

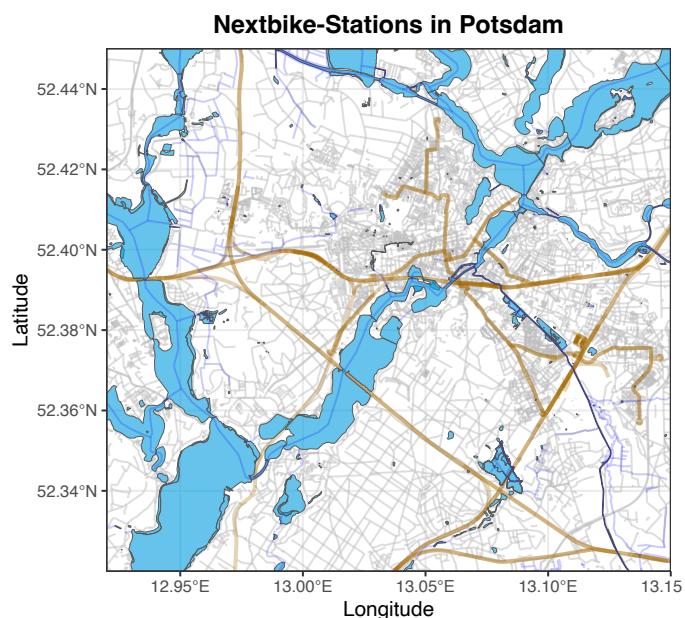
Case Study Master Data Science (Module 3)

Prof. Dr. Heike Trautmann
Prof. Dr. Pascal Kerschke
Moritz Seiler

May 2023



In the case study, data sets from the bike-sharing service *nextbike* will be analyzed. Depending on the city, users can rent and return the bikes either at the numerous stations of the provider or (similar to e-scooters) in predefined regions. In this case, the data of the city of Potsdam will be analyzed, whereby only trips between the stations are possible here. A map section of the considered region, including road, water, path, and rail network, can be found next to this text. The data and the associated R-code to reproduce the map can be found in the case study materials.



Should you need further information on the bike sharing service *nextbike*, have a closer look at the company's website (<https://www.nextbike.de/en/>).

Your Task:

Write a short report (pdf) describing your performed analyses and your findings. In addition, submit your R code as a meaningfully annotated separate R file.

1. Import the two data sets containing the trip information (trips-*.*rds*). Combine both data sets in a joint data set and augment the information contained therein with the associated day of the week and month.
2. Conduct an exploratory data analysis, i.e., get an overview of the number of observations and the type and range of variables, compute descriptive statistics, visualize the data, and examine it for noticeable patterns and/or correlations.
3. Check the data for univariate and multivariate outliers. If you find any conspicuous trips, decide how to deal with them (remove or keep them) and justify your decision.

4. In your materials, you will find another dataset (stations.rds) containing the coordinates of the *nextbike* stations. Add the locations to the map of the region shown above.
Hint: The map is a ggplot object, and consequently can be extended with ggplot commands.
5. Find a meaningful grouping of the bike locations. For this purpose, try out different cluster algorithms and choose a suitable number of clusters.
6. Visualize your cluster results by color-coding the locations (see 4.) according to their cluster affiliation (see 5.). Which cluster result seems most plausible to you?
7. Let's assume that *nextbike* wants to set up a separate service station for each cluster. Where should these service stations be placed in order to be accessible as well as possible from all stations of the respective cluster? Mark the recommended locations on your map as well!
8. Intuitively, one would expect a relationship between trip duration and the other variables (especially traveled distance). Examine this assumption by modeling the relationship using various regression techniques. Which procedure is best suited for this purpose? Interpret your result(s)!
9. For urban planning, one of the areas of interest is the traffic between different areas of a city. Consider the clusters identified in 5. as separate areas of the city and compare several classifiers that model the drop-off station cluster as a function of the other metrics in a benchmark study. Interpret your result.
10. Select one of the classifiers (from 9.) and perform a feature selection to determine which of the data set's features are most relevant.
Hint: Study the code chunks and associated slides in the feature selection chapter of your course materials and modify the code to match this task!