Data Wrangling with R

http://bit.ly/r_wrangling

Bernhard Piskernik

About me

Senior Data Scientist at Novomatic

Background in Humanities (Psychology)

R

programming language for data analysis and statistics

it is free - Open Source-Software with GNU General Public License (GPL)

it is also very popular in certain application areas (statistics, bioinformatics, …)

it is a dynamically typed interpreted language, typically used interactively

very extensible (>14k libraries at CRAN)

interfaces to add functions in Fortran, C, C++, …

What makes R special?

it is old (1992, but based on *S* which was developed 1972)

it is explicitly build for data analysis and statistics (do not try to use it for general purpose programming)

it is 1-indexed

syntax WAS obnoxious before *tidyverse*

in combination with RStudio (IDE) very pleasant data wrangling and analysis experience

The fundamental concept of R: vectors

```
a <- rep(1,10)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1
```

```
b <- 1:10
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
a+b
```

```
##  [1]  2  3  4  5  6  7  8  9 10 11
```

```
a+2*b
```

```
##  [1]  3  5  7  9 11 13 15 17 19 21
```

Note: arithmetic operations act element-wise

Building k-nearest neighbor classification from scratch

DEMO

Quick overview of kNN-classification (naive)

kNN is a non-parametric method for classification

- compute the distances from the test examples to all stored examples
- get k-nearest neighbors per case
- conduct majority vote on class membership