Assignment I: Importing, Cleaning, Merging, Regression Analysis

Data Science Project Management (DS400) | Winter Term 2024/25

Formal requirements

Submit your assignment earlier than **November 12, 2024, 12:00 noon**. The submission must include your code, results, and explanations for code and results, for both **R** and **python**.

Submit your assignment via ILIAS. Upload the `*.html`-files that were rendered from your `*.Rmd` and `*.ipynb`, or `.qmd` solutions. You can optionally upload the `*.Rmd` and `*.ipynb`, or `qmd` files as well. You must use the following naming convention: `<Lastname>_<Firstname>_Assignment##_<Language>.<extension>` (e.g., `Lovelace_Ada_Assignment01_R.html`).

You will receive up to **7.5 points** upon passing this assignment.

Code of conduct

By submitting the assignment, you are acknowledging that the submitted assignment is your own work. This implies that …

- the work you submit is your own
- the code you wrote is your own
- any comment or interpretation is in your own words
- you made clear whom you worked with

Disregarding this code of conduct will result in failure of the assignment!

Importing data

In this assignment, you will work with data on Airbnb listings. The data are adapted from: **https://insideairbnb.com/get-the-data/**. You will conduct your analysis on the city of Munich. The data files are provided in the `data.zip` file. The first table, *listings_muc.csv*, contains, for each listing, sixteen variables. The second table, *calendar_muc.csv*, contains all available dates for each listing, as well as some further information such as the price per night.

(1) Import the data into R and python, respectively. In R, try out all three functions discussed in the lecture: `utils::read.csv()`, `readr::read_csv()` and `data.table::fread()` on at least one of the two tables, and check whether the shape of the imported data is the same in every case. In python, use `pandas.read_csv()` to import the data.

(2) Print the first five rows and column types in R and python, respectively.

Profiling the import [R]

For this first task, we want to compare the time it takes to import the data using the three functions from the previous task.

(3) In R, find out about and use the {microbenchmark} package (Hint: Set `show_col_types = FALSE` for your calls of `readr::read_csv()`). Compare all three functions. Which function is the fastest? How much faster is it than the slowest function?

Tidying the data

In this task, we will tidy the data. If you have not done so before, please carefully read the tidy data paper by Hadley Wickham.

(4) Please write down (in your submitted notebook) the requirements for some data table to be tidy.

(5) Can any of the two tables be considered tidy? If not, what would you have to do to make them tidy? Please perform all necessary steps to make the data tidy. Are there any other changes to the data necessary? Print the first five rows of the resulting tables.

(6) What is the shape of the tables after making the data tidy?

Merging the data

(7) In the calendar data, change the name of the `listing_id` column to `id`.

(8) Aggregate the calendar data to the listing level. The variable `available` should be transformed into the number of available days per listing. The variable `date` should be discarded. The variables `price`, `minimum_nights`, and `maximum_nights` should be aggregated to the listing level.

(9) Merge the aggregated calendar and the listings data by `id`. Call the resulting data frame `merged` and print its shape.

Analyzing the data

Finally we are able to perform some analysis.

(10) How many neighborhoods are there in Munich? Print the number of neighborhoods and all their names. How many and which room types are there?

(11) Regress the number of available days on the price and the minimum_nights. Interpret the results. What would you say – is this a good model?

(12) In a second model, regress the number of available days on the price and the minimum_nights, but include neighborhood and room type as fixed effects. Interpret the results. Do you see any problems with this model?