For the first step of the project, reviews of companies with over 10,000 reviews should have been extracted. The scraped information should contain related information such as the reviewer's name, country, review title, text, rating (in number of stars), and whether the company replied to the review.

For the web-scraping, the Trustpilot website was chosen as a source of reviews. To limit the number of reviews, the category "Travel & Vacations" was selected. The comparison analysis of the range of categories showed that "Travel & Vacations" contains a wide range of companies having a higher number of reviews and also the companies with a lower rating, which is important for data variability.

Extraction of the data was performed with the BeautifulSoup Python package for parsing HTML and XML data.

Here are some reasons why BeautifulSoup was chosen:

1) Truspilot uses clear and predictable pagination URLs (e.g.,`?page=2`, `?page=3`), which makes it easy to iterate through pages without the need to simulate clicks and AJAX requests.
2) The review data is already present in the raw HTML response. No JavaScript rendering is needed to reveal it, so there's no need for tools like Selenium or Playwright.
3) BeautifulSoup is lightweight and fast when dealing with static pages.

The data from the reviews page of the chosen category was extracted to get the list of companies with a review number higher than 10,000. It was extracted with the BeautifulSoup function, searching for specific elements and attributes. Further parts of the data will be extracted with the same method.

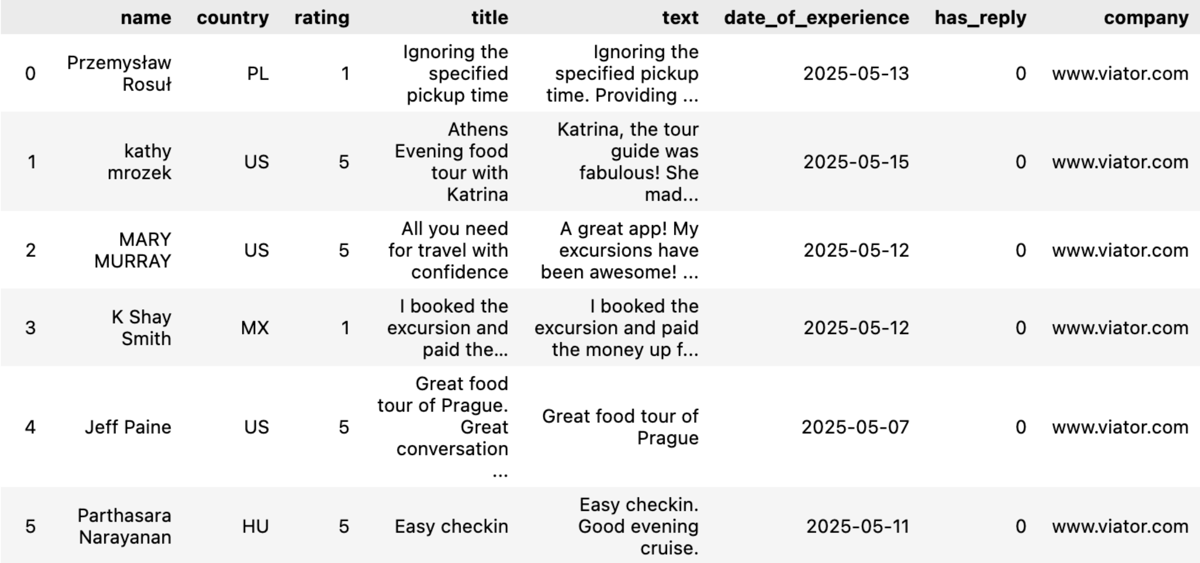
The related information, such as the company name, number of reviews, and the website (it will be used to construct the link to the page from which the reviews will be extracted).

At the end, we received the list of 59 companies that have more than 10,000 reviews.

The company with the biggest number of reviews, "Viator", has 260,646 reviews at the moment. It would take a couple of hours to scrape that much data, therefore, it was decided to cut the number of reviews that are going to be extracted. To make the data more variable for future sentiment analysis and avoid saving only positive comments, scraping was limited to 15 pages per company. Not only does it increase variety, but it also filters the comments for the most recent.

For every company, the reviews with the name, country, review title, text, rating (in number of stars), and whether the company replied to the review were collected. The information resulted in a CSV file with 21,000 reviews. That number of reviews was the most optimal, as Trustpilot blocks the requests if there are too many of them and they are too quick to prevent automated scraping. This issue can be solved by suspending the execution of data extraction, but it makes the process time-consuming, so that it can lead to that the connection with the virtual machine can be lost.

In the following picture (pic.1), it can be seen what the gathered data looks like in a DataFrame. It can be easily linked to other tables with the "company" keyword.



	name	country	rating	title	text	date_of_experience	has_reply	company
0	Przemysław Rosuł	PL	1	Ignoring the specified pickup time	Ignoring the specified pickup time. Providing ...	2025-05-13	0	www.viator.com
1	kathy mrozek	US	5	Athens Evening food tour with Katrina	Katrina, the tour guide was fabulous! She mad...	2025-05-15	0	www.viator.com
2	MARY MURRAY	US	5	All you need for travel with confidence	A great app! My excursions have been awesome! ...	2025-05-12	0	www.viator.com
3	K Shay Smith	MX	1	I booked the excursion and paid the...	I booked the excursion and paid the money up f...	2025-05-12	0	www.viator.com
4	Jeff Paine	US	5	Great food tour of Prague. Great conversation ...	Great food tour of Prague	2025-05-07	0	www.viator.com
5	Parthasara Narayanan	HU	5	Easy checkin	Easy checkin. Good evening cruise.	2025-05-11	0	www.viator.com

Picture 1. Data in DataFrame.