Coursera Course: *Applied Data Science Capstone*

Peer-graded Final Assignment Deliverable: *Capstone Project - The Battle of Neighbourhoods*

Where to Open Your Pizza Place - With the Power of Clustering

Submitted By: *Saumya Gupta*

6th September 2021

Introduction (Business Problem)

Problem

You are so passionate about Pizza and Pizza making that you decide to make money out of it. You decide to open your brand-new Pizza place in Stockholm County. Stockholm said to be the capital of Scandinavia, is also one of the leading food-tech hubs in the world. You have almost figured out everything you want to do in the restaurant; menu items, prices, weekend specials, even your Christmas offers.

Nevertheless, there is one thing that keeps bugging you- the location of your Pizza place. Locations, as they say, can make or break businesses. Moreover, that very well goes for the places where people choose to eat. Should it be at the centre of the city? Should it be where all the other Pizza places are? You are confused as to where to start looking at, for the least.

This solution helps you that this solution will give you a great starting point for locations with excellent business potential. In this solution, we help find a group of similar neighbourhoods (or districts) in Stockholm County with Pizza places as the most common venue type for most neighbourhoods. How does that help, you ask? Neighbourhoods with well-established Pizza place businesses should be good for your new business too. However, if there are too many similar businesses already in the neighbourhood, you would want to know other neighbourhoods that are very similar to this one in terms of their Pizza businesses and could accommodate more such places. That is why this solution gives you not one neighbourhood but also find similar neighbourhoods. We do this by clustering Stockholm County neighbourhoods together based on the types of different venues present in the neighbourhood. So, if you have too many well-established and flourishing Pizza place businesses in a neighbourhood, you can go for a similar neighbourhood (based on your other specific requirements), which does not have too many of such joints but does have a potential for profits for a new joint if you open one.

Target Audience

As can be interpreted from earlier texts, this solution helps businesses find locations to set up their branches or people trying to set up their first restaurant. In this solution, we investigate the possibilities of business setup in Stockholm County. However, in future works with the availability of high definition geocoordinate data and other demographic data such as population and mass affordability, this can be extended to any settlement with better results due to more information.

Data

In this analysis, we use two datasets. One is the geo-coordinate data of the districts of Stockholm County. The second is the venue categories information of the same districts of Stockholm County, retrieved with the help of the Foursquare API using the geocoordinate information present in the first dataset. Both these datasets are combined and processed further to create the numbers used for the clustering analysis. The following two sections describe each dataset in detail.

Geo-coordinate Data

This dataset contains the latitude and longitude information of the districts of Stockholm County. This study initially intended to use the geo-coordinate data corresponding to each postal code, but given the limitations due to chargeable usage to numerous APIs, the author resorted to using the data readily available on the internet. The website http://www.geonames.org/ provides postal code data for nearly 100 countries along with their geo-coordinate information. For this analysis, we download the zip for Sweden available here and make our way from there. The .zip contains a tab-delimited text file enlisting all postal codes of Sweden, districts names, latitudes, and longitudes information along with some other data such as municipal name and municipal code. However, for our analysis, we only read the postal code, district name, county name (to be used for filtering districts of Stockholm County), latitude and longitude information. Fig. 1 shows the first five rows of the data read into the pandas' data frame. It is, of course, after applying the Stockholm County filter.

	Postal Code	Neighbourhood	County	Latitude	Longitude
0	186 00	Vallentuna	Stockholm	59.5344	18.0776
1	186 01	Vallentuna	Stockholm	59.5344	18.0776
2	186 03	Brottby	Stockholm	59.5632	18.2403
3	186 21	Vallentuna	Stockholm	59.5344	18.0776
4	186 22	Vallentuna	Stockholm	59.5344	18.0776

Figure 1. Postal code data after being read into a pandas data frame. We have Postal Code, Neighbourhood (or district) of the postal code, County of the district, Latitude and Longitude for the postal code. Note that one district can have multiple postal codes.

Unfortunately, we observe that the latitude and longitude information for all the districts within a neighbourhood or districts are the same. It points to the inaccuracy in the estimation of geocoordinates. It is evident in rows 0, 1, 3 and 4 of Fig. 1. Due to the unavailability of any alternative data, we decide to move forward with this and introduce consistency by removing all duplicates. We have now the geocoordinate data (1 row each) for each of the 121 districts in the County. Fig. 2 shows the first five rows result of geo-coordinate data.

	Neighbourhood	Latitude	Longitude
0	Vallentuna	59.5344	18.0776
2	Brottby	59.5632	18.2403
43	Ingmarsö	59.4675	18.7494
44	Åkersberga	59.4794	18.2997
45	Ljusterö	59.5275	18.6211

Figure 2. The final neighbourhood geo-coordinates data after processing.

Foursquare Venue Data

Now that we have the latitude and longitude information of neighbourhoods, we use the geo-coordinate information of each neighbourhood (or district) to find the information of the venues in the 500 meters radius of each neighbourhood. The number of venues in each category for each neighbourhood constitutes the data for clustering analysis. To get this data, we need to create a developer account with Foursquare and then use the generated credentials for development usage (client_id, client_secret) to call the explore API. For a neighbourhood, apart from the geo-coordinate data (ll) and the user credentials, we also need to mention the radius (radius) within which we want to retrieve all venues, a limit (limit) that defines the number of results the API should return and the version (v) of the API, which is a date and accesses a particular version of the API. Given the geo-coordinate information, changing these extra parameters would result in varied results, especially the API version.

For example, for v=20210816, ll=59.6352,17.9125, radius=500 and limit=100, the explore API returns the JSON response given in Fig. 3.


```
[{'reasons': {'count': 0,
   'items': [{'summary': 'This spot is popular',
     'type': 'general',
     'reasonName': 'globalInteractionReason'}]},
  'venue': {'id': '524ee0bb498e1f86894b4765',
   'name': 'Spottingplats 19R Banänden',
   'location': {'lat': 59.63156937426219,
    'lng': 17.908217057824203,
    'labeledLatLngs': [{'label': 'display',
      'lat': 59.63156937426219,
      'lng': 17.908217057824203}],
    'distance': 470,
    'cc': 'SE',
    'country': 'Sverige',
    'formattedAddress': ['Sverige']},
   'categories': [{'id': '4bf58dd8d48988d165941735',
     'name': 'Scenic Lookout',
     'pluralName': 'Scenic Lookouts',
     'shortName': 'Scenic Lookout',
     'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/parks_outdoors/sceniclookout_',
      'suffix': '.png'},
     'primary': True}],
   'photos': {'count': 0, 'groups': []}},
  'referralId': 'e-0-524ee0bb498e1f86894b4765-0'},
 {'reasons': {'count': 0,
   'items': [{'summary': 'This spot is popular',
     'type': 'general',
     'reasonName': 'globalInteractionReason'}]},
  'venue': {'id': '52308c5911d2bc85ba9b2ddf',
   'name': 'Commuter platform Arlanda',
   'location': {'lat': 59.639211,
    'lng': 17.9155,
    'labeledLatLngs': [{'label': 'display', 'lat': 59.639211, 'lng': 17.9155}],
    'distance': 477,
    'cc': 'SE',
    'country': 'Sverige',
    'formattedAddress': ['Sverige']},
   'categories': [{'id': '4bf58dd8d48988d1fc931735',
     'name': 'Light Rail Station',
     'pluralName': 'Light Rail Stations',
     'shortName': 'Light Rail',
     'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/travel/lightrail_',
      'suffix': '.png'},
     'primary': True}],
   'photos': {'count': 0, 'groups': []}},
  'referralId': 'e-0-52308c5911d2bc85ba9b2ddf-1'}]
```


Figure 3. The venue data JSON received from the Foursquare Explore API for a neighbourhood (Stockholm-Arlanda). Note that we receive only two venues for this neighbourhood.

Fig. 3 shows two venues returned for the geo-coordinates. From the JSON returned for every venue, their latitude, longitude, and venue information are extracted and made into the form

in Fig. 4. Notice that these two tuples retain only certain information for the two venues from the JSON object. These are neighbourhood names, neighbourhood attitude, neighbourhood longitude, venue name, venue latitude, venue longitude and most importantly, venue category.

```
[('Stockholm-Arlanda',
  59.6352,
  17.9125,
  'Spottingplats 19R Banänden',
  59.63156937426219,
  17.908217057824203,
  'Scenic Lookout'),
 ('Stockholm-Arlanda',
  59.6352,
  17.9125,
  'Commuter platform Arlanda',
  59.639211,
  17.9155,
  'Light Rail Station')]
```

Figure 4. The list of tuples versions of the JSON present in Fig. 3. Each venue gets formulated into one tuple.

We do the same to all 121 neighbourhoods. That is to say that we pass the 121 rows of geo-coordinate information of districts in the first dataset to Foursquare's explore API and receive 1081 rows (list of tuples converted to a data frame) in return because of multiple venues per neighbourhood. Fig. 5 shows the first five rows of our data.

	Neighbourhood	NeighbourhoodLatitude	NeighbourhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Vallentuna	59.5344	18.0776	Vallentuna Stenugnsbageri	59.534369	18.077338	Bakery
1	Vallentuna	59.5344	18.0776	Gym & Sim	59.533476	18.084472	Gym
2	Vallentuna	59.5344	18.0776	Lidl Vallentuna	59.532625	18.083467	Supermarket
3	Vallentuna	59.5344	18.0776	Vallentuna (L)	59.533849	18.079585	Light Rail Station
4	Vallentuna	59.5344	18.0776	Vallentuna Centrum	59.534321	18.078444	Plaza

Figure 5. The final processed Foursquare venue data. We have Neighbourhood's names and geo-coordinates along with the venues' names, geo-coordinates, and categories.

Methodology

This analysis aims to find clusters (or groups) of similar neighbourhoods (or districts) in Stockholm County based on the number of venues in the district in each category. We use the K-Means clustering algorithm to find these clusters. We then examine the clusters and propose the cluster we think would represent the best set of districts for our perspective Pizza place owner to start looking at locations and analyzing further. The subsections that follow describe the step to achieving the aim.

Preparing data for clustering analysis

After initial pre-processing (talked about in detail in the Data Section), we have the data shown in Fig. 5 with us. The data enlists 121 districts of Stockholm County along with the various categories of venues within their radius. Fig. 6 shows the total number of venues in the whole of Stockholm County in each category. It would be ethical to pass on the information to our prospective client that Pizza places with 32 venues in total are the 4th most common venues in the County. 32 sounds very small and unreal but remember that this number is based on the results from the Foursquare explore API, and the actual number might be higher than this one.

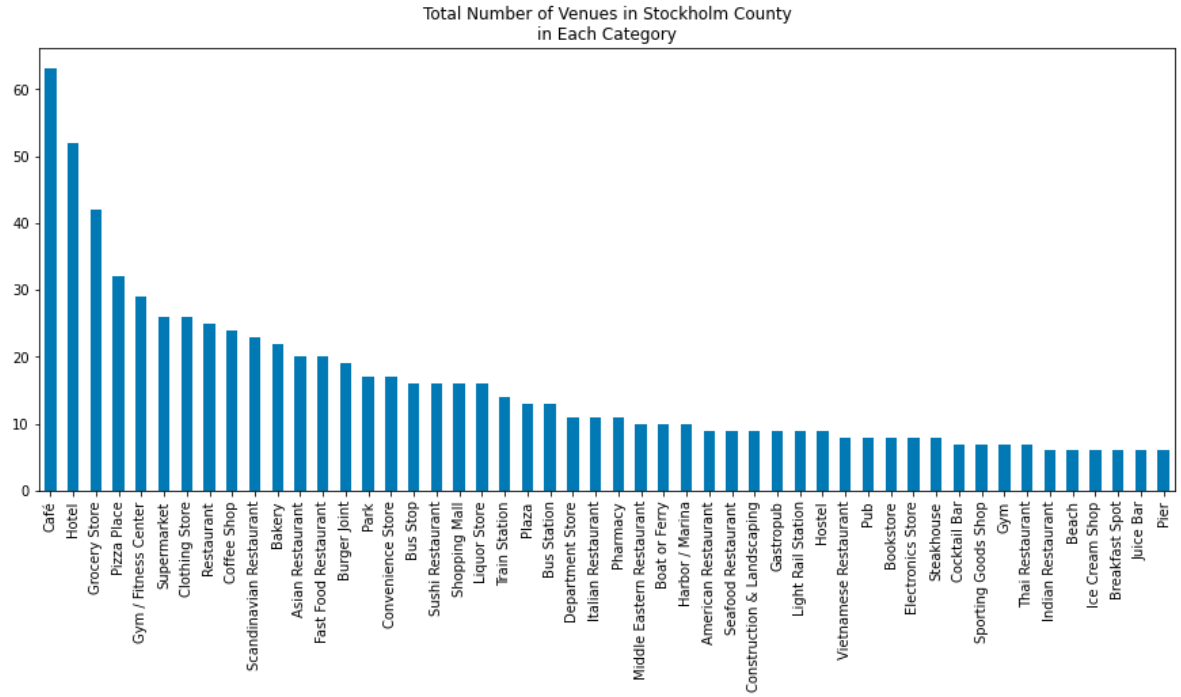

Figure 6. The total number of venues in the whole of Stockholm County in each venue category.

Which neighbourhood and what venue categories are two things that we want to use to prepare our clustering data? Hence, we try to assign each neighbourhood scores for each venue category (type of venues such as Pizza Place, Hotel, and Coffee Shop). A high score for a venue category will represent a higher frequency within that neighbourhood. We do this by presenting the 'VenueCategory' column of the data represented in Fig. 5 to pandas one-hot encoding function (get_dummies()) and then group the number obtained by the neighbourhood name using the mean aggregate. Fig. 7 shows the first five rows of the resulting data. The neighbourhoods name, along with their scores for different venue categories, represent the clustering data. Clusters created using this data would be similar in terms of the type of venues they have.

	Neighbourhood	American Restaurant	Arts & Crafts Store	Asian Restaurant	Bakery	Boat or Ferry	Bookstore	Breakfast Spot	Burger Joint	Bus Station	Bus Stop	Café	Clothing Store	Cocktail Bar	Coffee Shop	Construction & Landscaping	Convenience Store
0	Arholma	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
1	Bagarmossen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.250000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
2	Bandhagen	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.076923	0.0	0.076923	0.0	0.0	0.0	0.0	0.076923
3	Blidö	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
4	Bro	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000

Figure 7 Neighbourhoods and their scores for each venue category based on the number of venues in each category. It serves as the data that we later feed to the clustering machine.

After preparing this data for clustering, we want to be able to evaluate our clusters. We use the data in Fig. 7 and find the top ten most common venue categories for each neighbourhood. How is it done? Well, we order the venues categories in descending order based on the fraction values for the category. Fig. 8 shows the result. We will be using this data post clustering to find the cluster we want to propose to our client.

	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
0	Arholma	Harbor / Marina	Grocery Store	American Restaurant	Plaza	Italian Restaurant	Light Rail Station	Liquor Store	Middle Eastern Restaurant	Park	Pharmacy
1	Bagarmossen	Gym / Fitness Center	Grocery Store	Bus Station	Pizza Place	American Restaurant	Plaza	Italian Restaurant	Light Rail Station	Liquor Store	Middle Eastern Restaurant
2	Bandhagen	Light Rail Station	Fast Food Restaurant	Train Station	Hotel	Bus Station	Gym	Café	Grocery Store	Pizza Place	Playground
3	Blidö	Bakery	American Restaurant	Pub	Indian Restaurant	Italian Restaurant	Light Rail Station	Liquor Store	Middle Eastern Restaurant	Park	Pharmacy
4	Bro	Burger Joint	Fast Food Restaurant	American Restaurant	Pub	Italian Restaurant	Light Rail Station	Liquor Store	Middle Eastern Restaurant	Park	Pharmacy

Figure 8. Neighbourhoods along with their top ten most common venues categories. It serves as the data used later in the cluster evaluation.

Clustering Analysis

We use K-Means clustering for clustering (or making groups) of our neighbourhoods (or districts) based on their venue frequencies.

Why not any other algorithm, you ask? Other clustering algorithms, such as connectivity-based clustering (hierarchical clustering) or density-based clustering (DBSCAN), tend to be more expensive with better features where all that we need to do here with K-Means is choosing an optimal K for the number of clusters. In this way, K-Means proves to be the simplest and the easiest and fastest to run. If you have an idea about the number of clusters, you are good, but if you do not, you still can find it through repetitive training and testing.

Now, how exactly can you find out the optimal K if you have no idea how many clusters there could be? For example, in our case, we do not know the exact number of clusters in our data, but we do know that it should be around 4 – 7 clusters. Nevertheless, let us find the optimal K-value using a very popular heuristic called the elbow method. Here, we repeatedly cluster with different K values, measure distortions or inertias (costs), and then plot them as a function of the number of clusters. At last, we select the one where the cost plot forms an elbow (considered the optimal point). In our case, Fig. 9 shows the elbow plot showing model

distortion and inertia values for K-values from 1 to 10. Both the plots in Fig. 9 shows that the elbow of the curve is at K = 5. Hence 5 becomes the number of clusters parameters with which we will run our final clustering function.

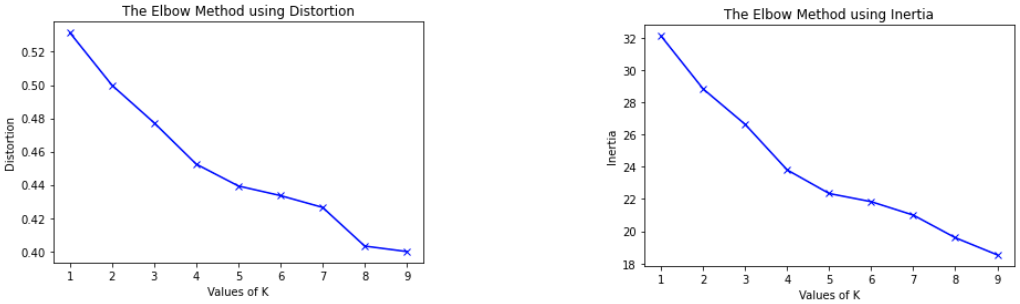

Figure 9. Elbow plots showing model distortions (left) and model inertias (right) as a function of K (number of clusters).

After clustering, we must examine our clusters. To examine we, combine the cluster number information (which cluster a neighbourhood belongs to) with the data in Fig. 8. Fig. 10 shows the resulting data ready for analysis for location cluster selection. For cluster selection, we are looking at the first most common venues for each neighbourhood of the cluster.

	Neighbourhood	Latitude	Longitude	ClusterLabels	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
2685	Arholma	59.8500	19.1000	1	Harbor / Marina	Grocery Store	American Restaurant	Plaza	Italian Restaurant	Light Rail Station	Liquor Store	Middle Eastern Restaurant	Park	Pharmacy
1799	Bagarmossen	59.2556	18.1167	1	Gym / Fitness Center	Grocery Store	Bus Station	Pizza Place	American Restaurant	Plaza	Italian Restaurant	Light Rail Station	Liquor Store	Middle Eastern Restaurant
1651	Bandhagen	59.2968	18.0313	1	Light Rail Station	Fast Food Restaurant	Train Station	Hotel	Bus Station	Gym	Café	Grocery Store	Pizza Place	Playground
2680	Blidö	59.6150	18.8917	1	Bakery	American Restaurant	Pub	Indian Restaurant	Italian Restaurant	Light Rail Station	Liquor Store	Middle Eastern Restaurant	Park	Pharmacy
3474	Bro	59.5167	17.6333	1	Burger Joint	Fast Food Restaurant	American Restaurant	Pub	Italian Restaurant	Light Rail Station	Liquor Store	Middle Eastern Restaurant	Park	Pharmacy

Figure 10. Cluster information merged into the data in Fig. 8 for cluster evaluation.

Results

Fig. 11 shows two maps; the first shows all the neighbourhoods in Stockholm County, and the second shows the same clusters. A different colour can identify each cluster. Notice how specific dots (representing neighbourhoods) are present in the first but have disappeared in the second map. It is because of removing rows containing venues with less than five frequencies in Stockholm County, resulting in losing certain less critical districts.

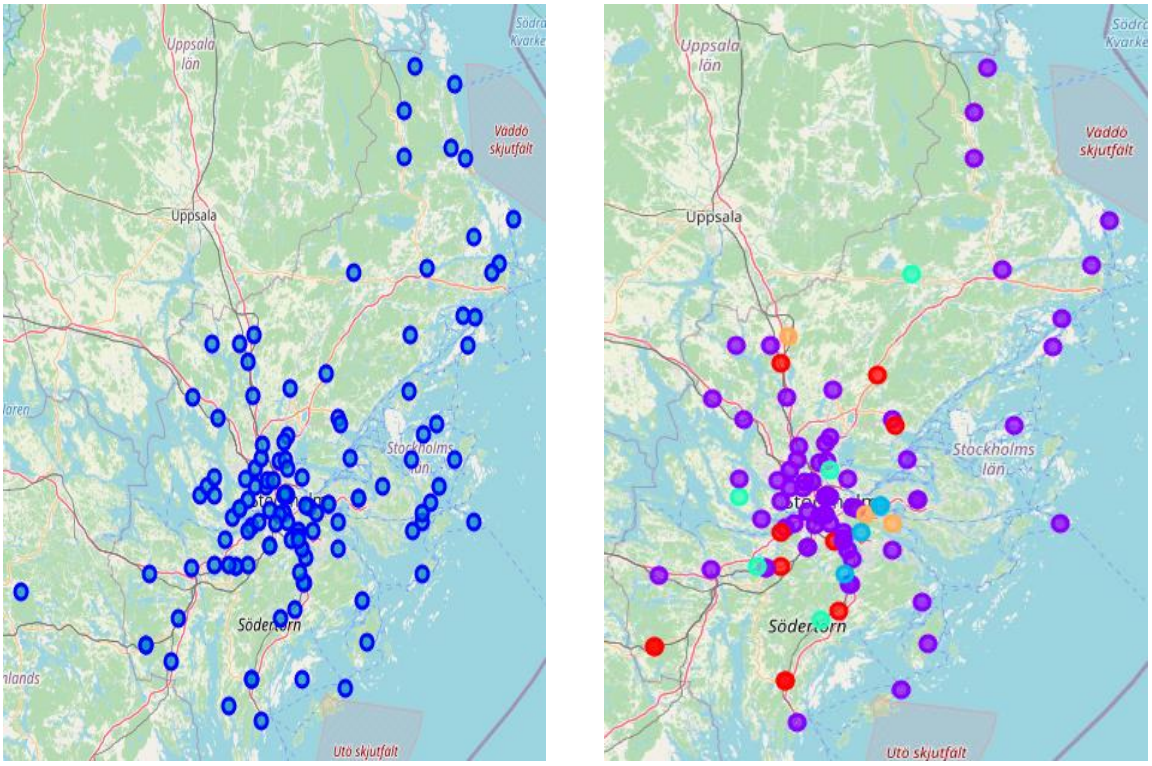

Figure 11. Neighbourhoods in Stockholm County (left). Neighbourhoods in Stockholm County clustered (right).

Now let us examine the clusters. First, let us look at cluster distribution. Cluster 1 has the most districts (77 districts), and clusters 0, 2, 3, 4 have 6, 5, 3 and 3 districts. Since our client wants to open a Pizza place, we try to see if any cluster has a Pizza place as the most common venue category for most of the districts present in that cluster. Yes, read that again! Cluster 2, 3 and 4 do not have any neighbourhood with Pizza places as their first most commonplace. So let us ignore them. Now let us look at clusters 0 and 1 closely. Fig. 12 shows the number of districts with their first most common venues in both clusters.

For cluster 1, only two districts have Pizza place as their most common venue. Pizza Places do not seem to be the driving factor for all these districts belonging to cluster 1, but Hotels seem to be. Suppose these two districts already have sufficiently many Pizza places. In that case, there is not much certainty of a beneficial outcome for the client's Pizza place in the other districts since Pizza place is not a very common venue in other districts of the cluster.

In this cluster 0, 3 districts have Pizza places as their first most common venue. Our client could choose to go for a location in any of these five districts. Suppose, for some reason (maybe, after looking at other factors such as population to Pizza place venue ratio or population affordability), the client does not want to open in these districts and decides to open in any other district in this cluster. In that case, there is much certainty of the success of the Pizza place. These districts belong to the same group because of their similar frequencies of Pizza places. So, if it works so well in A, but A has sufficient venues already, it will also work so well in B, which could use one more Pizza place.

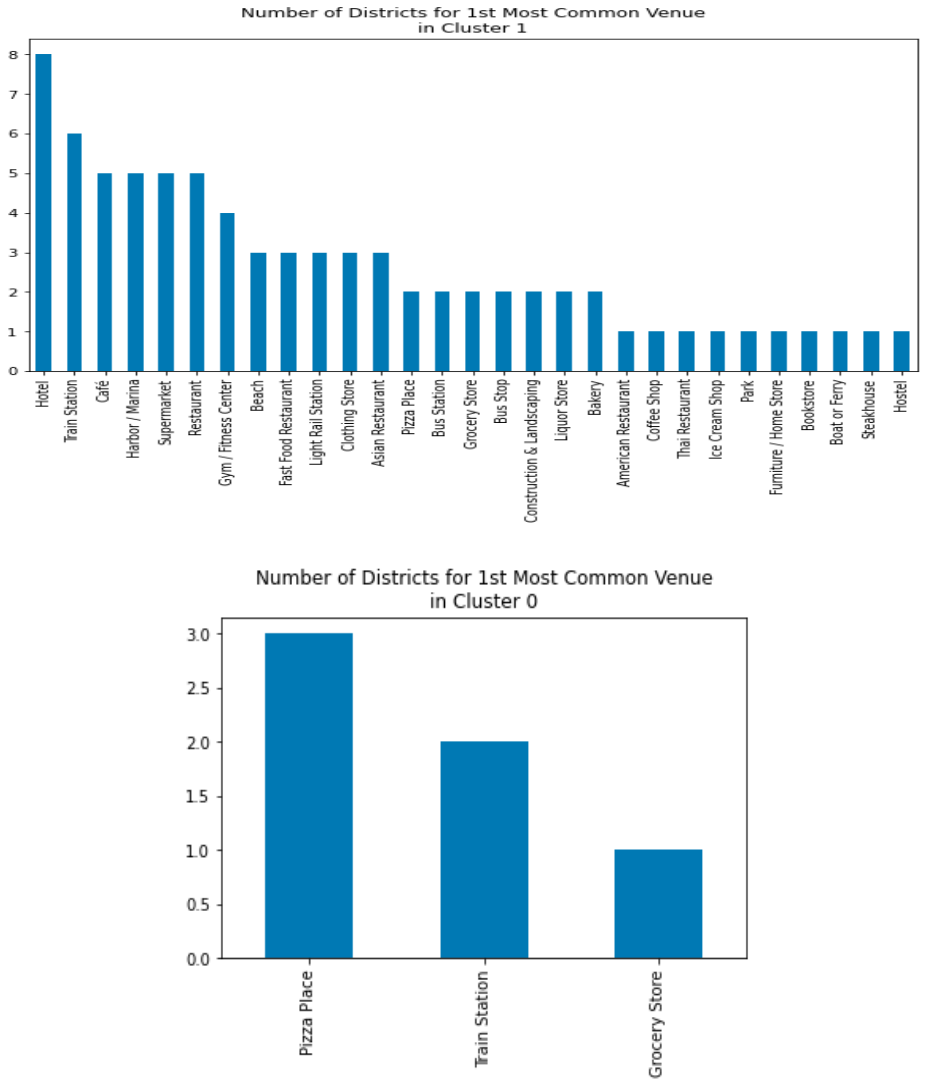

Figure 12. The number of districts with their first most common venues in cluster 1 (top) and cluster 0 (bottom).

Voilà, I give you not just some districts whose most common venue is a Pizza place, but with the power of clustering give you a group of similar districts which share the possibility of a potentially successful new Pizza place. It is now in the client's hands how he/she wants to examine the districts in this cluster.

Conclusion

Through K-Means clustering performed on the venue data of Stockholm County, we can produce a cluster for clients with six districts in the County that are similar based on their venues. Most of their similarity is because of the number of Pizza places they hold, and hence this gives us a chance to explore further in the correct direction.