Capstone Project
– Neighbourhoods of Denmark

By Kenneth C. Kleissl
September 2019

1 INTRODUCTION

This project was completed as part of the IBM Data Science Professional Certificate and makes up the Applied Data Science Capstone Project "The Battle of Neighbourhoods".

In this project, location data from Foursquare must be leveraged to explore and compare neighbourhoods or cities in a geographical location and in the end come up with an answer to a made-up business-related problem.

1.1 THE BUSINESS PROBLEM

This section introduces the considered business problem and the target audience that would be interested in the project.

A newly started Danish restaurant franchise are experiencing great success with their first restaurant located in the city of Kongens Lyngby[1]. The franchise is so successful that they are now considering opening another restaurant. The question however is, in which other cities are they most likely to experience similar success? To answer such question, a comparison analysis of the various Danish cities/neighbourhoods are needed. The client wants to know which Danish cities resembles Kongens Lyngby the most and thereby gives them the best chances of repeating their initial success.

After thorough discussions with the client it was clarified that in this relation resembling means cities with a similar selection of local venues.

The scope of this neighbourhood analysis is for now limited to other Danish cities, as the franchise is still very young and because the client believe part of their success is related to the Danish nationality.

Carrying out this analysis is crucial for this young franchise as a poor initial decision could easily ruin the future of the franchise.

The target audience of this analysis are stakeholders of the existing restaurant and possible future owners/investors.

[1] Kongens Lyngby is with its population of approx. 55,500 one of the larger suburbs to Copenhagen (the capital of Denmark) and an important shopping destination among the northern suburbs [1].

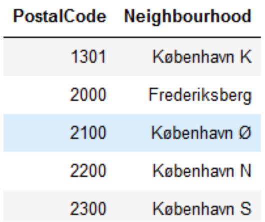
2 DATA DESCRIPTION

This section contains a description of the data, together with its sources, and how it will be used to solve the problem.

In this project the Foursquare location data are the primary source of information. However, to know which geographical locations to lookup a secondary dataset, containing all of Denmark's postal codes and region names, is used in combination with the GeoPy library leveraging OpenStreetMap data.

The Foursquare location data is retrieved by use of their free to use Places API with a personal developer account [2], while the Danish postal codes are downloaded from Edemann.dk [3].

A snippet of the dataset containing the Danish postal codes and region names are shown below.

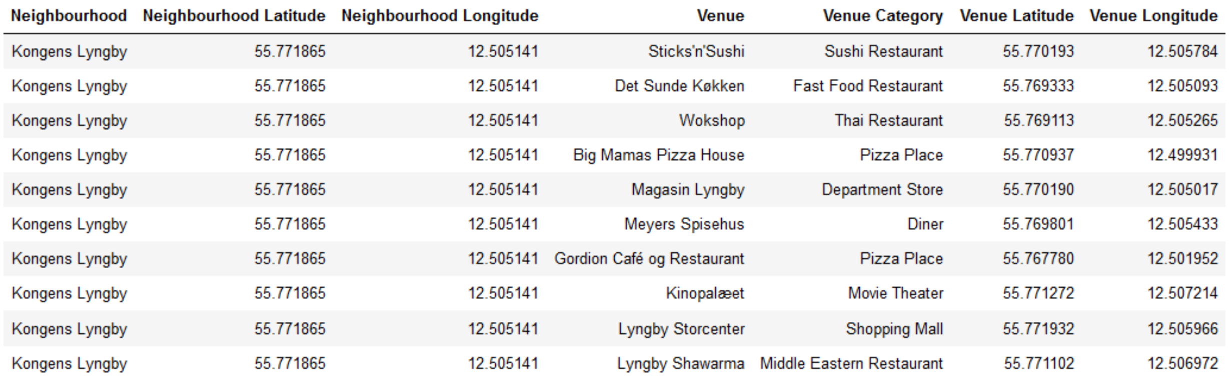


PostalCode	Neighbourhood
1301	København K
2000	Frederiksberg
2100	København Ø
2200	København N
2300	København S

Figure 1 A few samples from the dataset containing the Danish postal codes and region names.

The neighbourhood/region names are being used to look up their respective coordinates using the GeoPy library. Finally, the neighbourhood coordinates are used to look up all the local venues and their respective categories using the Foursquare Places API.

A snippet of the resulting dataset containing the neighbourhood name and location together with the venue name, category and location are shown below.



Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category	Venue Latitude	Venue Longitude
Kongens Lyngby	55.771865	12.505141	Sticks'n'Sushi	Sushi Restaurant	55.770193	12.505784
Kongens Lyngby	55.771865	12.505141	Det Sunde Køkken	Fast Food Restaurant	55.769333	12.505093
Kongens Lyngby	55.771865	12.505141	Wokshop	Thai Restaurant	55.769113	12.505265
Kongens Lyngby	55.771865	12.505141	Big Mamas Pizza House	Pizza Place	55.770937	12.499931
Kongens Lyngby	55.771865	12.505141	Magasin Lyngby	Department Store	55.770190	12.505017
Kongens Lyngby	55.771865	12.505141	Meyers Spisehus	Diner	55.769801	12.505433
Kongens Lyngby	55.771865	12.505141	Gordion Café og Restaurant	Pizza Place	55.767780	12.501952
Kongens Lyngby	55.771865	12.505141	Kinopalæet	Movie Theater	55.771272	12.507214
Kongens Lyngby	55.771865	12.505141	Lyngby Storcenter	Shopping Mall	55.771932	12.505966
Kongens Lyngby	55.771865	12.505141	Lyngby Shawarma	Middle Eastern Restaurant	55.771102	12.506972

Figure 2 Example of the venue data obtained from Foursquare.

3 METHODOLOGY

This section discusses and describes the data cleaning, exploratory data analysis carried out, feature selection and the chose of machine learning algorithm.

3.1 Data Cleaning

The postal code data scraped from html was checked for any possible null values and inconsistent data types.

The location data obtained via GeoPy has also been checked. For unknown reasons an empty response was sometimes received. This was solved by re-requesting invalid responses. During the exploratory data analysis, it was also noted that for similar named neighbourhoods, like Copenhagen E and NE, GeoPy sometime return identical geo locations coordinates. This was solved by merging these incidences into one neighbourhood.

3.2 Exploratory Data Analysis

All the geo location coordinates, both for neighbourhoods and venues, are explored by visualizations on a map of Denmark using the Python Folium library. This allowed for easy evaluation of if all parts of Denmark were represented (see also Figure 5) and if all the obtained venues in a certain neighbourhood looked reasonable with within the specified radius (see Figure 3).

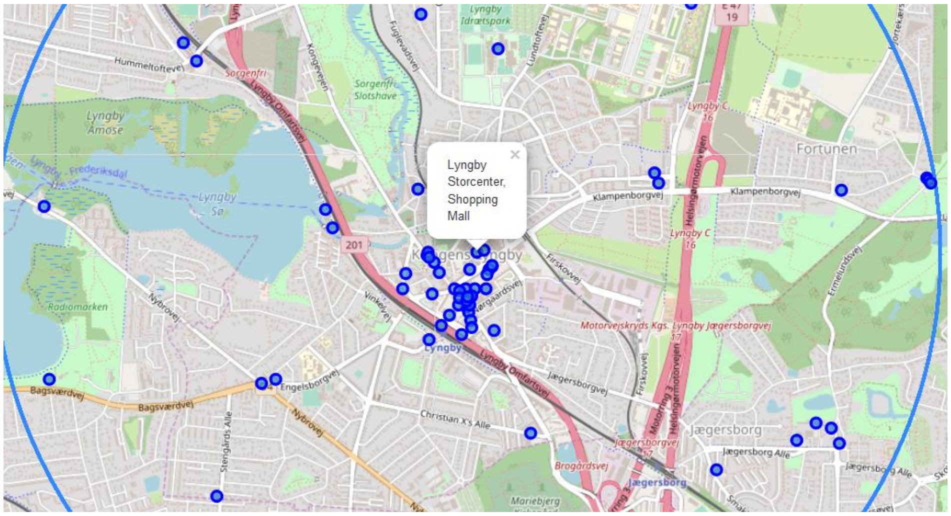


Figure 3 Map visualisation of venues in Kongens Lyngby. The large circle reflect the radius within which venues was requested.

3.3 Feature Selection

During the explorative data analysis, it was noted that several neighbourhoods had an extreme low venue count within the assigned radius of 2500 meters. Constructing a venue profile for these locations on such sparse data is not considered reliable. In the further analysis only locations with more than 10 venues are consider.

In Figure 4 below, the Python Seaborn library is used to visualise the density of venues across the 129 considered neighbourhoods with enough venues for further analysis. Note how the number of venues is limited to 100 venues per neighbourhood via the Foursquare request.

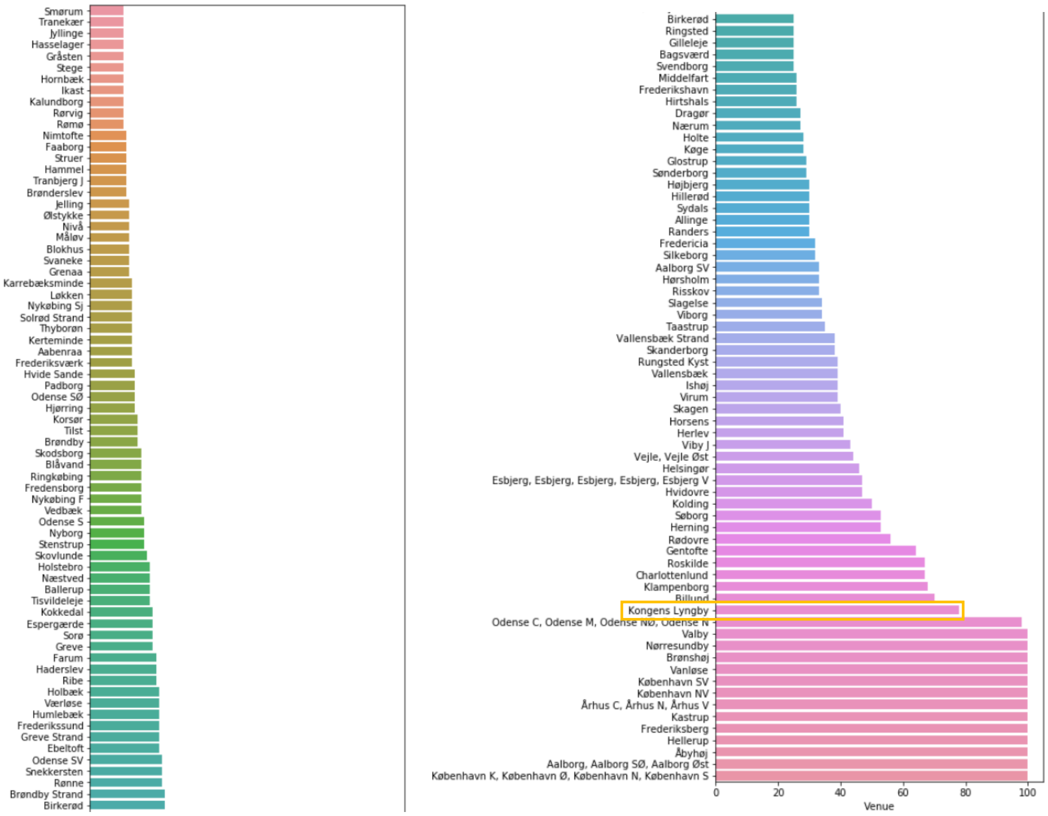


Figure 4 Horizontal bar plot visualization of venue density.

Note that we're looking for neighbourhoods similar to Kongens Lyngby that has relative many venues (approx. 80). Not considering neighbourhoods with very few venues is therefore not expected to rule out any neighbourhoods similar to Kongens Lyngby.

This filtering of the neighbourhoods with enough venues are also visualised in Figure 5, where all the green dots represent the Danish neighbourhoods, while those highligthed with a lightblue circle had enough venues for further analysis. The filtered out neighbourhoods are noted to all be located far our into the countryside, why the low number of venues appear realistic.

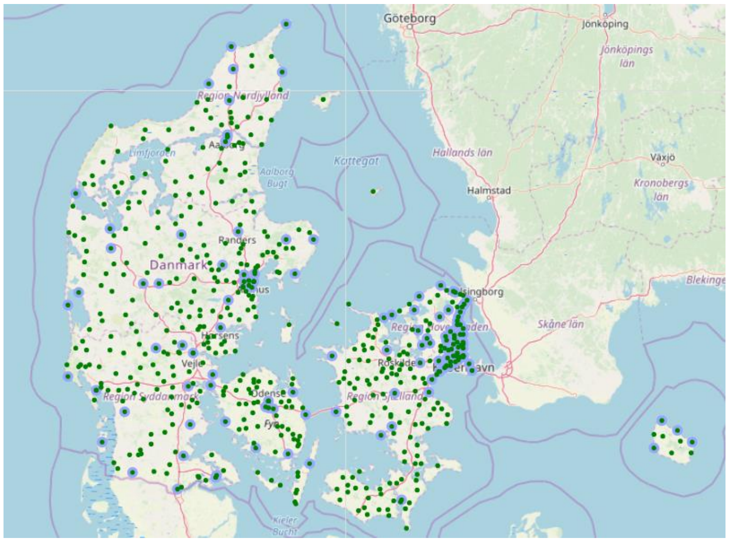
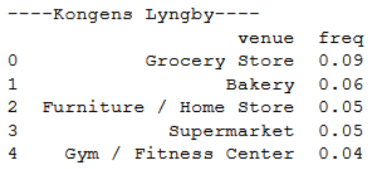


Figure 5 Map visualisation of all the Danish neighbourhoods. The neighbourhoods with enough Foursquare data for further analysis are highlighted with light blue circles.

The venue categories obtained from Foursquare are now used to build a venue frequency profile for each of the neighbourhoods. This is done by evaluating the frequency of each of the 279 unique venue categories for each of the neighbourhoods. As this is also normalising the data, no further normalisation is nessesary before model training.

The frequency profile of the Kongens Lyngby neighbourhood are shown in Figure 6. From this can be seen that 9% of the venues are grocery stores and 6% bakeries etc.



```
----Kongens Lyngby----
                  venue  freq
0         Grocery Store  0.09
1                Bakery  0.06
2  Furniture / Home Store  0.05
3           Supermarket  0.05
4    Gym / Fitness Center  0.04
```

Figure 6 Venue frequency profile for Kongens Lyngby. Note that only the top five frequencies are shown.

For the following analysis two neighbourhoods are considered similar if the venue frequency profile are similar.

3.4 Machine Learning Algorithm

This is a classic example of unlabelled data and therefore in the category of unsupervised machine learning.

To determine neighbours that are similar to Kongens Lyngby (see business problem), a cluster analysis is very suited. In this case the neighbourhoods are clustered using the K-Means algorithm, the most commonly applied method for clustering analysis.

As the model requires the number of clusters ‘K’ to be given beforehand, this parameter can be optimised for by use of the elbow method. How well a clustering performs can be evaluated by the total intra-cluster variation or total within-cluster sum of square (WSS), both measures describing how compact the clusters are (smaller is better). As an increasing number of clusters in the end always go towards more compact clusters, the Elbow method states that the ideal number of clusters ‘K’ should be taken as when significant diminishing returns are reached i.e. when the WSS-K curve bends.

The WSS over number of clusters ‘K’ curve for this specific model are shown in Figure 7. No pronounced bends can be observed but something in the order of five clusters appear recommendable here.

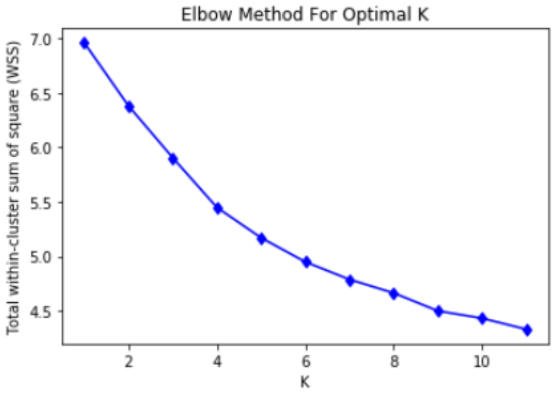


Figure 7 WSS as a function of the number of clusters 'K'.

While the K-Means algorithm are quite efficient it does not garatee finding the global optimum. To overcome this a certain number of ‘restarts’, with randomised starting states, are needed. Here the number of restarts are set based on when a reasonable level of reproduceability can be obtained.

Once the K-Means model have been trained with the optimal number of clusters ‘K’, the cluster containing Kongens Lyngby and its similar neighbourhoods can be studied in more depth.

4 Results

This section discusses the results obtained by the above described modelling approach.

With the ideal number of clusters K=5 (see Section 3.4) a clustering as illustrated in Figure 8 is found. While this result shows which of the Danish neighbourhoods are more similar to Kongens Lyngby than others, the number of similar neighbourhoods are found too large to help answering the stated buisness problem.

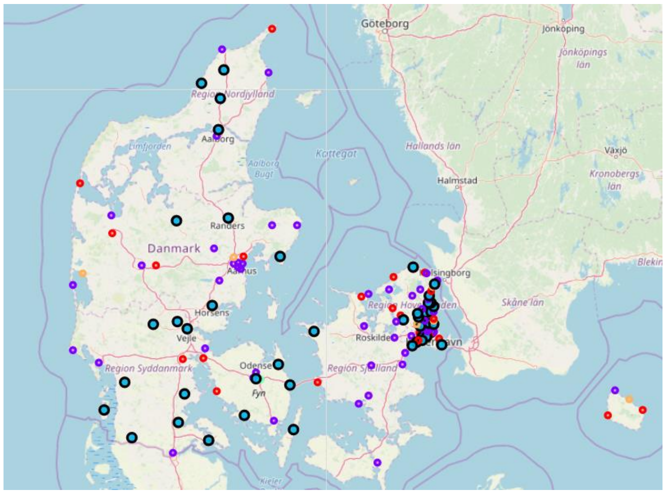


Figure 8 Map visualisation of the of neigbourhood clustering with K = 5. The cluster containing Kongens Lyngby is highligthed.

To reduce the cluster size to a handfull of neighbourhoods the number of clusters had to be increased significantly (K = 90). The resulting clustering are shown in Figure 9, where there's zoomed in on the cluster containing Kongens Lyngby.

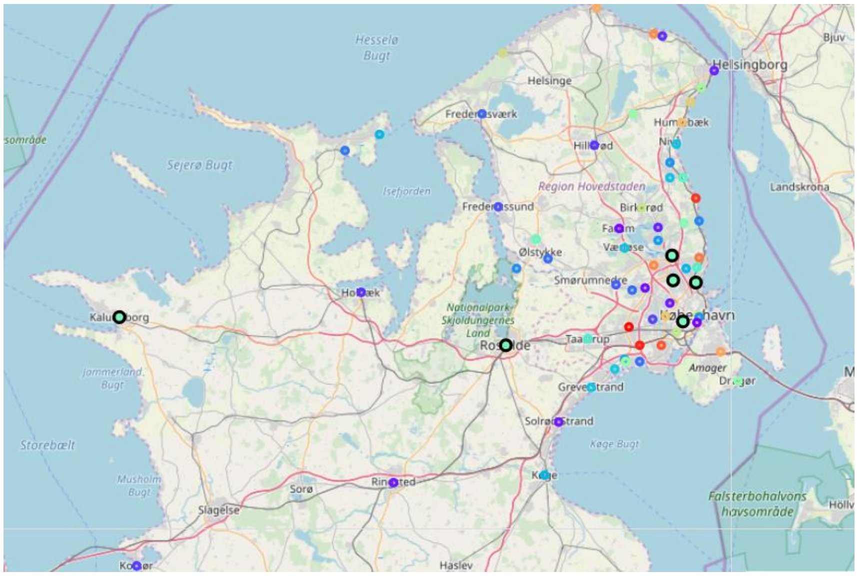
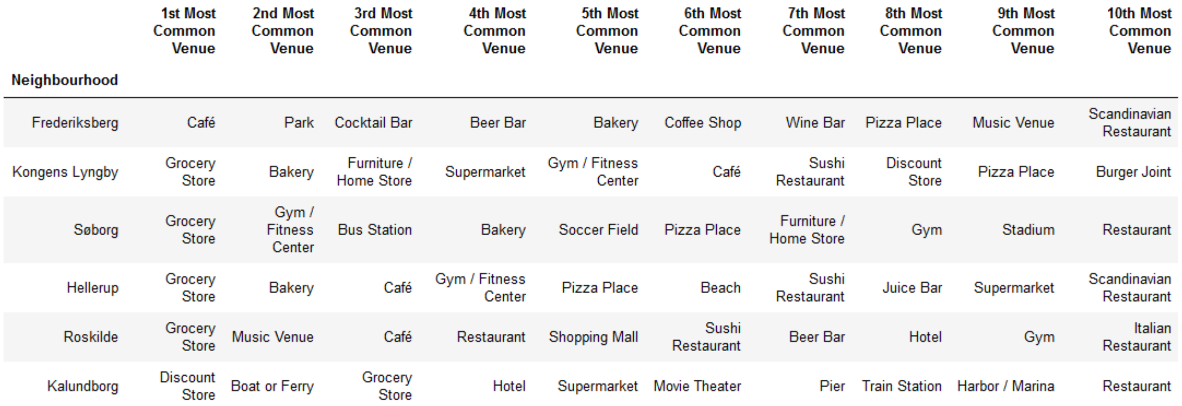


Figure 9 Map visualisation of the of neighbourhood clustering with K = 90. There's zoomed in on the cluster containing Kongens Lyngby.

Finally the Top 10 most common venues for each of the neighbourhoods similar to Kongens Lyngby are compared in Figure 10.



Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Frederiksberg	Café	Park	Cocktail Bar	Beer Bar	Bakery	Coffee Shop	Wine Bar	Pizza Place	Music Venue	Scandinavian Restaurant
Kongens Lyngby	Grocery Store	Bakery	Furniture / Home Store	Supermarket	Gym / Fitness Center	Café	Sushi Restaurant	Discount Store	Pizza Place	Burger Joint
Søborg	Grocery Store	Gym / Fitness Center	Bus Station	Bakery	Soccer Field	Pizza Place	Furniture / Home Store	Gym	Stadium	Restaurant
Hellerup	Grocery Store	Bakery	Café	Gym / Fitness Center	Pizza Place	Beach	Sushi Restaurant	Juice Bar	Supermarket	Scandinavian Restaurant
Roskilde	Grocery Store	Music Venue	Café	Restaurant	Shopping Mall	Sushi Restaurant	Beer Bar	Hotel	Gym	Italian Restaurant
Kalundborg	Discount Store	Boat or Ferry	Grocery Store	Hotel	Supermarket	Movie Theater	Pier	Train Station	Harbor / Marina	Restaurant

Figure 10 Top 10 most common venues for each of the neighbourhoods similar to Kongens Lyngby.

5 Discussion

This section discuss the observations and recommendations that can be made based on the results.

First it is observed that the neighbourhoods most similar to Kongens Lyngby are all located on Zealand[2], i.e. a much smaller geographical region than all of Denmark. The geographical distance appear to somehow play a role.

While almost all of the neighbourhoods in the considered cluster belongs to the high venue density part of Figure 4, Kalundborg only just passed the filtering criteria (in Figure 11 the high density part is reproduced with corresponding highligths). Despite a good frequency profile similarity, the large difference in venue density (not included in the modelling approach) indicate that Kalundborg might not be the best choice.

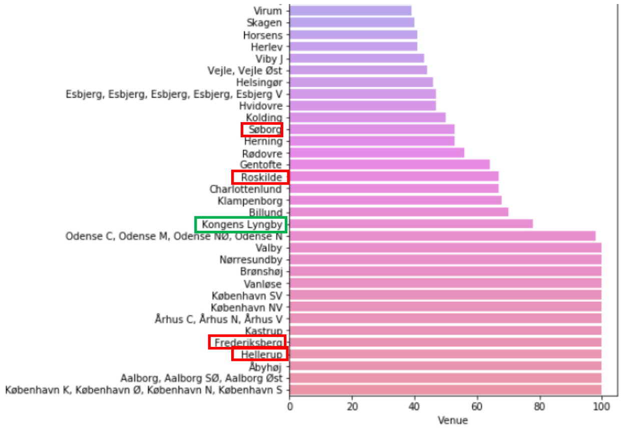
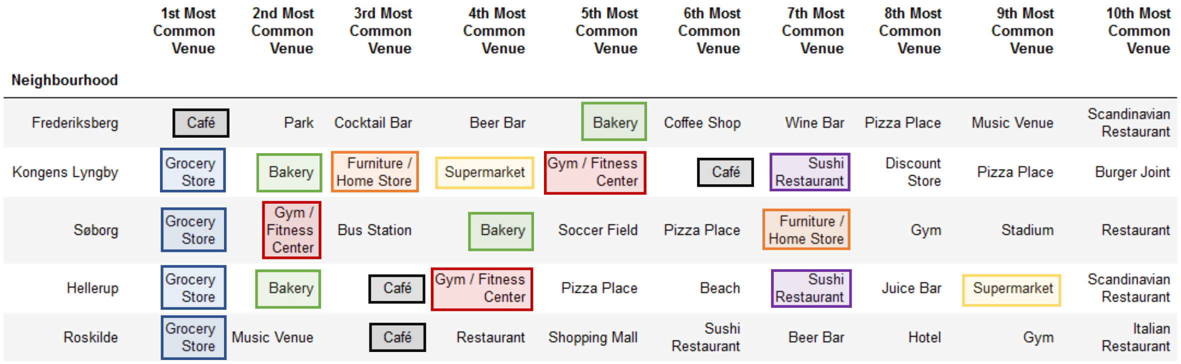


Figure 11 Snippet of the venue density plot from Figure 4 with the neighborhoods of interest highlighted.

Beside similar general venue density and venue frequency profile, one could consider giving increased weights to the most common venues in Kongens Lyngby. A comparisoon of how many of Kongens Lyngby most common vendues are present in the similar neighbourhoods' Top 10 are shown in Figure 12.

[2] The largest and most populous island of Denmark including the capital Copenhagen [4].



Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Frederiksberg	Café	Park	Cocktail Bar	Beer Bar	Bakery	Coffee Shop	Wine Bar	Pizza Place	Music Venue	Scandinavian Restaurant
Kongens Lyngby	Grocery Store	Bakery	Furniture / Home Store	Supermarket	Gym / Fitness Center	Café	Sushi Restaurant	Discount Store	Pizza Place	Burger Joint
Søborg	Grocery Store	Gym / Fitness Center	Bus Station	Bakery	Soccer Field	Pizza Place	Furniture / Home Store	Gym	Stadium	Restaurant
Hellerup	Grocery Store	Bakery	Café	Gym / Fitness Center	Pizza Place	Beach	Sushi Restaurant	Juice Bar	Supermarket	Scandinavian Restaurant
Roskilde	Grocery Store	Music Venue	Café	Restaurant	Shopping Mall	Sushi Restaurant	Beer Bar	Hotel	Gym	Italian Restaurant

Figure 12 Comparison of Kongens Lyngby's most common venues with the remaining similar neighbourhoods in consideration.

Among the neighbourhoods sharing cluster with Kongens Lyngby, Hellerup generally appear as the neighbourhood the most similar to Kongens Lyngby.

6 Conclusion

Based on the analysis and findings of this report, Hellerup have been identified as the neighbourhood most similar to Kongens Lyngby. It is therefore recommended that the newly established franchise investigates if Hellerup could be the location of their second restaurant, as it is presumed that this location will give them the best chances of repeating their initial success.

6.1 Future Improvements

One could improve the predictions by expanding the model with other types of relevant data such as e.g. average regional income.

7 References

[1] Wiki page https://da.wikipedia.org/wiki/Kongens_Lyngby

[2] Foursquare Developer site https://developer.foursquare.com

[3] Edemann site https://edemann.dk/liste-danske-postnumre-og-byer/

[4] Wiki page https://en.wikipedia.org/wiki/Zealand