A new restaurant in Copenhagen

In which region would a new restaurant most likely survive?

1. Introduction

1.1 Background

A client wants to open a new restaurant in big Copenhagen area in Denmark. There are 51 regions in this area based on the postal codes. Before the client starts to think about the costs in different areas, he would like to firstly know which regions are more suitable for a new restaurant considering the competitions and economic conditions in a specific region. There are too many things that could have an impact on the survival of a restaurant. For example, local policy, tax rate, rental, transport, public infrastructure, shops, schools, etc. However, with the help of Foursquare data, we could easily get a dataset of surrounding venues which basically covers all kinds of assets around a region center.

1.2 Problem

The relations of all the venues around a region may not be explicit, but the structure of these venues could largely reflect the economy conditions of the region. How can we utilize all these venue data to get a view of how similar, from economic conditions point of view, these different regions can be? And based on the conclusion, can we decide whether or not should we open a new restaurant? Or in which regions should we suggest the client to open a restaurant? In another word, would a new restaurant generally survive the competition in a specific region?

1.3 Interest

Based on the venue dataset, we could group the 51 regions into clusters with similar economy conditions. Assume that the economy is stable, and all existing restaurants in each region are running in a good condition, we could then reasonably tell that a region with relatively less restaurants (compared with another region in the same cluster) could have more new restaurants. For example, assume that region A and B are in the same cluster, meaning these two regions have similar economic conditions (similar venue structures). If the restaurants in Region A takes 5% of all venues, and the number for region B is 2 %, we can say that the economy or competition in region B would allow more restaurants, as region A actually has proved that restaurants as 5% of all venues could survive in this region cluster.

There can be two phases to get the final conclusion. Phase 1: group the 51 regions into clusters and investigate which ones to choose for further selection. Phase 2: analyze each cluster and find the regions which is more suitable for opening new restaurant.

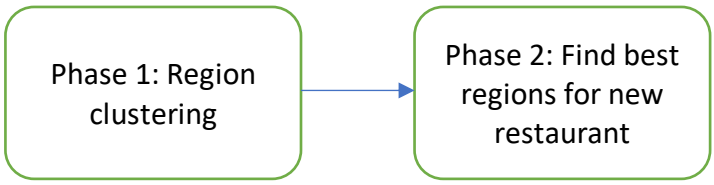


Chart1. phase chart

2. Data

2.1 Location data acquisition

The postal code data can be easily found from Wikipedia, together with the region name, the coordinates of the region centers can be fetched using Google map geocoding API. The python library *geopy.geocoder* can also help to get the coordinates, but some addresses are missing in this case, thus geocoding API is used here.

2.2 Venue data acquisition

As the coordinates data for all 51 regions in Copenhagen is available, the Foursquare API can be used in order to get the surrounding venues, with a radius of 500 meters and venue limits of 200. The venue categories are also extracted and stored for further analysis. The venue dataset should essentially contain the following fields: *region name, postal code, region coordinate, venue name, venue category, etc*.

	Region	Postcode	Region Latitude	Region Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Copenhagen K	1000	55.684199	12.579543	Kongens Have	55.684361	12.580099	Park
1	Copenhagen K	1000	55.684199	12.579543	Jazzfestival Kgs Have	55.683420	12.579783	Performing Arts Venue
2	Copenhagen K	1000	55.684199	12.579543	Møller & Mammen	55.682335	12.582285	Hardware Store
3	Copenhagen K	1000	55.684199	12.579543	ALDI	55.682531	12.581922	Discount Store
4	Copenhagen K	1000	55.684199	12.579543	Netto	55.682102	12.576530	Supermarket

Table 1, an example of the venue information

2.3 Feature selection

As described in the introduction part, there can be two phases to solve the problem. There could be different features that need consideration in these two phazes.

For phase 1, the venue structures in all regions need to be compared with one another, and the rate of occurrence of each venue category in each region could be used as the feature to compare, in order to calculate the similarity of the regions. For phase 2, as the purpose is to analyze and find the best regions for new restaurant, the venues with category of restaurant or food related should be all considered. The occurrence rate of the restaurants in each region of the same cluster need to be calculated and used as the feature for comparison.

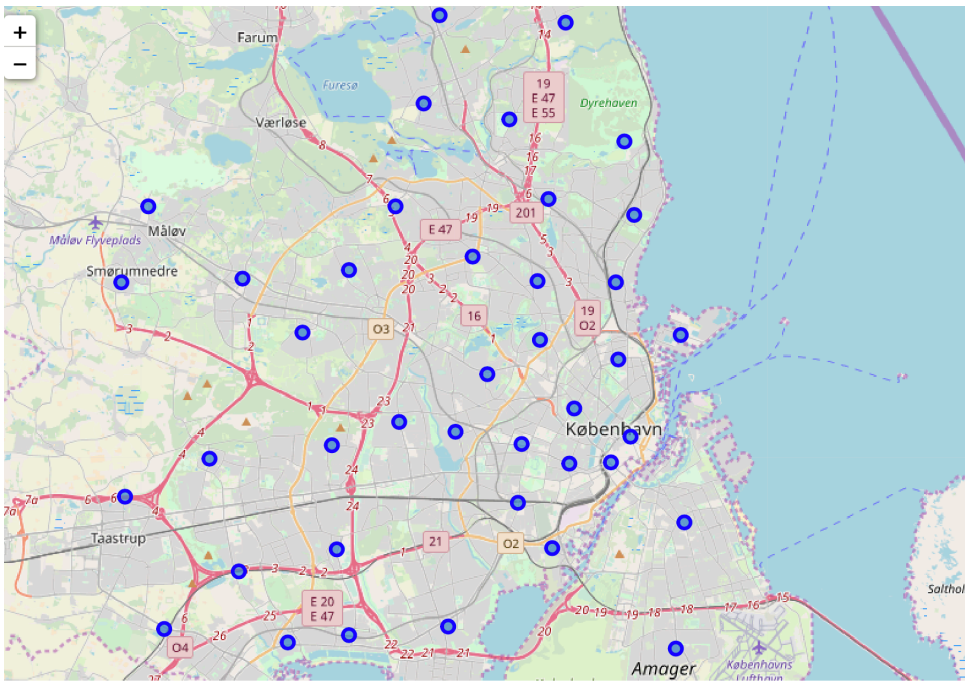
3. Methodology

3.1 Coordinate Data Analysis

A basic region table is created to contain the coordinates of all 51 regions in Copenhagen area (see table 1). The postal codes are extracted from Wikipedia, and the coordinates from Google map geocoding API.

	Postal Code	Area	latitude	longitude	address
0	1000	Copenhagen K	55.684199	12.579543	København K, København, Denmark
1	1500	Copenhagen V	55.676097	12.568337	Copenhagen, Denmark
2	1800	Frederiksberg C	55.675666	12.545006	Frederiksberg C, Frederiksberg, Denmark
3	2000	Frederiksberg	55.681845	12.517944	2000 Frederiksberg, Denmark
4	2100	Copenhagen Ø	55.708533	12.572776	2100 København Ø, Denmark

Table 2. Copenhagen region center coordinates



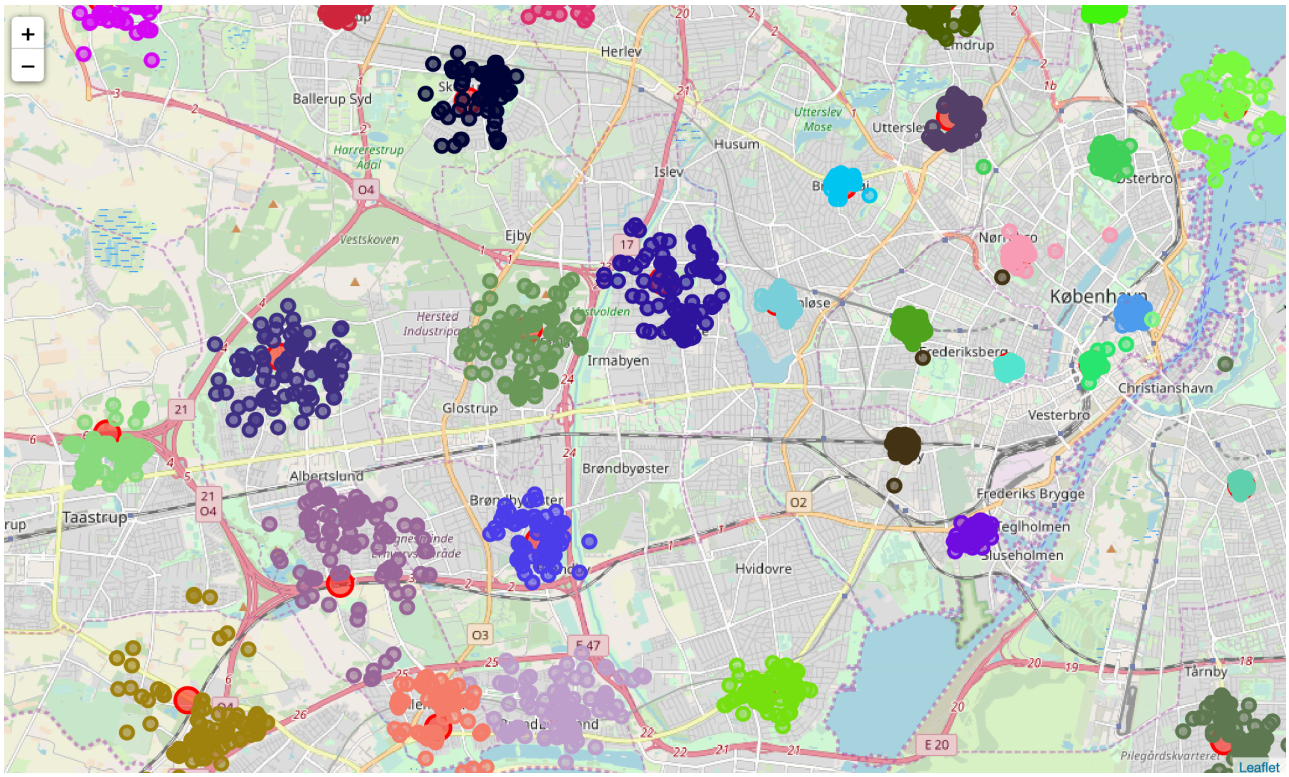
Pic1. Region center distribution

In order to validate the correctness of the coordinates data, a map is drawn with all region centers displayed in the map (pic 1). From the region center distribution, we can see that the region points are basically evenly spaced on the map, thus we can make a further analysis with this coordinate dataset.

3.2 Venue Data Analysis

The venue data can be fetched using Foursquare API, with the region center coordinate as input. To get a list of surrounding venues around a specific point on the map, two other parameters also need to be specified: the data point limit and the radius.

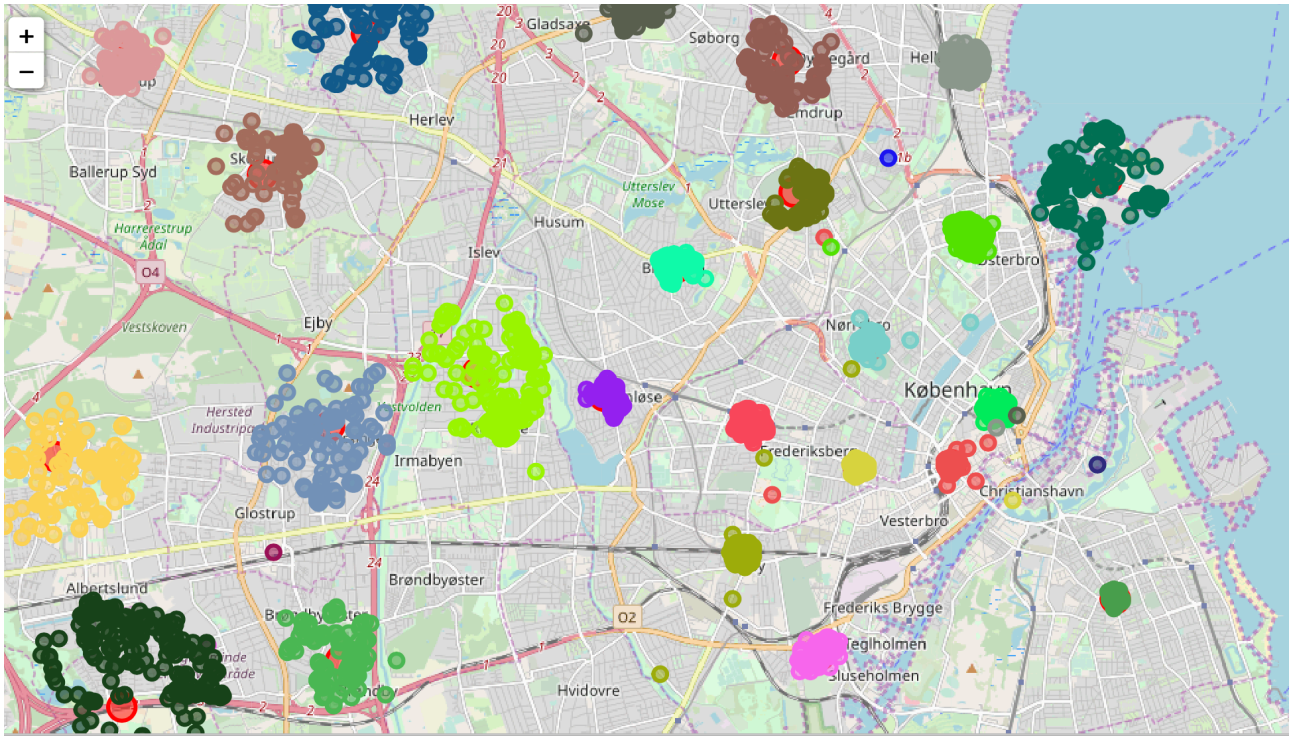
The data point limit specifies the maximum data points around a given coordinate that FourSquare would send back with one API call. And the radius specifies the distance limit between the venues and the region centers. Picture 2 below shows a distribution of all regions along with the surrounding venues, when the data point limit is 100, and radius is 1km.



Pic 2, venues distribution with limit 100, radius 1000

The map above shows the venue distribution in remote regions are more spread out, some region even overlapped to each other, while region close to city center are more centralized as the venues locate closer to each other. Hence it makes sense to narrow down the radius so that remote areas focus more on the center of the town, and to increase the data points limit to get more samples for the regions in the center.

Next step, the data point limit is set to 200 with a radius 500. The venue distribution is drawn on map below (Pic 3). It seems from the map that the distribution of venues is now a bit more centralized with more density. This means the parameter change is leading to a better result.

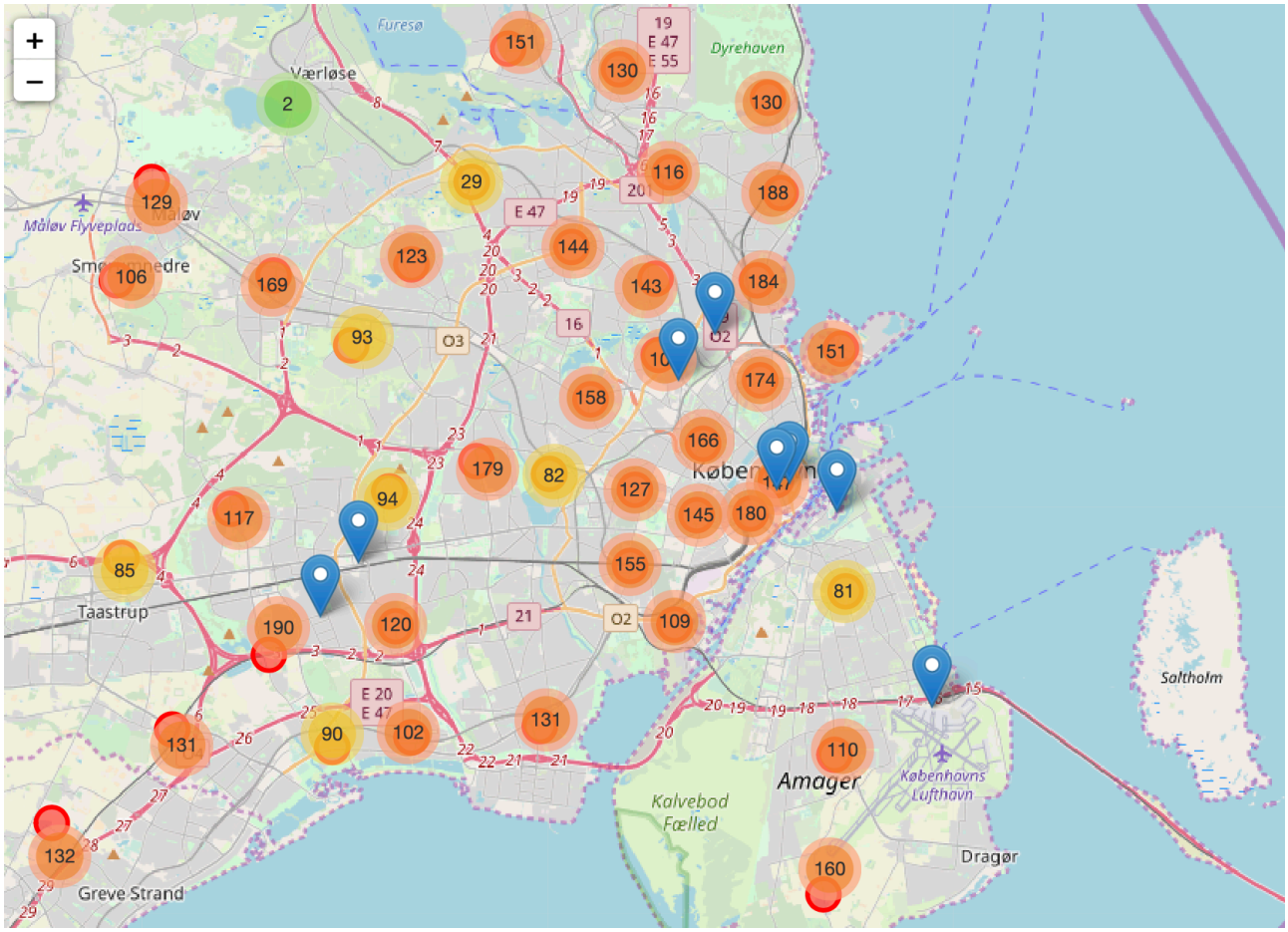


Pic 3, venues distribution with limit 200, radius 500

As previous analysis basically shows that a bigger data point limit and smaller radius could actually lead to a better venue distribution on the map. It would be interesting to know if we can make the distribution even better by further manipulating these two parameters. Let's start with checking the current number of venues of each venue group. Using folium class *MarkerCluster*, the number of venue data points of each region is put into the map in picture 4 below. The data shows that even the data point limit is set to 200, we can rarely see any regions which actually reach this limit. It means that the data point sample is limited in *FourSquare* database with current parameter, and the only way to get more data samples is to expand the radius. However, the venues distribution map in picture 3 indicates that many regions are already reaching to the boundary to another region, or to the rural areas, thus expanding the radius would make no sense in this case. Thus further analysis will be based on the venue data which is fetched with data point limit 200 and radius 500. Table 2 below shows a detailed data of the venue and the region. The *Venue category* will be used in the next step

	Region	Postcode	Region Latitude	Region Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Copenhagen K	1000	55.684199	12.579543	Kongens Have	55.684361	12.580099	Park
1	Copenhagen K	1000	55.684199	12.579543	Jazzfestival Kgs Have	55.683420	12.579783	Performing Arts Venue
2	Copenhagen K	1000	55.684199	12.579543	Møller & Mammen	55.682335	12.582285	Hardware Store
3	Copenhagen K	1000	55.684199	12.579543	ALDI	55.682531	12.581922	Discount Store
4	Copenhagen K	1000	55.684199	12.579543	Netto	55.682102	12.576530	Supermarket

Table 3. Region and Venue data



Pic 4. Number of venue data point in each region

3.3 Data processing for clustering

Now that we have got a table with all different regions and related venues (table 3), it is time to use these revenue data to group those regions into clusters, based on their similarity. The first question here is how we should define the similarity. As the similarity in this case is referring to the economy structure, and this could be to some extent reflected by the structure of venue categories. Based on the data from region and venues dataset that have been collected from last session, we could calculate and use the occurrence rate of each venue category in each region as the feature to measure the similarity of the regions.

The venue categories was first transformed into columns using one hot coding, then the dataset was grouped based on the region, and the mean value of all categories were calculated as the occurrence rate. See table 4 as example below.

	Region	Accessories Store	Adult Education Center	Advertising Agency	Airport	American Restaurant	Animal Shelter	Antique Shop	Apres Ski Bar	Arcade	...
0	Copenhagen N	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...
1	Copenhagen NV	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...
2	Copenhagen S	0.0	0.0	0.0	0.0	0.012346	0.0	0.0	0.0	0.0	...
3	Copenhagen SV	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...
4	Copenhagen V	0.0	0.0	0.0	0.0	0.005525	0.0	0.0	0.0	0.0	...

Table 4, the occurrence rate of venue categories in each region

Based on this occurrence rate table, we can get a list of top venue categories with the highest occurrence rate for each region. This list can show a direct impression on the features of each region. An example is shown in table 5.

	Region	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
0	Copenhagen N	Residential Building (Apartment / Condo)	Food Truck	Doctor's Office	Bar	Bike Shop	Office	Thrift / Vintage Store	Bus Line	Salon / Barbershop	Garden
1	Copenhagen NV	Residential Building (Apartment / Condo)	Medical Center	Cemetery	Building	Café	Bus Line	Doctor's Office	Sushi Restaurant	Pizza Place	Grocery Store
2	Copenhagen S	Pizza Place	Salon / Barbershop	Coffee Shop	General Entertainment	Arts & Crafts Store	Sushi Restaurant	Flea Market	Pet Store	Café	Rock Club
3	Copenhagen SV	Office	Residential Building (Apartment / Condo)	Convenience Store	Plaza	Other Great Outdoors	Café	Professional & Other Places	Building	Bank	Pizza Place
4	Copenhagen V	Office	Bar	Music Venue	Building	Convenience Store	Art Gallery	City Hall	Coffee Shop	Fast Food Restaurant	Courthouse

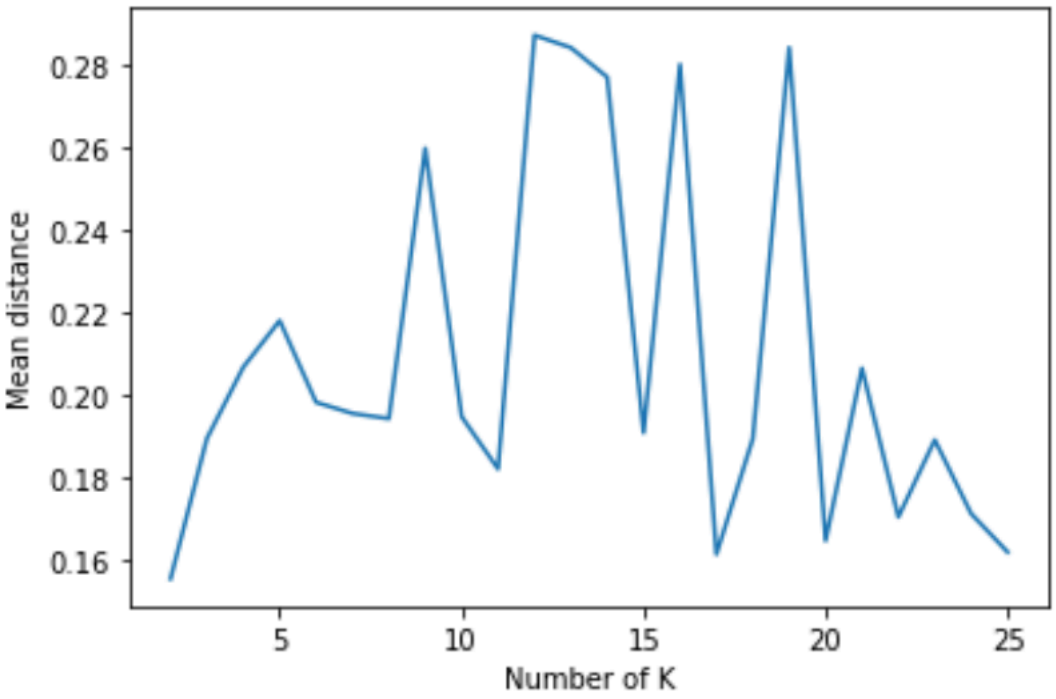
Table 5. top venue categories in each region

From the data above, the region ‘Copenhagen N’ and ‘Copenhagen NV’ can be best categorized as ‘residential area’, and region ‘Copenhagen SV’ and ‘Copenhagen V’ are more like ‘Office area’. This feature along with others can be used to make clusters for all regions in the next step.

3.4 Determine the best K for K-means clustering

Clustering as a machine learning method will be used to build the cluster model. Considering that the dataset contains large amount of data points (51 rows and 424 columns with 6643 venues), and the similarity or distance between each region’s venue data cannot be effectively pictured, the clustering algorithm used in this case will be K-means clustering, which is a relatively efficient method compared to other clustering algorithms like DBSCAN and hierarchical clustering.
K-means is partition-based clustering algorithm, and the number of clusters need to be pre-defined. As there are 51 different regions, we will try to find a most suitable number of clusters (K value) between 2 and 25. We will calculate the mean value of Euclidean distance between each cluster center and its cluster points, for all clusters. Afterwards, the mean distance will be

compared to find the number of clusters (K value) which leads to a minimum distance. The distance values long with relative K values are plotted as shown in picture 5.



Pic.5 mean distance with different K value

From the plot above, it is obvious that the mean distance has the smallest value when the K value is 2, selected from a range between 2 and 25. This means making 2 clusters for the 51 regions would actually lead to a better clustering result. Below table shows the top 5 K values which have the least distance value.

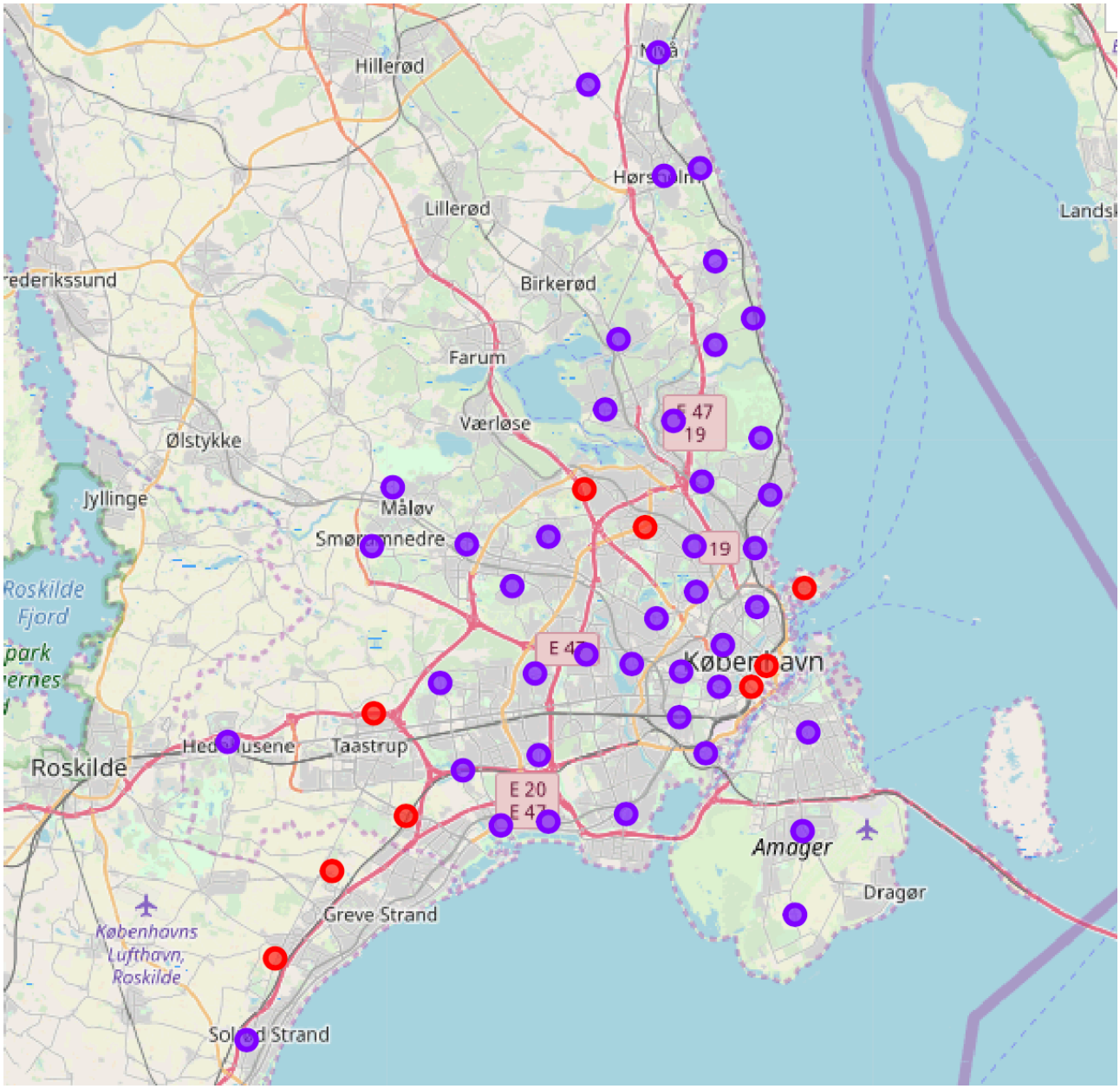
K	2.000000	17.000000	25.000000	20.000000	22.000000
Dist	0.155476	0.161444	0.161987	0.164745	0.170498

Table 6. Top 5 K values with the least distance value

As the dataset contains 51 regions, grouping them into too many clusters would not be a good idea, as the number of regions (data point samples) in each cluster could be quite limited (2.5 regions in each cluster on average for 20 clusters). Here we will take 2 clusters as the K value for further analysis, as the average distance are the smallest and enough regions are covered in each cluster.

3.5 Create and analyze model

Taking 2 clusters as a predefined parameter, using K-means clustering algorithm, the grouped regions can be displayed on a map as picture 6 below. 9 regions out of 51 are clustered into one group (the red points), and the rest are clustered into another (the blue points).



Pic 6. K-means clustering with 2 clusters for 51 regions

Since we now have all 51 regions clustered into two groups, next we need to calculate the percentage of the restaurant venues in all venues in each region, and then find out the maximum value of this number. We will take this maximum percentage value as the measurement of full competition for restaurants in a region in the same cluster with.
The result is displayed below in table 7.
The formula below is for all regions in the same cluster:

restaurant_rate = restaurant_count/ Venue_count
max_rest_rate = max(restaurant_rate)

where *restaurant_rate* stands for restaurant venues percentage of all venues in a region, and *max_rest_rate* stands for the maximum *restaurant_rate* among all regions in a cluster.

	Cluster Labels	Area	Postal Code	Venue_count	restaurant_count	restaurant_rate	max_rest_rate
0	1	Copenhagen N	2200	166	22.0	0.132530	0.184615
1	1	Copenhagen NV	2400	104	6.0	0.057692	0.184615
2	1	Copenhagen S	2300	81	9.0	0.111111	0.184615
3	1	Copenhagen SV	2450	110	3.0	0.027273	0.184615
4	0	Copenhagen V	1500	181	14.0	0.077348	0.077348

Table 7. labeled regions with restaurant rate

Now we can calculate how many new restaurants can be opened for each region, considering the region is with full competition. The formula is:

*New_Rest = Venue_count * max_rest_rate - restaurant_count*

The result is put into the same dataframe, shown as table 8.

	Cluster Labels	Area	Postal Code	Venue_count	restaurant_count	restaurant_rate	max_rest_rate	New_Rest
0	1	Copenhagen N	2200	166	22.0	0.132530	0.184615	8
1	1	Copenhagen NV	2400	104	6.0	0.057692	0.184615	13
2	1	Copenhagen S	2300	81	9.0	0.111111	0.184615	5
3	1	Copenhagen SV	2450	110	3.0	0.027273	0.184615	17
4	0	Copenhagen V	1500	181	14.0	0.077348	0.077348	0

Table 8. the number of new restaurant when full competition

As discussed previously, the number of new restaurants can be considered as the feature to determine if a region is suitable for a new restaurant and how well a new restaurant can survive in this region, considering the competition. Now let's list the top 5 areas which should be best choices for a new restaurant. See table 9 below.

Cluster Labels	Area	Postal Code	Venue_count	restaurant_count	restaurant_rate	max_rest_rate	New_Rest
1	Vallensbæk	2625	190	6.0	0.031579	0.184615	29
1	Vedbæk	2950	190	7.0	0.036842	0.184615	28
1	Rødovre	2610	179	6.0	0.033520	0.184615	27
1	Dragør	2791	161	4.0	0.024845	0.184615	25
1	Copenhagen Ø	2100	174	8.0	0.045977	0.184615	24
1	Holte	2840	159	6.0	0.037736	0.184615	23

Table 9. Top 5 regions best for opening a new restaurant

4. Results

From table 9 in previous part, we can see that the top 5 regions which are probably best choice for opening a new restaurant are: *Vallensbæk, Vedbæk, Rødovre, Dragør, Copenhagen Ø* and *Holte*.
As these 5 regions are all with label 0, they have a similar economic structure considering the surrounding venue categories. However, they are all extremely under sufficient competition, when compared with the region which is in full competition. Below table shows the two regions with full competition for each cluster, where the *new restaurant* features are actually 0, meaning they are not a good choice for opening a new restaurant, as there's no comparison and evidence showing that a new restaurant will actually survive.

Cluster Labels	Area	Postal Code	Venue_count	restaurant_count	restaurant_rate	max_rest_rate	New_Rest
1	Klampenborg	2930	130	24.0	0.184615	0.184615	0
0	Copenhagen V	1500	181	14.0	0.077348	0.077348	0

Table 10. regions with full competition

5. Discussion

In the result, we can see that the top 5 regions are within the same cluster with label 0, but it doesn't mean that cluster 0 is better for a new restaurant. In this case we select the K value as 2 to build the clustering model, because the mean distance is the best. However the second best mean distance is actually quite close to the chosen one, and the reason it is not chosen in this case is because 17 clusters could simply be too many for the 51 regions, and making 17 groups will simple lead to many clusters with only 1 or 2 regions in the cluster, which is quite limited numbers of samples. Less sample would probably lead to less accuracy in this case.
Besides, the venues in Copenhagen which are registered on *FourSquare* seems not quite adequate for a better clustering modeling, as there are basically less than 200 venues for each region. And this as inadequate data points would also be a key factor that affect the accuracy.
Furthermore, the result we have is based on the venue structure in a limited area in each region center, it may not accurately reflect the whole economy. However, the result does show a picture of the competitive conditions, which might not be easily foreseen by the investor. Based on this result, the client could also further consider other factors, like cost, nature, in order to make the final decision.

6. Conclusion

This report described a problem, to select one region out of 51 regions in Copenhagen area to open a new restaurant. The analysis was done based on venues information around each region area. The coordinate info was acquired using Google Map's Geocoding API, and the venue info was fetched from FourSqure's API.
Firstly, I made a pre-processing of the venue dataset, in order to get the valuable data feature 'venue category'. Then I use this feature to group all 51 regions into 2 clusters by building up machine learning model with K-means algorithm. Having these 2 clusters with different regions in each cluster, I was able to define the full competition level of each cluster, and then calculate the number of new restaurants that can be opened in each region. With this result sorted for each region, we can finally see a clear picture of which regions have the best potential to open a new restaurant.