Einführung in Data Science und maschinelles Lernen mit R

Einführung



- Vorstellung
- Kursinhalte
- Was ist Data Science?
- Warum R?
- Wozu Rstudio?
- Datenstrukturen in R

 Am Ende der Veranstaltung immer die Anwesenheitsliste unterschreiben

Bitte Bescheid sagen, wenn ihr aussteigt!

Einzelvorstellung

Name Interesse

Kurzfragebogen

bit.ly/oc-datascience



Vorstellungsrunde

5 Minuten jeweils 3 sich unbekannte Personen

Vorstellungsrunde

5 Minuten jeweils 3 sich unbekannte Personen

29.10.	Einführung
18: 00- 20: 00	Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
05.11.	Grafische Darstellung von Daten
18: 00- 20: 00	Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
12.11.	R-Projekte und Versionierung mit Git und GitHub
18: 00- 20: 00	Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
19.11.	Datenaufbereitung mit Tidyverse
18: 00- 20: 00	Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
07. 01.	Einführung in das maschinelle Lernen
18: 00- 20: 00	Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
14. 01.	Support Vector Maschinen
18: 00- 20: 00	Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
21. 01.	Neuronale Netze und Deep Learning
18: 00- 20: 00	Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
28. 01.	Präsentation der Auswertungsprojekte
18: 00- 20: 00	Starterkitchen, Kuhnkestraße 6, Wissenschaftspark



"What's the difference between data science, machine learning, and artificial intelligence?"

(http://varianceexplained.org/r/ds-ml-ai/)

Data science produces insights.

Machine learning produces predictions.

Artificial intelligence produces actions.

DATA SCIENCE

Anwendung statistischer Auswertungsverfahren

mit deren Hilfe Ergebnisse für einen Bericht oder Artikel erzeugt werden.

- Deskriptive Statistiken
- Statistische Schlussfolgerungen
- Visualisierung von Daten
- Design von Experimenten

MASCHINELLES LERNEN

Anwendung eines statistisches Auswertungsverfahren

zur Schätzung von Modellparametern (auf Basis einer Kreuzvalidierung)

mit deren Hilfe eine Funktion erstellt wird,

die Prognosen liefert und die Modellparameterschätzung durch Hinzufügen neuer Daten ständig verbessert.

- Prognose zukünftigen Verhaltens
- Prognose unbeobachteten Verhaltens
- Prognose aktueller Zustände (etwa, ob ein Bild einen Vogel enthält)

KÜNSTLICHE INTELLIGENZ

"an autonomous agent executes or recommends actions"
(Poole, Mackworth, & Goebel, 1998)

"Systeme mit einem 'intelligenten' Verhalten, die ihre Umgebung analysieren und mit einem gewissen Grad an Autonomie handeln, um bestimmte Ziele zu erreichen."

(Europäische Kommission, 2018)

"Unter Künstlicher Intelligenz verstehen wir hochentwickelte Softwaresysteme, welche lernfähig und trainierbar sind, um komplexe Aufgaben bewältigen können."

(KI-Strategie des Landes Schleswig-Holstein, 2019)

KÜNSTLICHE INTELLIGENZ

- Programme, die Spiele spielen (Deep Blue, AlphaGo)
- Programme zur Steuerung von Robotern
- Programme zur Optimierung (Google Maps)
- Programme zur Verarbeitung natürlicher Sprache (Bots)

EIGENSCHAFTEN VON R

- Von Statistikern entwickelte Auswertungssprache (Python etwa kommt aus dem IT-/Tech-Umfeld)
- Konsolen-Programm ohne grafische Benutzeroberfläche

```
# Prepare Environament
 3 #####
    setwd("C:/Users/Steffen/Arbeit/opencampus/99_source-code/elearning-statistics/00_functions")
    source("02_LimeSurvey-API.R") # import LimeSurvey API
    setwd("../../auswertung pre-challenge")
    library(inspectdf)
    library(dplyr)
    library(readr)
    library(stringr)
12
14 - #####
    # Import Answer Data
    introData <- as_tibble(ls.getResponses(615549, sHeadingType = "code")) # import of the Introductory Questionnaire
    closingData <- as_tibble(ls.getResponses(722958, sHeadingType = "code")) # import of the Closing Questionnaire
21 - #####
22 # Prepare Table for Scoring
23 - #####
24 # Shrink Datasets to Knowledge Items and join them
    introItems <- select(introData , matches("Q2|Q3"))</pre>
    closingItems <- select(closingData , matches("Q2|Q3"))</pre>
    # shorten name to exclude questioannire ID
    names(introItems) <- substr(names(introItems),3, 10)</pre>
    names(closingItems) <- substr(names(closingItems),3, 10)</pre>
    # join items according to question IDs
    items <- full_join(introItems, closingItems)</pre>
    # contruct table including all given anwers, how often they occured (percentage and count) and include corresponding score categories if provided
    scoreTable <- get.scoreTable(items)</pre>
    write.csv2(file="ScoringGuide_new.csv", scoreTable, row.names = F, na="")
    scoreTable <- read_csv2(file="ScoringGuide_new.csv")</pre>
```

NACHTEILE VON R

- Man muss programmieren
- Im Bereich des maschinellen Lernens weniger verbreitet als Python
- Weniger Schnittstellen (APIs) als Python zu speziellen Anwendungen im IT-Bereich

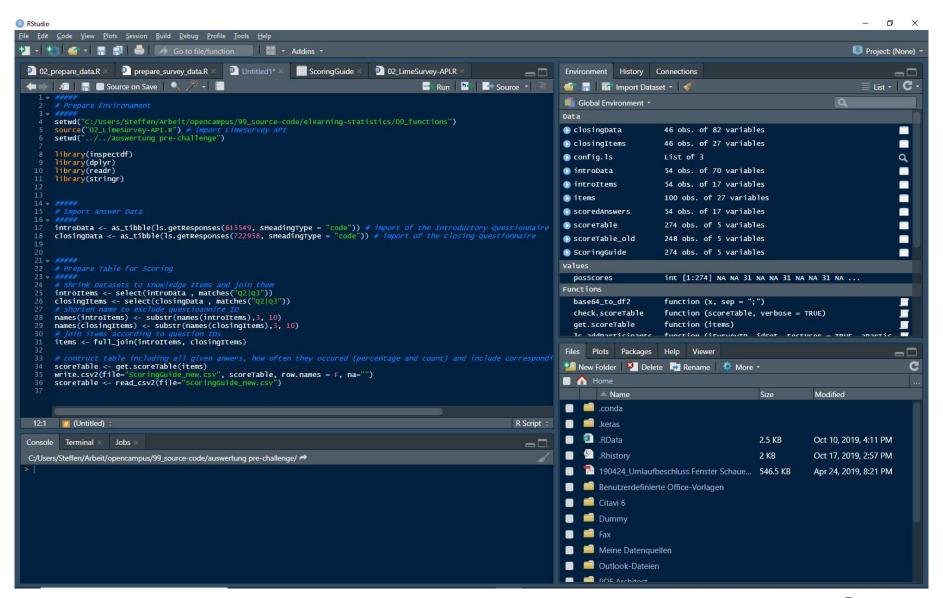
VORTEILE VON R

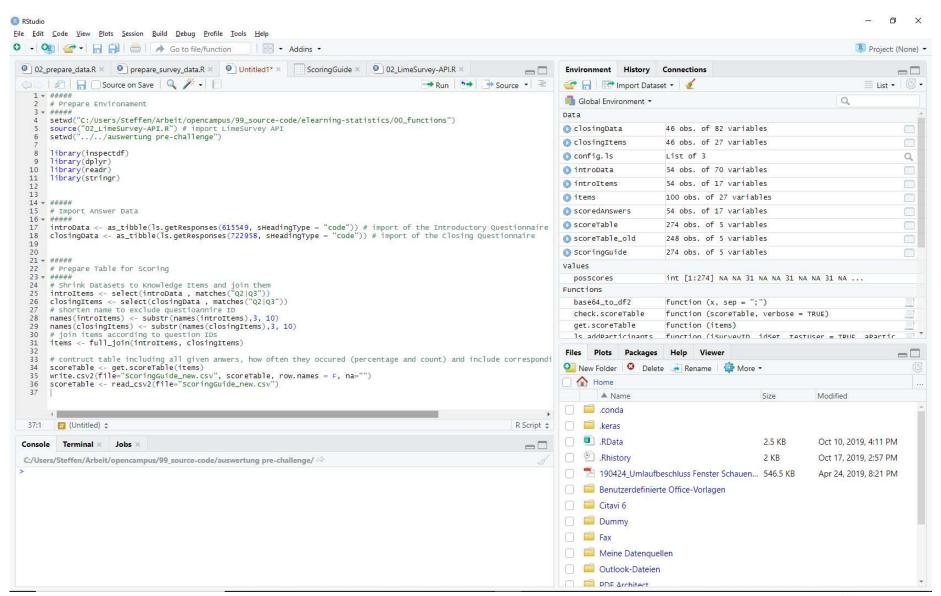
- Open Source
- Sehr große Community
- Im wissenschaftlichen Bereich zunehmend Standard für statistische Auswertungen
- Umfangreichste Funktionssammlung im Bereich Data Science / Statistik
- Leichte Integration von anderen Programmiersprachen

RSTUDIO

(Beliebteste) Umgebung zur Erstellung und Ausführung von R-Programmen

- Komfortable Bearbeitung und Ausführung der Syntax,
- Zugriff auf Hilfe-Dateien,
- Anzeige aktuell im Speicher geladener Daten und Funktionen,
- Anzeige grafischer Ausgaben,
- Installation zusätzlicher Funktionspakete ("Packages")
- und vieles mehr...





AUFGABEN

- Finde heraus wie man ein R-Package installiert und installiere das Package "fortunes"
- Finde heraus, was es macht.
- Berechne mit Hilfe von R den Mittelwert der folgenden beiden Zahlen:
 9,87654321 und 1,23456789

AUFGABEN

- Google!
- Stackoverflow
- Cheatsheets
- Dokumentation der Funktionen in R

DATENSTRUKTUREN

Es gibt drei Grundtypen für ein Datum

```
Boolean (TRUE / FALSE)
```

- Numeric (1.1392)
- String ("Text")

und zusätzlich abgeleitete, speziellere Typen (integer, date, ...)

```
Integer (Untertyp von Numeric; 12)
```

- Date (Untertyp von Numeric; "2019-04-11")
- Factor (Untertyp von String; "female"/"male")

VEKTOREN

Alle Elemente eines Vektors haben den gleichen Typ

→ Jeder Vektor hat einen eindeutigen Typ

```
■ v1 <- c(FALSE, TRUE, TRUE)
```

- ort <- c("kiel", "hamburg", "berlin")</pre>
- v2 <- c(112, 343, 235)
- alter <- 12

MATRIZEN

m <- matrix(c(1,2,3,4), 2, 2)</pre>

entspricht:
$$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

LISTEN

Können beliebige Elemente enthalten:

- Vektoren und Matrizen unterschiedlichen Typs
- sonstige R-Objekten (etwa auch Funktionen)
- → Listen haben keinen eindeutigen Typ
 - 11 <- list(v1, alter, c(4,7,9))</pre>
 - 12 <- list(name="Fred", child.ages=c(4,7,9))

SPEZIELLE LISTEN

Datentabellen

bestehen aus Vektoren gleicher Länge (aber potentiell unterschiedlichen Typs)

→ Datentabellen haben keinen eindeutigen Typ

- ☐ Data Frame (base Package)
- ☐ Tibble (tidyverse Package)
- ☐ Data Table (data.table Package)

SELEKTION VON DATEN

Am Beispiel des Data Frames zu mtcars

Selektion einer Spalte als Vektors über den Namen

mtcars\$mpg

Selektion einer Spalte als Vektor über die Position

mtcars[,1] (oder mtcars[[1]])

Selektion einer Zeile als *Data Frame* über die Position

mtcars[1,]



SELEKTION VON DATEN

Am Beispiel des Data Frames zu mtcars

Selektion einer Spalte als Vektors über den Namen

mtcars\$mpg

Selektion einer Spalte als Vektor über die Position

mtcars[,1] (oder mtcars[[1]])

Selektion einer Zeile als Data Frame über die Position

mtcars[1,]



SELEKTION VON DATEN

Selektion von Daten über einen Vektor vom Typ Boolean

[1] Konstruktion des Vektors

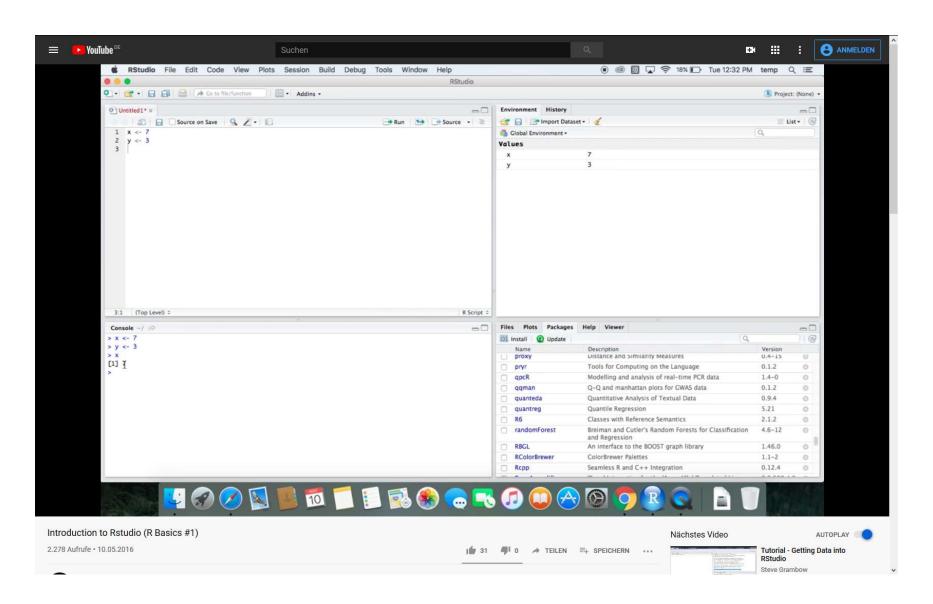
- mtcars\$hp < 100</pre>
- mtcars\$gear == 5

[2] Selektion der Fälle (Zeilen) mit dem Wert TRUE

- mtcars[mtcars\$hp<100,]</pre>
- mtcars[mtcars\$gear==5,]

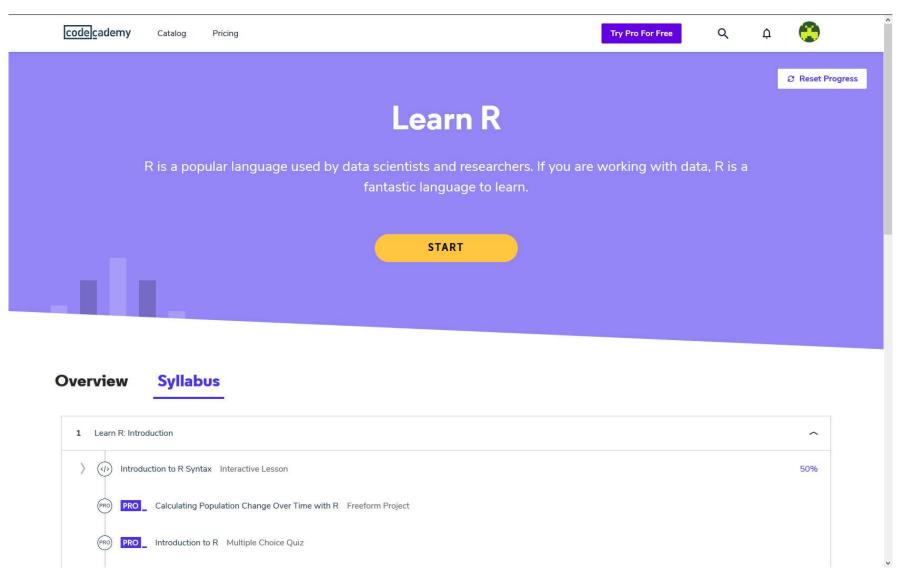
AUFGABEN

- Speichere den Datensatz airquality in der Variable airQuality.
- Berechne die Gesamtdurchschnittstemperatur.
- Berechne die Durchschnittstemperatur für den Monat Juli.
- Vergleiche, ob die Monate Juli und Mai sich signifikant in ihrer Durchschnittstemperatur unterscheiden.



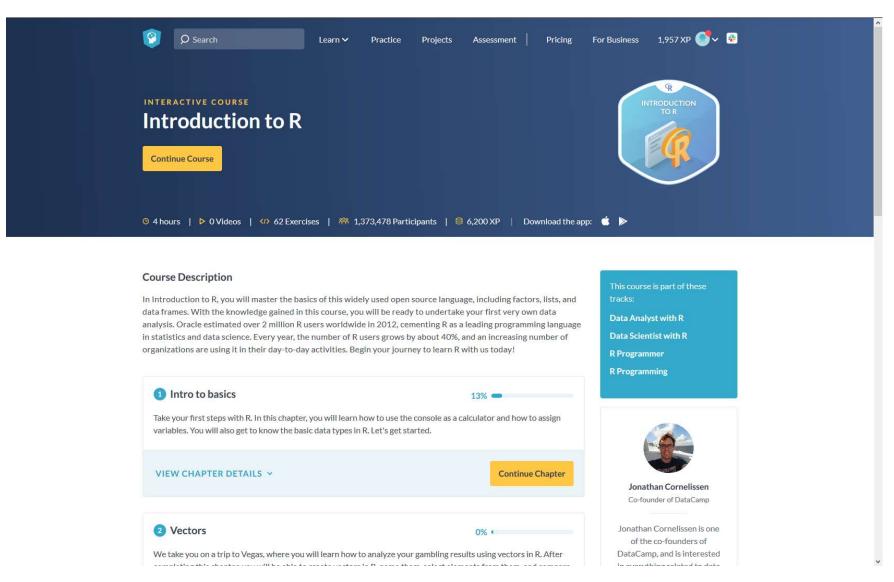
https://www.youtube.com/watch?v=QvBo1RBptvY





https://www.codecademy.com/learn/learn-r





https://www.datacamp.com/courses/free-introduction-to-r

