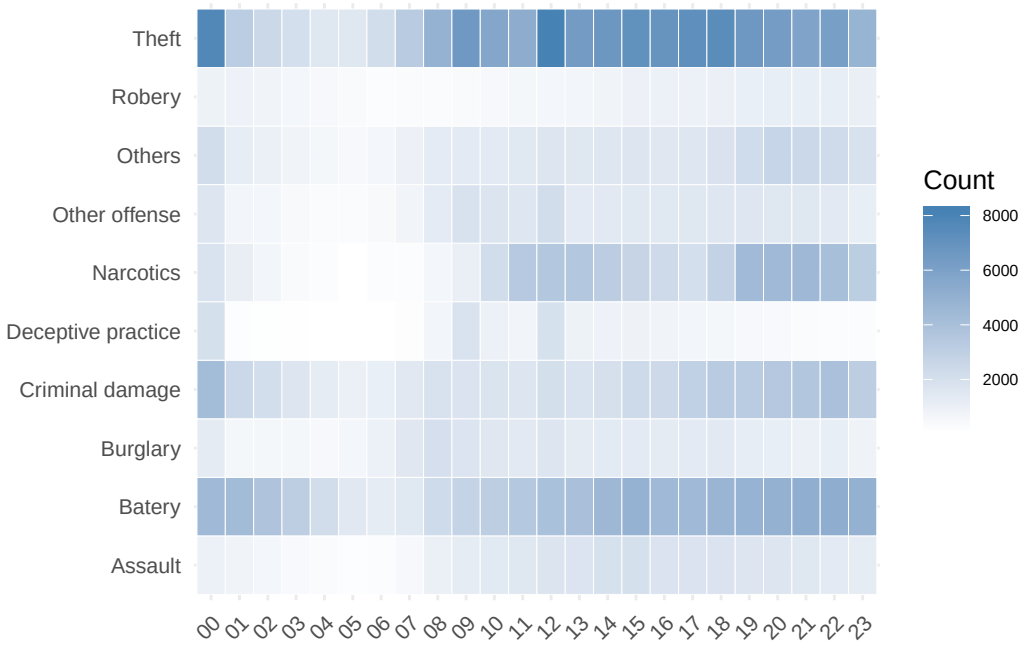

Figure 4: Type of crime vs hour

3.2.3 In which locations is crime more likely to happen?

These crimes are concentrated in Streets, give percentage.

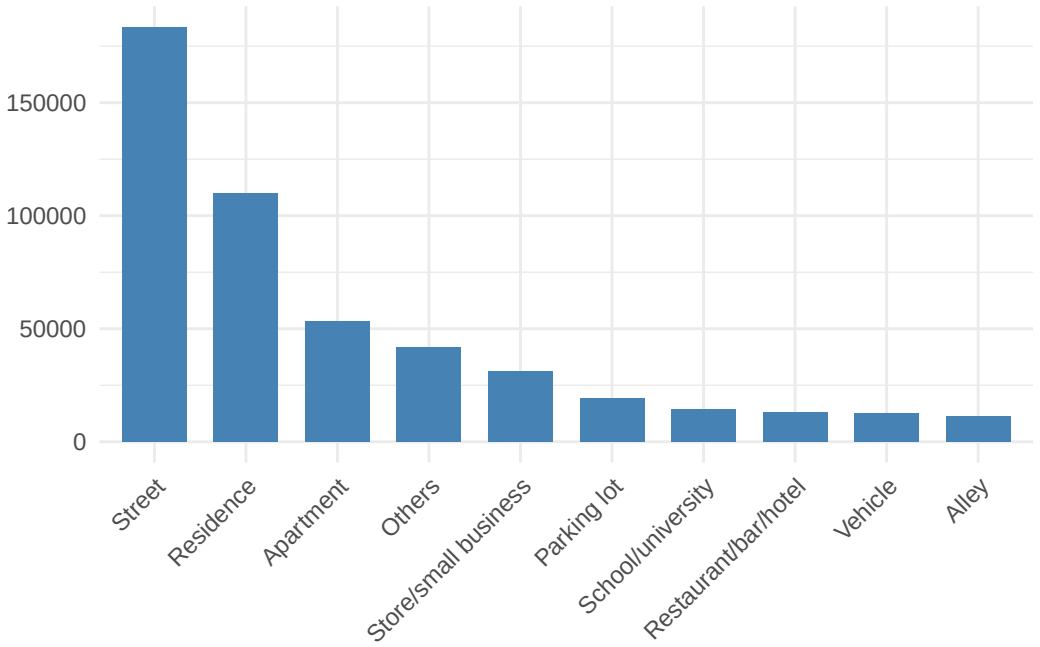

Figure 5: Crimes per location

Figure 4 shows that some locations are more likely per type of crime. For theft for instance it is better this or that.

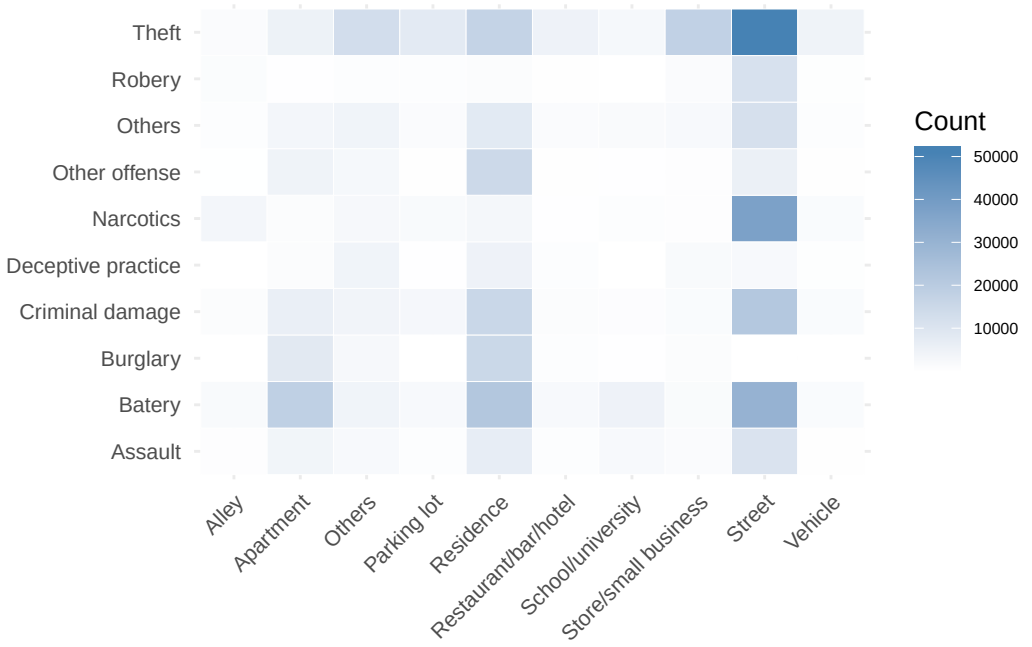

Figure 6: Type of crime vs location

4 Conclusions