Course Summary and Further Resources

Martin Burger

STATS PROGRAMMING TUTOR

www.r-tutorials.com

Summary

Useful resources for further studies

Tidyverse: A collection of R libraries dedicated to data pre-processing

- E.g.: Libraries dplyr and tibble

Planning your learning path and its pitfalls

Course summary and resources

About Learning R

Optimizing your learning path is key when learning R programming

Avoid focusing on a narrow field of application

- Do not limit your knowledge only on certain problems
- Data science challenges vary a lot

Start with a broad foundation: Basic skills required in all scientific disciplines

- Data pre-processing/wrangling

The Elementary R Skillset

Data import

Exploratory analysis

Data class selection

Data visualization

Missing value imputation

Querying

The Elementary R Skillset

A collection of R libraries were developed to effectively perform foundational data science tasks

- Tidyverse: The universe of tidy data

Once the foundation is built, you can branch out to specific fields

- Machine learning, time series analysis, econometrics or other sub-disciplines

What Is the Tidyverse?

Tidyverse

A collection of R libraries that work together in order to achieve clean and tidy data.

Basic Principles of the Tidyverse

Coding with the pipe operator (%>%)

Functional, easy to understand code

Memory efficiency

Installing the Tidyverse

Installing and loading the whole Tidyverse is not recommended (>70 libraries)

Get only the libraries you need to avoid conflicts and improve performance

```
library(tidyverse)

    >ggplot2

    >dplyr

    >tidyr

    >readr

    >purrr

    >tibble

    >stringr

    >forecats
```

- The core libraries of the Tidyverse
- Calling the Tidyverse activates its core packages only
- Other libraries of the Tidyverse must be activated individually

Data Visualization with Library ‘ggplot2’

Complex, high quality, publication-ready data visualizations

Graphs are coded in a sub-language of R built around the pipe operator

Data Manipulation with Library ‘dplyr’

Data frame (tibble) manipulation

Joining tables

Sorting data

Running queries

Rearranging data

Summary statistics

Data Pre-Processing with Library ‘tidyr’

A toolbox to clean and tidy up datasets

Conversion between wide and long table formats

Splitting and merging data on demand

Data Import and Custom Functions

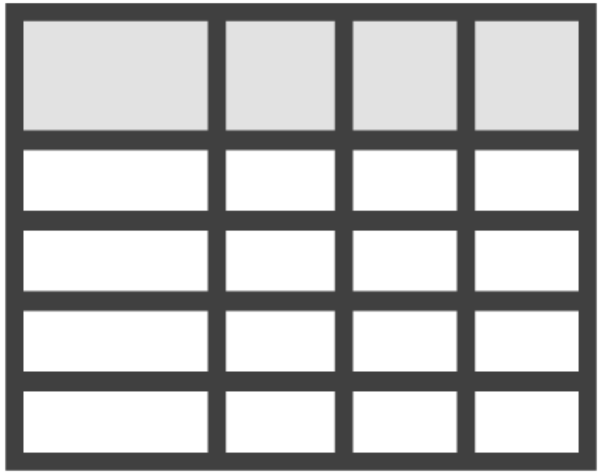
Library ‘readr’

Data import toolset that warns for data irregularities and unintended transformations

Library ‘purrr’

Improved functional programming toolkit for working with functions and vectors

Improved Data Structures with Library ‘tibble’



Class tibble is an updated version of data_frame (deprecated)

- Alternative class to data.frame and data.table

Improves data.frame functionalities:

- Recycling can be controlled
- No unintended type conversion
- Clean layout with data type information

String Manipulation with Library ‘stringr’

- Working with character data
- Upper- and lowercase conversion
- Splitting and concatenation
- Finding letter combinations in text
- Improvement on gsub operations

Factor Manipulation with Library ‘forecats’

Counting observations of a category

Fusion of categories

Relabeling categories

```
library(tidyverse)

    >ggplot2

    >dplyr

    >tidyr

    >readr

    >purrr

    >tibble

    >stringr

    >forecats
```

- The core libraries of the Tidyverse
- Calling the Tidyverse activates its core packages only
- Other libraries of the Tidyverse must be activated individually
- Date and time related secondary packages: lubridate, hms

Does the Tidyverse offer a solution to all data science tasks and challenges?

Focus Points of the Tidyverse

Data import

Data cleaning

Data visualization

Custom functions

Course Summary

Understanding Dataset Structures and Formats

Getting familiar with RStudio, a widely used graphical user interface for R programming

Table-like structures: data.frame (R Base), data.table (data.table), tibble (dplyr)

Selecting and Converting Data Types

Numeric (double, float)

Integer

Character (string)

Factor

Boolean (binary)

Date time (POSIXt)

Numeric and integer values: Continuous measures and counts

Character: Text with unlimited possibilities of character combinations

Factor: Grouping variable with a given number of categories

Boolean: True and false values used for binary classification of observations

Date and time values:

- Classes POSIXt, chron and Date
- Focus on format and time zone

Querying and Filtering Data

Extracting parts of the data based on index positions or logical conditions

Query systems in R:

- Data.frame, data.table, tibble

Equal efficiency, but user preferences may differ

Further Resources and Course Summary

Setting up a learning path

Exploring the Tidyverse

Resources for Further Studies

Resources and Further Studies

Where to get help and information

Further steps of learning R

Programmer and developer community at stackoverflow.com

- Discuss specific coding problems

General topics on R programming at r-bloggers.com

- Aggregator of R related blogs

R learning paths at Pluralsight:

- Managing Data in R Using Data Frames

Good Luck in Your Career

Time spent on learning R is well invested

Keep on learning: See you in another Pluralsight course