Communicating Data Science Results

Rolf-Dieter Kaschke

April 6, 2016

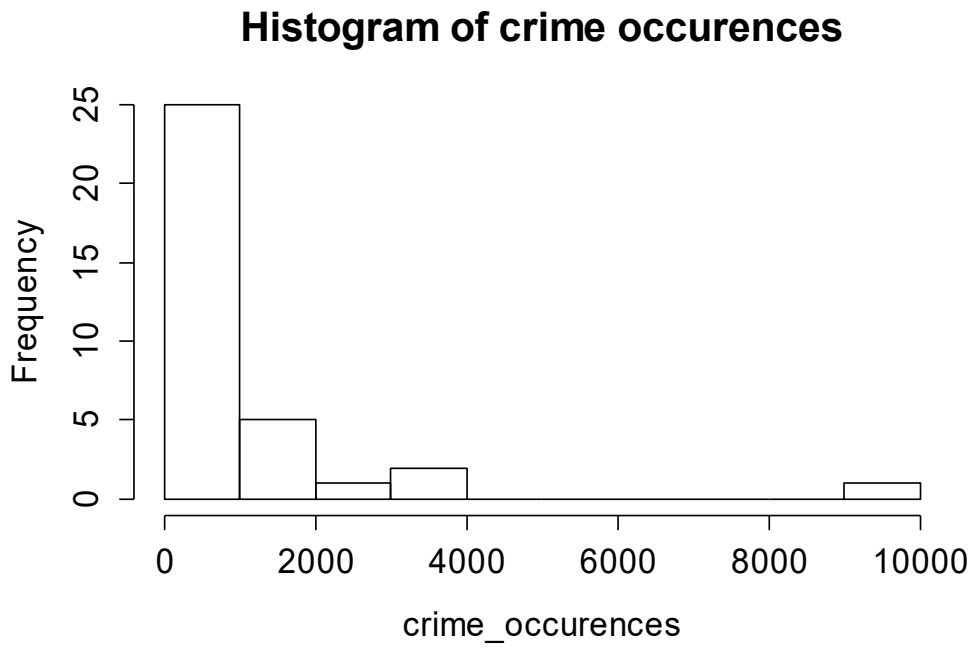
„In this assignment, you will analyze criminal incident data from Seattle or San Francisco to visualize patterns and, if desired, contrast and compare patterns across the two cities.“

For this assignment I take the San Francisco Dates as provided by coursera. My opinion is, that unless the examples in the course, it makes mostly no sense to provide vizualisations with many dimensions of data, unless you color them or not. Visualization always is strongly related with the goal I want to express. Maybe in scientific work it helps to find patterns unseen in the data, but this isn't asked in the assignment. So my approach is to bring some different, low dimensional data graphs with a clear direction.

When I looked first at the data, it seems that all common interesting data are ordinal (or nominal). This includes the crime category „Category“, the associated police department „PdDistrict“, the result „Resolution“ and all adress related data, like „Address“ or „Location“. The other interesting data are related to date and time. If we don´t want to identify a single event, like f.i. The crime rate at Christmas, it is better to cluster time to time frames. That is done by cluster them the categorial variables „Monday“, „Tuesday“ … „Sunday“ and time intervals of 4 hours.

To fulfill my assignment I used R, mainly to prepare the data and to plot them.

The first diagram I made is a histogram of crime occurences:



To be a little bit tech here – the plot was done in R by:

```
crime_occurences = table(data$Category)
hist(crime_occurences, main = "Histogram of crime occurences",
axes=TRUE)
```

This graph shows the number of crimes per crime type as a histogram. What is wrong for the casual viewer. A viewer expects to understand what is expressed in the graph. So he want to see as labels of the x-axis the name of crime and the sum of occurences. But this is not was a histogram expresses. Here you can see, that crime types occuring all in all 10000 times, occures 1 times, while crime types that are in the range from 0 to 1000 occurrences occures most, 25 times. For a scientific view it may help, for a casual user this gives us not much information. To improve that I create a short list for explanation:

Number of different crime types:	34
Number of all crimes:	28993

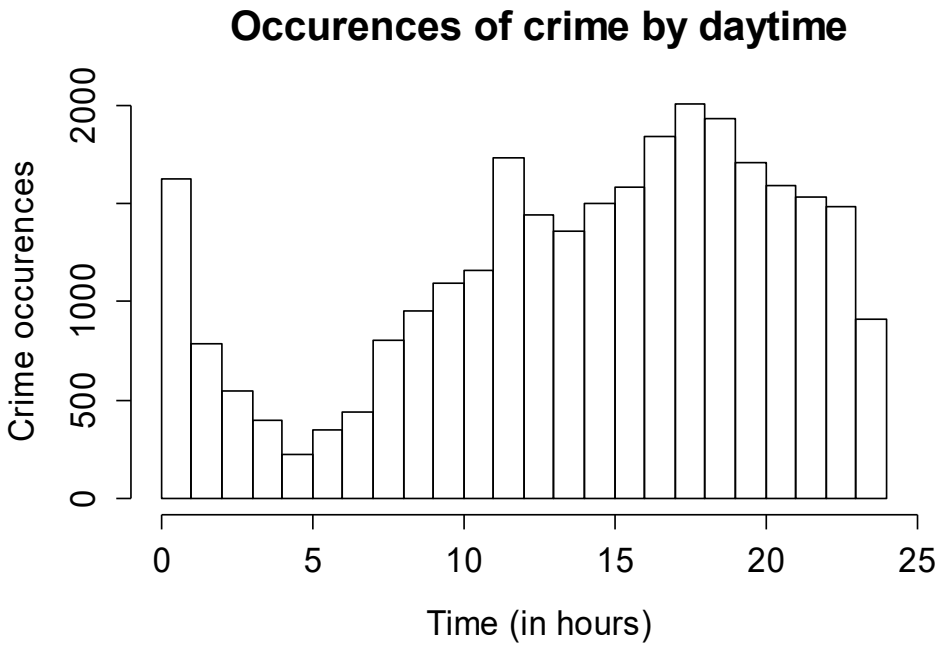
The number of all crimes can be found by

```
summary(data$Category)
```

and the number of all crime types („Category“) by

```
str(factor(data$Category))
```

No it is more clear what our graph says: We have 34 distinct crime types and 28993 occurences all in all, one crime type with 10000 occurences occur one time.



But I think that is not so interesting, so I decided to plot the crime type versus time. As I statet above I converted the time/date combination to an interval of 1 hours independent for of the days of a week. This gives a nice histogram where you can see the peaks of crimes in the time intervals. Of course this can be plotted more differentiated when we use a time frame of one hour, but the result lacks in clarity.

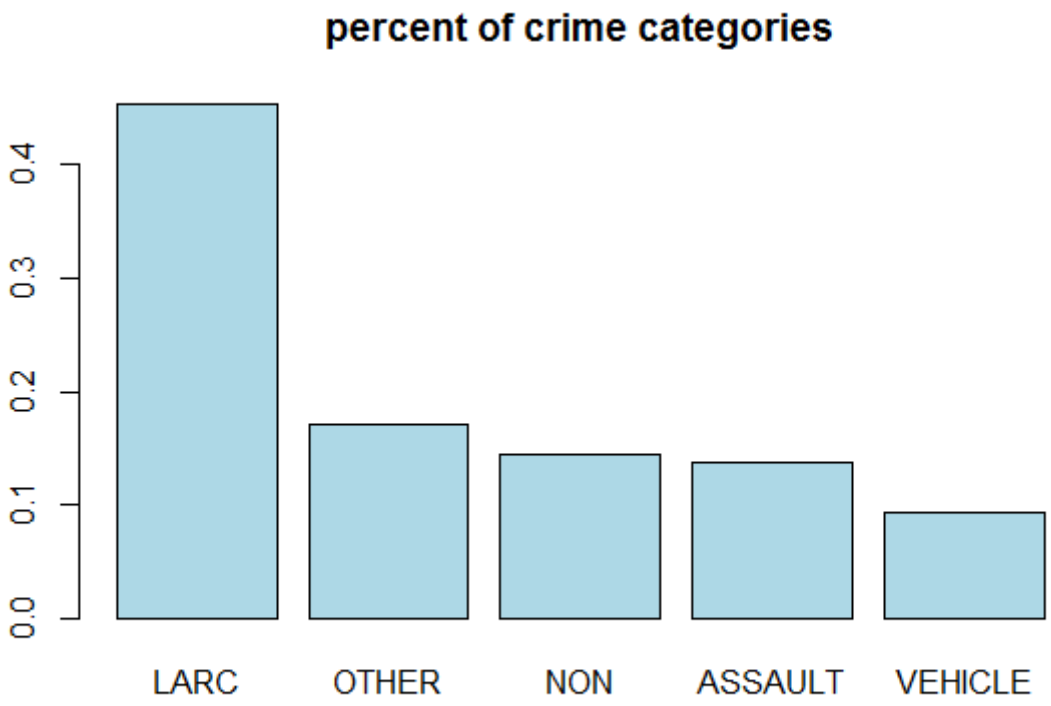
Did you realize the peak at 5PM and 0:00AM. You are more secure if you went home at 4AM.
This plot was done by the R command:

```
hist(data$time_num, breaks=24, xlim=c(0, 25),
     xlab="Time (in hours)", ylab="Crime occurences",
main="Occurences of crime by daytime")
```

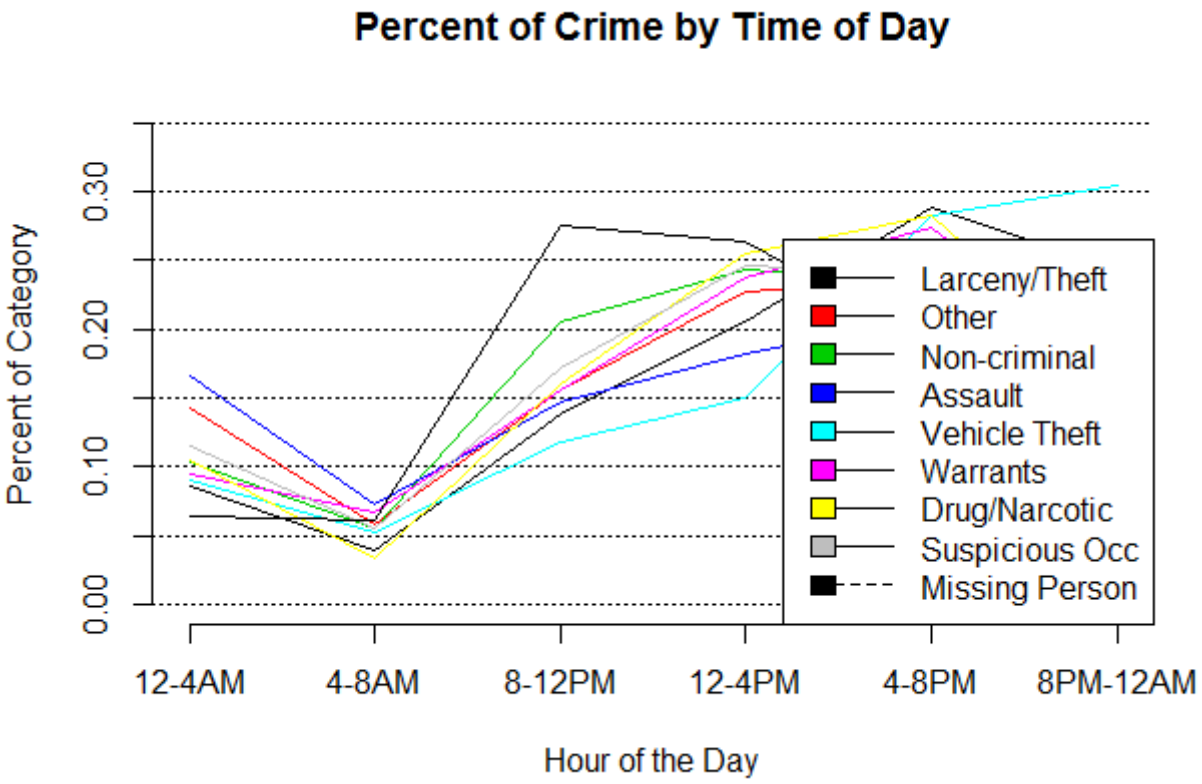
while data$time_num was created by a split at „:“ of the time and normalization to an interval of an hour.

This diagram can be made more interesting by colors, but I don´t like to do so as I think it is clear enough.

Can we find more information from the data and plot them?
I think yes.
So I think perhaps it is useful to know the percent of occurences of the top 5 crime types. I abbreviated the categories to get them plotted apropriate.



Next I clustered as mentioned above the crime timestamp into intervals of 4 hours and plotted 9 more often crime types (normalized) as a line plot. With this graph you can see the occurences of different crime types at specific time. So you can distinguish the different crimes. I plotted only some crime types. With more types it gets more unclear.
As you can see there is a general time depencence of crimes with some little differences.



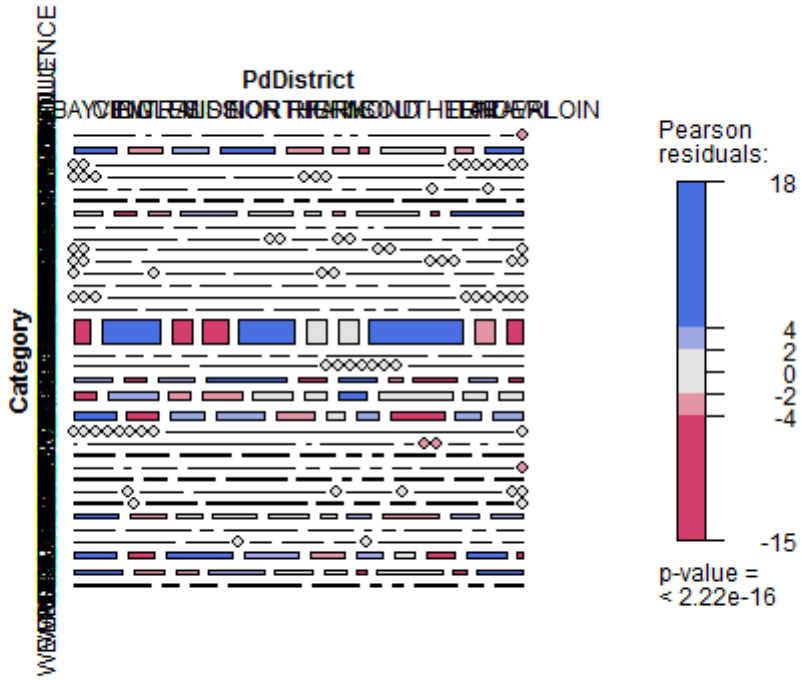
This plot was done by normalizing the different crime types as f.e.

```
larcency <- table(data[data$Category=='LARCENY/THEFT',]
$time_cat)/sum(data$Category=='LARCENY/THEFT')
```

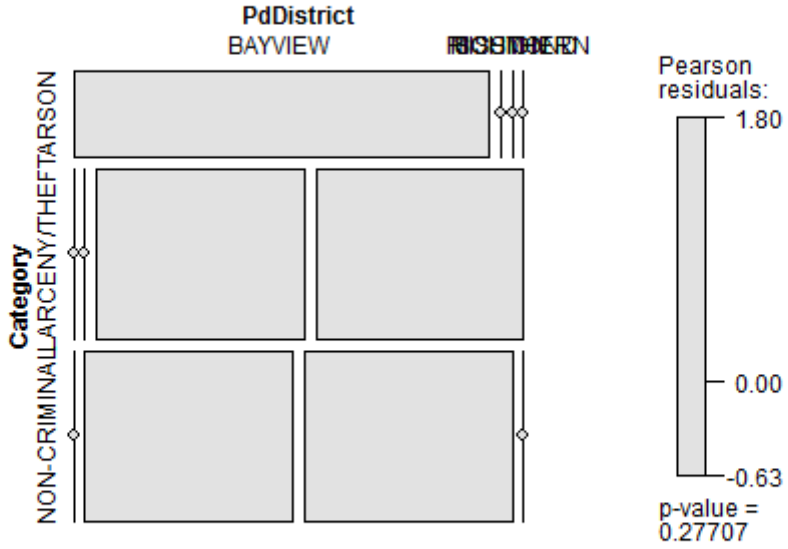
and the plot

```
plot(0:max, larcency, type="l", col=1, axes=FALSE,
      ylab="Percent of Category", ylim=c(0,0.35),
      xlab="Hour of the Day", main="Percent of Crime by Time of
Day")
axis(2, at=(0:7)/100*5, tick=TRUE)
```

The last diagram I produced is a mosaic plot. The variables I useed are the categories of crime, their occurencies, the PdDistrict and the occurences in these districts:



It is really not possible to make a suitable graph with all PdDistricts and the categories together with the pearson residuals (as calculated by the mosaic). Als if I only 5 combinations of the data above the diagram



is not very helpful. The information inside is very interesting, but to understand the graph as presented needs to much explanations to gain insights. But for scientific purposes it is, maybe, very interesting and useful. If we look for insights of partial combinations of Category vs. PdDepartment (by filtering out these combinations) would be a better choose.

The corresponding R code for the mosaic is:

```
comb <- data[,c("Category","PdDistrict")]
mosaic(table(comb1), shade=TRUE)
```

or the cutted dataframe by `comb[1:5,,]`

As usual there are a lot of other possibilities to investigate the data, but as this is a visualization assignment and not a R course I stop here. The graphs offered are useful, but for a dashboard for management I wouldn't use R as it seems to be too scientific in presentation.