

Web scarping the restaurants of Copenhagen

Group 20

Exam in Social Data Science, Økonomisk Institut, Københavns Universitet

Sofie Dalum,

Naja Algreen Suhr, GRZ240

Rasmus Habel Svärd,

Michael Fogh Emcken,

## Abstract

People have different opinions when it comes to food. Flavours, ingredients, setting of the place your eating and price. Everything around a meal, places a role. This can be seen on the reviews from TripAdvisor, that rates restaurants all over the world, based on user’s experience.  
In this project, we

First we eat, then we do everything else. – M.F.K. Fisher

## Introduction to TripAdvisor and TripAdvisor Denmark

TripAdvisor Inc is a website company in the industry of travel services. The multiplatform company enables users to book hotels, vacation rentals, flights, find restaurants and activities, and brands itself as a travel guide. The platform also serves as a user review site of all on-site products, available in 28 languages and on 49 markets, according to the website.[[1]](#footnote-1) For this project we have limited the scope to the Danish subpage ‘tripadvisor.dk’ where the user interface and review language is Danish.

Users of TripAdvisor have the opportunity of finding, booking and reviewing products of the site, such as hotels or restaurants. The company revenue comes from commercial advertising, such as display based, click-based and subscription-based advertisement and from booking fees when users book through the website.[[2]](#footnote-2)

All registered users are able to review an element on TripAdvisor, but the reviews are not verified. Users can review as many or as few as they like, and the user can choose how much information alongside the ratings they want to disclose. Users are invited to review bookings they have made on TripAdvisor after the end of the booking, and some companies may ask their customers to review them online, sometimes motivated by participation in a contest or drawing of prices, if the customers submits a review. Because of this behaviour concerning TripAdvisor reviews cannot be trusted solemnly in general to be a true evaluation of a customer’s opinion.

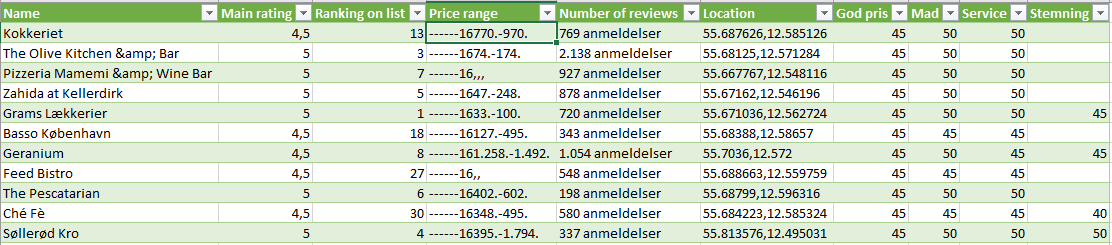
An example of this is the restaurant in London called ‘The Shed at Dulwich’ which at one point the highest rated restaurant on TripAdvisor in London. The problem was though that the restaurant didn’t exist and had never existed. The high rating came exclusively from fake reviews of family and friends, and after only 92 great fake review the restaurant entered first place.[[3]](#footnote-3) From this story it is to be learned that is it rather easy to manipulate TripAdvisor rating and that TripAdvisor as a platform doesn’t validate reviews or ratings. This is an important point to have in mind when moving forward using TripAdvisor data. We have chosen to use TripAdvisor data despite is disadvantage, because of its multiplatform nature and overall volume of Copenhagen based restaurants.

## Data generation and gathering

The data for our analysis has been retrieved from TripAdvisor.dk on dd-mm-yy. After choosing TripAdvisor as our area of interest we build a web scraping script in Python. The full scraping process has been logged with a script made by Snorre Ralund, provided in class. The script delays the requests of TripAdvisor by 0.5 seconds and logs the scraping in a text file for documentation and process control.

The scraping process of TripAdvisor consists of several parts. When searching for restaurants on TripAdvisor in Copenhagen the result is a number of overview pages containing the individual restaurants, as seen below.



The search result for restaurants in Copenhagen contains 78 overview pages. The script scrapes the overview pages for links on the individual restaurant pages and after removing duplicates, the script creates a csv file, ‘individual\_urls.csv’, with all individual restaurant links. The next layer of the scraper loops through ‘individual\_urls.csv’ loads all the restaurants individual data. The individual data is as follows: name, location, price range, price class, number of reviews, ranking, rating on value for money, food quality, service, atmosphere and overall rating. The individual data is here seen as an example.

REPLACE WITH FIANL DATA FRAME

The scraping script is attached as a Jupyter Notebook called “TripAdvisor\_scraper.ipynb”. This script creates **X** 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 HTML of TripAdvisor is structured in a way that makes the site somewhat easy to scrape their data. All links are absolute and stable links, and the data on the individual restaurant is stored the same way for all restaurants on site. Not all restaurants have the same data available on the individual pages, but the HTML structure is still preserved with values then being hidden. We have noted that data is missing on a limited number of restaurants, but after looking into it, we have concluded that it is because of the restaurants having a few or no reviews at all.

Before beginning the process of building the web scraper script we discovered that TripAdvisor has an API with restricted access. One can apply for a key to the content API, but the key is not granted for the use of content on data analysis, academic research and B2C application. Overall the API wouldn’t be effective for us in this project as the limitation on the API is 1000 calls per day, which means the data collecting process would have been stretched over several days. Furthermore, the time scale on the processing of the application on a key is unknown, but possible longer than this project. We therefore chose to build the web scraping script.

## Ethics of data collection and gathering:

If we had chosen to apply for the API key and had been successful in the scale of the project, we would have broken the terms and conditions when analysing the data. We could have deceived our purpose in our application and thus gotten access, but chose not to, as this would have been ethically wrong.

By scraping any web page under Danish jurisdiction is crucial to reflect on ethics and hacking, to avoid breaking the law. The scarping process we have scripted only collects data otherwise available when using the web page. All data collected from TripAdvisor is visible when clicking on and around the web page. The script is build using open source tools that are able to read HTML in a legal manner. Most, maybe even all, restaurants have website with the corresponding information on TripAdvisor, why we could have collected the data from the individual private web pages of the restaurants if needed, though we chose not to. Restaurants are public companies with an interest of sharing basic information, such as telephone number, addresses and other data, such as menu and pricing, why most restaurants have sample menus with pricing available online. Companies such as restaurants often benefits from being online as it creates more online reach and in the end more revenue. We find that scraping public restaurant data is ethically sound.

We have chosen to scrape only aggregated data, why we have only collected means and bundled data of restaurant reviews and therefore have no scarped data of individual physical or online persons, such as usernames, profile photo or other data. If we had chosen to scrape data of individual persons, we would have had to discuss the application use and the ethics thereof further. By limiting our self to only public restaurant data from TripAdvisor we have concluded that our scraping and analysis is ethically sound.

## Data processing

After extracting the data from TripAdvisor, the data needed processing and cleaning before being ready for analysis and modelling. The processing script reads in the scraping script output, the raw web scraping csv. The column ‘Distance from Kgs. Nytorv (m)’ is created by measuring the distance from Kgs. Nytorv to the coordinates of the restaurants with the geopy great\_circle function and later rounded. For cleaning all badly formatted &-signs are replaced with correctly formatted ones, and the columns ‘Good price’, ‘Food’, ‘Service’, ‘Atmosphere’ are divided by 10 to regain format.

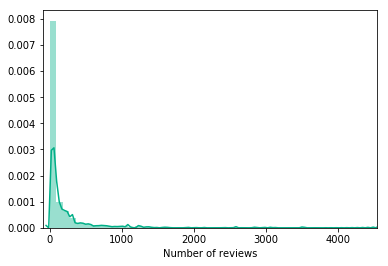
The processing script then handles the price class in dollar format and translates them into a numeric value by looping through the list and checking the values and translating with a key. E.g. $ translates into 1, $$ translates into 2 and so on. The translation is formatted to a dataframe then merged with the main dataframe. Next the script replaces the word ‘anmeldelser’ (reviews) with nothing (‘’) from the column ‘Number of reviews’, so that the column now only contains integers. After that the script reverts the column ‘Ranking on list’ into negative float as a new column ‘Reverse ranking on list’, sorts the values by main rating, reverse ranking and then ranking to construct a new order for ranking called ‘Full ranking’ which is then merged with the main dataframe. The latter is done to make sure that the two separate rankings of the communes Frederiksberg and Copenhagen is not conflicting, which they are in raw format.

#color formatting

For the Folium map on rankings the column ‘Ranking\_color’ is created. Based on the value from the column ‘Full ranking’ a colour is assigned. If the value is high the colour is green, if the value is low the colour is red.

## Analysis

The dataset contains 2320 restaurants from the greater area of Copenhagen and 23 columns. The columns are ['Restaurant', 'Main rating', 'Ranking on list', 'Price range', 'Price class', 'Location', 'Good price', 'Food', 'Service', 'Atmosphere', 'Type of food', 'Number of reviews', 'Address', 'Type of food link', 'New price class', 'Price class numeric', 'Reverse ranking on list', 'Full ranking', 'Distance from Kgs. Nytorv (m)', 'Postal code', 'Latitude', 'Longitude', 'Ranking\_color']. Out of the 2320 restaurants all of them have location data, but only 2173 have a main rating. 2170 have a ranking and around 1300 have ratings on good price, food, service, and only 612 have on atmosphere, all with standard deviations between 0.57 and 0.59. The main rating mean is 4.0 and a standard deviation of 0.7 meaning that restaurants in Copenhagen generally have a high rating, and that most restaurants are captured in the range between 3.3 and 4.7 in main rating. 2170 out of 2320 restaurants have as a minimum one review, but the average is 144 reviews, the median number is 26 reviews and a standard deviation of 370. The maximum one restaurant has is 4551 reviews.



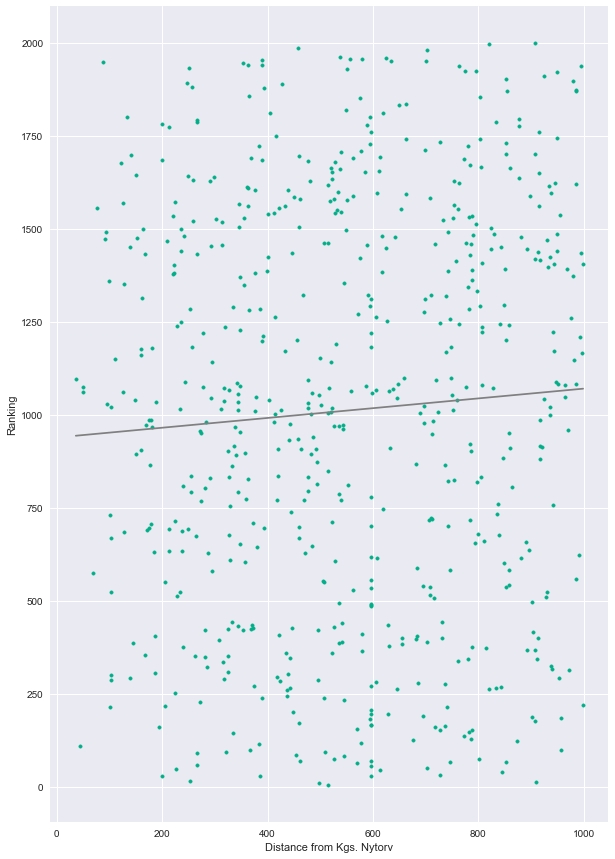
## Interactive Folium Map

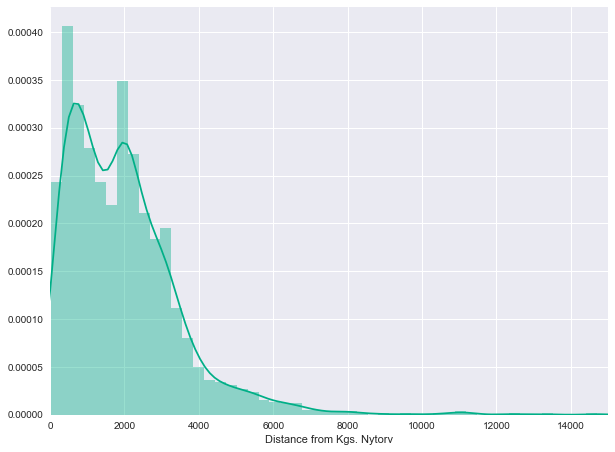
The Python Folium package has been used for showing distribution of restaurants on different variables in Copenhagen in a simple map visualisation format. The Folium package has the benefit of being interactive, so it is possible to zoom in and out while using Jupyter Notebook. On the maps each restaurant in the dataset has its own point based on its coordinates. Some maps have clustering so when zoomed in restaurants are gradually more and more clustered.

See attachment 1 for map visualisation and colour scale. On the two maps in attachment one the individual points or clusters are coloured by rankings of the restaurants. The ranking is a calculated value of the main rating, rank on list, number of reviews. The scale goes from 2000 to 0, with 2000 being the highest and best ranking. Based off this calculated ranking a colour is calculated. If the rank is between 2000 and 1800 the colour of the point or cluster will be green. If between 200 and 0 the colour will be red. Black illustrates restaurants without ranking data. See attachment one for colour scale. As seen on attachment one, restaurants of all ranks are spread out over Copenhagen. All neighbourhoods have high, medium and low ranked restaurants. Clusters and collections of low-ranking restaurants is centred around the central station, near Vesterport, Vesterbro and in Tivoli. Other areas of low rankings are on Højbro Plads, Kultorvet, Kongens Nytorv, Nyhavn, Islands Brygge and on Østrebrogade. Clusters of high ratings are present in many areas of Copenhagen, but especially the area around Frederiksborggade, Dronning Louises Bro and Nørrebrogade. Medium ranked restaurants are likewise present all over Copenhagen.

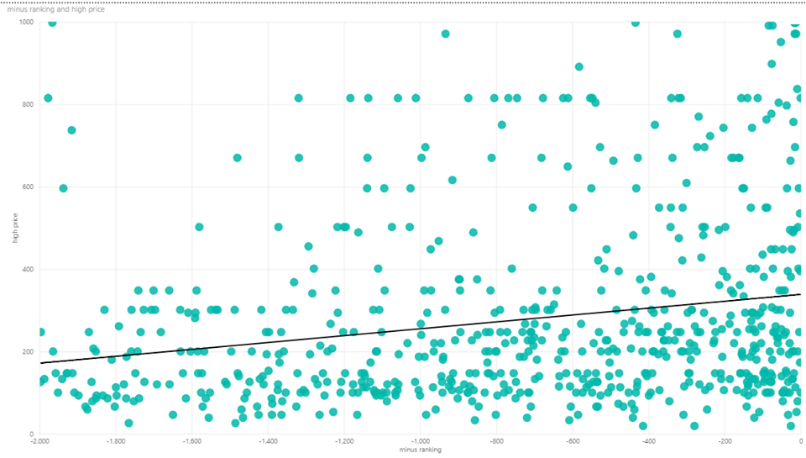
## Analysis

As seen on the following graph, most restaurants in the dataset are located less than 4000 meters away from Kgs. Nytorv in the centre of Copenhagen. Close to half of the restaurants are less than 2000 meters away. As seen on the graph to the right, there is a small positive correlation between distance between distance and ranking: the further away from Kgs. Nytorv the higher and better ranking. As seen the correlation is clearly small.





As seen on the graph below, there is a small positive correlation between the highest price and the reverse ranking: restaurants with higher prices tend to have higher rankings in Copenhagen. The correlation is small, but present. The graph shows that prices of restaurants with ranking between zero and two hundred ranges on the full scale from close to zero and up to a thousand Danish kroner.



## Conclusion

## Web pages

TripAdvisor:

Abouts us: “om os”, <https://tripadvisor.mediaroom.com/us-about-us>, 22.08.2019

Main page in English: [www.tripadvisor.com](http://www.tripadvisor.com), 22.08.2019-30.08.2019

Main page in Danish: [www.tripadvisor.dk](http://www.tripadvisor.dk), 22.08.2019-30.08.2019

Developer page about the TripAdvisor Api, <https://developer-tripadvisor.com/content-api/>, 22.08.2019

Vator:

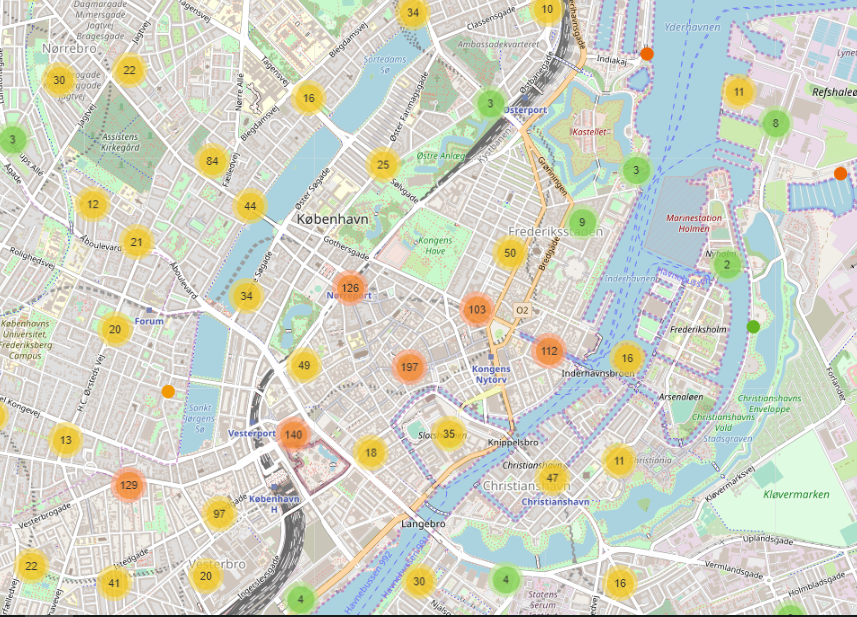
Article about the business model of TripAdvisor <https://vator.tv/news/2018-04-13-how-does-tripadvisor-make-money>, 23.08.2019

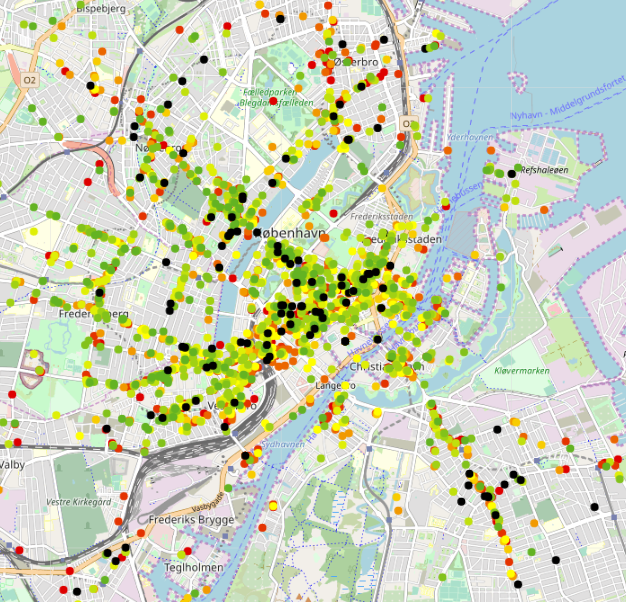
Independent:

The Shed at Dulwich: <https://www.independent.co.uk/life-style/food-and-drink/the-shed-at-dulwich-was-london-s-top-rated-restaurant-just-one-problem-it-didn-t-exist-a8107791.html>, 27.08.2019

# Attachment 1

Colour scale for mapping ranking of restaurants. The scale covers from rank 2000 to 0 with 200 in each step colour group. First green colour from left is 2000-1800 and the last red colour at the very opposite the rating under 200. Black illustrates restaurants without data.





1. TripAdvisor, about us: ”om os” [↑](#footnote-ref-1)
2. Vator TV: “How does TripAdvisor make money?” [↑](#footnote-ref-2)
3. Independent: ”The Shed at Dulwich” [↑](#footnote-ref-3)