Kunden:ZHAW:2015:00067.03-Titelblatt-Bachelorbuch:Versionen:Institute:wmf:2015-04-27-ZHAW-Titelblatt-Bachelorbuch-IDP.wmf

**Bachelorarbeit (Informatik)**

Analyse von Umsatzzahlen aus dem Gastronomiebereich

|  |  |  |
| --- | --- | --- |
| **Autoren** |  | Manu Paul Kunnumpurathu  Muhammed Fatih Solmaz |
| **Hauptbetreuung** |  | Dr. Martin Frey |
| **Nebenbetreuung** |  | Dr. Reto Bürgin |
| **Industriepartner** |  | Prognolite |
| **Externe Betreuung** |  | Marco Wirthlin |
| **Datum** |  | 10.06.2022 |

**Erklärung betreffend das selbstständige Verfassen einer Bachelorarbeit an der School of Engineering**

**Erklärung betreffend das selbstständige Verfassen einer Bachelorarbeit an der School of Engineering**

Mit der Abgabe dieser Bachelorarbeit versichert der/die Studierende, dass er/sie die Arbeit selbständig und ohne fremde Hilfe verfasst hat. (Bei Gruppenarbeiten gelten die Leistungen der übrigen Gruppenmitglieder nicht als fremde Hilfe.)

Der/die unterzeichnende Studierende erklärt, dass alle zitierten Quellen (auch Internetseiten) im Text oder Anhang korrekt nachgewiesen sind, d.h. dass die Bachelorarbeit keine Plagiate enthält, also keine Teile, die teilweise oder vollständig aus einem fremden Text oder einer fremden Arbeit unter Vorgabe der eigenen Urheberschaft bzw. ohne Quellenangabe übernommen worden sind.

Bei Verfehlungen aller Art treten die Paragraphen 39 und 40 (Unredlichkeit und Verfahren bei Unredlichkeit) der ZHAW Prüfungsordnung sowie die Bestimmungen der Disziplinarmassnahmen der Hochschulordnung in Kraft.

**Ort, Datum:** **Name Studierende:**

**Zusammenfassung**

**Abstract**

**Preface**

During the course of our bachelor studies in computer science at ZHAW School of Engineering, various concepts and methods were introduced, and applied in diverse software projects – ranging from Java software applications to machine learning projects.

One of the domains, which intrigued us, were machine learning and data science. By going through the advertised bachelor theses, we came across a project proposed by Dr. Martin Frey – in cooperation with the industry partner, Prognolite – which met our interests. The notion to work with an industry partner on a machine learning and data science project was the determining factor for our decision on conducting this research.

First and foremost, we would like to express our sincere gratitude to our supervisors Dr. Martin Frey and Dr. Reto Bürgin, for giving us the opportunity to conduct this research, for their valuable guidance and patience throughout the entire project, and for their profound knowledge and insights in the field of statistical data analysis. They have taught us the tools and methods to perform and present research of this kind.

Beside our supervisors, we would like to thank Marco Wirthlin from Prognolite, for his participation, assistance, and for his availability. He provided us with the data required for the research. We are very grateful for his support, without which the completion of this thesis would not have been possible.

Last, but not least, we would like to thank our families and friends for providing us with love and support, and for accompanying us throughout our journeys so far.

Contents

[1 Introduction 1](#_Toc105494020)

[1.1 Goal and scope of the thesis 2](#_Toc105494021)

[2 Theoretical principles 4](#_Toc105494022)

[2.1 Correlation coefficients 4](#_Toc105494023)

[2.1.1 Peason product-moment correlation 4](#_Toc105494024)

[2.1.2 Spearman rank correlation 5](#_Toc105494025)

[2.1.3 Interpretation of the correlation coefficients 5](#_Toc105494026)

[2.2 Time series decomposition 6](#_Toc105494027)

[2.2.1 Example on a real-world dataset 7](#_Toc105494028)

[3 Methodology 9](#_Toc105494029)

[3.1 Hardware 9](#_Toc105494030)

[3.2 Software 9](#_Toc105494031)

[3.3 Data 9](#_Toc105494032)

[3.3.1 Restaurant revenue data: Prognolite 9](#_Toc105494033)

[3.3.2 Restaurant review data: Tripadvisor and Google 10](#_Toc105494034)

[3.4 Data acquisition: Web scraping 11](#_Toc105494035)

[3.4.1 Tripadvisor 11](#_Toc105494036)

[3.4.2 Google 13](#_Toc105494037)

[3.5 Data processing 15](#_Toc105494038)

[4 Results 17](#_Toc105494039)

[4.1 Analysis of the overall rating 17](#_Toc105494040)

[4.1.1 Tripadvisor overall rating 17](#_Toc105494041)

[4.1.2 Google overall rating 18](#_Toc105494042)

[4.2 Correlation between Google restaurant rating and Prognolite turnover data 20](#_Toc105494043)

[4.2.1 Average restaurant rating against average turnover 20](#_Toc105494044)

[4.2.2 Overall restaurant rating development against average turnover 20](#_Toc105494045)

[4.3 Correlation between Tripadvisor restaurant rating and Prognolite turnover data 20](#_Toc105494046)

[4.3.1 Average restaurant rating against average turnover 20](#_Toc105494047)

[4.3.2 Overall restaurant rating development against average turnover 20](#_Toc105494048)

[4.4 Correlation between Tripadvisor and Google restaurant rating 20](#_Toc105494049)

[4.4.1 Average restaurant rating Tripadvisor against Google 20](#_Toc105494050)

[4.4.2 Overall restaurant rating development Tripadvisor against Google 20](#_Toc105494051)

[5 Conclusion 21](#_Toc105494052)

[6 Lists 22](#_Toc105494053)

[6.1 Bibliography 22](#_Toc105494054)

[6.2 List of figures 22](#_Toc105494055)

[6.3 List of tables 22](#_Toc105494056)

[7 Appendix 23](#_Toc105494057)

[7.1 JSON schema of the restaurant review data for a restaurant 23](#_Toc105494058)

[7.2 Comparison between Google review as JSON and the actual Google review 24](#_Toc105494059)

# Introduction

The technological advances over the past few decades have introduced new, more efficient, and effective opportunities of interactions between customers and restaurants. On online review platforms, such as Yelp, Tripadvisor and Google, restaurant visitors can post and share their experiences and opinions in form of online reviews – known as electronic-word-of-mouth (eWOM) – about the meals and services they have purchased or consumed. With a simple mouse click, one can produce information about a restaurant – ranging from food quality and variety to service, hygiene, and atmosphere – or acquire them from a myriad of other diners.

The vast amount of information provided by the online review platforms have enabled customers to learn from each other’s experiences and have helped them to make better decisions on visiting a particular restaurant. Restaurateurs, on the other hand, can utilise eWOM to improve customer needs and maximize their revenues. Hence, it is important to understand how these platforms affect the economy in the restaurant sector. **(Quelle: The Effects of Online Review Platforms on Restaurant Revenue)**

Extant studies conducted on food service, adopting online reviews and sale performances data, have found, that the number of online reviews and overall rating can increase the popularity of a restaurant, resulting in having a positive effect on restaurant revenue – especially of restaurants with an excellence certificate. **(Source: A data-driven approach to measure restaurant performance by combining, The impact of social media reviews on restaurant performance)**

A Harvard business study by Michael Luca (2016) supports these findings. Luca explored the influence of online review platform Yelp on the restaurant data from the Washington State Department of Revenue. The study arrived at the following conclusions: First, a one-star increase in customer rating on Yelp leads to an increase of 5-9 % in revenue. Second, the ratings have a greater impact on restaurant revenues than the reviews, on the basis that consumers do not use all information available to them, since many restaurants on Yelp receive hundreds of reviews, making it time-consuming to read them all. Third, a greater number of restaurant reviews translates into a greater causal impact on that restaurant’s revenues. **(Source: Reviews, Reputation, and Revenue, Source: FROM RATINGS TO REVENUES)**

Another study by Limin Fang concluded that doubling consumers’ exposure to Yelp increases the revenue of a high-quality new independent restaurant by 8-20 % and decreases that of a low-quality one by a similar amount. Other platforms have shown similar effects but in smaller magnitude. In contrast, online review platforms do not aﬀect the revenues of chains and old independent restaurants. **(Source: The Effects of Online Review Platforms on Restaurant Revenue)**

Studies not related to the gastronomy sector also verified a positive correlation between eWOM and firm performances. Xie et al. (2014) found that customer ratings are positively associated with revenues of hotels listed on Tripadvisor. Torres et al. (2015) observed that the number of online reviews has a positive effect on hotel booking on Tripadvisor. **(The signaling and reputational effects of customer ratings on hotel, A data-driven approach to measure restaurant performance by combining)** The paper written by a Finnish research group also highlights the effects of eWOM on the sales of mobile applications in Google Play. The results show that higher values of overall rating correlate statistically with higher sales, and the number of ratings correlates positively with sales in the long term but negatively in the short term. **(Source: Busting Myths of Electronic Word of Mouth)**

## Goal and scope of the thesis

The primary goal of the thesis is to investigate, whether there is any statistical relationship between the online review data and the revenue data for a selected set of restaurants in Switzerland. The revenue data is provided by the industry partner [Prognolite](https://prognolite.com), a firm, which helps restaurants, bakeries, and businesses in the food sector to optimize their processes and resource management based on their past data e.g., revenue, weather, events, and holidays. The review data are obtained from Tripadvisor and Google, because these two platforms have become one of the dominant sources of consumer reviews **(Quelle: The Eﬀects of Online Review Platforms on Restaurant Revenue, Survival Rate, Consumer Learning and Welfare)** The secondary aim is to examine the correlation between the ratings across platforms where revenue data exists. The review data consists of customer ratings and reviews: a customer rating refers to the numerical star value given by a customer to express their satisfaction, while the costumer review is a verbally written message by a customer. The focus is on the customer rating. The verbal dimension is not within the scope of the thesis.

Based on the objectives and the reviewed literature, three research questions are formulated:

1. Is there any statistical relationship between Google and Tripadvisor restaurant review data?

Although restaurants are rated differently across platforms, the correlation results between Google and Tripadvisor presented by Limin Fang show moderate correlation, which indicates that the two platforms are different. Tripadvisor targets primarily travellers or tourists whose rating standards for restaurants may systematically differ from those for Google, which is widely used by locals and tourists. **(Source: The Effects of Online Review Platforms on Restaurant Revenue)** On the basis of the study undertaken by Fang, we hypothesize that there is a weak correlation between Google and Tripadvisor restaurant review data, i.e., customer satisfaction.

1. Is there a correlation between Google restaurant review data and Prognolite restaurant revenue data?

Based on the finding from the literature research, we believe, there is a positive correlation between the Google restaurant review data and the Prognolite restaurant revenue data. We expect that higher ratings will result in attracting both regular and new costumers which in turn promotes the increases in turnover of those restaurants.

1. Does a correlation exist between Tripadvisor restaurant review data and Prognolite restaurant revenue data?

We assume that Tripadvisor – like Google – has a similar impact on revenues, but in smaller magnitude, since Tripadvisor is not widely used by the majority of the population.

On the basis of the above-mentioned objectives, research questions, and hypotheses, which have been formulated and discussed with Prognolite, the scope can be divided into three main parts:

1. Data acquisition:
   * The Google and Tripadvisor review data needed for the data analysis shall be fetched from an API, if possible. Otherwise, a scrapping tool shall be developed for the review data extraction.
2. Data processing:
   * In order to perform data analysis, the gathered review dataset and the Prognolite restaurant revenue data have to be processed and organised – a clean implementation which facilitates this process needs to be thought through.
3. Data analysis:
   * By applying statistical or machine learning methods, the correlation between the datasets ought to be investigated.

The outline of the subsequent chapters is as follows: **Chapter 2** provides the theoretical background in statistics needed to interpret the data analysis performed in the latter chapters. In the following **chapter 3**, the methodology is described. This chapter contains the technical approach how the data was acquired and processed. **Chapter 4** describes the results of our study. Finally, in **chapter 5**, a conclusion about the entire work as well as an outlook is discussed.

# Theoretical principles

This chapter introduces the basic knowledge of the two most commonly used correlation coefficients – the Pearson coefficient and the Spearman coefficient – required to comprehend the correlation analysis performed in **chapter 4**. We focus on how they should and should not be used and interpreted. **Section XY** covers the fundamentals of seasonal decomposition, a concept we used to decompose the restaurant revenue time series.

## Correlation coefficients

Correlation is a measure of a monotonic relationship between two variables in a correlated data, where the increase of the value of one variable tend to result in either an increase (positive correlation) or a decrease (negative correlation) of the value of the other one, and vice versa. **(source: CorrelationCoefficients-AppropriateUseandInterpretation C 2.1)**

### Peason product-moment correlation

A special case of a monotonic association is a linear relationship between two variables. Most often, the term correlation is used in conjunction with such a linear relationship, known as Pearson product-moment correlation, commonly abbreviated as *r*. This coefficient is a dimensionless measure and ranges from -1 to 1.

The **figure** below depicts scatterplots of sample data with different Pearson correlation coefficients.



Figure 1: Scatterplots of sample data with different Pearson correlation coefficients

**Figure XY** A illustrates a perfect correlation of -1. A perfect correlation of -1 or 1 implies that all the data points lie exactly on a straight line. In **Figure XY** B and F, the scatterplot approaches a straight line as the coefficient tends towards -1 or 1, whereas in **Figure XY** D there is no linear relationship, as the coefficient is 0. **Figure XY** E displays that the correlation depends on the range of the assessed value, a wider range leans towards higher correlation than the smaller range in the shaded area.

### Spearman rank correlation

In contrast to a Pearson correlation, a Spearman correlation – generally abbreviated as *ρ* (rho) or *rs* – can be used to analyse nonlinear monotonic relationships. Furthermore, it is relatively robust against outliers. The Spearman correlation also ranges from -1 to 1, whereas *ρ* = 0 implies that there is no association, while *ρ* = -1 or 1 implicate a perfect correlation.

### Interpretation of the correlation coefficients

The scatterplots in the following **figure XY** illustrate the two correlation methods – Pearson and Spearman – on a sample dataset. Note, that the correlation coefficient should always be assessed by a visual representation of the data. For example, in **figure XY** A, both coefficients are close to 0, which connotes that there is no association between the x-axis and y-axis variables, when in fact, the plot suggests a strong quadratic relationship. Another interesting observation is, that despite the same Pearson correlation coefficient values *r* in **figures XY B through D**, the data is quite different in each of the panels. **Figure XY B** reveals, on the one hand, the robustness of the Spearman coefficient against outliers and on the other hand, its notable influence on the Pearson coefficient. In **figure XY** C, a sinusoid relationship – neither linear nor monotonic – is depicted, both correlation methods are unable to capture it. This can be further observed in **figure XY D**.

Over the course of years, several threshold values to translate a correlation coefficient into descriptors such as “weak”, “moderate” or “strong” relationship – which are arbitrary and inconsistent – have been proposed. While most researchers would agree that a correlation less than 0.1 indicates a negligible and one greater than 0.9 a strong relationship, values in between are disputable and therefore should be interpreted within the context of the posed research question.



Figure 2: Scatterplots of sample data with both correlation coefficients

## Time series decomposition

Time series decomposition is a method that splits a time series, which may exhibit a variety of patterns into several components, each representing an underlying pattern category, “trend”, “seasonality”, and “residual”. **(Source:** [**https://otexts.com/fpp2/decomposition.html**](https://otexts.com/fpp2/decomposition.html)**,** [**https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2**](https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2)**)** This is often employed to help improve understanding of the time series or to improve forecast accuracy. **(**[**https://otexts.com/fpp2/decomposition.html**](https://otexts.com/fpp2/decomposition.html)**)**

The decomposed pattern components are defined as follows:

* **Trend** describes whether the time series is decreasing, increasing or constant over time. (**Source:** [**https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2**](https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2)) It does not have to be linear. Sometimes we refer to a trend as a “change of direction”, e.g., when it might go from an increasing to a decreasing trend. **(**[**https://otexts.com/fpp2/tspatterns.html**](https://otexts.com/fpp2/tspatterns.html)**)**
* **Seasonality** describes the periodic signal in the time series. (**Source:** [**https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2**](https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2)) This pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality has always a fixed and known frequency. **(**[**https://otexts.com/fpp2/tspatterns.html**](https://otexts.com/fpp2/tspatterns.html)**)**
* **Residual** is what remains behind the separation of seasonality and trend from the time series. It is the variability in the data that cannot be explained. (**Source:** [**https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2**](https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2))

A time series can be considered as a combination of these components, either additively or multiplicatively. **(Source:** [**https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/**](https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/)**, paraphrase this)**

An additive decomposition model is defined as

,

where is the data; , the seasonal component; , the trend component; and , the residual component, at period . Alternatively, a multiplicative decomposition is formulated as (**Source:** [**https://otexts.com/fpp2/components.html**](https://otexts.com/fpp2/components.html))

.

To identify whether the problem is additive or multiplicative, a review of a plot of the time series can be regarded as a good starting point. **(Source:** [**https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/**](https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/)**)** A rule of thumb for selecting the right model is to see if the trend and seasonal variation are relatively constant over time, i.e., linear. When this is the case, an additive model can be chosen. Otherwise, if the trend and seasonal variation increase or decrease over time, a multiplicative decomposition shall be used. (**Source:** [**https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2**](https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2))

### Example on a real-world dataset

For better understanding, let us look at a real-world dataset **(see figure XY)**, which describes the total number of airline passengers from 1949 to 1960. The horizontal axis represents the number of monthly observations during that period, the vertical axis, the number of airline passengers in thousands. **(Source:** [**https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/**](https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/)**)**

The line plot may suggest a linear trend. A seasonality can also be observed; however, the amplitude appears to be increasing, indicating a multiplicative problem. Hence, a multiplicative decomposition is applied as shown in **figure XY**. **(Source:** [**https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/**](https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/)**)**

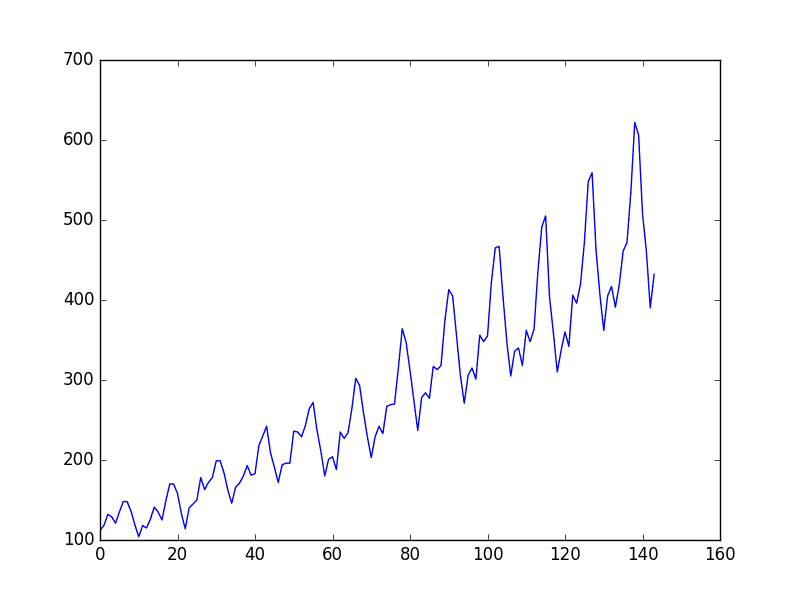


Figure 3: Airline passengers dataset

There are also datasets for which a naive or classical decomposition fails, as illustrated in **figure XY**, where it was not able to separate the noise from the linear trend. For such scenarios, caution and scepticism is required. **(Source:** [**https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/**](https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/)**)**

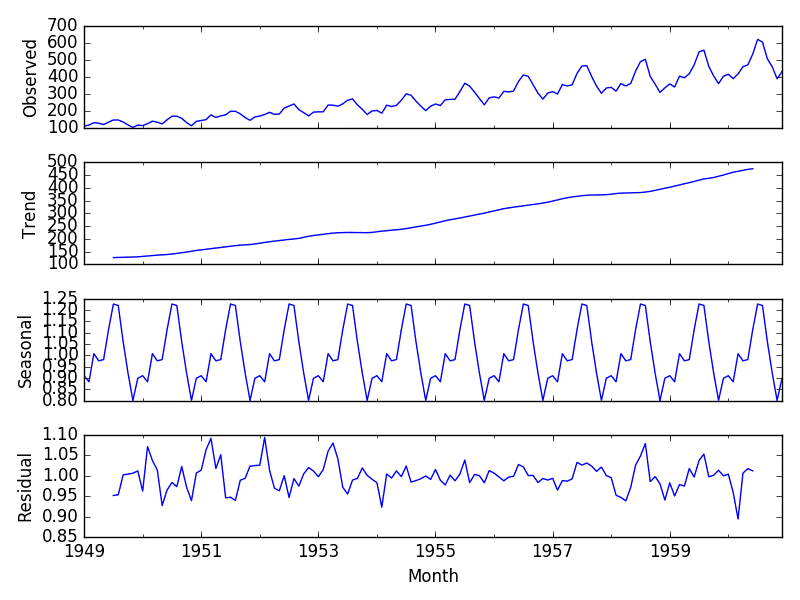


Figure 4: Multiplicative decomposition of airline passengers dataset

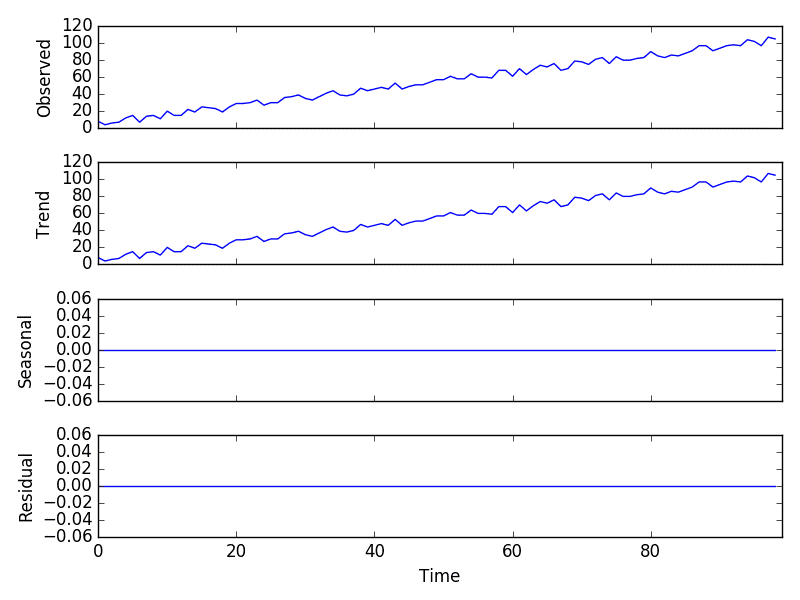


Figure 5: Additive decomposition of linearly increasing trend with random noise

# 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. **(source: https://www.ibm.com/cloud/blog/python-vs-r)** 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 **table** below gives an overview of the gathered review data in numbers. Just by observing the numbers, we can tell that, in generall, more reviews are written on Google than on Tripadvisor. This can be justified by the fact that Tripadvisor targets primarily travellers or tourists, while Google is widely known and used by the majority, including locals and tourists.

|  |  |  |
| --- | --- | --- |
| **Restaurant** | **Number of reviews** | |
| *Tripadvisor* | *Google* |
| The Butcher |  |  |
| Uster | 13 | 368 |
| Aarbergergasse | 146 | 1233 |
| Zug, Metalli | 152 | 960 |
| Zürich West | 63 | 906 |
| Missu Miu |  |  |
| Europaallee | 120 | 1272 |
| Negishi |  |  |
| Zug, Metalli | 395 | 910 |
| Pilatusstrasse | 140 | 1379 |
| Steinen | 137 | 644 |
| Nooch |  |  |
| Aarbergergasse | 368 | 930 |
| Basel, Barfi | 188 | 646 |
| Mall of Switzerland | 57 | 772 |
| Mattenhof | 50 | 448 |
| Richti, Wallisellen | 195 | 951 |
| Uster | 28 | 369 |
| Outback |  |  |
| Stadelhofen | 403 | 1627 |

Table 1: The scraped review data in numbers

The review data come from two different sources; hence it had to be processed and organised and brought to the same format. For this purpose, we use the schema defined in **Appendix XY**, which 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

Initially, we contacted Tripadvisor via email, asking if they could provide us the review data through an API. Unfortunately, they did not respond to our request – a creative solution had to be devised. We decided to scrape the data off the webpage **(see figure XY)**, which is publicly accessible.

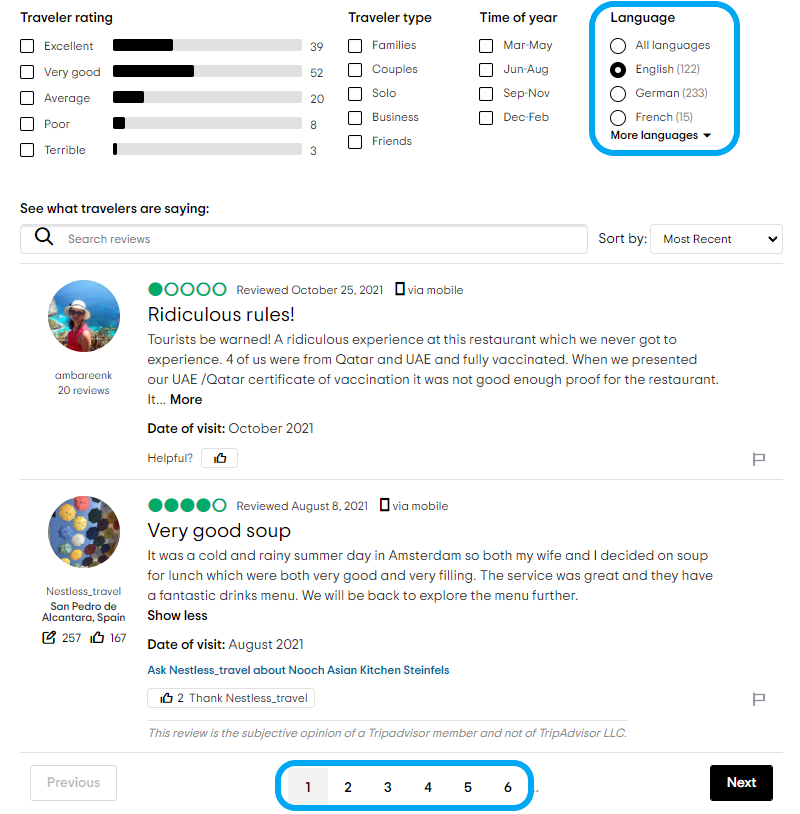


Figure : Tripadvisor webpage with restaurant reviews. Note that for purpose of presentation, only two reviews are listed.

This is the part of the webpage where the reviews for a single restaurant are shown. First, we tried scraping the review data using two Python packages called Requests and BeautifulSoup. These packages are only capable of loading the static HTML page; interactions – framed in blue, in **figure XY** – such as button clicking, are not possible. However, these interactions are very much needed to scrape all the reviews.

Since Tripadvisor has designed its pages in a manner, that only 10 to 15 reviews are displayed per site, many page switches are necessary to gather all the data. Furthermore, the website initially loads reviews only in the web browser's default language, which also has to be changed by clicking the option “All languages”; a lot of reviews are filtered out if this option is not selected.

We needed a technology to perform these tasks successfully. We found Selenium to be a great tool to overcome these challenges. Using this framework, one can fetch a page, select specific elements, and perform the preferred actions, such as clicking on a button on the chosen element, using JavaScript commands. In addition, it can also preserve the state it is in after performing an action, e.g., after switching to another page, it stays on that page, available for further actions.

Using this technology as a baseline, we developed a “ScrapingTool”, which can be found in our GitHub repository. (https://github.com/fatihsolmaz22/ScrapingTool) This tool simplified the data extraction from Tripadvisor.

**Tripadvisor review container**

The customer reviews are bundled in HTML elements with HTML tag “div” and attribute name “review-container”, as depicted **in figure XY**. Accordingly, we collected all the review elements for each page and extracted the required information from them, for instance, the individual ratings, stored inside the “review-container” with HTML tag “span” and attribute name “ui\_bubble\_rating.bubble\_40“; and the dates within the element tag “span” and attribute name "ratingDate ".

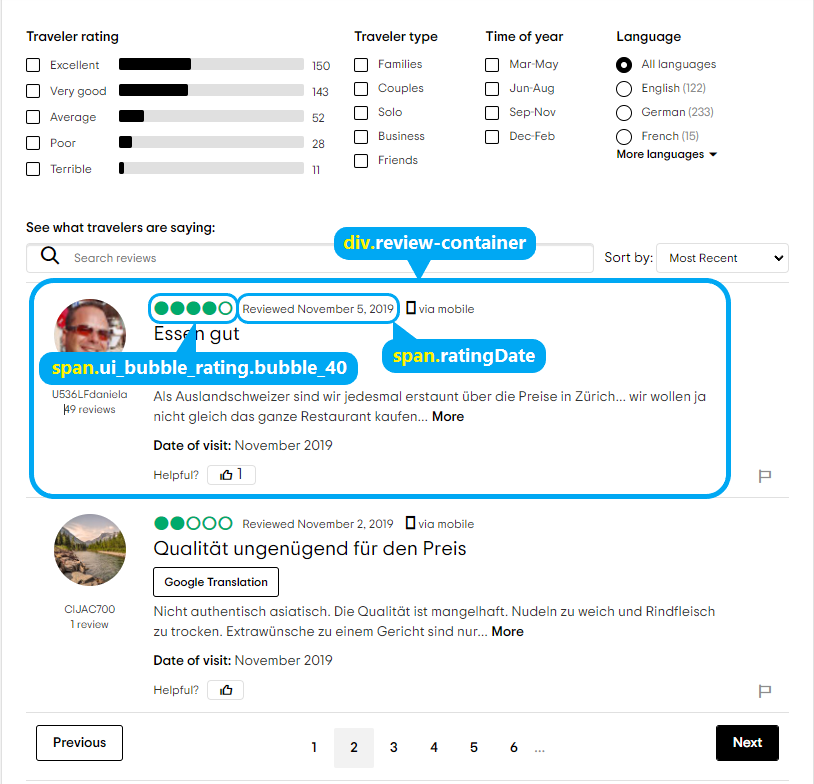


Figure : Review container, which comprises the review data of an individual customer.

### Google

While scraping Tripadvisor, it occurred to us that there was not sufficient review data for the correlation analysis. For this reason, we decided to scrape Google reviews as well. The following **figure** illustrates a restaurant review section on Google Maps.

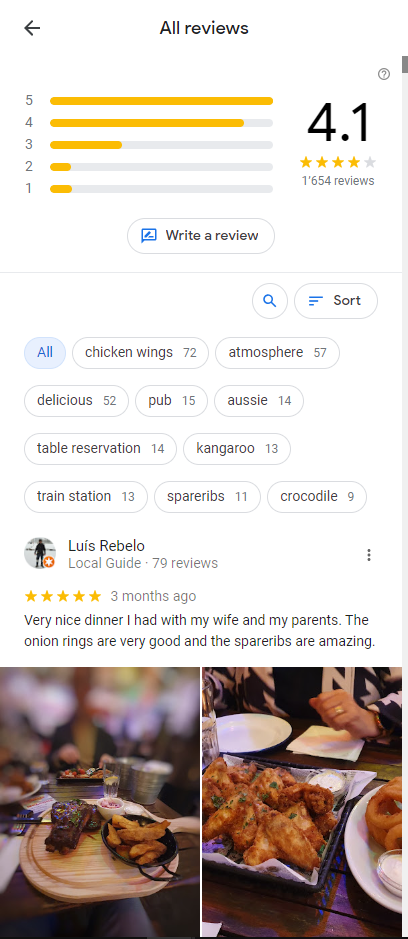
  
This is the initial page when visiting a restaurant's review section on Google Maps; only the first 10 reviews are visible to the user. To view further reviews, one would have to scroll down to the bottom of the page. There is no issue in acquiring the first few reviews but after loading a set amount of reviews, the webpage becomes unresponsive and shows an error **(see figure XY)**.

Figure : Example of a restaurant review section on Google Maps.

We started to think, it is impossible to extract the review data from Google. But then, another idea occurred. Usually, when data is dynamically requested, it has to be separately downloaded by the browser and injected into the webpage. Based on this realisation, we started observing the network traffic in the browser, as shown below.

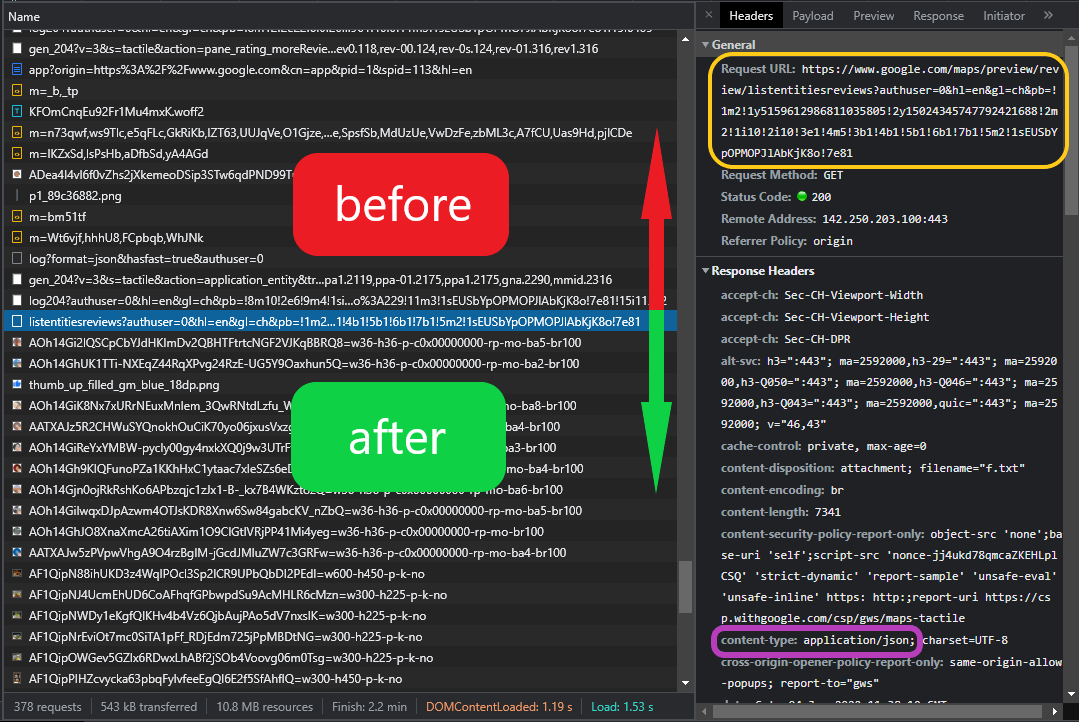


Figure : Network traffic before and after scrolling to the bottom of the review section

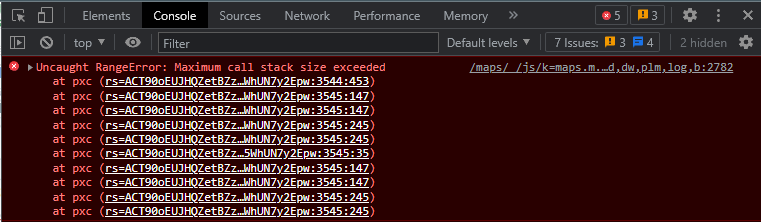


Figure : Error message in the browser after attempting to load too many reviews

After scrolling down to the bottom of the page for the first time, to load the next set of reviews, the browser shows inside the network traffic section, a new list of files or information, which is highlighted in green and labelled with “after” in **figure XY** – to be able to present the freshly loaded reviews on the webpage. By inspecting the first link, which is emphasised in blue, one can see on the right side, the entire request link – framed in yellow; more notably, the “content-type” of the response to that request, which is a JSON file – highlighted in purple. Consequently, we decided to take a look at this file, since it might contain the information, we require.

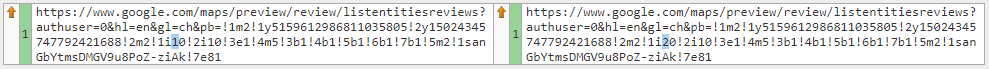
The JSON file only provided restaurant review data of 10 reviewers. Therefore, we decided to scroll down one more time and compare both request links that granted us access to the review data **(see figure XY)**.

Figure : Comparison of the request links before scrolling down (left) with the one after scrolling down (right).

Note, that both links only differ in one digit as emphasised in blue. The left link provides reviews with the indices 10 to 19, whereas the right one, the reviews with the indices 20 to 29. In order to request all the review data, one would only have to replace the highlighted digit with 0, 10, 20, 30 etc.

However, this approach leads to sending a lot of requests to the Google server, which in turn asks us to complete a CAPTCHA – a type of challenge-response test used in computing to determine whether the user is human. **(**[**https://en.wikipedia.org/wiki/CAPTCHA**](https://en.wikipedia.org/wiki/CAPTCHA)**)** The tricky part was to realise that after completing the CAPTCHA manually, the browser will store the interaction between the Google server and the user as a cookie, which then again can be saved and later used to prevent Google servers from blocking the execution of further requests.

## 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.

# Results

In this chapter we present the results to answer the initially formulated research questions and to verify the hypotheses. First, we discuss our findings of how the overall rating is calculated for restaurants on Tripadvisor and Google. The subsequent sections are devoted to the correlation analyses – the primary goal of our thesis – where we utilize different tools, e.g., line and scatter plots; the correlation coefficients, Pearson and Spearman; and time series decomposition.

## Analysis of the overall rating

The notion behind this analysis is that nowadays, popular online review platforms provide an enormous amount of information. We assume that the first thing a person might do to learn about a restaurant on such platforms, is to look at the overall rating, often displayed at the beginning of the website. We believe that a good overall rating can increase the popularity of a restaurant, which in turn can have a positive effect on restaurant revenue. Furthermore, we hypothesize that the reckoning of the overall rating varies from platform to platform, since each platform is built differently.

### Tripadvisor overall rating

Tripadvisor states that the overall bubble rating takes into account the quality, quantity and age of individual tourists’ ratings and reviews. The “bubble” rating ranges from one to five – one meaning “terrible” and five meaning “excellent”. The overall rating can be broken down by distribution of ratings; type of travel, such as business or family; or aspects of the business, such as service or cleanliness. **(Source: https://www.tripadvisorsupport.com/en-GB/hc/traveler/articles/438)**

Since the formula, Tripadvisor use to calculate the overall rating, is not publicly accessible, an alternative solution approach had to be devised. Initially, the plan was to employ multiple linear regression. We therefore scraped additional restaurant review information **(see Section XY)**, we thought, were necessary to build and train a model to predict the overall rating. Before deciding for this approach, we had also considered – contrary to Tripadvisor's statement – whether the overall rating might just be the average of the individual costumer ratings.

Thus, we scraped the review data from 1720 restaurants in canton Zurich. We then computed the average rating for each restaurant and compared it with the overall rating. The below scatterplot in **figure XY** depicts that the computed and rounded average of the individual costumer ratings is equivalent to the overall rating. For all the 1720 restaurants, the data points lie on a straight line. The scatterplot above in **figure XY** showcases the link between the overall rating and the computed overall rating. Notice, that despite all the data points being scattered in a horizontal direction, they remain within the bounds specified by the vertical dashed red lines.

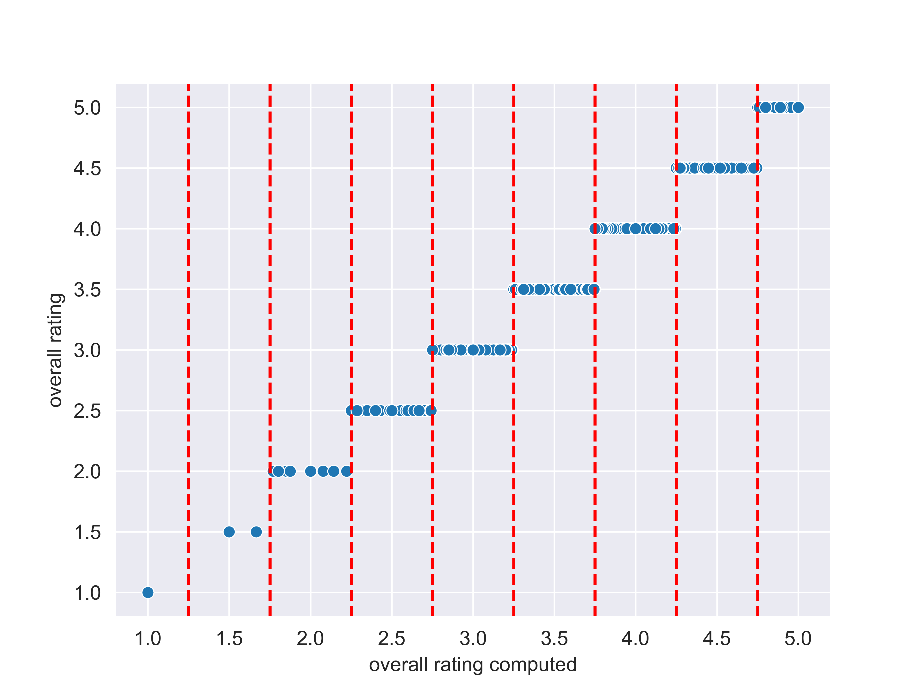
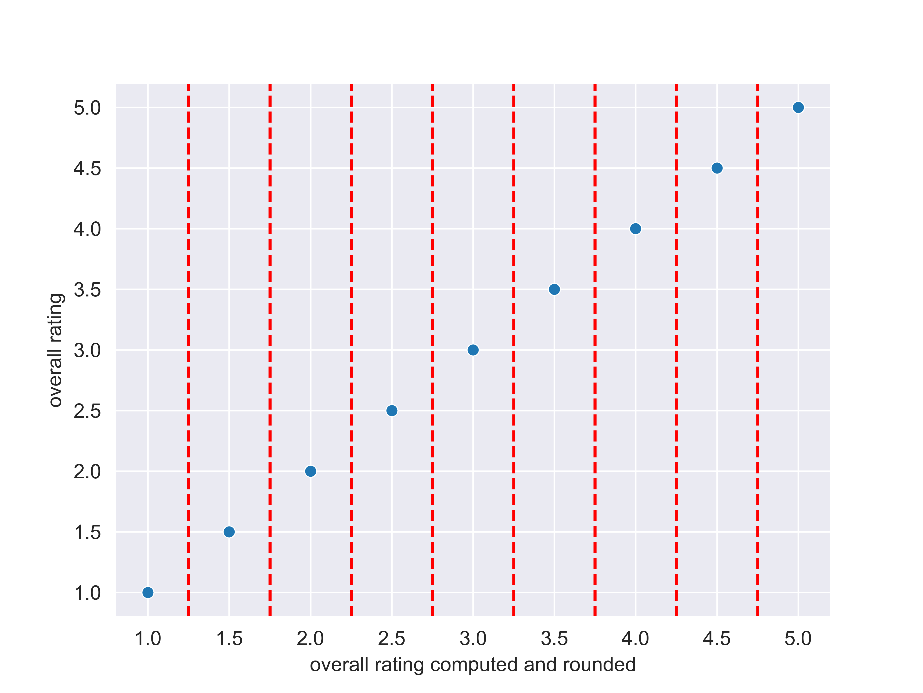


Figure 12: Scatterplots which describe the relationship between Tripadvisor overall rating and the computed overall rating. The vertical red dashed lines represent the bounds.

### Google overall rating

The restaurant overall rating on Google is calculated from user ratings and a variety of other factors. **(Source:** [**https://support.google.com/business/answer/4801187?hl=en-GB**](https://support.google.com/business/answer/4801187?hl=en-GB)**)** Nevertheless, we tried the same approach as previously described, with the only difference, we did not use a large dataset, but rather the review data of the 15 restaurants, where we had the revenue data, because the scraping of a large Google review dataset is sophisticated. In addition, we did not have time to investigate this further. The scatterplots in **figure XY** demonstrate that for the 15 restaurants, the computed and rounded average of the individual user ratings is equivalent to the overall rating and the data points remain within the bounds.

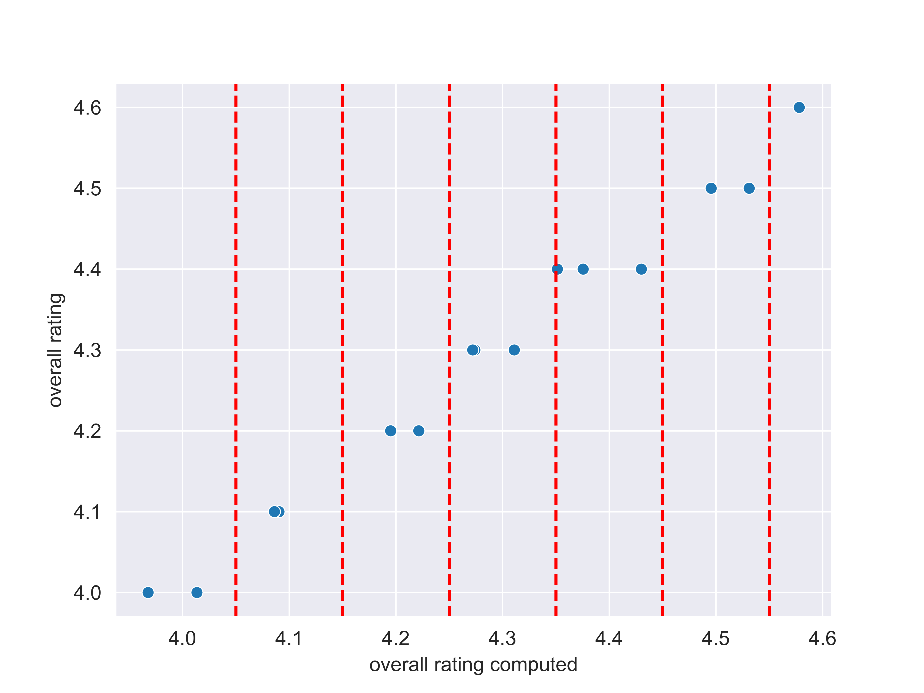
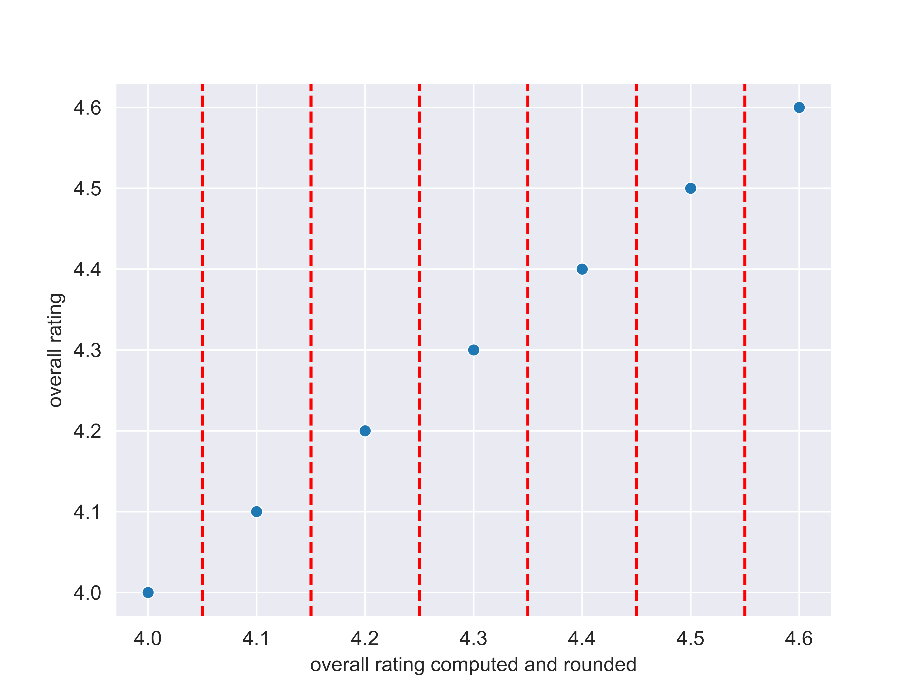


Figure 13: Scatterplots which describe the relationship between Google overall rating and the computed overall rating. The vertical red dashed lines represent the bounds.

## Correlation between Google restaurant rating and Prognolite turnover data

### Average restaurant rating against average turnover

### Overall restaurant rating development against average turnover

## Correlation between Tripadvisor restaurant rating and Prognolite turnover data

### Average restaurant rating against average turnover

### Overall restaurant rating development against average turnover

## Correlation between Tripadvisor and Google restaurant rating

### Average restaurant rating Tripadvisor against Google

### Overall restaurant rating development Tripadvisor against Google

# Conclusion

# Lists

## Bibliography

## List of figures

## List of tables

# Appendix

## JSON schema of the restaurant review data for a restaurant

**{**

**"restaurant\_name": string,**

**"overall\_rating": int,**

**"reviews\_count": int,**

**"all\_reviews": [**

**{**

**"author\_data": {**

**"author\_level": int,**

**"author\_member\_since": int,**

**"author\_stats": {**

**"contributions": int,**

**"cities\_visited": int,**

**"helpful\_votes": int,**

**"photos": int**

**},**

**"author\_distribution": {**

**"review\_value\_5": int,**

**"review\_value\_4": int,**

**"review\_value\_3": int,**

**"review\_value\_2": int,**

**"review\_value\_1": int**

**}**

**},**

**"review\_data": {**

**"date": string,**

**"title": string,**

**"rating": int,**

**"content": string,**

**"likes": int**

**}**

**}, ...]**

**}**

## Comparison between Google review as JSON and the actual Google review



Figure : Formatted JSON file preview of Google review



Figure 15: The actual Google review