Chicago Crime Data Analysis

201081646

1 Introduction

This is the first assessment for the **Statistical Theory and Methods module**. Its objective is to (1) summarise a sample of Chicago Data Crime dataset and (2) to highlight the key findings.

2 Data and methods

The dataset we use is sample of 500,000 rows of the original data which come from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.

First, we prepare the data, then we explore it through a univariable analysis and a multianalysys based on heat maps. in Counclusions we sumarise the main findings.

The report has been done explaining the most essential code. However, due to space limitarion couldn´t add. Youall the code which can be analysed in the following github link.

3 Results

3.1 Data preparation

First, we load the data into the `R` environment.

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

Then, we have a look at the variables.

```
# show the dataset variables
names(dd)
```

```
##  [1] "X"                    "ID"                   "Date"
##  [4] "Block"                "IUCR"                 "Primary.Type"
##  [7] "Description"          "Location.Description" "Arrest"
## [10] "Domestic"             "Beat"                 "District"
## [13] "Ward"                 "Community.Area"       "FBI.Code"
## [16] "Year"                 "Latitude"             "Longitude"
```

We chose 5 of them (`Date`, `Primary.Type`, `Location.Description`, `Arrest` and `District`) and drop the rest.

```
# Drop all variables we are not interested in
dd <- dd[, -c(1:2, 4:5, 7, 11, 13:18)]
```

Secondly, we clean the dataset of missing values.

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

Third, we create new variables (`count`, `hour`, `Month_Yr`, `Month`, and `weekday`) based on the existing ones, and give them the right format for later explotation.

Next, we simplify the variables `Primary.Type` and `Location.Description` grouping their categories and call them `Type_grouped` and `Location.Description` respectivelly.

Finally, the data is ready for the explotation.

```
head(dd[dd$VAR1==4,],6)
```

```
##  [1] Date            Arrest          Domestic        District
##  [5] count           hour            Month_Yr        mon
##  [9] weekday         Type_grouped    Location_grouped
## <0 rows> (or 0-length row.names)
```

3.2 Data exploration

3.2.1 Univariable analysis

3.2.1.1 Crimes evolution

The number of crimes in Chicago has decrease dramatically per year from 200x until 2015.

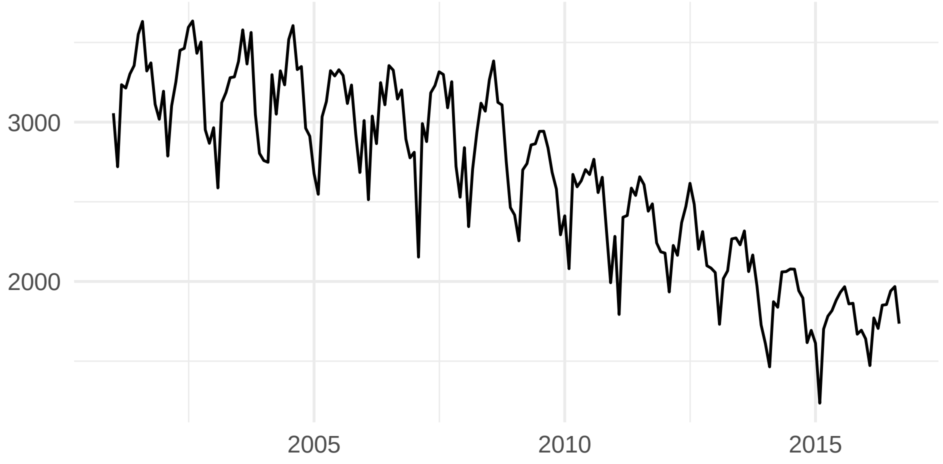

Figure 1: Crimes evolution

Except for the deceptive practice, all the crimes have decresead in more or less grade.

3.2.1.2 Crime per Hour

The crimes are concentrated in hours

3.2.1.3 Crime per weekday

Friday concentrated most of the crimes, percentage?

3.2.1.4 Crime per month

Summer is in difference the period with more crimes recorded.

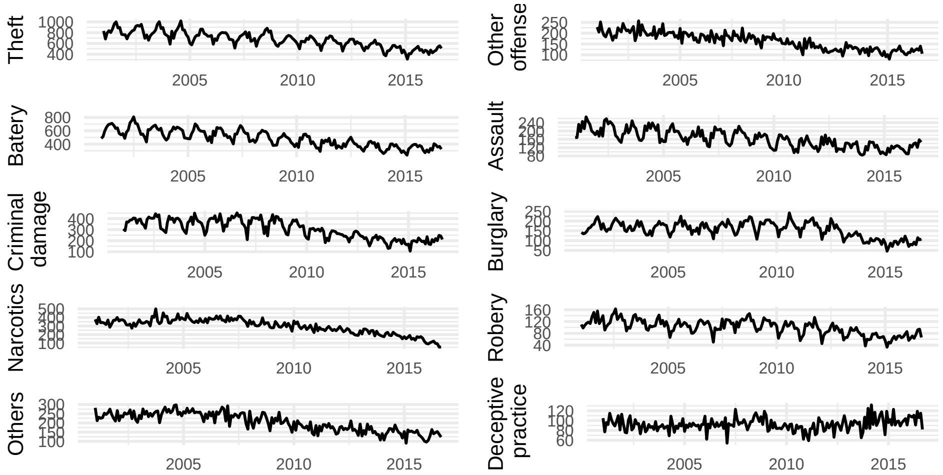

Figure 2: Evolution per type of crime

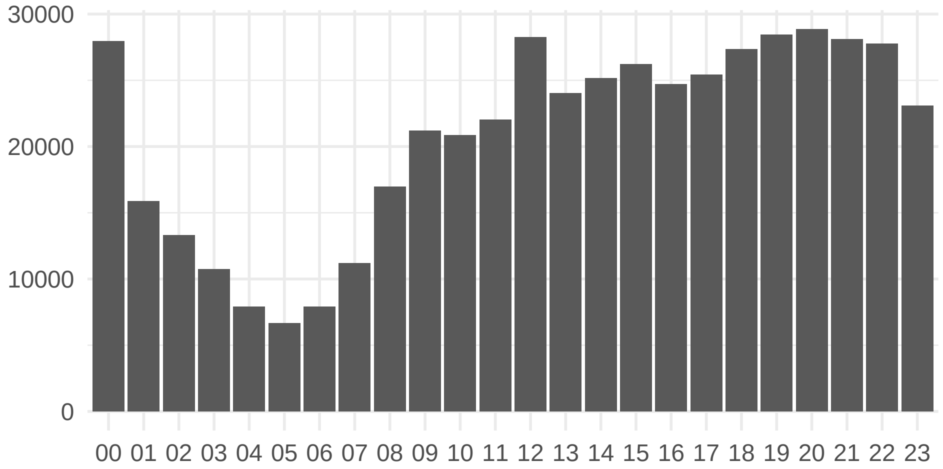

Figure 3: Crimes per hour

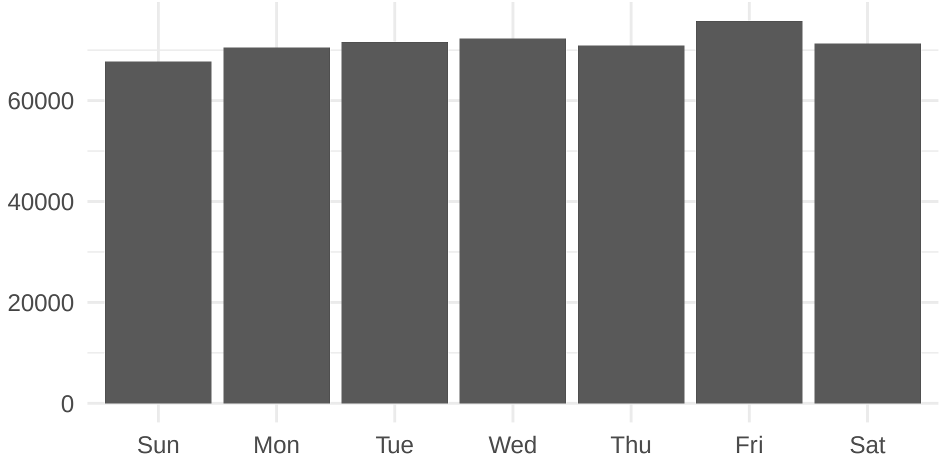

Figure 4: Crimes per weekday

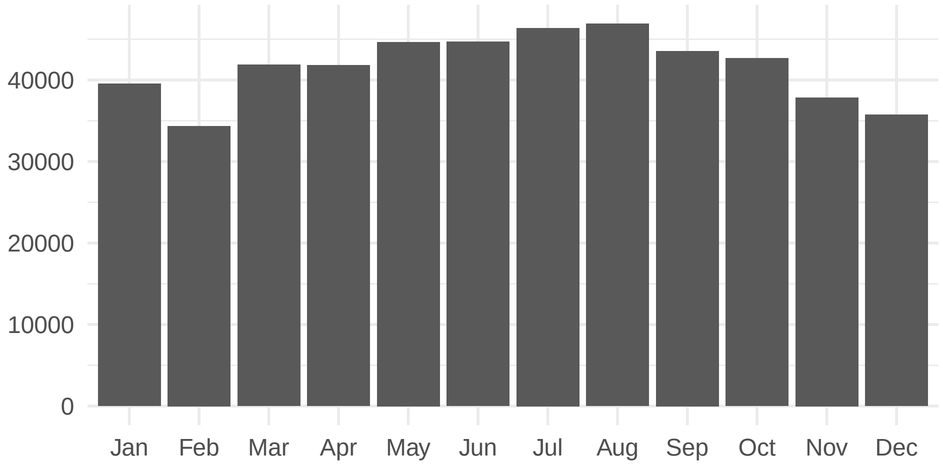

Figure 5: Crimes per month

3.2.1.5 Type of crimes

Per type of crime Theft is in difference the biggest number. Change the scientifyc number.

Change scientific numbers!

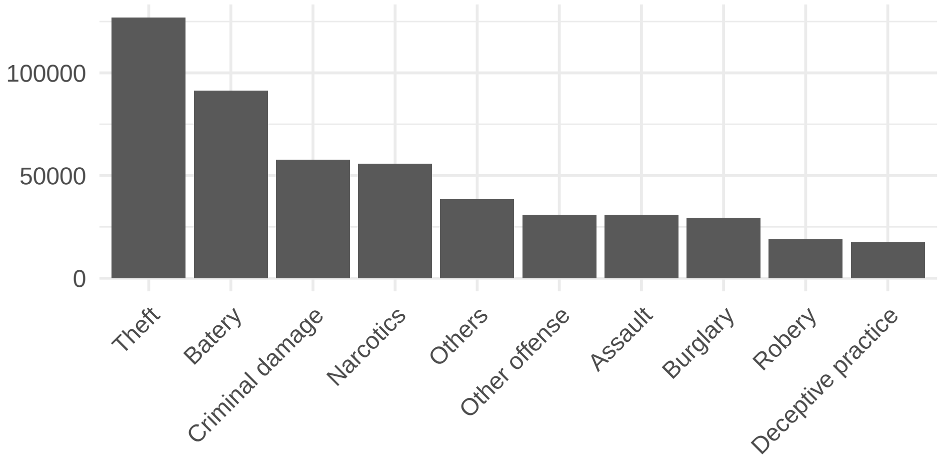

Figure 6: Crimes per type

3.2.1.6 Location of crimes

These crimes are concentrated in Streets, give percentage.

3.2.1.7 Crimes per districts

Per districts the most dangerous are 8.

3.2.2 Multivariable analysis

The multiple analysis focuses on type of crime crossed with hour, location and district.

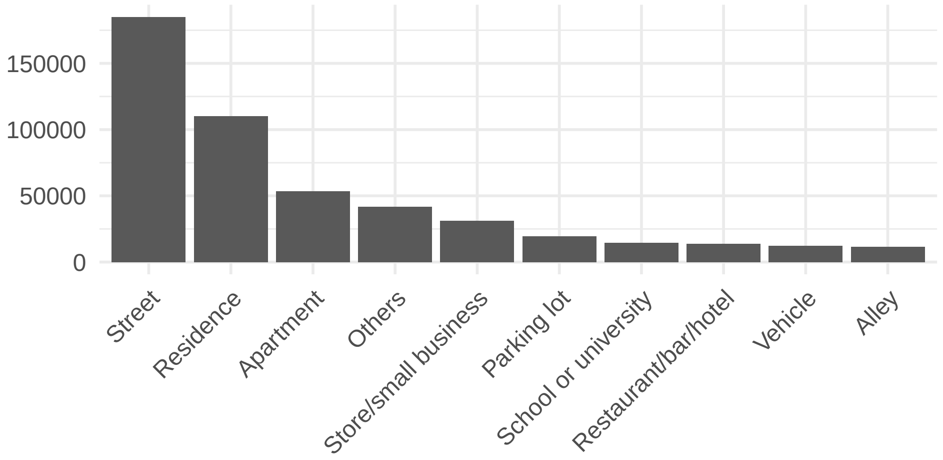

Figure 7: Crimes per location

3.2.2.1 Type of crime vs hour

The most dangerous hours per Thefth are 00 and 12.

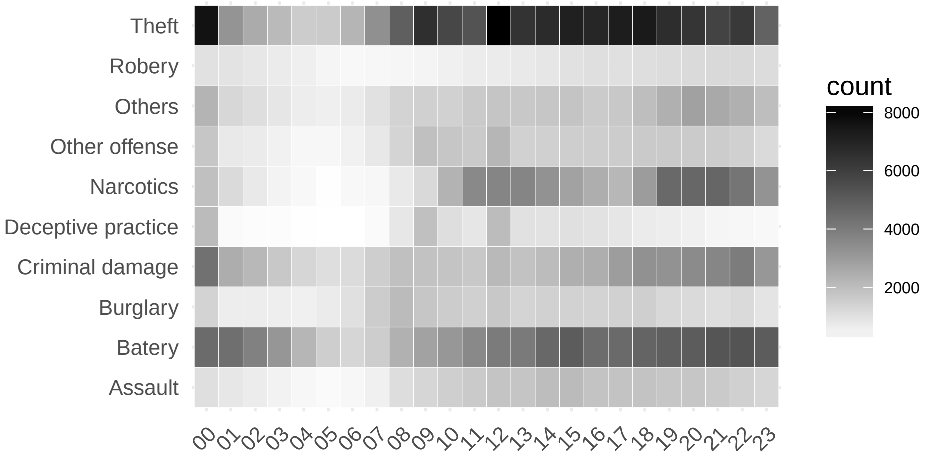

Figure 8: Type of crime vs hour

3.2.2.2 Type of crime vs location

Street is particularly important for Theft.

3.2.2.3 Type of crime vs district

Narcotics in district 11 is crealy a problem.

4 Conclusions

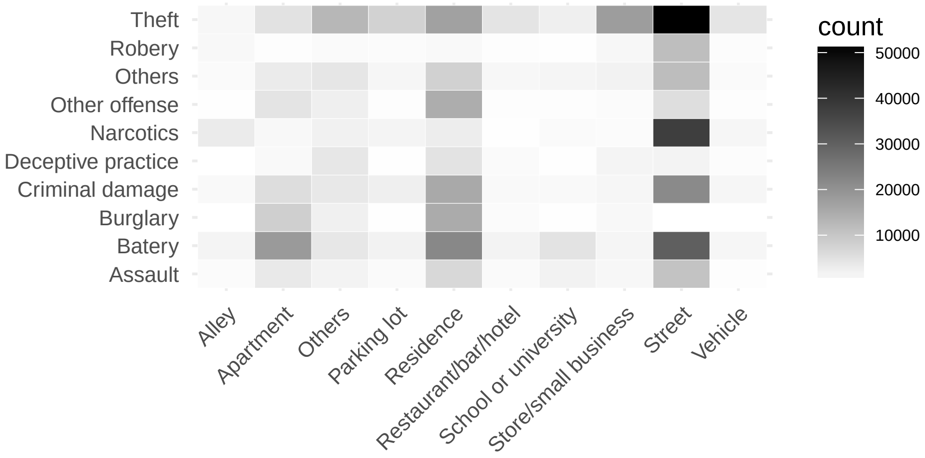

Figure 9: Type of crime vs location

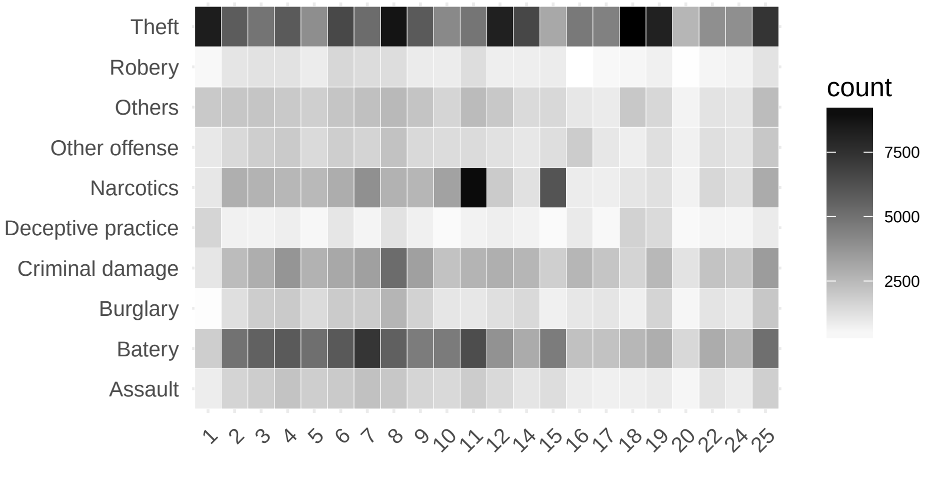

Figure 10: Type of crime vs district