Unsupervised Learning: PSet 1

Felipe Alamos

10/14/2019

Exploration & Computation

I will work with the data set of crimes in the city of Chicago during 2018. Downloadable here

Visualization

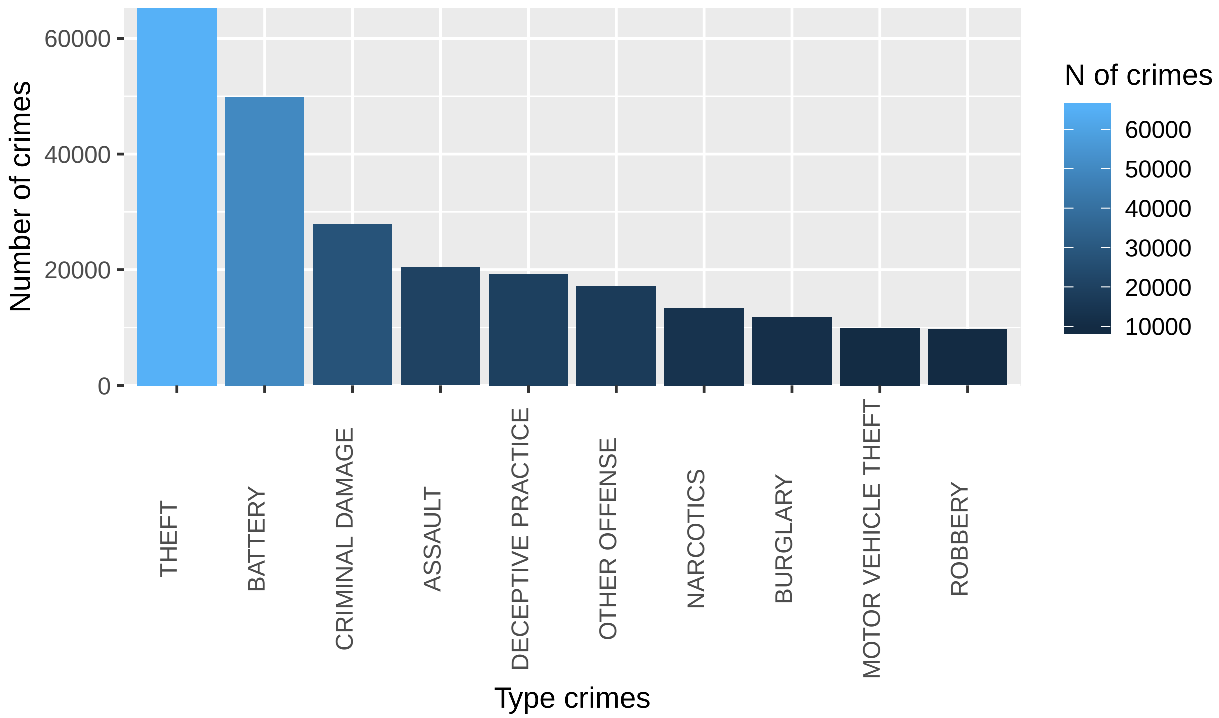
```
library(ggplot2)
library(tidyverse)
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 3.5.3
```

```
crimes_data <- read.csv("crimes-2018.csv")

#Count how many crime of each type occurred during 2018
counts_by_type <- crimes_data %>%
  count(Primary.Type)

#Order by type
counts_ordered <-counts_by_type[order(-counts_by_type$n),]

#Select top 10
top_10 <- head(counts_ordered, 10)

#Plot geom bar
ggplot(data=top_10, aes(x=reorder(Primary.Type, -n), y=n, fill=n))+
  geom_bar(stat = "identity")+
  scale_y_continuous(
    name="Number of crimes",
    expand=c(0,0)
    )+
  theme(axis.text.x = element_text(size=9, angle = 90, vjust=-0.001))+
  labs(title = "Top 10 types of crimes commited in Chicago during 2018",
      subtitle = "Most common type is thefts, followed by battery",
      caption = "Source: Chicago Data Portal",
      x = "Type crimes",
      fill = "N of crimes")
```

Top 10 types of crimes commited in Chicago during 2018

Most common type is thefts, followed by battery



Source: Chicago Data Portal

Description

We can observe the top 10 most common types of crimes in the city of Chicago during 2018. In particular, we can observe that theft is the most common type of crime, with over 60,000 occasions. Battery closely follows with around 50,000 number of crimes. Interestingly, there is a significant jump in the amount of the next type of crime. As a matter of fact, we observe that there not a huge difference in numbers between the 4th most common type of crime - assault - and the 10th - robbery.

Common measures of central tendency and variation

```r
skim(counts_by_type)
```

```
## Skim summary statistics
##  n obs: 32
##  n variables: 2
##
## -- Variable type:factor ----------------------------------------------------------------------------
##      variable missing complete  n n_unique                     top_counts
##  Primary.Type       0       32 32       32 ARS: 1, ASS: 1, BAT: 1, BUR: 1
##  ordered
##    FALSE
##
## -- Variable type:integer ---------------------------------------------------------------------------
##  variable missing complete  n    mean       sd p0   p25    p50      p75
##         n       0       32 32 8359.91 14979.33  1 170.5 1214.5 10424.25
##   p100     hist
##  65234 <U+2587><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
```

5. Description of the numeric output

- We observe that 32 different type of crimes occur in Chicago during 2018
- An average type of crime occurred around 8,000 times. But the most frequent type of crime occurred more than 65,000 times, and the least only once.
- We observe a significant spread in the quantities of crimes for the different types of crimes. This is easy to observe, for example, comparing the different first quartile (p25) and the third quartile (p75). This might occur because of the simple reason that some crimes are easier to execute or more profitable for criminals. Nevertheless, it could also be because of the complexity or granularity of different types of crimes (for example, “thefts” may involve a variety of crime activity, whereas “armed car robbery” is very specific)
- Observing the frequency of the most important crimes in the city is important so that, for example, crime polices can be designed to address with more priority the most common types.
- It is also useful to know the different ways crimes are aggregated, and if this aggregation is sparse or dense.

Critical Thinking

1. Describe the different information contained in/revealed by visual versus numeric exploratory data analysis. (Hint: Think of different examples of each and then what we might be looking for when leveraging a given technique).

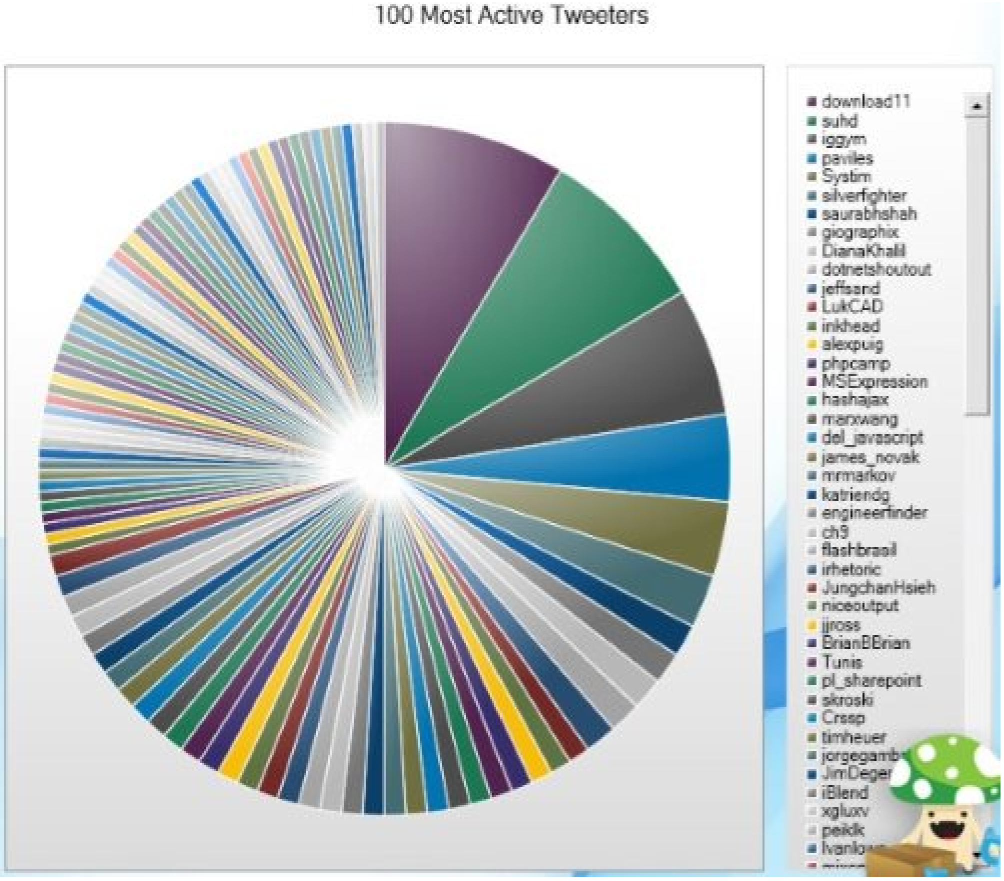
Numeric exploratory data analysis usually reveal numbers that aim to represent the data being explored. These numbers tend to give the audience a quick glance of measurements that are already familiar to him/her.A classic example would be all the information revealed by the `summary` function in R: means, quartiles, standard deviations, etc. Numeric exploratory is usually useful to give a quick impression of a big picture of a data set.

Visual explanatory analysis, on the contrary, intend not to reveal information through numbers but by the use of images. Visualizations usually allow users to have a more intuitive and broader representation of the data: rather than focusing in a few numbers, visualizations allow to grasp a lot of information from a simple glance. A simple example would be a histogram, showing the distribution of values that certain numeric variable receives.

Anscombes Quartet example shows us the power of visualizations and how numeric exploratory data analysis can sometimes *lie* to us.

2. Find (and include) two examples of “bad” visualizations and tell me precisely why they’re bad.

Bad Visualization 1

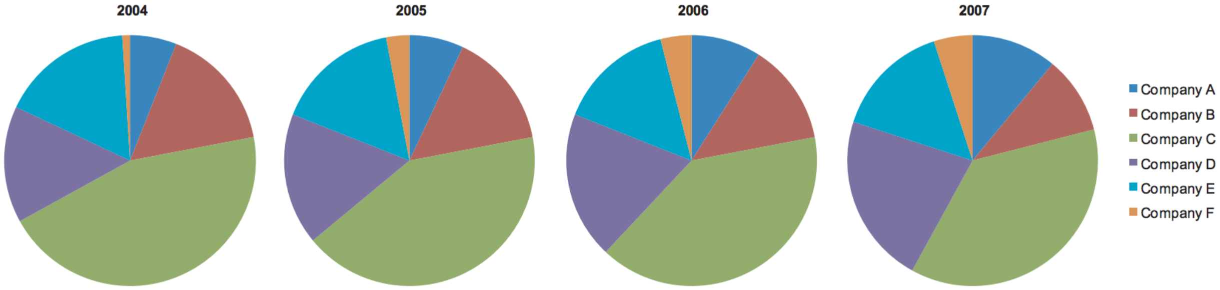


:

Source: http://livingqlikview.com/the-9-worst-data-visualizations-ever-created/

There are a couple of terrible problems with this visualization. The most obvious one, its impossible to differentiate in the pie chart which piece is related to each label. This due to several reasons: there are too many different categories, colors between the different labels are too similar, etc. In addition, sizes of the different slices are usually hard to compare in pie charts, and particularly in this one.

Bad Visualization 2



Source: Alex C Englers presentations.

The biggest problem with this visualization is trying to show a time-series with pie charts. The problem is that - as seen in the picture - it is extremely difficult to understand how the categories evolve in time. This is particularly challenging because the positions of the boundaries of each section *move*, making it hard to compare them between pie charts.

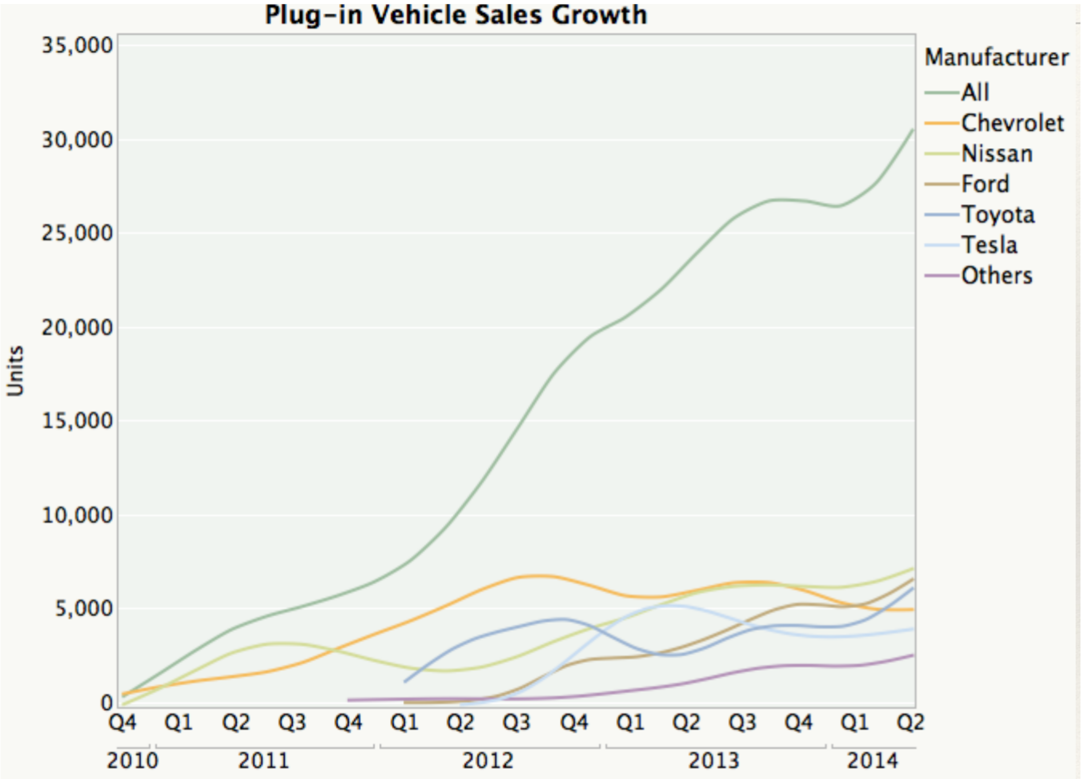


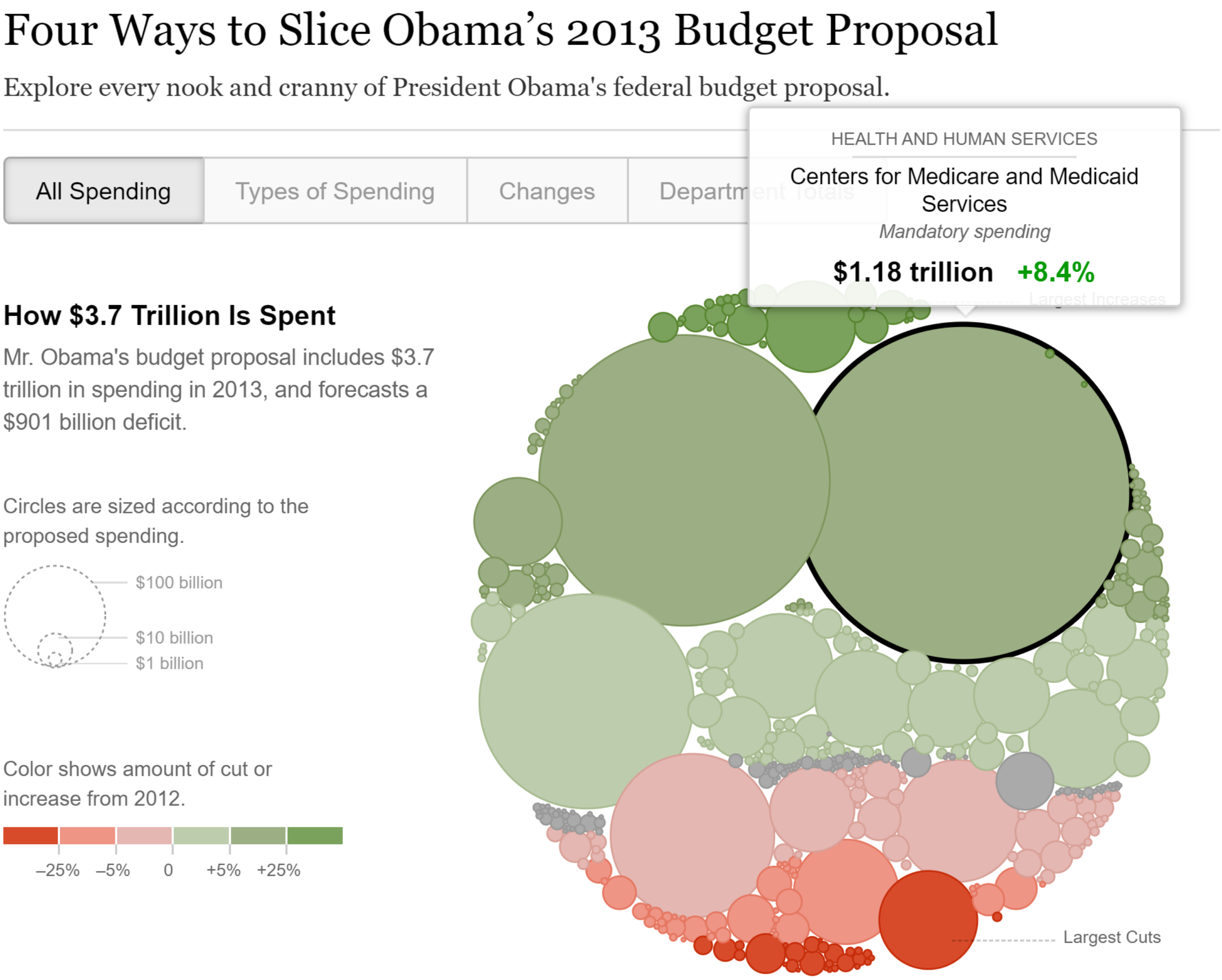
Figure 1:

3. Find (and include) two examples of "good" visualizations and tell me precisely why they're good.

Good Visualization 1

This is a good and simple visualization. First, it contains the basic elementary elements: title, legends, and axis names. It is particularly useful and distinctive how the author used two x-axis lines two give different levels of granularity in time. Colors are easy to differentiate. Also, legends are ordered in the same order as they appear in the graph (top bottom wise), and hence its easy to connect the lines with the legends.

Source: Solomon Messing's Blog



Four Ways to Slice Obama’s 2013 Budget Proposal

Explore every nook and cranny of President Obama's federal budget proposal.

How $3.7 Trillion Is Spent

Mr. Obama's budget proposal includes $3.7 trillion in spending in 2013, and forecasts a $901 billion deficit.

Circles are sized according to the proposed spending.

Color shows amount of cut or increase from 2012.

HEALTH AND HUMAN SERVICES

Centers for Medicare and Medicaid Services

Mandatory spending

$1.18 trillion +8.4%

Figure 2:

Good Visualization 2

Source: https://archive.nytimes.com/www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html?_r=0

This interactive visualization has the great virtue of exposing a lot of information in a easy-to-consume and intuitive way. First o fall, colors are easy to distinguish and understood thanks to the legend. Secondly, size of circles are a great way of getting a fast impression of the amounts they represent. There are also interesting annotations that come with the graph, helping the reader. Lastly, the interactivity allows the user to explore details of each particular category in a manageable way.

4. When might we use EDA and why/how does it help the research process?

We use EDA when we want to summarize the main characteristics of a data set. In contrast with a statistical model, EDA is used to understand what can we learn from the data beyond an explicit model or test.

In particular, EDA can help the research process by:

- Providing a quick and first impression of the data set being worked with
- Giving intuitions or suggestions on hypothesis about why an observed phenomena is occurring
- Support the selection of further techniques for understanding the data, such as appropriate statistical tools [1]

5. What did John Tukey mean by "confirmatory" versus "exploratory"? Give me an example for each.

Confirmatory data analysis - or statistical hypothesis - is the exercise of testing a hypothesis based on sample data and a proposed model of the data generation process. So, for example, given that we know the probability distribution of kids weight in a school, we might ask what is the probability of finding randomly choosing a kids whose weight is above or below certain amount.

John Tukey suggested that confirmatory data analysis competed with exploratory data analysis in the sense that, in many occasions, too much interest was put on the former rather than on the latter. He argued that EDA could be used to suggest hypotheses that could be then tested during confirmation analysis. Nonetheless, he also warned of the risk of confusing the two types of analysis, and that one could lead to bias on the other creating a circular dependency. [^ https://en.wikipedia.org/wiki/Exploratory_data_analysis]

Typical graphical techniques used in EDA are scatter plots, visualizations where two variables are plotted along two axes, and where the resulting points on the graph might reveal any correlations between the variables.

A classic confirmatory analysis would be a z-test, where we aim to reject a given hypothesis testing assuming a normal distribution of the variable of interest. It is particularly common in cases were we have a sample of data , and want to know if this data is sufficient evidence to establish certain conclusion about the population the sample represents.

[1] Behrens-Principles and Procedures of Exploratory Data Analysis-American Psychological Association-1997. http://cll.stanford.edu/~willb/course/behrens97pm.pdf