Exploratory Data Analysis *or* 'Data Mining'

Marc A.T. Teunis, PhD

2019-08-28

Contents

Chapter 1

Prerequisites

1.1 Purpose of this website

This website is the accompanying website to the "Exploratory Data Anlysis" Course for the Master of Informatics of the Univeristy of Applied Sciences. It contains seven chapters that correpond to the seven so-called 'labs' or 'master-classes' of the course. The course content and further structure of the course are explained in **??**. All methods described in this website and used in the course relate to the Language and Environment for Statistical Computing called **R** (**?**). To work more effeciently in R we will be using the most commonly used Integrated Development Environment for R: **RStudio**(**?**).

1.2 R Packages

When you install R for the first time, you only get a basic installation. This so-called base R includes the R core of the language but does not get you very far if you really want to do more complex data analytics. There are many so-called 'packages' to add on to this base-R installation and for this course we use many of them. The code below retrieves the current amount of packages published on the Comprehensive R Archiving Network, which is a main resource for R-packages.

```
library(rvest)
```

```
## Loading required package: xml2
```

```
pkgs <- read_html("https://cran.r-project.org/web/packages/available_packages_by_name.html")
mylines <- pkgs %>%
    html_nodes("tr") %>%
```

```
    xml_text()

nb_pkgs <- length(which(sapply(mylines, nchar)>5))
```

As can be discovered from the code above

Other important resources for R-packages are:

- BIOCONDUCTOR.
- Github
- Bitbucket
- ROpenSci

A full list of packages that are needed for this course is available in the package appendix **??**

1.3 Getting the materials

To compile this website locally, you can clone the website repository from

https://github.com/uashogeschoolutrecht/edamoi_site.

Click the Build button to build the website on you local computer. This repository can also be used to acces the course materials directly in RStudio during the classes.

1.4 Getting R and RStudio

During the course we will be working on a Cloud Computing Environment which provides webaccess to R and RStudio via a Virtual Machine. This machine runs a server edition of RStudio and has the latest R version and all the required packages available. In order to access the server you will need credentials, which you will recieve before the course starts. This is a convenient way to use R in a course and during the course we will only use this environment. This is to ensure reproducibility and prevents a lot of trouble shooting from your side and the teacher's side.

If you want to use R and RStudio locally on your laptop (the teacher will not support this during the course), this is where you can download the software from:

RStudio

R

If you want to use R after the course, this is what you will need to do because the accessibility to the RStudio server will only be guaruateed untill a month

after the course has finished. During the course, I will show you how to manage getting files to and from the server using an FTP client.

1.5 Bring your own or BYOD!

During the course we will use a lot of different datasets that are available directly from R, in R-packages or from open data sources on the internet. If you want to bring you own data to practice with you can and this is encouraged! Please be aware that I may want to share your data and/or analysis (issues) with the rest of the participants, so please bring only data for which this is allowed.

1.6 The {bookdown} package

This website was created using the {bookdown} package written by Yihui Xie. The package can be downloaded from CRAN. For more information see the documentation.

The {bookdown} package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Chapter 2

Introduction

2.1 Exploratory Data Analysis

The process of Exploratory Data Analysis (EDA) is not a formal strict process. It involves interative cycles of `loading`, `cleaning`, `wrangling`, `visualizing`, `communicating` data and patterns in the data. The process of EDA is directed towards gaining insight in the data in basically any way you can. It can involve a number of statical approaches that are sometimes collectively called IDA (Initial Data Analysis), which is the the description and check if undelying assumptions for any formal statitical modelling are met. Usually the formal statistical inference is out of scope of the EDA process, although we will see examples in which statical modelling can help us perform EDA better.

Because it is a not-so-strict inerative process there is no formal manual on which steps to peform in the EDA process. You could consider doing EDA more as being in a certain state-of-mind. To help you overcome this rather abstract way of looking at this process, several authors have created a check list to aid performing EDA in a more structured way. Here I will go over such a check list, but bare in minf that is only an aid and no formal manual:

When doing EDA you should always keep an open mind to deviate from the checklist, skip a check-box or add one yourself.

2.2 EDA checklist

"If a checklist is good enough for pilots to use every flight, it's good enough for data scientists to use with every dataset." A quote from Daniel Bourke on Towards Data Science

When starting EDA consider:

1. What question(s) are you trying to solve (or prove wrong)?
2. What kind of data do you have and how do you treat different types?
3. What's missing from the data and how do you deal with it?
4. Where are the outliers and why should you care about them?
5. How can you add, change or remove features to get more out of your data?

I will go over each in more detail in @ref(lab5_eda). If you want to get on with a first example to have some practice go directly to @ref(eda_case)

2.3 The `{tidyverse}` packages

2.4 Inference and Modelling

2.5 R: A Language and Environment for Statistical Computing

2.6 Getting help in R

2.7 Resources for learning

Chapter 3

Course Contents

3.1 Data Mining Course

This course is about "Exploratory Data Analysis and Initial Data Analysis"

Wikipedia definition

```
In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize
```

3.2 Course aims

- EDA is not a formal procedure, getting the right mind set
- Learn tools in R to GET, CLEAN, EXPLORE and MODEL data
- Acquire R skills for the complete EDA cycle; object oriented and functional programming

- Visualize data
- Explore assumptions (IDA)
- Using R for Reproducible Research

3.3 Course contents

Each course masterclass or so-called 'lab' will carry forward a specific theme related to Exploratory Data Analysis. In this course you will exclusively work with the Programming Environment `R` in the Integrated Development Environment `RStudio`. There will be no need to install any software on your laptop, because we will be using a cloud computing solution. This offers flexibility and

speed. I will show you where to get the required software. In case that you would be wanting to install your own environment locally.

In this course you will get in-depth knowledge on how to use R in conjunction with RStudio to `IMPORT`, `INVESTIGATE`, `CLEAN`, `VISUALIZE`, `EXPLORE`, `MODEL` and `COMMUNICATE` data and conclusions about data.

To this end, I devided an number of logically connected topics together in seven `labs`. Each lab has several interative cycles of theory and exercises. During each lab, I will explain a small amount of topics after which the course participants will have the opportunity to practice with exercises.

3.4 BYOD; Bring Your Own Data

During this course you will have the opportunity to bring your own data as case example. Please think about which data you will be able to (freely) share with me and your fellow paticipants.

3.5 Lab Contents

Below, I will shortly summarize what we will be covering in each lab (seven in total for the complete course).

3.5.1 Lab 1: Introduction to R and RStudio

1. Getting the course materials from github.com
2. Creating objects in R
3. RStudio introduction
4. Object Class; vectors, dataframes, lists
5. Vector types
6. Getting help
7. Plots
8. Data examples

3.5.2 Lab 2: Visualize & Explore Data

1. Build in datasets
2. Using the grammar of graphics with `{ggplot2}`
3. Plot types
4. `geom_...` and `aes()`
5. Adding dimensions to a plot
6. Solving overplotting

7. Plot annotation and labels
8. Saving plots
9. Examples of `{shiny}` apps

3.5.3 Lab 3: Data Wrangling & Functional Programming

1. Using `{dplyr}` for data wrangling
2. The `{dplyr}` verbs
3. Loops and `map` family of functions
4. Scripts and the `source` function

3.5.4 Lab 4: Importing data & Getting (Open) Data

1. Open data sources
2. Finding data
3. Importing data into R (.csv, .tsv, .txt, .json, .xml, .xls(x)
4. Using APIs (Twitter, Kaggle, Google, CBS) https://medium.com/@traffordDataLab/querying-apis-in-r-39029b73d5f1

3.5.5 Lab 5: Exploratory Data Analysis *or* The PPDAC Cycle

1. The process of EDA
2. Problem-Plan-Data-Analysis-Conclusion (PPDAC) Cycle
3. Missingness
4. Distibutions
5. Finding patterns
6. Graph types
7. Multidimensional data
8. Principal Components

3.5.6 Lab 6: Exploring Assumptions & Models

1. Why Assumptions?
2. Distributions (Gaussian, Poisson, Uniform, Binomial)
3. Regression
4. Model output (`{broom}`)
5. Managing many models
6. Linear and Generalized Linear models
7. A few machine learning examples in R (`{caret}`)

3.5.7 Lab 7: Communication: Reproducible Research, RMarkdown and R-packages

1. RMarkdown
2. The 'RMarkdown First' Principle
3. Using Git/Github together with R/Rstudio
4. Open Science
5. `{bookdown}` / `{pagedown}` and `{blogdown}` packages
6. `{usethis}` for creating R-packages

Chapter 4

EDA Case Example - The Curious Case of the H. Shipman Murders

"Thanks to David Spiegelhalter for pointing me to this nice example for my teching. I hope you allow me to 'borrow' it from you"(**?**)

4.1 The case

4.2 The checklist

1. What question(s) are you trying to solve (or prove wrong)?
2. What kind of data do you have and how do you treat different types?
3. What's missing from the data and how do you deal with it?
4. Where are the outliers and why should you care about them?
5. How can you add, change or remove features to get more out of your data?

4.3

Chapter 5

Appendix

5.1 Packages

The packages that you will need for the course can be installed with the following command

```
## read list of course packages:

pkgs <- read.csv(here::here("course_packages.txt"), stringsAsFactors = FALSE)
list <- as.list(pkgs$X...list.of.course.packages)
lapply(list, install.packages)
```

Chapter 6

References