# Methodology

In this chapter, we discuss the components needed for the analysis in **chapter 4** – beginning with a brief overview of the hardware and software setup, followed by an outline of the three datasets: the restaurant review data from Tripadvisor and Google, and the restaurant revenue data from Prognolite. The chapter concludes with an in-detail description of the web scraping process – a technique, we applied to extract and store the restaurant review data from websites – followed by a section about data processing.

## Hardware

The web scraping, depending on the amount of review data, is a time-consuming procedure, which can take several days to complete. Hence, it was conducted on a ZHAW server with Intel Core processor (Broadwell- microarchitecture) with 2.5 GHz, 16 cores and 128 GB main memory (RAM). The server runs the Linux kernel 4.15. All the other tasks could be performed without the application of the server.

## Software

This thesis mainly involves data science tasks. The two common programming languages suited for such tasks are R and Python. Despite the popularity of R among data scientists and the broad variety of libraries it provides for statistical analysis and visualisation, we decided to use Python, because, firstly, we didn’t have any prior knowledge of R and secondly, over the course of our bachelor studies, we were exposed to several Python projects and were therefore already familiar with this language. Like R, Python also offers several libraries for data science, such as NumPy and SciPy for scientific computing, Pandas for data analysis and manipulation, Matplotlib and Seaborn for creating data visualizations. Besides the data science libraries, packages like Requests and Selenium are used for review data acquisition, which will be discussed in-depth in **section 3.4**.

The project, which can be found [here](https://github.com/fatihsolmaz22/Bachelorarbeit_FS22), is set up with Poetry, a tool for dependency management and packaging in Python. As a development environment, PyCharm is used.

## Data

In the following subchapters, the data required for the correlation analysis are presented individually.

### Restaurant revenue data: Prognolite

The restaurant revenue data obtained from our industry partner Prognolite, is a large CSV file comprising information on 15 restaurants in Switzerland, some of which are chain affiliated. The information includes details such as date with timestamp, turnover in CHF, local (school) holidays, local temperature in degree Celsius, etc. In this work, we use only the date with timestamp and the corresponding turnover information for each restaurant.

### Restaurant review data: Tripadvisor and Google

To study the effectiveness of eWOM on restaurant revenues, review data needs to be collected. Our first consideration was to acquire the restaurant review data effortlessly and efficiently, if possible, from an API, provided by popular online review platforms. An important criterion was, that the review data for each restaurant where revenue data exists, should contain sufficient ratings. After comprehensive research by consulting extant literature, the three platforms Yelp, Tripadvisor, and Google emerged. Yelp was ineligible, since there was no review data for some of the restaurants we wanted to investigate. The other two platforms, however, provided the relevant information we needed, although, it had to be extracted through web scraping (see **section 3.4**), as neither platform offer APIs. As a result, we decided to scrape the restaurant review data from Tripadvisor first and then from Google, because the latter proved to be more complex and arduous. The gathered review data come from two different sources; hence it had to be processed and organised and brought to the same format.

The schema (see **Appendix XY**) is derived from the review information presented on the Tripadvisor website for one restaurant. Note, that not all the available information on the website were scraped and therefore presented. The important ones for the analysis are the overall rating (from 1 to 5), the number of reviews and the review data, which includes e.g., the individual user ratings with the corresponding dates when they were published. This schema is also applicable to Google restaurant reviews, as it shares certain similarities to the ones on Tripadvisor.

Many restaurants on online review platforms receive up to hundreds of customer reviews, making it time-consuming to read them all. In such cases, we assume, that the first thing a person may do to learn about the quality of a restaurant, is to look at the overall rating, which might take into account several factors, for instance the author data – information about a costumer, who composed the review – or the number of likes, a review received from other customers. This additional information was also extracted, with the goal of finding how the overall rating is calculated. The results of this analysis are stated in **chapter XY, section XY**.

## Data acquisition: Web scraping

### Tripadvisor

### Google

## Data processing

After performing the data acquisition, the extracted restaurant review and revenue data have to be processed, in order to provide the required variables needed for the correlation analysis, namely:

* **Average rating per time period** of a restaurant, which describes the average customer satisfaction and is computed by adding up the individual costumer ratings per time period and then dividing by the number of ratings per time period. Note, that time period is a placeholder for month, quarter or year.
* **Overall rating development over time period**, which characterizes how the overall rating of a restaurant has developed over the course of months, quarters or years since the first rating, the restaurant received, till the recent rating. This measure is calculated by adding up the values from the first rating to the latest rating, divided by the number of ratings during that period.
* **Average turnover per time period** of a restaurant, a measure, which is the sum of turnover over days during a time period, divided by the number of days on which revenue was generated during that same period. Again, time period refers to month, quarter or year.

For time series data, that originate from a restaurant which has strong recurring seasonal fluctuations and rising or falling trends, the decomposition of these patterns is crucial. For example, a restaurant is more likely to have a high revenue during summer due to better weather conditions or on holidays such as Christmas. However, these circumstances probably do not influence user ratings nor are they related to the restaurant rating. As a result, there will be an upwards trend in revenue in the summer or on holidays, but no visible upwards trend in the costumer ratings. Therefore, we certainly need to normalize the revenue data to make it comparable to the ratings of the restaurant. This is achieved through decomposing the time series with Python into several components: trend, seasonal and residual. For the correlation analysis between the restaurant review and revenue data, the original time series and the decomposed results: trend and residual, are used.