

An Introduction to Programming in R (Part 3)

Philipp Buschmann

Hochschule Fresenius - Market Research and Empirical Research Methods

Winter term 2024

Contents

- 1 Importing Data
- 2 Data processing with tidyverse

Contents

1 Importing Data

2 Data processing with tidyverse

How to import csv and txt files?

- csv-files:


```
read.csv(file, header, sep)
```

- txt-files:

```
read.table(file, header, sep)
```

- file is the path which the data are to be read from. You have to insert it as character string.
- header is a logical value indicating whether the first row of the data set contains column names.
- sep specifies how the data points in the file are separated. In a csv-file it is typically a comma, in a txt-file it is usually the space character. You have to enter the separator as character string.

Task 1

- a) Download the file `NetflixOriginals.csv` from Ilias.
- b) Import the data set into  and store it in an object `Netflix`.
- c) Try the commands `str(Netflix)`, `dim(Netflix)` and `head(Netflix)`.

Contents

1 Importing Data

2 Data processing with tidyverse

Installing R-packages

- R-packages are extensions of the built-in commands in R.
- A package is a collection of functions, codes and data.
- For using a package, you have to install and load the package.
- Installation is done by

```
install.packages("packagename")
```

- You can load the package by

```
library(packagename)
```

packagename is a place-holder for the name of some package.

The tidyverse package

- To be precise, tidyverse is not a single package. It is a collection of packages providing tools for data manipulation, visualization, and analysis.
- If you install `tidyverse`, you install all of these packages.
- We will consider the `dplyr` package. It contains functions allowing you to easily filter, select, mutate, arrange, and summarize data.

Data manipulation with dplyr

List of some basic dplyr-functions:

- `filter()`: Selects rows of data based on specified conditions.
- `select()`: Picks specific columns or variables from a data frame.
- `mutate()`: Creates new variables or modifies existing variables of a data frame.
- `summarize()`: Computes summary statistics.
- `group_by()`: Groups the data for subsequent operations.

The pipe operator `%>%` can be used to chain together multiple functions in a sequence.

Task 2

Import the data set `NetflixOriginals.csv` before starting.

- a) Create a data set that only contains the variables `Title`, `Runtime` and `Genre`.
- b) What is the longest and the shortest runtime in the data set?
- c) Compute the average runtime for each movie genre.

Task 3

Import the data set `NetflixOriginals.csv` before starting.

- a) How many German movies are included in the data set? What is their average IMDB Score?
- b) Which movie has the longest running time? What is the language of this movie?
- c) Extend the data set by the variable `RuntimeHour`, which measure the runtime in hours.