Data Analysis and Visualization in R

Ilaria Battiston *

Winter Semester 2019-2020

*All notes are collected with the aid of material provided by J. Gagneur. All images have been retrieved by slides present on the TUM Moodle Course Webpage.

Contents

1 Introduction

Data science is an interdisciplinary field about processes and systems to extract knowledge or insight from data (structured or unstructured). It is a continuation of fields such as statistics, data mining and predictive analytics.

Data science is useful to discover new phenomena, make new hypotheses from observations, analyze results and extract conclusions.

1.1 R

R is an Open Source programming language aimed for statisticians to interactively explore data, making easier the process of thinking and understanding before computing results using a machine.

It is a high level language, loosely typed (variables are not declared). Running time, on the other hand, might be slow. R has thousands of packages which can be used in analytics or developed by users, and an integrated markdown language.

Data is imported and manipulated using most kinds of text and tabular files, and stored in R using a data frame. It can be plotted with many techniques such as scatterplots, dimensional reduction and clustering.

Data exploration leads to hypotheses which can be proven (with R, of course) using for instance Fischer test, t-test and Wilcoxon.

R works on command line, and eventually its main IDE RStudio along with R markdown.

```
1 + 1 # this is a comment
x <- x + 1 # assignment
exp(2) # exp(0) = 1
ls() # variables in the environment
norm_vec <- rnorm(100, 5, 2) # normal distributed vector of random numbers
hist(norm_vec) # histogram
```

RStudio displays assigned variables in the workspace in the top right corner. The documentation of each function is available with the ? sign before their name.