## Data gathering process:

The data for our analysis has been retrieved from TripAdvisor.com on dd-mm-yy.   
We have coded a HTML scrapper that gets all the data for the restaurants classified as being located in the Copenhagen area.

The HTML scraper is attached as a Jupyter Notebook called “TripAdvisor\_scraper.ipynb”. This notebook creates 3 CSV files that contains the overview links (overview\_urls.csv) the individual restaurant links (individuals\_urls.csv) and one with all the data for analysis (data.csv). The first part of the notebook is code we have borrowed from slides from Snorre Raslund during the Social Data Science course.

We start by getting a list containing 78 overview links with all the restaurants in Copenhagen. We feed these links into another layer of our scraper, that retrieves the individual restaurant links, from the 78 pages. We then remove duplicates, as there were a few, and create a list with all 2319 individual restaurant links.

The list with all the individual links is then fed into our last layer of our scraper, that loads all the restaurants individual data. That data we end up with, is a name, address, long and latitude, total rating and individualized on category and lastly price range. We end up with ---- individual restaurants that has the data we wanted.

The way that TripAdvisor has structured their site makes it somewhat easy to scrape their data. They are consistent when creating links, and the data is stored the same way for all restaurants. We have experienced missing data for some restaurants, but this is simply due to the restaurants not having enough reviews.

We did find that TripAdvisor had an API, that probably could have used for this project. But there was some ethical reason for not doing so. This will be discussed in the section below. Another thing was the limitations to the API with only 1000 calls per day. This meant we had to do the data gathering over several days and potentially not be able to get all the data we wanted.

## Ethics of data gathering:

We found TripAdvisor to have an API, which we likely could have gotten access. But the explicit stated on their API documentation that, access would not be granted to academic research. We could have deceived our purpose in our application and thus gotten access, but chose not to, as this would have been ethically wrong.

Also, when HTML-scrapping one also has to talk about ethics of the data scrapped. We found no problem in scraping the data, as all the data we gathered was about the restaurants. This could have been gathered on multiple other sites and combined. No person’s data was involved, meaning we did not gather any personal information on the users that have reviewed the data.