

A scenic view of the Elbe river in Magdeburg, Germany. In the foreground, several boats are on the water, including a small tugboat, a barge, and two larger passenger ferries. The river flows towards a bridge in the distance. On the right bank, there are lush green trees and a prominent church with two tall, dark spires. The sky is clear and blue.

# 00 - Course Introduction

## Data Science with R · Summer 2021

Uli Niemann · Knowledge Management & Discovery Lab

<https://brain.cs.uni-magdeburg.de/kmd/DataSciR/>



# Welcome to Data Science with R!

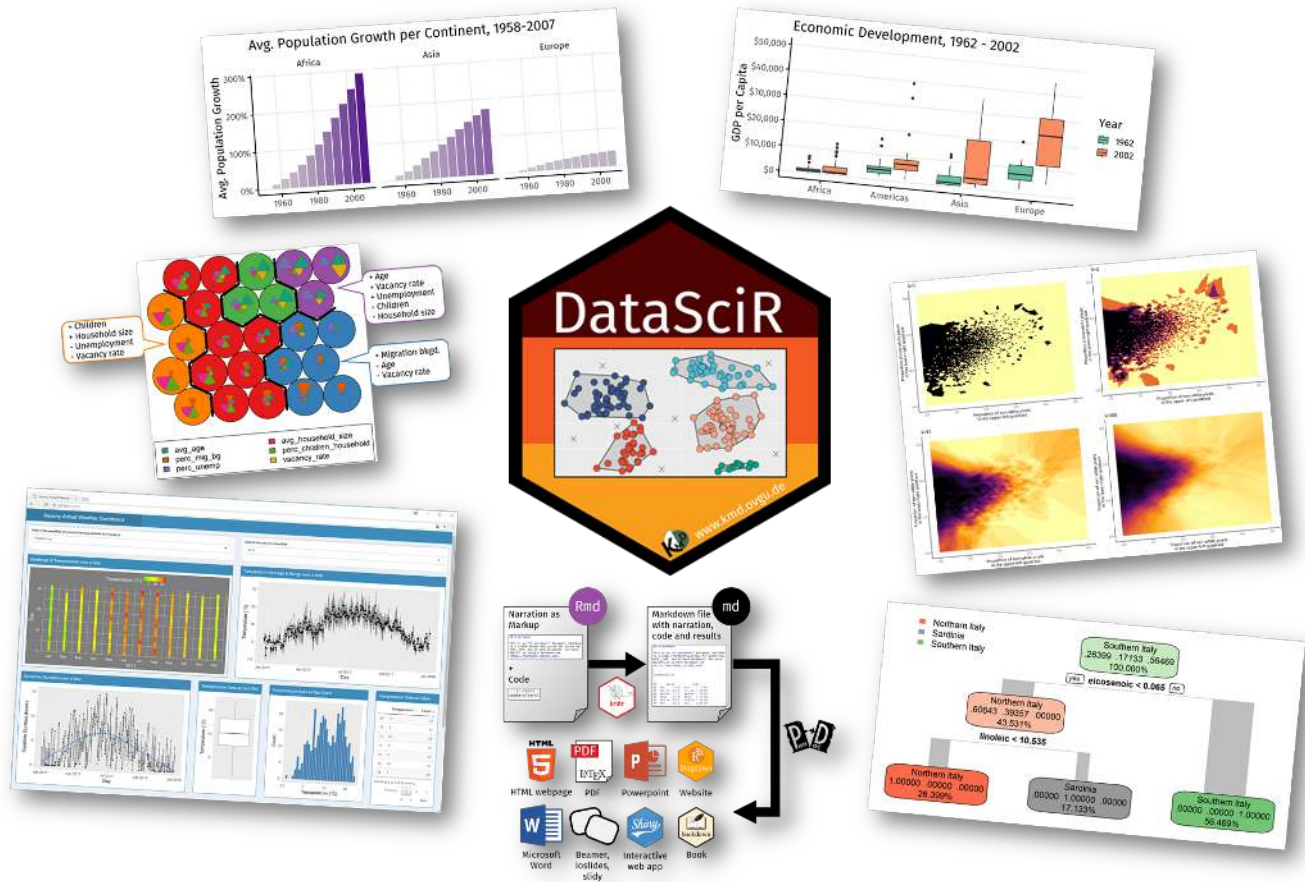
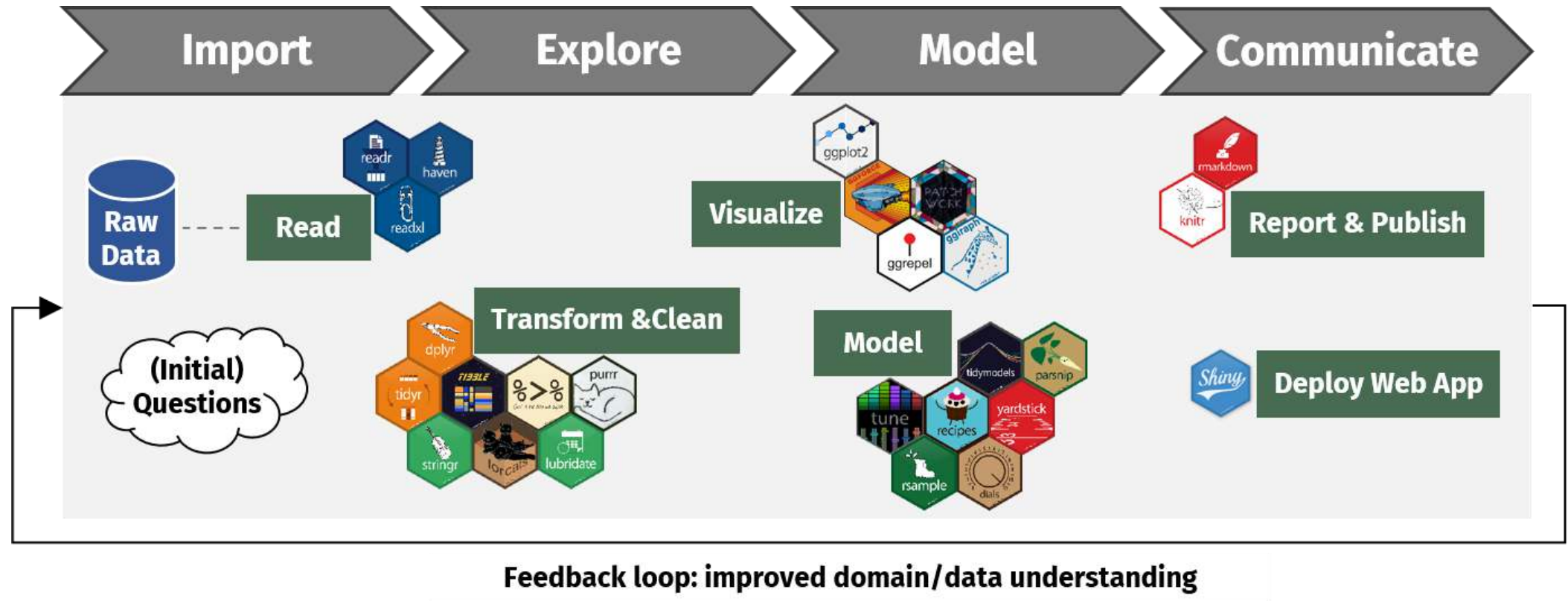




Photo by [Amelie & Niklas Ohlrogge](#)

# Data Science with R



# Course structure

## Part 1: R "crash course"



Photo by [NeONBRAND](#)

## Part 2: Data science project



Photo by [Annie Spratt](#)

# Website & syllabus

News

Administrative Information

Application & registration

Course description

(Tentative) Schedule

Prerequisites

Recommended Reading

Software

## Data Science with $\mathbb{R}$ (DataSciR)

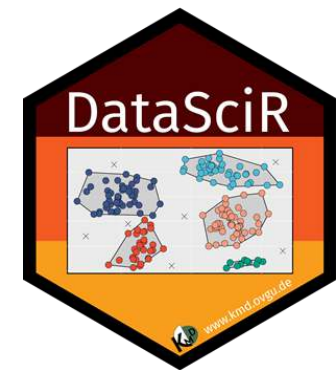
This is the website for the Data Science with  $\mathbb{R}$  (DataSciR) course offered in 2021. Click here (<https://brain.cs.uni-magdeburg.de/kmd/DataSciR/2020>) if you are looking for the 2020 course website.

### News

- **05.04.2021:** Notifications regarding course admission have been sent by email.
- **22.03.2021:** Application to the course is possible from today.

### Administrative Information

- Course day/time: Fridays, 9:15-10:45 (via Zoom)
- Instructor: Uli Niemann



(<http://www.kmd.ovgu.de/Team/Academic+Staff/Uli+Niemann.html>)



# Resources

- Website: <https://brain.cs.uni-magdeburg.de/kmd/DataSciR/>
- GitHub: <https://github.com/unmnn/datascir21>
- Moodle: <https://elearning.ovgu.de/course/view.php?id=9799>
- Zoom: <https://ovgu.zoom.us/j/92337441611>

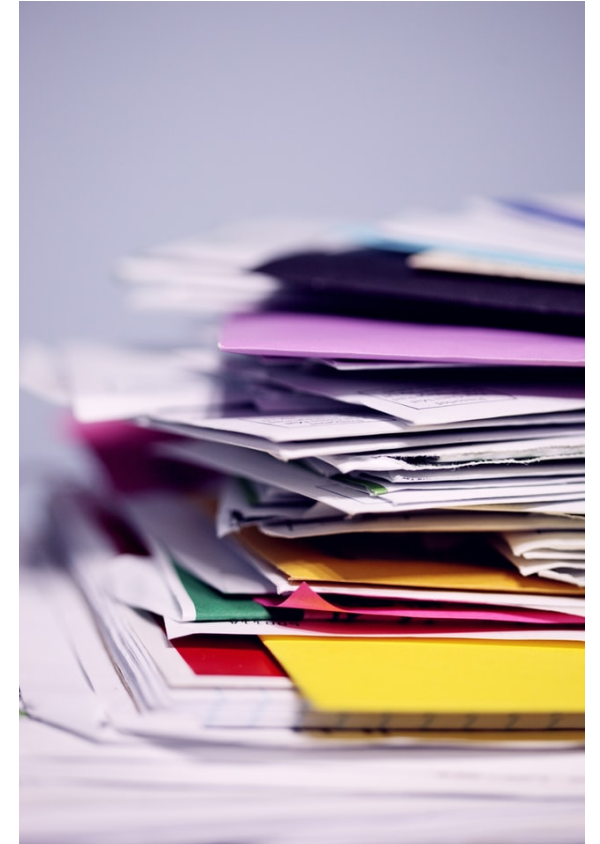


Photo by [Sharon McCutcheon](#)

# Quizzes

- Handled on E-Learning (Moodle)
- 7 quizzes in total with 5 questions each
- Each quiz open on Mondays 00:00 and is due on the following Sunday at 23:59 German time
  - The first quiz is available between 12.04 00:00 and 18.04 23:59
- Work on them individually
- Lowest score is dropped, i.e., only your 6 best results count
- Total score of correctly answered questions gives you up to 30 pt for your final grade
- Questions must be answered in a fixed order, which means you **cannot** return to previous questions



## Question 1

Not yet answered

Marked out of 1.00

Using the "bookings" data, write a pipeline that computes the median review score per weekday, resulting in the following output:

```
# A tibble: 7 x 2
  checkin_day median_review_score
  <chr>         <dbl>
1 sat             7.67
2 fri             7.66
3 sun             7.45
4 wed             7.18
5 thu             7.18
6 mon             7.14
7 tue             7.06
```

(Note that the weekdays are sorted by median review score in descending order.)

↵

A ▼

B

I

≡

≡

🔗

🔗

🖼️

## Question 1

Not complete

Marked out of 1.00

Which statement is correct given the below code snippet?

```
ggplot(penguins, aes(x = bill_depth_mm, y = bill_length_mm)) +
  geom_point(aes(color = species)) +
  labs(
    title = "Bill depth and length",
    subtitle = "Measurements for Adelie, Chinstrap, and Gentoo penguins",
    x = "Bill depth (mm)", y = "Bill length(mm)",
    color = "Species",
    caption = "Source: Palmer Station (LTER) | palmerpenguins R package"
  ) +
  scale_color_brewer(palette = "Set2") +
  geom_smooth(method = "lm")
```

- ☐ a. The mapping of the variable **species** to the color aesthetic is an example of "global aesthetical mapping".
- ☐ b. In this plot, the x- and y-aesthetics can be mapped globally or locally. The visual appearance of the plot would be the same.
- ☐ c. To change the "local aesthetical mapping" of **species** to a global one, we have to add **local = FALSE** to **geom\_point()**, i.e., the second line becomes **geom\_point(aes(color = species), local = FALSE)**.
- ☐ d. A separate regression line is shown for each penguin species.

Check

# Project

In a nutshell: ***Find an interesting dataset and do something with it!***

- analyze a dataset of your choice
- demonstrate your proficiency in the techniques we have covered in this course (+ desirably beyond)
- **quality over quantity**: don't compute every possible statistic, but :
  - show that you can **ask meaningful questions and answer them using R**
  - display that you are **skilled at interpreting and presenting your results**
  - critically scrutinize your approach (**adequacy**) and your results (**reliability, validity**) and make suggestions for improving your analysis
- emphasize on **compelling visualizations** and a **convincing, coherent and clear narrative**
- all analyses must be done in R; you are encouraged to use tidyverse packages but you don't have to

# The project data

- choose any dataset which is *not analyzed to death* (no Kaggle, UCI ML repo, or similar)
- the dataset may already exist or you may collect your own data via a survey or an experiment
- the dataset may be based on your personal interests or based previous projects in other courses

Project Team
Project Milestones
Team formation and project proposal
Topic
R Markdown process notebook
Code
Project website
Project screencast
Final presentation
Grading
Mapping of points to grades
Project Submission Instructions
Example final projects

# Project & Grading

Your grade is determined by (1) an individual component and (2) a team project component.

For the individual component, you will demonstrate your proficiency of the presented topics in multiple choice quizzes in E-Learning (<https://elearning.ovgu.de/course/view.php?id=9799>).

For the team project component, you will work on a semester-long **data science project** using `R`. The goal of the project is to go through the complete data science process to answer questions you have about some **topic of your own choice**. You will acquire and preprocess the data, design your visualizations, run machine learning algorithms, and communicate your results.

## Project Team

You work in a **team of 3 to 5 students**. In general, the grades for each group member will most likely be the same. However, if one team member evidently (a) did not contribute a fair share of the team's work, (b) delivered poor or incomplete work, (c) missed deadlines, (d) did not assist team mates and/or (e) threatened to quit if the work became difficult, this team member will receive a lowered grade.

## Project Milestones

There are a few milestones for your final project, see the table below. Please note that **no extensions** will be given for any of the project due dates for any reason. Projects submitted after the final due date will not be graded. Mandatory deliverables submitted after due date will be



# Deliverables & Hard Deadlines

Date	Description
20.05. <del>16.05.</del>	Team formation & project proposal submission; registration
24.-28.05.	Project proposal feedback
06.07.	Final project submission due
09.07.	Final presentations, exact time and location to be announced



Photo by [Kevin Ku](#)

# Your final score in the course

Your overall course grade will be determined by the following components:

1. **30 pt:** Weekly quizzes

2. **70 pt:** Project:

- 8 pt: Project proposal
- 25 pt: Quality of R Markdown notebook with respect to correctness, comprehensibility and reproducibility
- 10 pt: Complexity and level of difficulty of the project
- 5 pt: Completeness and overall functionality of the repository and website
- 8 pt: Screencast
- 14 pt: Final presentation

# Zoom meetings

We will have weekly Zoom meetings. They are completely optional.

The Zoom meetings are mainly for:

- code-alongs
- Q&A

Expectations for Zoom meetings:

In larger sessions you should

- have your microphone muted by default
- use the *raise hand* feature or type questions and comments in the chat

In small team sessions you should

- have your camera turned on as much as possible
- engage with your team mates via voice and text chat
- take turns sharing your screens when necessary

# Collaboration regulations

- Only work in the context of the team project should be completed collaboratively.
- Exercises should be completed individually.
- Quizzes must be completed individually.
- You are welcome to discuss problems in general and ask for advice.




# Sharing and reusing code policy

- You can use any online resources, but you must explicitly state where you got the code that you use directly or use as inspiration for your solutions.
- Any reused code that is not explicitly cited will be treated as plagiarism, regardless of the source.

# Late work policy

- Submissions after the deadline will not be considered.
- General advice: Don't wait until the last minute.
- But: Email me if there are reasons beyond your control that prevent timely submission.



A photograph of a modern university courtyard. In the foreground, there is a large, leafy green tree on the left and a paved path leading towards the center. In the middle ground, there are several metal benches arranged in a circle on a grassy area. In the background, there is a large, multi-story white building with many windows. The sky is blue with some clouds.

**Thank you! Questions?**