HOCHSCHULE OSNABRÜCK
UNIVERSITY OF APPLIED SCIENCES

Syllabus

Data Analytics with R

Instructor: Prof. Dr. Nicolas Meseth

Module number: 44B0586

Module description: https://www.hs-osnabrueck.de/module/44b0586/

Course Description

This course provides an introduction to modern data analytics with a focus on using R and the Tidyverse. It is designed for students to gain proficiency in the key steps of the **exploratory data analysis cycle**, as proposed by Wickham (2023). The curriculum covers essential techniques in data handling, including the loading of data from various sources and effective data transformation strategies.

Students will learn the principles of data visualization, developing the ability to create meaningful graphical representations of data using R and the Tidyverse. The course also emphasizes the importance of communicating data analysis results, guiding students in presenting their findings in clear and appropriate formats.

An integral part of the course is understanding what can go wrong during the EDA cycle. Students will be exposed to common sources of error and pitfalls in data analysis, such as data quality issues, biases in data collection and interpretation, misguided visualizations or misrepresentation of results. This aspect of the course aims to develop critical thinking skills and foster a meticulous approach to data analysis, ensuring that students are well-prepared to identify and address these challenges in their future work.

Throughout the module, practical application is a priority, with hands-on exercises and projects that allow students to apply the concepts in real-world contexts. The course aims to build a solid foundation in data analytics, equipping students with the skills necessary to analyze, interpret, and present data effectively.

HOCHSCHULE OSNABRÜCK
UNIVERSITY OF APPLIED SCIENCES

Learning Objectives

- **Data Analysis Pitfalls**: Know and identify common challenges in data analysis and learn strategies to mitigate them.
  - Potential problems at the source of the data
  - False outcomes due to inappropriate data analysis methods or wrong application
  - Unclear or misleading messages through bad communication and visualization

- **Exploratory Data Analysis (EDA) Process**: Recall the essential steps involved in conducting an exploratory data analysis.

- **Conducting EDA with R & Tidyverse**: Apply R and the Tidyverse tools to perform exploratory analysis on unfamiliar data sets.
  - **Loading Data from Various Sources**: Import and manage data from diverse sources using R and Tidyverse tools, adapting techniques to accommodate different data formats and structures.
  - **Data Transformation with dplyr**: Apply dplyr functions within the Tidyverse for effective data transformation tasks.
  - **Visualization Decision-Making**: Analyze data scenarios to determine suitable visualization types.
  - **Chart Creation with R & ggplot2**: Design and implement various chart types using R and ggplot2.
  - **Communication of Results with Quarto**: Effectively communicate and present data analysis results, encompassing the creation of clear, well-structured, and visually appealing reports and presentations.

Learning Content

- **Loading data** with `readr` and `readxl`
  - Loading data from various sources (CSV, URL, Excel, R-Data) with the `read_*` functions
  - Changing column metadata (name, data type) with `rename` and `mutate`
  - Reordering of columns with `relocate`
  - Checking data after loading (dimensions, data types, missing values, data ranges) with `ncols`, `nrows`, `dim`, `glimpse`, `skimr`, and other functions and packages.
  - Tidying data (long form, atomic values, no summary rows) with `pivot_longer`

    - Removing duplicates

- **Data transformation** with `dplyr`
    - Picking columns with `select`
    - Reducing rows with `filter`
    - Sorting rows with `arrange`
    - Adding new columns and changing existing ones with `mutate` or `transmute`
    - Aggregate rows with `group_by` and `summarize`, using various aggregation functions such as `sum`, `mean`, `n`, `min`, `max`, `sd`, `quantile` and others.

- **Data visualization** with `ggplot2`
    - The grammar of graphics
    - Simple plots for trends & developments, amounts & proportions, distributions, and associations:
        - Line charts with `geom_line`
        - Bar charts with `geom_bar` and `geom_col`
        - Histogram with `geom_histogram`, box plots with `geom_boxplot`, ridgeline plots with the `ggridges` library
        - Scatter plot with `geom_point` and `geom_jitter`, heat maps with `geom_tile`
    - Tuning a visualization
        - Setting the title, axis title and labels, and ticks
        - Setting and defining color palettes
        - Applying predefined themes and creating your own theme

- **Communicating your findings** with Quarto
    - Using Markdown with Quarto
    - Blending code and text in automated reports
        - Options for hiding code, results, warnings etc.
        - Integration of plots
        - Using inline R for data values within text
    - Using cross-references
    - Adding a table of contents and bibliography
    - Rendering different output formats (HTML, Word, PDF)
    - Parameters for dynamic reports

HOCHSCHULE OSNABRÜCK
UNIVERSITY OF APPLIED SCIENCES

Required Text and Material

Wickham, Hadley, et al. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition, O'Reilly Media, Inc, 2023.

Wilke, C. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. First edition, O'Reilly Media, 2019.

Huntington-Klein, Nick. *The Effect: An Introduction to Research Design and Causality*. CRC Press, Taylor & Francis Group, 2022.

Course Schedule

Coming soon.

Assignments

Coming soon.

Attendance and Participation

Attendance Expectations

- **Expected Attendance**: Regular attendance is highly encouraged to facilitate learning and ensure you receive the full benefits of the course.
- **Minimum Attendance Requirement**: If fewer than 3 students are present for a live session, the session will not be conducted. This policy is in place to ensure a productive and interactive learning environment.

Participation Expectations

- Active Engagement: Participation is more than just attendance. You are encouraged to actively engage in discussions, ask questions, and contribute to group activities.
- Virtual Participation: For online or hybrid course formats, participation also includes engagement in virtual discussions, timely submission of assignments, and proactive communication.

Absence Policy

- Notification of Absences: If you are unable to attend a session, please notify the instructor in advance.

- Make-Up Work: In the case of absences, it is your responsibility to catch up on missed content. Lecture notes or recordings, if available, should be reviewed.
- Extended Absences: If you anticipate extended absences, please discuss this with the instructor as early as possible to make appropriate arrangements.

Academic Integrity Policy

In this course, we are committed to upholding the highest standards of academic integrity. Every member of our academic community is expected to demonstrate honesty, respect for others, and a commitment to ethical conduct in all learning and assessment activities.