Exploration of the Chicago-Crime Dataset

Tobias Machnitzki and Finn Burgemeister

12 November 2017

```
library(knitr)
library(markdown)
library(scales)
library(ggplot2)
```

1. Overview

First we want to load the data and get an overview of what we are dealing with.

```
crimes = read.csv("../../WORK/Blatt4/chicago_crime_sample.csv", head=TRUE, sep=",")

crimes$Date = as.Date(strptime(crimes$Date, '%m/%d/%Y %H:%M:%S %p'))

#summary(crimes)
```

For a short overview over the big dataset we use summary, which returns much useless information. Important are the row names. We can subdivide the rows in categories.

Table 1: Categorisation of Dataset

Category	Data Name
Official	"ID", "Case.Number", "IUCR"*, "FBI.Code"
Location	"Block", "Location.Description", "District", "Ward", "Community.Area" "X.Coordinate", "Y.Coordinate", "Latitude", "Longitude", "Location", "Beat" [i.e. police district], "Domestic"
Case	"Primary.Type", "Description", "Arrest"
Time	"Date", "Year", "Updated.On"

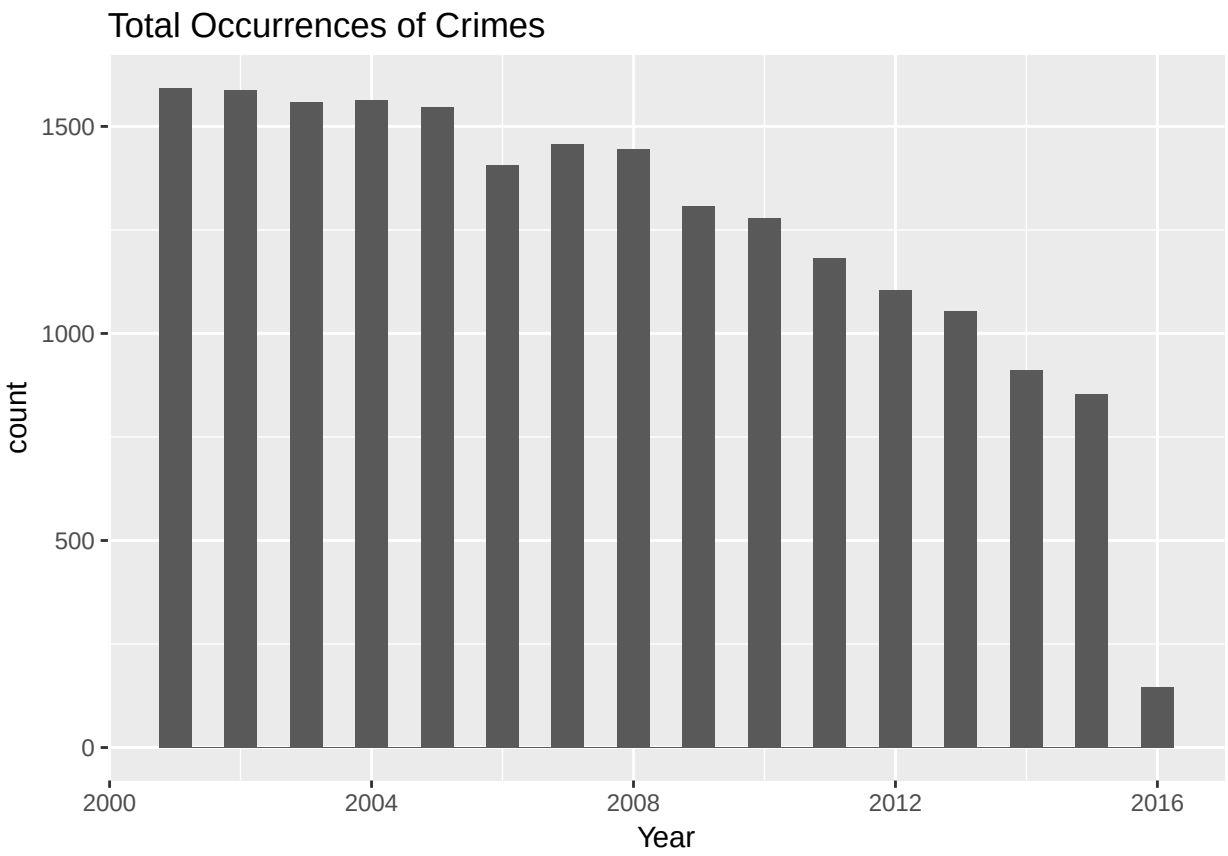
*Illinois Uniform Crime Reporting (IUCR) Codes

On first view we know, we have a data from the period 2001 to 2016. The total number of data is 20000. The most insecure date is 01/01/2005 at 12:01 with four parallel crimes. For 9240 crimes we don't have an exact location. After these 20000 crimes there were only 5795 arrests. The most common crime is theft. The next common crimes are battery [i.e. personal injury], criminal damage, narcotics, other offense and assault. The most crimes were commited on the street. 2564 crimes were domestic.

2. Time

Total Occurences of Crimes

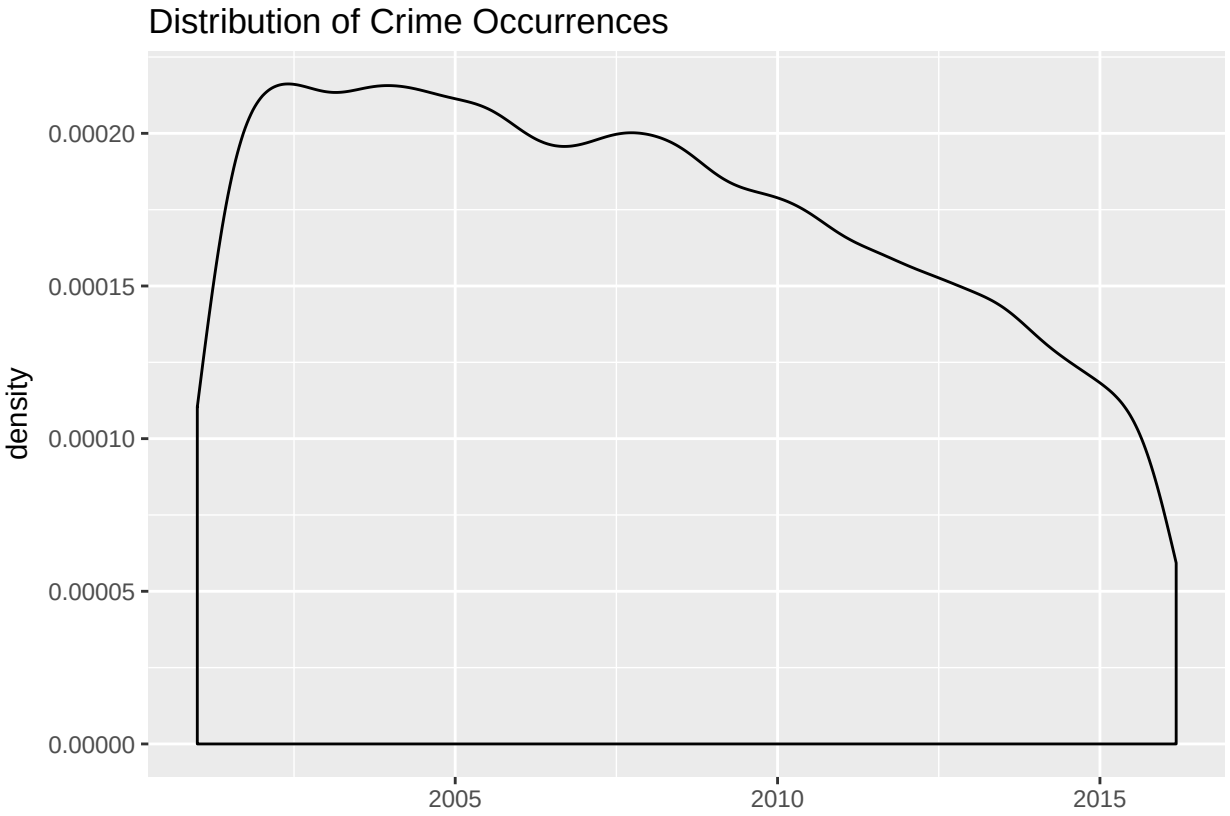
```
ggplot(crimes, aes(Year)) +
  geom_histogram(binwidth = 0.5) +
  labs(title="Total Occurrences of Crimes")
```



The histogram shows the total crime occurrences over the years. Obviously the dataset is for 2016 not complete. We can see a continous decreasing number of crime.

Distribution of Crime Occurrences

```
ggplot(crimes, aes(Date)) +
  geom_density(alpha=0.2) +
  scale_x_date() + xlab("") +
  labs(title="Distribution of Crime Occurrences")
```



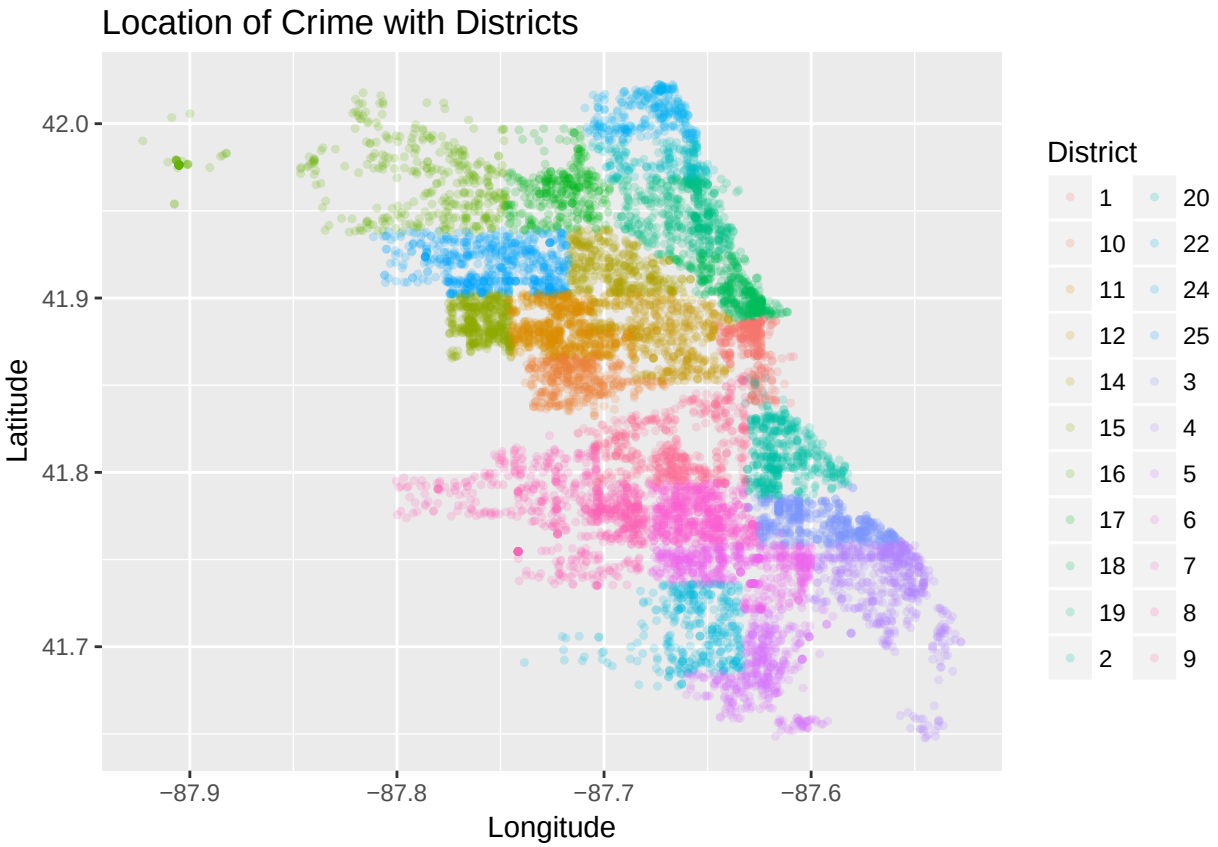
As expected the distribution of crime occurrences over the time is similar to shape of the histogram.

3. Location

Location of Crime in Districts

```
crimes$District = as.character(crimes$District)

ggplot(subset(crimes, !is.na(Longitude) & !is.na(Latitude)), aes(x=Longitude, y=Latitude, color=District)) +
  geom_point(alpha=.2, size=.9) +
  labs(title="Location of Crime with Districts")
```




```
crimes$District = as.numeric(crimes$District)
```

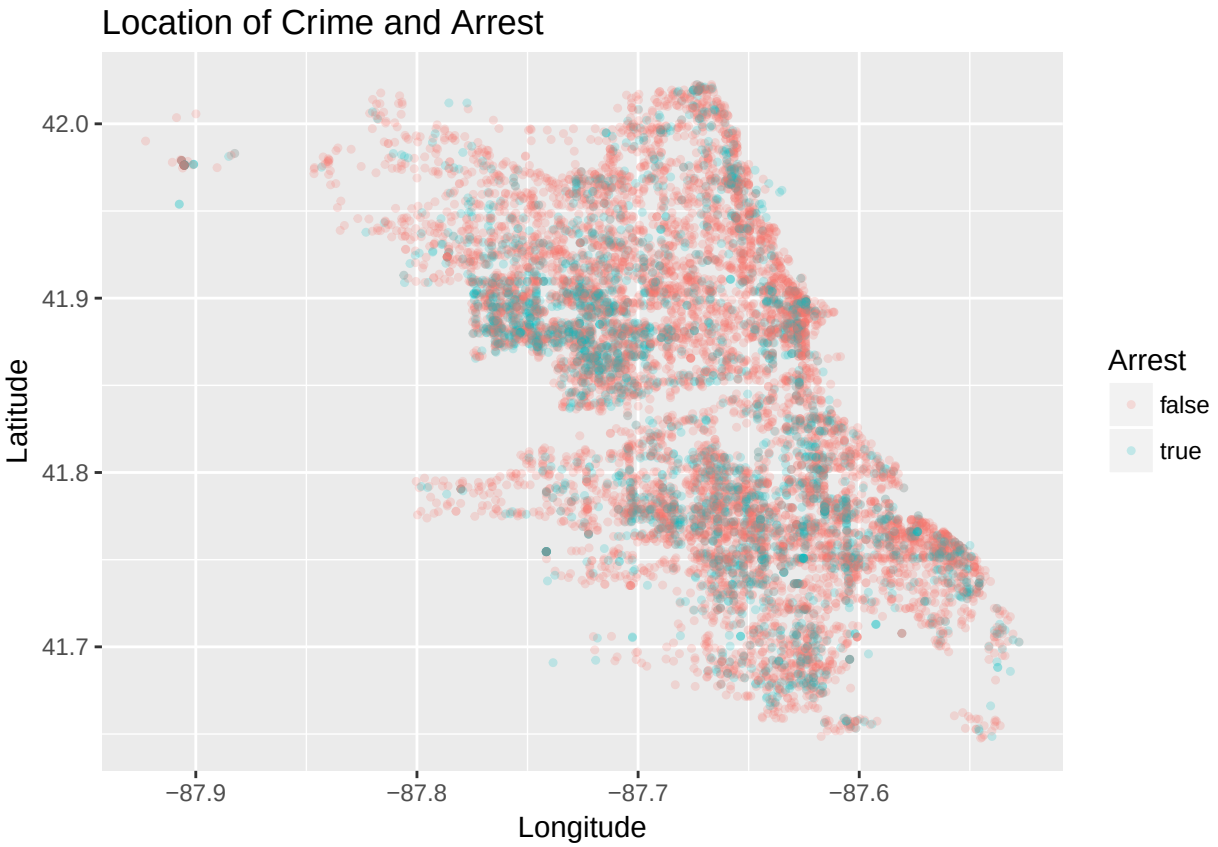
A map of Chicago defined by the crime locations with colored districts. We can see the location of the airport of Chicago in the northwest. It is out of town. The map is confusing because of the amount of districts.

3. Cross-Analysis

Location of Crimes and Arrests

```
crimes$District = as.character(crimes$District)

ggplot(subset(crimes, !is.na(Longitude) & !is.na(Latitude)), aes(x=Longitude, y=Latitude, color=Arrest)) +
  geom_point(alpha=.2, size=.9) +
  labs(title="Location of Crime and Arrest")
```

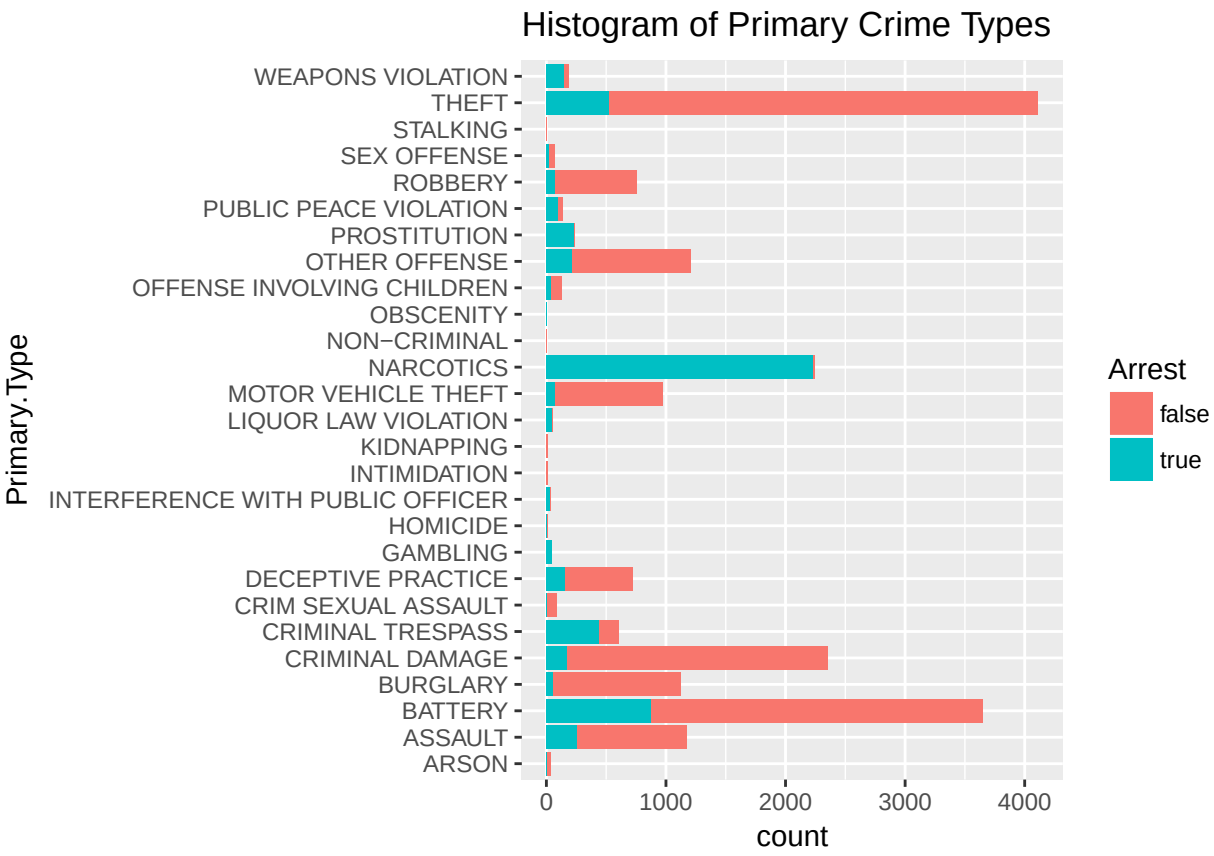



```
crimes$District = as.numeric(crimes$District)
```

There seems be some link between the location of the crime and arrest, in the central west north.

Which crime leads to jail?

```
ggplot(crimes, aes(Primary.Type)) +
  geom_bar(aes(fill=Arrest)) +
  labs(title="Histogram of Primary Crime Types") +
  coord_flip()
```



If you do drugs (narcotics) you will get arrested with a high probability (near 100 percent). Battery and assault lead also with a high amount in jail, over 50 percent don't got arrested.

Distribution of crime types over the years?

```
crimes$Year = as.character(crimes$Year)

ggplot(crimes, aes(Primary.Type)) +
  geom_bar(aes(fill=Year)) +
  labs(title="Histogram of Primary Crime Types") +
  coord_flip()
```