

Exercise Sheet 04 @‘Applied AI Using R’

Main objective of the current sheet is again to analyze some standard datasets by applying functions we have already discussed in the course. Notice that exercises like these prepare you for participating in projects and for internships in companies.

Aufgabe 19.

Study the remainder of ‘Advanced Data Wrangling’ in [slides04_Dumbledore](#) and solve the exercises on page 24 and page 33.

Aufgabe 20.

Solve the exercises on pages 36-38 in [slides04_Dumbledore](#).

Aufgabe 21.

We use the dataset `ATM.list` from Exercise 09 (load it via the last line in [R-Snippets02](#)).

1. Produce nice looking plots (one panel per year, use `facet_wrap`, mark holidays with a special color) of the timeseries and save them in a joint pdf (one page per unit).
2. One would expect that the withdrawn amounts are lower on holiday and higher on the days before holidays. Mark days before holidays in another color in the plots produced before.
3. We want to quantify a holiday effect in the following sense: if the day would be normal (=no holiday), then the amount would be x , since it is a holiday it reduces to px with $p \in [0, 1]$. Try to estimate the reduction p for every ATM - is it similar for all five ATMs?
4. Same question as the last one, but now focus on the days before holidays (in which case we expect $p \geq 1$, so on average the values should go up).

Aufgabe 22.

In this exercise we work with airbnb data.

1. Download the `listings.csv.gz` and `reviews.csv.gz` datasets for Vienna from <https://insideairbnb.com/get-the-data/> and load them into R
2. The column `neighbourhood_cleansed` of the `listings` dataframe contains the Bezirk in which the airbnb is located. For each neighbourhood calculate the average price of “entire rental units” in the dataset.
3. Calculate how many persons per bed each airbnb offers. What is the highest number of persons per bed in the listings?

4. Which guest returned to the same airbnb the most (i.e. wrote the most reviews for the same apartment)?
5. The `listings` dataframe contains columns `number_of_reviews`, `number_of_reviews_ltm`, `number_of_reviews_l30d`. Which of the three columns counts the number of reviews in the `reviews` dataframe?
6. Check that the columns `first_review` and `last_review` in the `listings` dataset are correct, i.e. check that for each airbnb the date of the first review in the `reviews` dataset is indeed equal to the `first_review` entry. Analogous for `last_review`.
7. Dive deeper in the data and produce at least one informative and pretty plot.

Aufgabe 23.

In this exercise we use the *tips dataset* sent out by email.

1. Parse the csv into R.
2. Choose your favourite machine learning model from the `{tidymodels}` package to predict the `tip` column from all other columns.
3. The dataset contains NAs. Define a preprocessor that imputes the missing data.
4. The dataset contains factorial/discrete columns. Add a preprocessing step to your recipe that encodes those columns as purely numerical ones.
5. Split the data into training (75%) and testing (25%). Fit your model to the data. You may try out different models and different preprocessing steps. The best performing model (rmsq) wins!
6. Repeat the previous step at least 10 times - which model would you finally choose and how good is its performance?