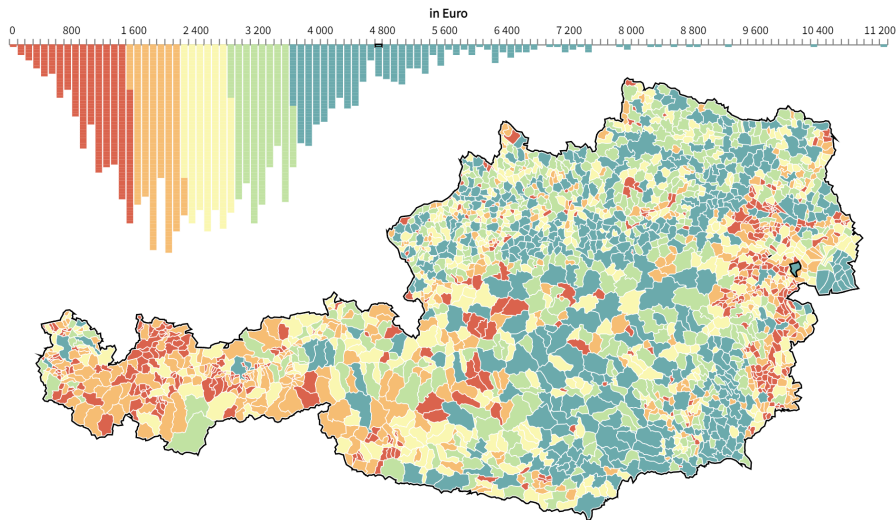


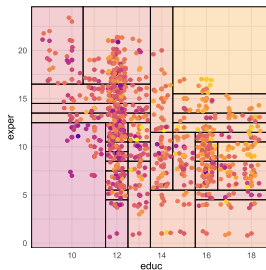
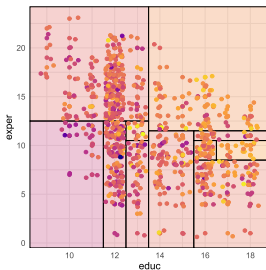
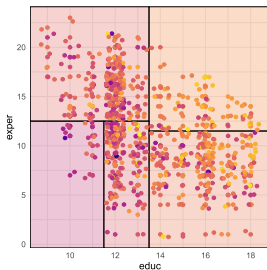
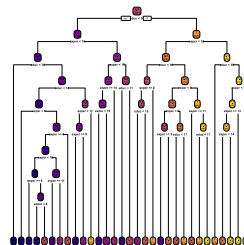
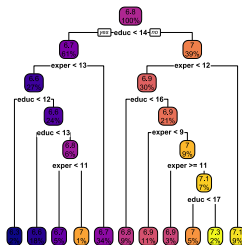
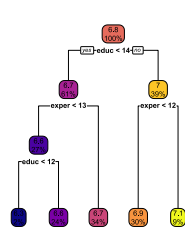
Data Science and Machine Learning

Info Session

Katharina Fenz, Lukas Schmoigl, Maximilian Thomasberger,
June 1, 2023

- ▶ Data Wrangling and Git
- ▶ Exploratory Data Analysis and Data Visualization
- ▶ Web Scraping and Geocoding
- ▶ Introduction into Machine Learning and Cross Validation
- ▶ Random Forest and Boosted Regression Trees
- ▶ Text Classification and Natural Language Processing
- ▶ Nearest Neighbor Learning
- ▶ Clustering Algorithms and Principal Component Analysis





Field Course:

- ▶ Lectures and in-class applications
- ▶ Grading
 - ▶ Exam (35%)
 - ▶ Project paper and presentation (35%)
 - ▶ Assignments and "Kaggle-style" competition (30%)

Does Space Matter?

🏠 Introduction

📁 Data Wrangling ▾

📊 Spatial Analysis ▾

📈 ML Analysis ▾

🔗 Outlook

Zillow Data

Census Data

Infrastructure Data

Join all datasets together

Impute Missing Values

Export

Data Import and Manipulation

Data is the fundamental and essential element for machine learning where models are built to learn the properties of training data. We split the data into training and test data where the model is trained on the training sample and evaluated using the test sample.

In this file we show how we read in data from different sources, clean them until we have a tidy dataset that does not contain any missing values since the random forest demands a complete dataset. We will then later split the full dataset into test and training data.

Zillow Data

The data for our dependent variable comes from the ZILLOW database (<https://www.zillow.com/research/data/>). Zillow is an American online real estate marketplace company that provides a large dataset on Home values in

- ▶ Coding
 - ▶ Examples and assignments will be given using R
 - ▶ Basic programming skills in R or a similar (e.g. python, julia) are required
- ▶ Mathematics and Statistics
 - ▶ Basic understanding of probability, statistics, linear algebra and calculus is necessary
 - ▶ However this is an applied class: no formal proofs
 - ▶ Ideal: practical understanding of statistical modelling and working with data