# Web scraping the restaurants of Copenhagen

Group 20

Exam in Social Data Science, Økonomisk Institut, Københavns Universitet

Sofie Dalum, XWJ215 – line: 0-5, 21-58, 170-203, 306-322

Naja Algreen Suhr, GRZ240 line: 6-10, 59-95, 204-237, 323-338

Rasmus Habel Svärd, LCP535 line: 11-15, 96-132, 238-271, 339-354

Michael Fogh Emcken, XHT111 line: 16-20, 133-169, 272-305, 355-370

Shared contribution – line: 371-410

Characters : 35.000

# Table of contents

# 1. Introduction

"TripAdvisor is to travel as Google is to search, as Amazon is to books, as Uber is to cabs – so dominant that it is almost a monopoly."[1] Every review on TripAdvisor is an expression of an opinion: how did the reviewer perceive an event? Was the overall impression of the food so great it deserves five out of five? Was the food worth the price? Did the restaurant feel nice and welcoming? Was the waiter behaving like you wished? Every little detail is boiled down to ratings on service, food, price value and atmosphere and ends up as a ranker of the restaurant in question: how well did it perform? In the later years restaurants are opening more and more often in Copenhagen. From 2013 to 2017 more than 570 new restaurants opened in Copenhagen. A growth of 35% - and a growth alone in 2018 of 4%[2] More restaurants means more competition: what does it take to survive as a restaurant in Copenhagen today? How does a restaurant succeed?

By using high ranking on TripAdvisor as a measure of success, this paper will investigate what the restaurant business of Copenhagen looks like, what trends are present and what it takes to get a high rank. This paper uses data scraped from TripAdvisor that is cleaned and processed and analysed as a dataframe.

# 2. Data generation and gathering

## 2.1. 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.[3] 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's 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.[4]

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 customer submits a review.

Because of this behaviour reviews cannot be trusted solemnly to be a true evaluation of a customer's opinion in general.

[1] The Guardian: How TripAdvisor changed Travel

[2] Berlingske: "Nu begynder udskilningsløbet"

[3] TripAdvisor, about us: "om os"

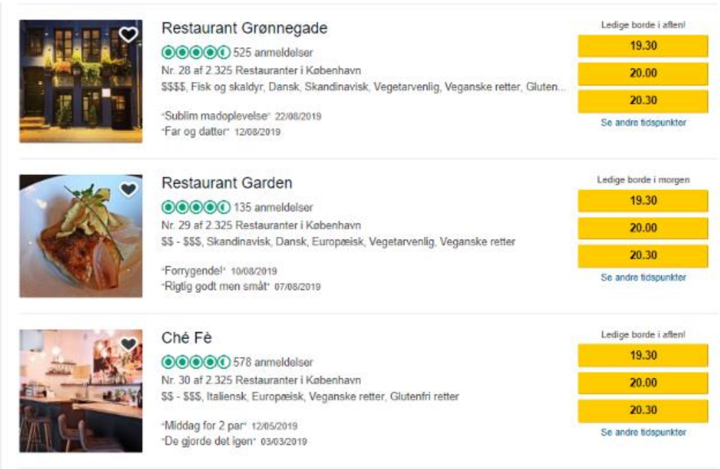
[4] Vator TV: "How does TripAdvisor make money?"

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.[5] 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 keep in mind when moving forward using TripAdvisor data. We have chosen to use TripAdvisor data despite is a disadvantage, because of its multiplatform nature and overall volume of Copenhagen based restaurants.

## 2.2. Generation and gathering

The data for our analysis has been retrieved from TripAdvisor.dk on 28.08.2019. 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 in Figure 1.

*Figure 1 - Restaurant overview example*



Restaurant Grønnegade
525 anmeldelser
Nr. 28 af 2.325 Restauranter i København
$$$$, Fisk og skaldyr, Dansk, Skandinavisk, Vegetarvenlig, Veganske retter, Gluten...
"Sublim madoplevelse" 22/08/2019
"Far og datter" 12/08/2019

Ledige borde i aften!
19.30
20.00
20.30
Se andre tidspunkter

Restaurant Garden
135 anmeldelser
Nr. 29 af 2.325 Restauranter i København
$$ - $$$, Skandinavisk, Dansk, Europæisk, Vegetarvenlig, Veganske retter
"Forrygende!" 10/08/2019
"Rigtig godt men småt" 07/08/2019

Ledige borde i morgen
19.30
20.00
20.30
Se andre tidspunkter

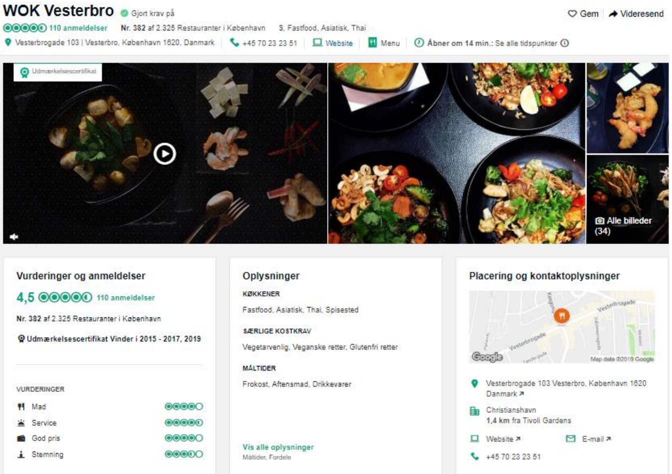
Ché Fè
578 anmeldelser
Nr. 30 af 2.325 Restauranter i København
$$ - $$$, Italiensk, Europæisk, Veganske retter, Glutenfri retter
"Middag for 2 par" 12/05/2019
"De gjorde det igen" 03/03/2019

Ledige borde i aften!
19.30
20.00
20.30
Se andre tidspunkter

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 of a page and of the dataframe (Figure 3 and Figure 2)

[5] Independent: "The Shed at Dulwich"

*Figure 3 - Individual restaurant site example*



WOK Vesterbro ✓ Gjort krav på ♡ Gem ➦ Videresend

●●●●◐ 110 anmeldelser Nr. 382 af 2.325 Restauranter i København $, Fastfood, Asiatisk, Thai

Vesterbrogade 103 | Vesterbro, København 1620, Danmark ☎ +45 70 23 23 51 Website Menu Åbner om 14 min.: Se alle tidspunkter ⓘ

Udmærkelsescertifikat

Alle billeder (34)

Vurderinger og anmeldelser

4,5 ●●●●◐ 110 anmeldelser

Nr. 382 af 2.325 Restauranter i København

Udmærkelsescertifikat Vinder i 2015 - 2017, 2019

VURDERINGER

| | |
|---|---|
| Mad | ●●●●○ |
| Service | ●●●●◐ |
| God pris | ●●●●○ |
| Stemning | ●●●◐○ |

Oplysninger

KØKKENER

Fastfood, Asiatisk, Thai, Spisested

SÆRLIGE KOSTKRAV

Vegetarvenlig, Veganske retter, Glutenfri retter

MÅLTIDER

Frokost, Aftensmad, Drikkevarer

Vis alle oplysninger
Måltider, Fordele

Placering og kontaktoplysninger

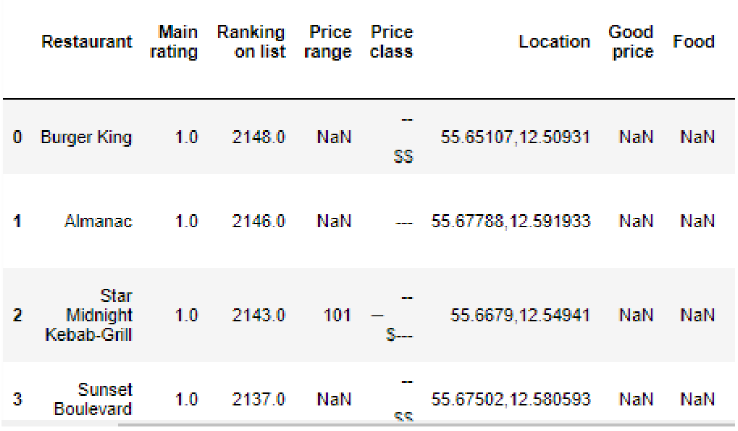
Vesterbrogade 103 Vesterbro, København 1620 Danmark ↗

Christianshavn
1,4 km fra Tivoli Gardens

Website ↗ E-mail ↗

☎ +45 70 23 23 51

*Figure 2 – Data frame example*



| | Restaurant | Main rating | Ranking on list | Price range | Price class | Location | Good price | Food |
|---|---|---|---|---|---|---|---|---|
| 0 | Burger King | 1.0 | 2148.0 | NaN | --<br>SS | 55.65107,12.50931 | NaN | NaN |
| 1 | Almanac | 1.0 | 2146.0 | NaN | --- | 55.67788,12.591933 | NaN | NaN |
| 2 | Star Midnight Kebab-Grill | 1.0 | 2143.0 | 101 | — --<br>$--- | 55.6679,12.54941 | NaN | NaN |
| 3 | Sunset Boulevard | 1.0 | 2137.0 | NaN | --<br>SS | 55.67502,12.580593 | NaN | NaN |

The scraping script is attached as a Jupyter Notebook called "TripAdvisor_scraper.ipynb". This script creates overview_urls and individual_urls CSV files that contains the overview links (overview_urls.csv) the individual restaurant links (individuals_urls.csv) and one with all the data for data processing, tripadvisordata_raw.

The HTML of TripAdvisor is structured in a way that makes the site somewhat easy to scrape. 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 with 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.

To avoid our CSV files to be overwritten in the exam folder, in case the notebook is run during examination, we have changed the names of the files that the scraper creates. TripAdvisor could make changes to HTML which causes other data to be loaded than what we initially loaded on the 28 August 2019. The raw file contained 2320 rows.
Our entire data processing and analysis to avoid different results than what we present in this paper.

## 2.3. Ethics of data generation 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 scraping 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 built using open source tools that are able to read HTML in a legal manner. Most, maybe even all, restaurants have websites 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 in sharing basic information, such as telephone numbers, 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 or restaurant level data, why we have only collected means and bundled data of restaurant reviews and therefore have no scraped 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 ourselves to only public restaurant data from TripAdvisor we have concluded that our scraping and analysis is ethically sound.

### 2.4. 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. Lastly we drop observations that have too much data missing, for us to use it. We end up with 2172 rows.
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.

We want to investigate if certain types of kitchens, pizza, Thai, seafood etc., are clustering together in some of the respective parts of the city. To be able to compare the different parts of the cities we have to transform the data.

First, we make a list of the top searched categories[6]. Then we group the different kitchens based on parts of the city the restaurants are located. And merge with a dataframe containing the total count of restaurants in the different city parts.[7]

We create a new data frame, with calculations of the percentage representation of different food types in the respective areas. The reason we do this is to make a relative scale, so we can compare the share of restaurant types across the areas. We also calculate the mean percentage representation across areas to give a quick overview of which restaurants cluster together in which areas.[8]

[6] Found on https://www.tripadvisor.dk/Restaurants-g189541-Copenhagen_Zealand.html#EATERY_OVERVIEW_BOX

[7] Appendix 3 chart 1

[8] Appendix 3 chart 2

# 3. The state of the restaurant business in Copenhagen

## 3.1. The dataset

The dataset contains 2172 restaurants from the greater area of Copenhagen and 22 columns. These 2172 restaurants are all restaurants in Copenhagen on TripAdvisor. The following data description therefore shows the state of the restaurant business. The dataset columns are ['Restaurant', 'Main rating', 'Good price', 'Food', 'Service', 'Atmosphere', 'Price range', 'New price class', 'Price class numeric', 'Type of food', 'Type of food link', 'Number of reviews', 'Address', 'Location', 'Distance from Kgs. Nytorv (m)', 'Postal code', 'Ranking on list', 'Reverse ranking on list', 'Full ranking', 'Latitude', 'Longitude', 'Ranking_color'].
A summary table of the data set can be found in appendix 1.

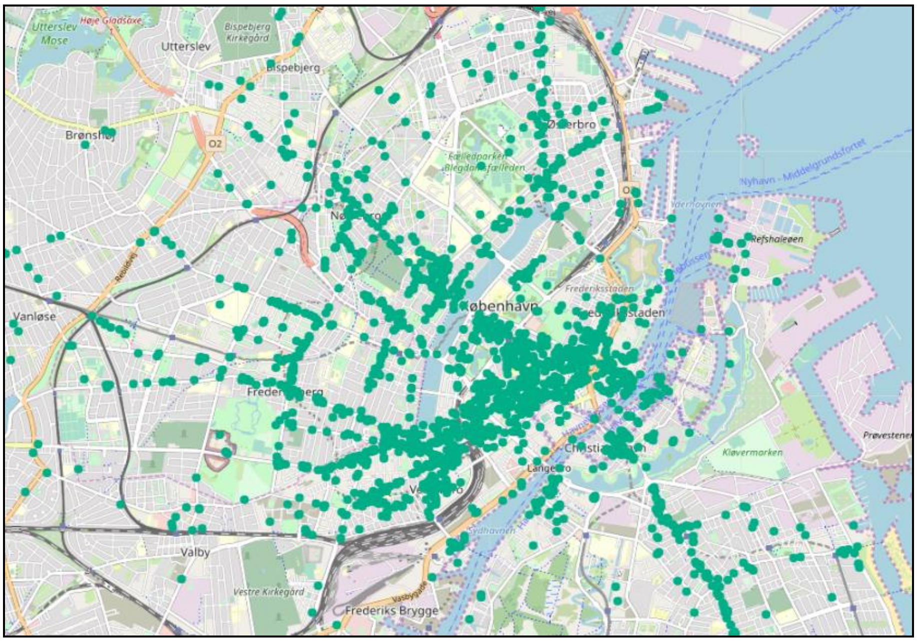
## 3.2. Plotting the restaurants of Copenhagen on maps

The Python Folium package has been used for showing distribution of restaurants on 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 appendix 2 for map visualisation and colour scale. On the two maps in appendix 2 the individual points and 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 in appendix 2, 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 Østerbrogade. 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. Especially the inner part of Copenhagen is dominated by medium ranked clusters, because of a great number of both high and low ranked restaurants.

## 3.3. Location of the restaurants

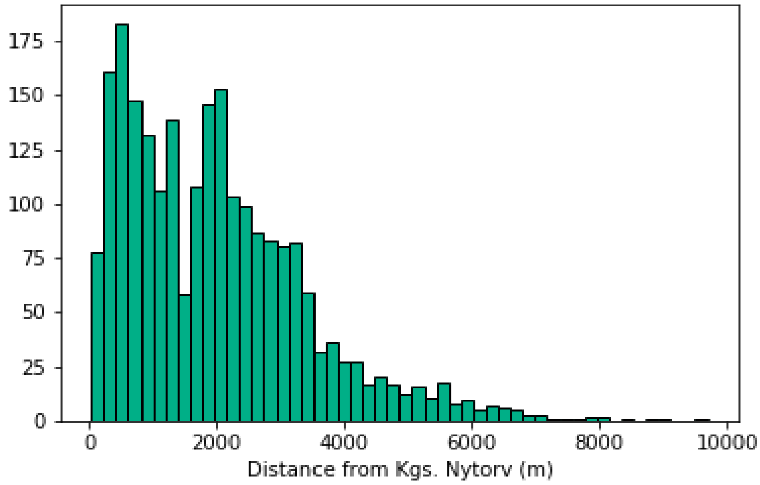
Out of the 2172 restaurants all of them have location data. This means that mapping all restaurants will actually show all restaurants in Copenhagen on TripAdvisor as seen below. As it is visible in figure 4, restaurants are spread out over the greater part of Copenhagen, but centered around the main streets of each neighborhood. On Østerbro the restaurants are placed on Østerbrogade, and likewise for Nørrebro on Nørrebrogade and Amager on Amagerbrogade. In the city center from the area around the main station and through to Kongens Have restaurants are closely clustered. In other words: the restaurants are competing with their next-door neighbour.

Figure 4 - Restaurants in Copenhagen



As seen in, 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. A few numbers of restaurants are located between 6000 and 8000 meters away for Kgs. Nytorv, but it is a small number.

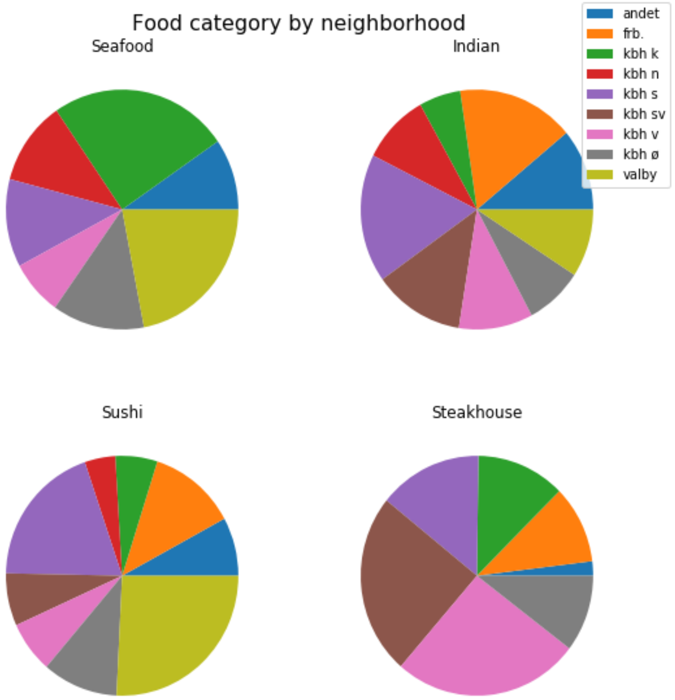
*Figure 5 - Density of distance to Kongens Nytorv*



### 3.4. Location of the kitchen types

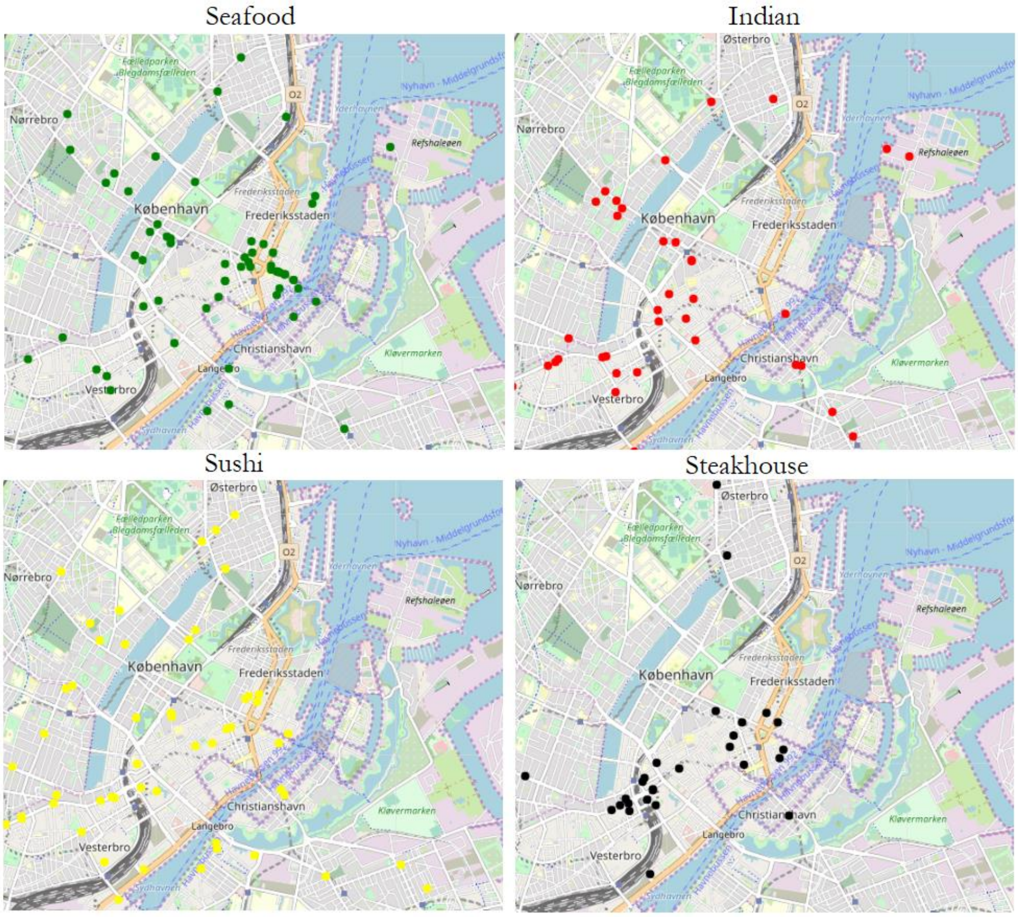
But are restaurants of different category placed differently across of Copenhagen? When looking at categories such as steakhouse, sushi, fish and Indian food, it is clear that different areas of Copenhagen attracts different restaurant types (Figure 6):

*Figure 6 - Food category by neighboorhood*



E.g Steakhouses are present across most of Copenhagen except for Nørrebro, where no Steakhouse is located. The steakhouses are centred around Kgs. Nytorv and the central station. More than half the steakhouse restaurants are located either in Copenhagen S, SV and V. Indian restaurants are not as present on Østerbro and in the city centre as on Vesterbro and Nørrebro. The Indian restaurants are also centred close to the central station and on Nørrebrogade. Seafood restaurants are not as present on Frederiksberg and on Nørrebro compared to the city centre. Sushi restaurants are spread out over all over Copenhagen, but about a quarter of the sushi restaurants are located in Valby. (Figure 7)

*Figure 7 - Kitchen types in Copenhagen*



See restaurant_seafood.html, restaurant_steakhouse.html mm for full maps.

From the maps above, we see evidence of agglomeration[9]. Meaning that restaurants cluster together, both in general and also when it comes to the different types of kitchens. The effect of agglomeration entails that when customers are seeking a specific type of food, they know exactly where to go to get that. We also see this in other types of retailers e.g the Jewish jewelers in New York or sales streets in Copenhagen. This agglomeration happens as restaurants want to compete for the same customers. When a new restaurant opens, potential customers are already passing by the restaurant if it places itself close to other restaurants. There are also scales of economic by being clustered, such as lower delivery costs of goods.

Looking at Figure 7, we see Steakhouses clustering right around Kgs. Nytorv and Copenhagen Central Station. Seafood is clustered around Nyhavn and Indian food hardly exist in the center of Copenhagen but clustering in groups of 2 and 3 spread out in Copenhagen. Lastly looking at Figure 6 we see Sushi places clustering in Valby and Copenhagen S, outside the maps above.

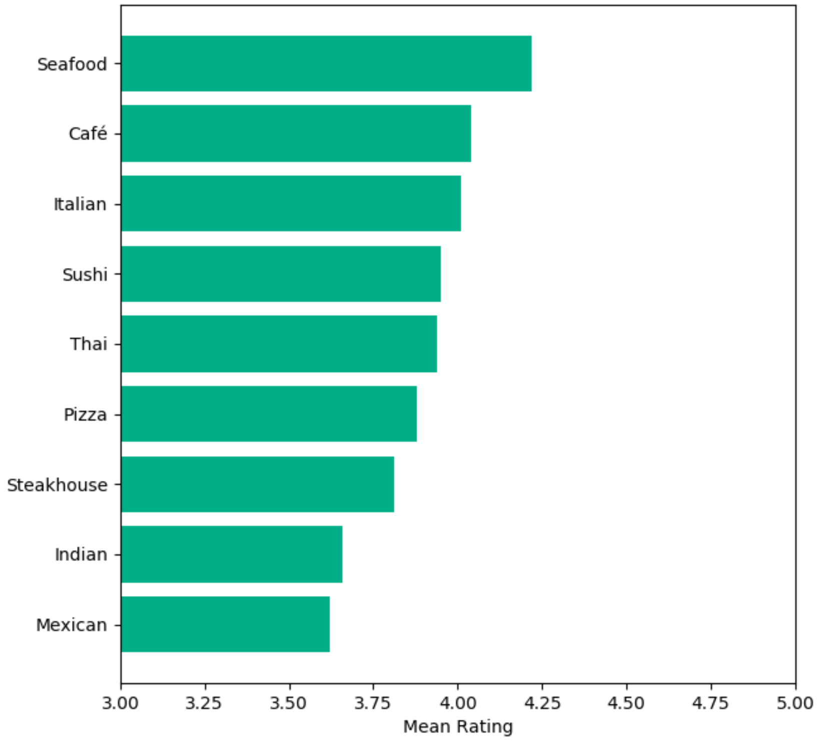
### 3.5. Restaurant ratings based on kitchen types

Are some types of kitchens rated better or worse than others? According to Figure 8 showing the mean rating of the restaurants belonging to the different kitchens then yes, there is a difference.

[9] Harvard Business School: The Logic of Agglomeration

The figure shows that restaurants serving Mexican food are being rated on a mean rating of 3.6 whereas the seafood restaurants are rated with the average rating of 4.2. A difference of 0.5 between the two categories.

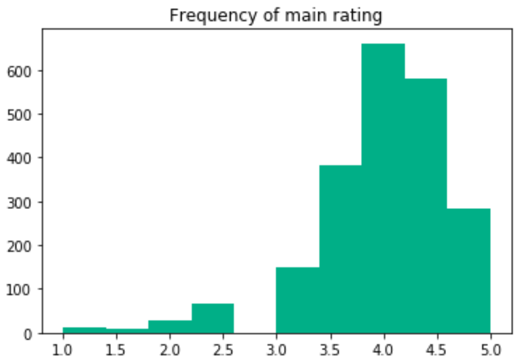
*Figure 8 - Mean restaurant ratings based on kitchen types*



3.6. Number of reviews and distribution of ratings

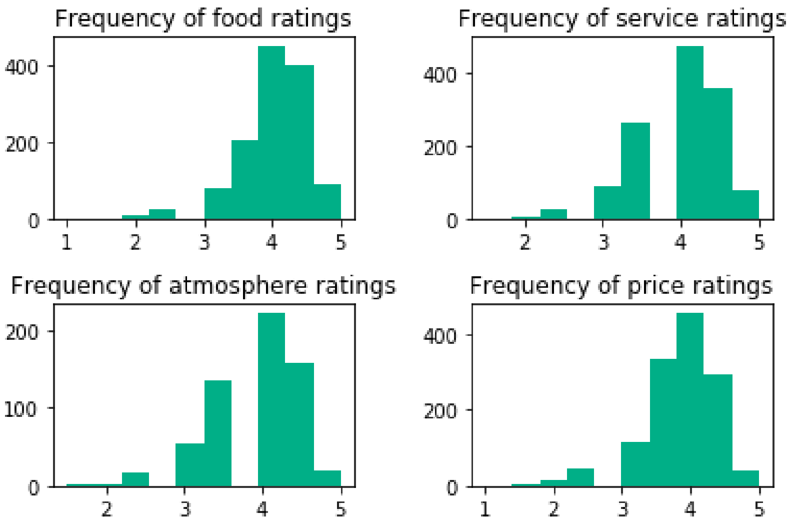
2170 out of 2172 restaurants have as a minimum one review, but the average is 144 reviews, the median number is 26 reviews and a standard deviation of 370. The maximum one restaurant has is 4551 reviews, which must be considered an outlier. Almost all restaurants have ranking data: out of 2172 a rank is available for 2170 restaurants. The remaining 150 restaurants have in general few or no reviews, why no ratings are available for generating a main rating and with that a rank. 2172 restaurants have a main rating. The main rating mean is 4.0 with 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. The mode of the main rating is also 4.0. As seen in Figure 9, most restaurants have a main rating between 3 and 5, centred around 4.

*Figure 9 - frequency of main rating*



1300 restaurants have ratings on good price, food, service, though only 612 have on atmosphere. All with standard deviations between 0.57 and 0.59. As seen in Figure 10 most restaurants have relatively high ratings between 3 and 5 for all categories. Only a few restaurants have lower ratings than 3.

*Figure 10 -frequency of specific ratings*



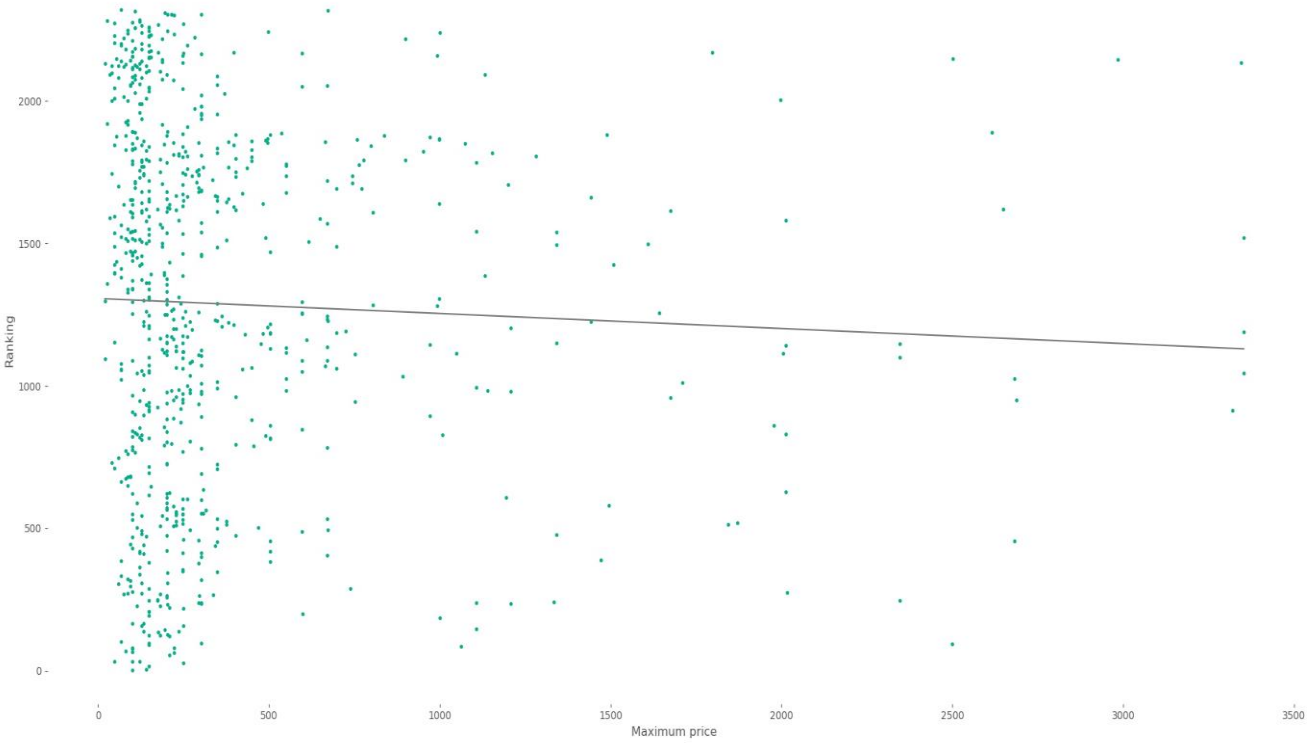
The restaurants in Copenhagen on TripAdvisor are in other words placed close to each other and rated very similarly overall on food, service, atmosphere and price. But how do the restaurants differ when separated by food category?

### 3.7.Ratings and price

Figure 11 shows a scatterplot of the maximum price of the restaurant's menu and the ranking on the list. We have limited the maximum price in this plot to 3500 DKK. We can see a slightly negative trend in the relationship, meaning that there is a tendency that more expensive restaurants are rated a little lower on the list.

However, we have only observations on maximum price for 762 restaurants (approximately one third of the sample) Therefore we choose to not use this measure in the further analysis.

*Figure 11 - Scatterplot of maximum price and ranking*

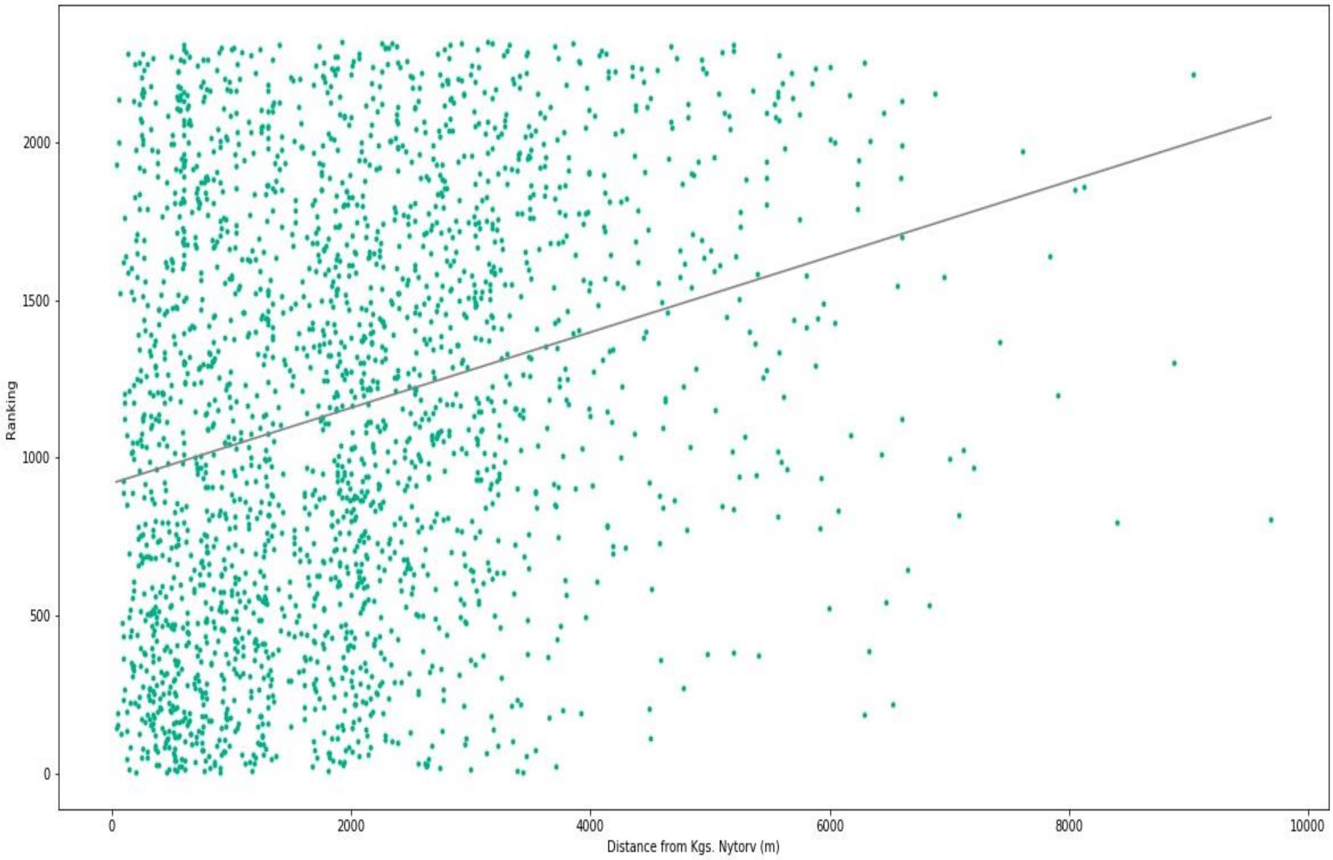


# 4. The OLS model

Figure 12 contains a scatterplot of the ranking on the list of restaurants as a function of the distance from the restaurant to Kongens Nytorv. By limiting the maximum distance from Kongens Nytorv to 10 km the scatterplot only shows restaurants actually located in Copenhagen. The fitted line on the scatterplot shows a positive relationship between distance to Kongens Nytorv and the ranking meaning that there could be a tendency that restaurants located further away from Kongens Nytorv are performing better.

As seen on the correlation graph below, there is a positive correlation between distance and ranking: the further away from Kgs. Nytorv the higher and better ranking.

*Figure 12 - Scatterplot of distance to Kongens Nytorv and Ranking*



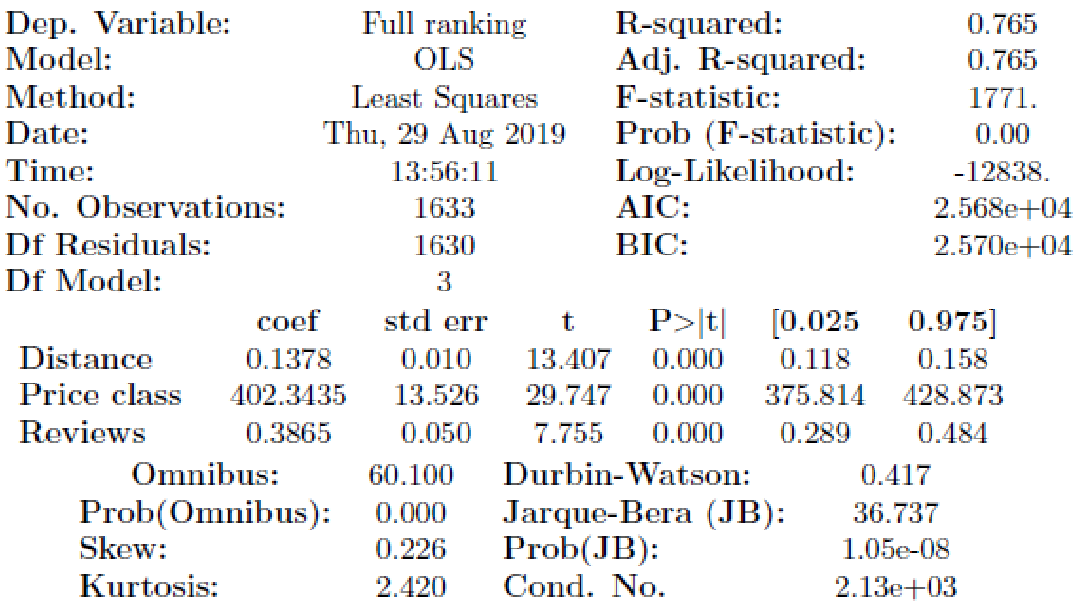
To be able to find out if there is a significant relationship between the distance to Kongens Nytorv and the rating of the restaurants we run a so called OLS model. The ordinary least squares (OLS) regression model minimizes the sum of the squared residuals.

The model we estimate can be written as:

$$Ranking_i = \beta_0 + \beta_{1,i} DistanceToKongensNytorv_i + \beta_{2,i} PriceClass_i + \varepsilon_i \text{ (Model 1)}$$

The results of the model are printed in Table 1 and shows that distance to Kongens Nytorv has a significant influence on the ranking of the restaurant. Also, the price class seems to have an impact on the ranking. However, the explanatory power of the model is very high and it seems counterintuitive that the distance to Kongens Nytorv and the price class should be able to explain almost 75 pct of the variation in the rankings of the restaurants. One explanation of this can be that the variables "Full ranking " and "Price-class" are constructed variables.
Full ranking is constructed directly from the ratings and we are not sure how the ratings are actually constructed, and there is a chance that the price class are a part of the main rating.

*Table 1 - OLS results of model 1*



```
Dep. Variable:          Full ranking        R-squared:              0.765
Model:                  OLS                 Adj. R-squared:         0.765
Method:                 Least Squares       F-statistic:            1771.
Date:                   Thu, 29 Aug 2019    Prob (F-statistic):     0.00
Time:                   13:56:11            Log-Likelihood:         -12838.
No. Observations:       1633                AIC:                    2.568e+04
Df Residuals:           1630                BIC:                    2.570e+04
Df Model:               3
                 coef       std err     t        P>|t|    [0.025     0.975]
 Distance        0.1378     0.010       13.407   0.000    0.118      0.158
 Price class     402.3435   13.526      29.747   0.000    375.814    428.873
 Reviews         0.3865     0.050       7.755    0.000    0.289      0.484
        Omnibus:            60.100   Durbin-Watson:        0.417
     Prob(Omnibus):         0.000    Jarque-Bera (JB):     36.737
     Skew:                  0.226    Prob(JB):             1.05e-08
     Kurtosis:              2.420    Cond. No.             2.13e+03
```

One solution could be to construct a new ranking based on the service level istad, to ensure that the ranking and the price class are not parts of the same measure. Therefore we estimate the model:

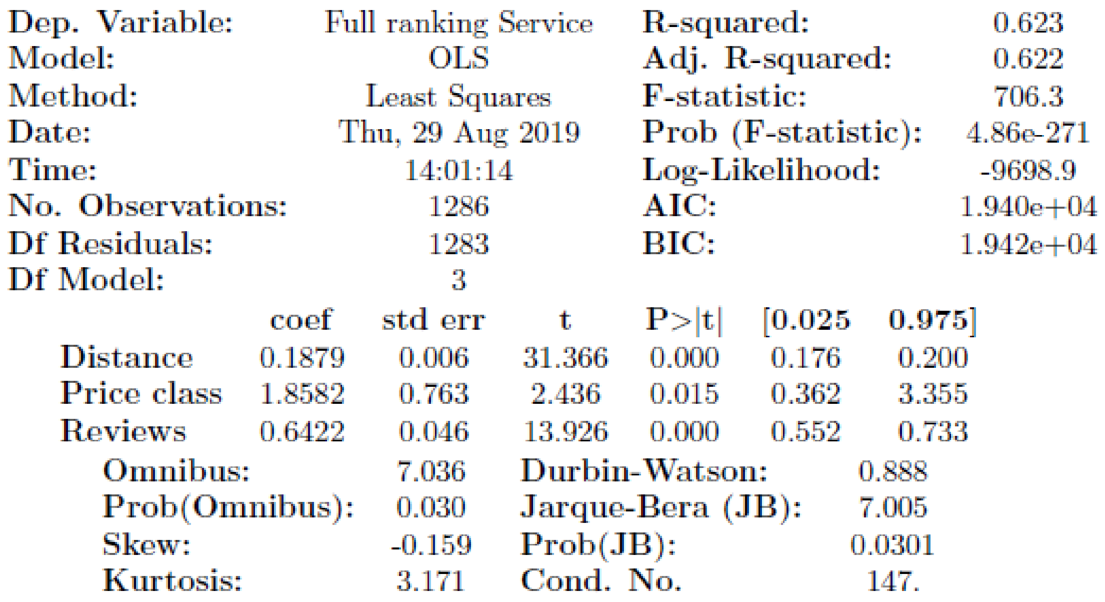
$$ServiceRanking_i = \beta_0 + \beta_{1,i}DistanceToKongensNytorv_i + \beta_{2,i}PriceClass_i + \varepsilon_i$$
(model 2)

The results are printed in Table 2. The regression on the ranking of the service level still shows significant results but almost 15 percent points smaller R^2.
The interpretation of the results is that a one-meter increase in the distance from Kongens Nytorv are approximately resulting in a 0.18 better ranking on average. If the price class increases with one the service-ranking is on average two places better.

An important thing to have in mind when working with regression models are the question of causality. We can in our model not be sure if a better ranking makes the customers view on the restaurant more positive from the beginning. It is well-known that people affect each other when it comes to consider the quality of eg. art or music, so maybe the good ratings are self-reinforcing mechanism. In addition, it is also plausible that good rankings could have affected the price-levels in the way that if a restaurant becomes very popular due to good reviews, the costumes would maybe be willing to pay a higher price and the restaurant can increase the prices without losing profit.

However, due to uncertainty on how our data are actually constructed, the causality and in addition the knowledge of the many factors we are not able to take into account, we will be wary of putting too much weight on the exact results but only conclude that there seems to be a positive relationship between the distance from the main touristic area of Copenhagen and the customers view on the restaurants and that there seems to be a positive relationship between price and quality in the restaurants of Copenhagen.

*Table 2 - OLS results of model 2*



| | | | |
|---|---|---|---|
| Dep. Variable: | Full ranking Service | R-squared: | 0.623 |
| Model: | OLS | Adj. R-squared: | 0.622 |
| Method: | Least Squares | F-statistic: | 706.3 |
| Date: | Thu, 29 Aug 2019 | Prob (F-statistic): | 4.86e-271 |
| Time: | 14:01:14 | Log-Likelihood: | -9698.9 |
| No. Observations: | 1286 | AIC: | 1.940e+04 |
| Df Residuals: | 1283 | BIC: | 1.942e+04 |
| Df Model: | 3 | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Distance | 0.1879 | 0.006 | 31.366 | 0.000 | 0.176 | 0.200 |
| Price class | 1.8582 | 0.763 | 2.436 | 0.015 | 0.362 | 3.355 |
| Reviews | 0.6422 | 0.046 | 13.926 | 0.000 | 0.552 | 0.733 |

| | | | |
|---|---|---|---|
| Omnibus: | 7.036 | Durbin-Watson: | 0.888 |
| Prob(Omnibus): | 0.030 | Jarque-Bera (JB): | 7.005 |
| Skew: | -0.159 | Prob(JB): | 0.0301 |
| Kurtosis: | 3.171 | Cond. No. | 147. |

## 5. Predicting new restaurant ratings:

A further developmental wish for this project would be a supervised machine learning model. The model in question would be a model that is able to predict what rating a newly opened restaurant would get, given its price class, type of food and distance to popular areas, like theatres or tourist attractions. The scope of the machine learning model would be limited to the available data: The model input would be the extracted data from TripAdvisor, though we are aware that restaurants success can be measured from other variables. The model would then in conclusion be a predictor of whether or not a newly opened restaurant will be successful. The model could be implemented in the following way:

So, our thought was to implement a ML model that could predict what types of restaurants, in which areas and what price range that would do well. Given that all this data is available on TripAdvisor, we would split our dataset into two sets, one training set and one testing set. The model would learn from the training set or also called estimate the model from this set. Using the training set, we try to teach the model, to be able to predict out-of-sample predictions, by changing the bias in the model. We then evaluate the model using our untouched test set. We would have to do different model estimators, where OLS, Ridge and Lasso would be good choices. After that we would calculate the RMSE for each model and compare which would predict the most accurate result.

# 6. Conclusion

Overall this paper has assessed the restaurant business in Copenhagen based on web scraped data from TripAdvisor pages from the 28th of august 2019. The paper has investigated the current state and the trends of the market. Furthermore, ranking as a measure of success has been examined. On TripAdvisor a total of 2320 restaurants in Copenhagen is online and available, of them 2172 have enough reviews for generating a ranking. The restaurants are located all over Copenhagen, spread over all neighbourhoods. In the same way restaurants of all rankings, high and low, are spread over Copenhagen. Certain types of restaurants are located in different areas: more seafood restaurants are located in the centre of Copenhagen and more steakhouses are located near the main station. This agglomeration trend is visible for some categories of food, but not for sushi restaurants on the same scale in the centre of Copenhagen.

The linear regression model shows a positive relationship between the restaurants ranking on the list and the distance to Kgs. Nytorv. However, precaution needs to be taken when interpreting the model estimations, because of the possible misspecifications. But a cautious conclusion of model results is that there is a tendency for lower ranked restaurants to be located closer to Kgs. Nytorv. Also, the price class and number of reviews have a positive effect on ratings, but the model suffers from the ratings and price level possibly being directly correlated. To test for the model to be robust to this we regressed on the ranking service ratings instead and found significant and positive coefficients.

## Web pages

TripAdvisor:

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

Main page in English: www.tripadvisor.com, 22.08.2019-30.08.2019

Main page in Danish: 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

The 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

The Guardian:

How TripAdvisor changed travel:
https://www.theguardian.com/news/2018/aug/17/how-tripadvisor-changed-travel,
29.08.2019

Berlignske:

Kendt cafékonge slår alarm om mange københavnske restauranter: »Nu begynder udskilningsløbet«:
https://www.berlingske.dk/virksomheder/kendt-cafekonge-slaar-alarm-om-mange-koebenhavnske-restauranter-nu 29.08.2019

Harvard Business School:

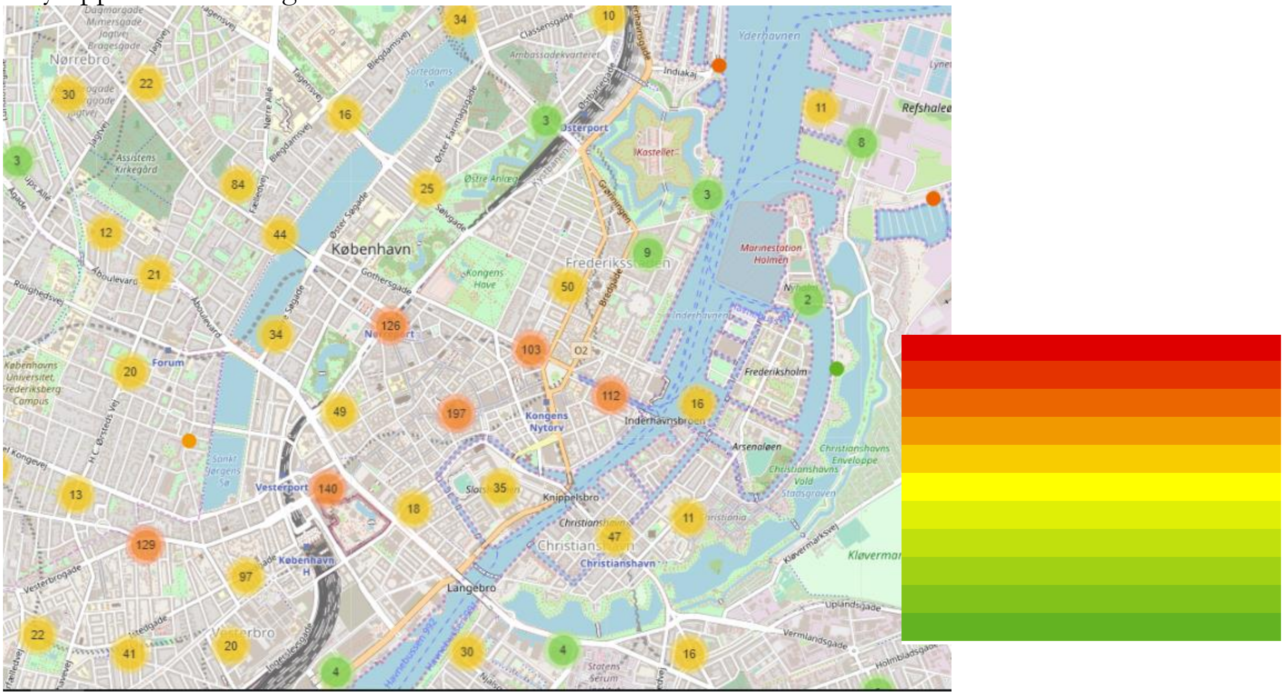
The Logic of Agglomeration
https://www.hbs.edu/faculty/Publication Files/16-037_eb512e96-28d6-4c02-a7a9-39b52db95b00.pdf?fbclid=IwAR32LNKs0XY2Pa_VlQSnDmQi7aA6icQ24xU4zkKgbT4h2Nq0cdDUv7l8PKQ

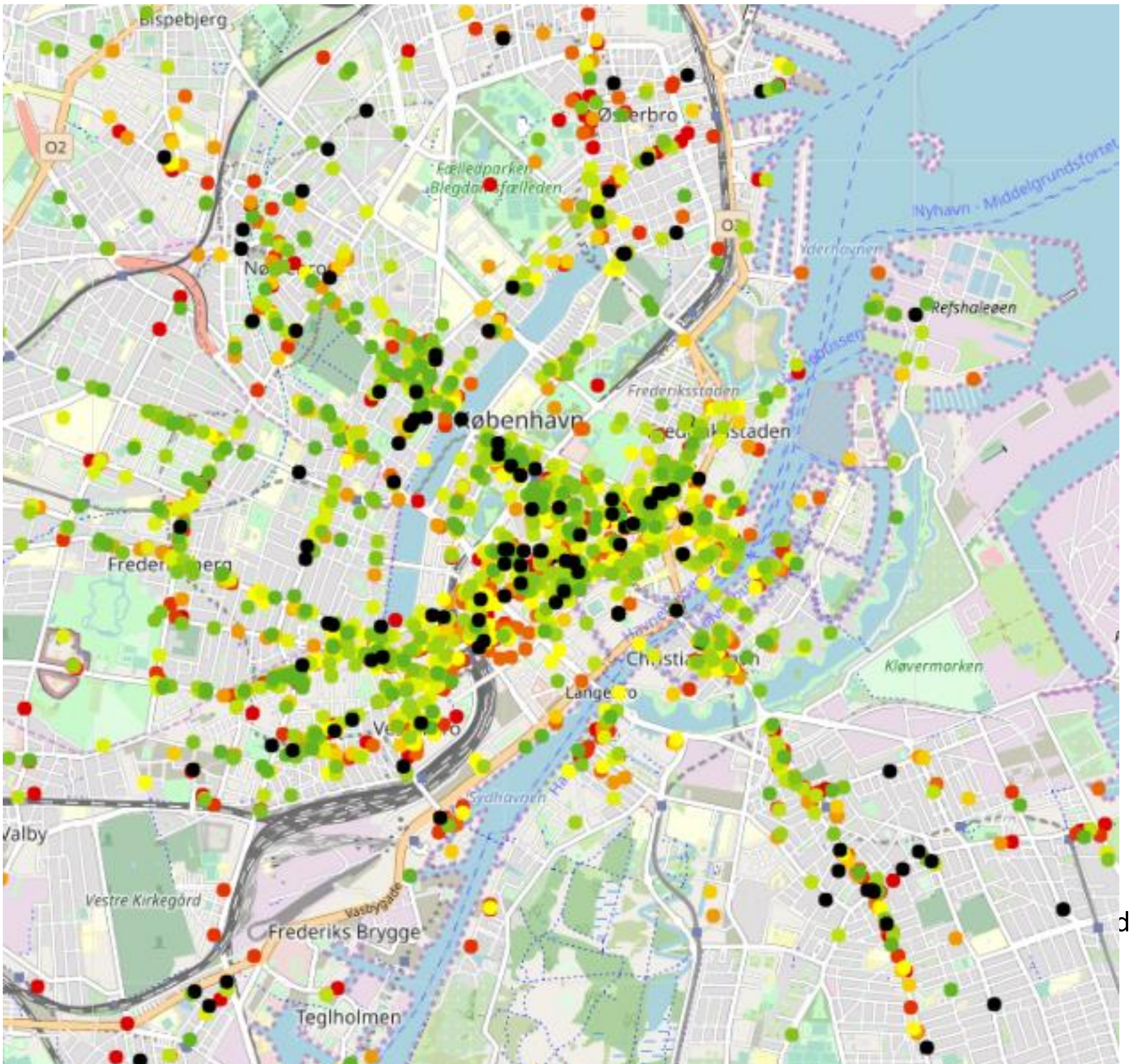
# 7. Appendix 1

| | *Main rating* | *Good price* | *Food* | *Service* | *Atmosphere* | *Number of reviews* | *Price class numeric* | *Full ranking* | *Distance from Kgs. Nytorv (m)* |
|---|---|---|---|---|---|---|---|---|---|
| ***count*** | 2.173 | 1.295 | 1.270 | 1.298 | 612 | 2.317 | - | - | 23.200 |
| ***mean*** | 4,0 | 3,9 | 4,0 | 4,0 | 3,9 | 14 | - | - | 4.913 |
| ***std*** | 7,1 | 0,6 | 0,6 | 0,6 | 0,6 | 37 | - | - | 727 |
| ***min*** | 1,0 | 1,0 | 1,0 | 1,5 | 1,5 | 10 | 1 | 1 | 36 |
| ***25%*** | 3,5 | 3,5 | 3,5 | 3,5 | 3,5 | 60 | 2 | 580 | 832 |
| ***50%*** | 4,0 | 4,0 | 4,0 | 4,0 | 4,0 | 260 | 2 | 1.160 | 1.854 |
| ***75%*** | 4,5 | 4,5 | 4,5 | 4,5 | 4,5 | 1.090 | 3 | 1.740 | 2.814 |
| ***max*** | 5,0 | 5,0 | 5,0 | 5,0 | 5,0 | 4.551 | 3 | 2.320 | 2.020.183 |

## 8. Appendix 2

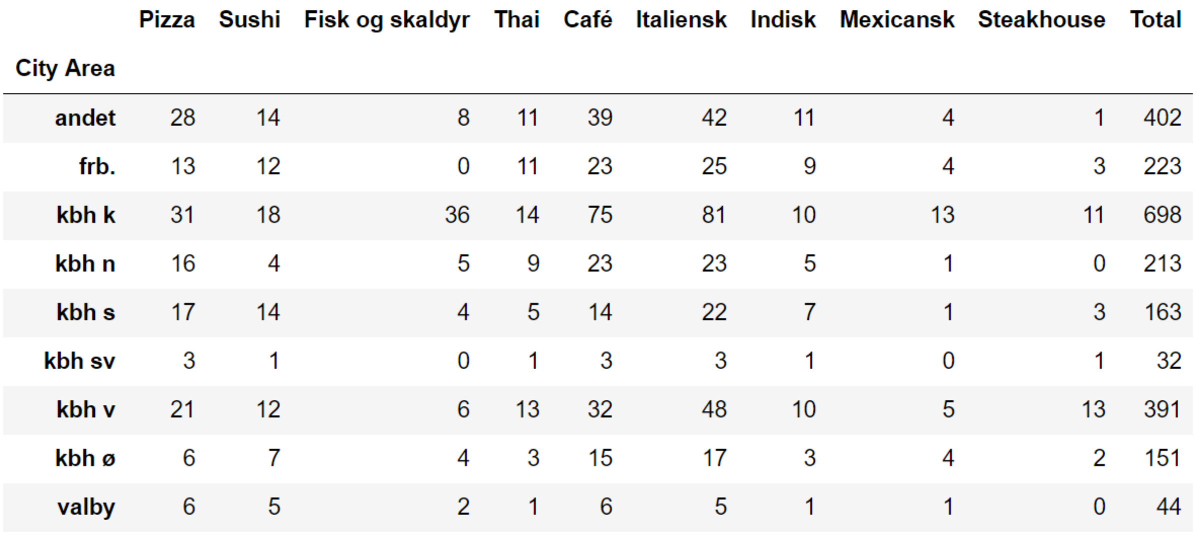
Colour scale for mapping ranking of restaurants. The scale covers from rank 2000 to 0 with 200 in each step colour group. First green colour from left is 2000-1800 and the last red colour at the very opposite the rating under 200. Black illustrates restaurants without data.





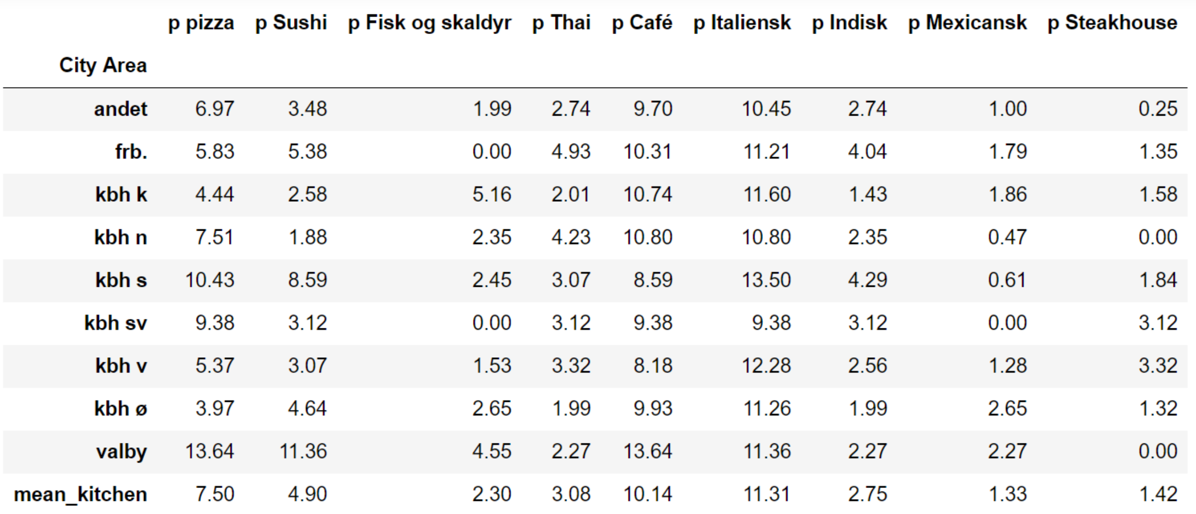
# 9. Appendix 3

Chart 1:



| | Pizza | Sushi | Fisk og skaldyr | Thai | Café | Italiensk | Indisk | Mexicansk | Steakhouse | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| City Area | | | | | | | | | | |
| andet | 28 | 14 | 8 | 11 | 39 | 42 | 11 | 4 | 1 | 402 |
| frb. | 13 | 12 | 0 | 11 | 23 | 25 | 9 | 4 | 3 | 223 |
| kbh k | 31 | 18 | 36 | 14 | 75 | 81 | 10 | 13 | 11 | 698 |
| kbh n | 16 | 4 | 5 | 9 | 23 | 23 | 5 | 1 | 0 | 213 |
| kbh s | 17 | 14 | 4 | 5 | 14 | 22 | 7 | 1 | 3 | 163 |
| kbh sv | 3 | 1 | 0 | 1 | 3 | 3 | 1 | 0 | 1 | 32 |
| kbh v | 21 | 12 | 6 | 13 | 32 | 48 | 10 | 5 | 13 | 391 |
| kbh ø | 6 | 7 | 4 | 3 | 15 | 17 | 3 | 4 | 2 | 151 |
| valby | 6 | 5 | 2 | 1 | 6 | 5 | 1 | 1 | 0 | 44 |

Chart 2:



| | p pizza | p Sushi | p Fisk og skaldyr | p Thai | p Café | p Italiensk | p Indisk | p Mexicansk | p Steakhouse |
|---|---|---|---|---|---|---|---|---|---|
| City Area | | | | | | | | | |
| andet | 6.97 | 3.48 | 1.99 | 2.74 | 9.70 | 10.45 | 2.74 | 1.00 | 0.25 |
| frb. | 5.83 | 5.38 | 0.00 | 4.93 | 10.31 | 11.21 | 4.04 | 1.79 | 1.35 |
| kbh k | 4.44 | 2.58 | 5.16 | 2.01 | 10.74 | 11.60 | 1.43 | 1.86 | 1.58 |
| kbh n | 7.51 | 1.88 | 2.35 | 4.23 | 10.80 | 10.80 | 2.35 | 0.47 | 0.00 |
| kbh s | 10.43 | 8.59 | 2.45 | 3.07 | 8.59 | 13.50 | 4.29 | 0.61 | 1.84 |
| kbh sv | 9.38 | 3.12 | 0.00 | 3.12 | 9.38 | 9.38 | 3.12 | 0.00 | 3.12 |
| kbh v | 5.37 | 3.07 | 1.53 | 3.32 | 8.18 | 12.28 | 2.56 | 1.28 | 3.32 |
| kbh ø | 3.97 | 4.64 | 2.65 | 1.99 | 9.93 | 11.26 | 1.99 | 2.65 | 1.32 |
| valby | 13.64 | 11.36 | 4.55 | 2.27 | 13.64 | 11.36 | 2.27 | 2.27 | 0.00 |
| mean_kitchen | 7.50 | 4.90 | 2.30 | 3.08 | 10.14 | 11.31 | 2.75 | 1.33 | 1.42 |