22/10/2019

Comparing Neighbourhoods' in København and Malmö

Applied Data Science Capstone

Pravin Karthick Murugesan
AALBORG UNIVERSITY, DENMARK

Pravin Karthick Murugesan
22-10-2019

Comparing Neighbourhoods' in København and Malmö

1) Introduction

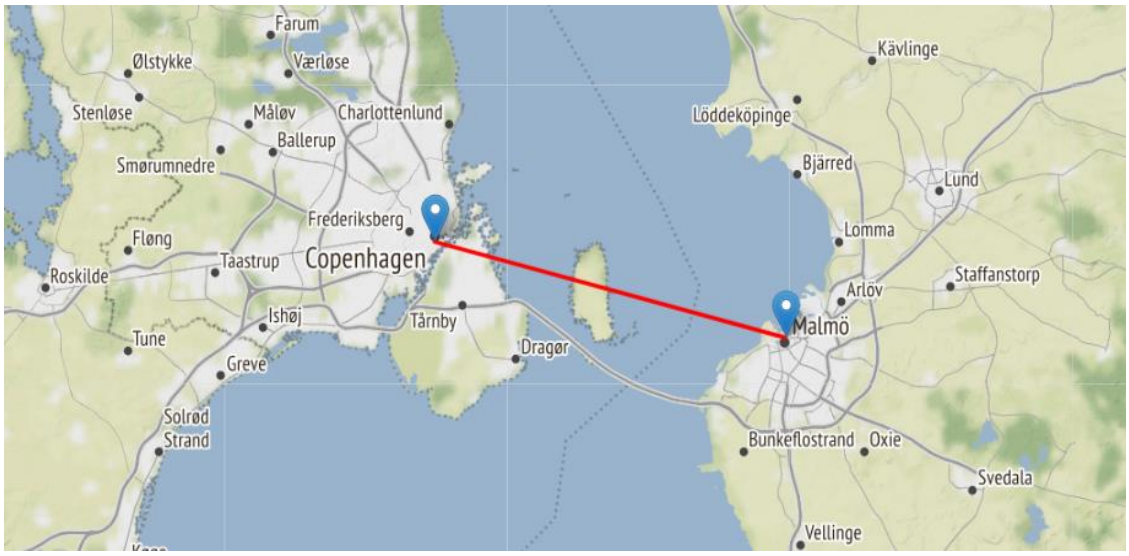
This report is compiled based on an assignment required to complete the course Applied Data Science Capstone which is part of the Applied Data Science Specialization program.

a) Problem Description

People relocate to different cities due to work, relationships, change in their kid's school or for other personal reasons. When they move to a new neighbourhood, wouldn't it be useful if they could compare this neighbourhood to the neighbourhood they previously lived in? I certainly will find such a comparison handy to make a decision. So over here, I have shown the similarities between the neighborhoods' in Copenhagen and Malmö, the capital city of Denmark and Sweden. There is also constant relocation within a city. One such reason is when a person buys a new apartment/house in a different neighbourhood.

Apart from moving between nearby cities, a lot of people have migrated to Scandinavia in the past two decades and continue to do so. As of 2017, the total immigrant population in Denmark and Sweden was 768,275 and 1,877,050 respectively. This data is also helpful to them to see how the neighborhoods' differ from each other and what the attractive venues the neighbourhood can provide to them and their family.

Additionally, the results of this project can be used by business owners to see which venues are a shortage or doesn't exist in a particular neighbourhood. They can then make an informed decision on creating an appropriate business in that neighbourhood to always have customers and guaranteed profits.



b) Data Used

The data contains the list of neighbourhoods' in Copenhagen and Malmö along with their respective geographical coordinates, list of venues (obtained by Foursquare APIs).

2) Methodology

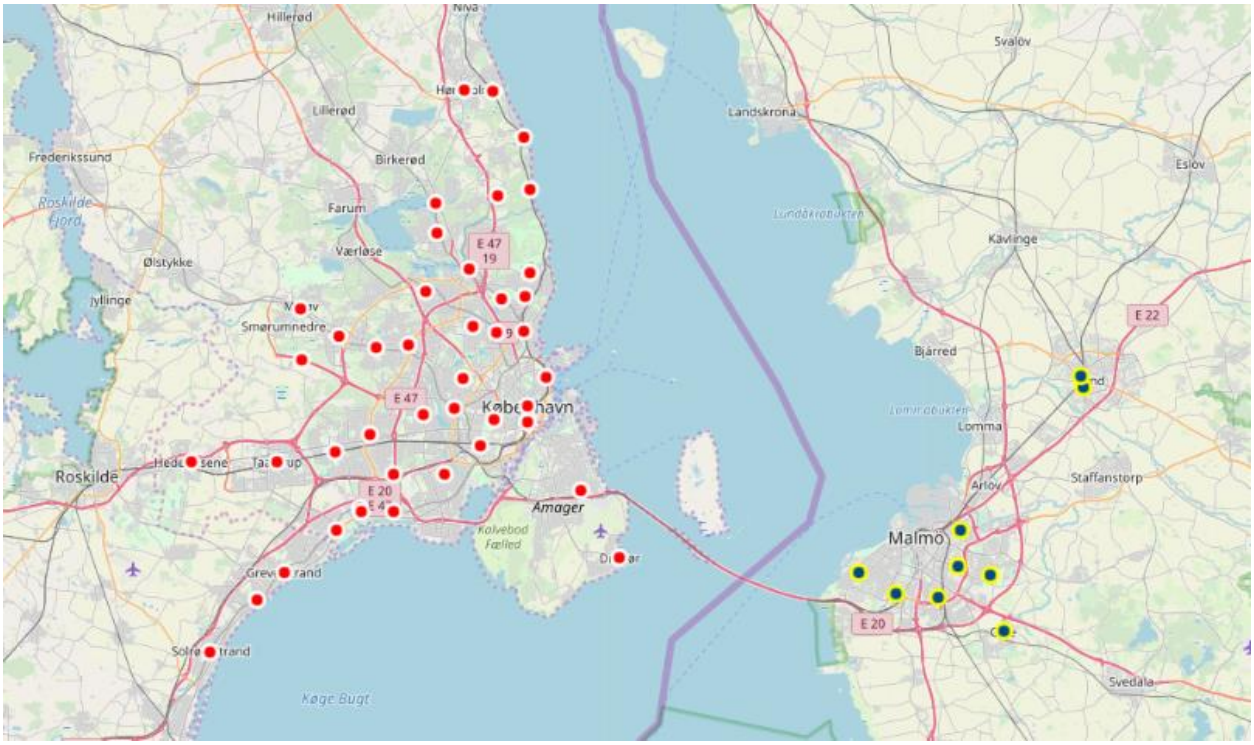
First, we obtain the essential data from a combination of reading from excel and web scraping from Wikipedia. Then this data is cleaned by renaming the columns and removing other unwanted columns. Next, the geographical coordinates are obtained from using the library geopy. An address is converted into latitude and longitude values. Then the outliers are removed. Some neighbourhoods' have completely wrong geographical coordinates. Having these neighbourhoods' causes problems when we plot with folium. These neighbourhoods' are later shown on a map. Top 20 rows of the data are displayed.

:

```
Size is (53, 4)
```

	District	Neighbourhood	Latitude	Longitude
0	Copenhagen	Frederiksberg	55.678	12.5326
1	Copenhagen	Copenhagen Ø	55.6867	12.5701
2	Copenhagen	Nordhavn	55.705	12.591
3	Copenhagen	Copenhagen SV	55.6766	12.5695
4	Copenhagen	Valby	55.6618	12.517
5	Copenhagen	Glostrup	55.6692	12.396
6	Copenhagen	Brøndby	55.6441	12.4217
7	Copenhagen	Rødovre	55.6812	12.4547
8	Copenhagen	Albertslund	55.6586	12.3563
9	Copenhagen	Vallensbæk	55.623	12.3845
10	Copenhagen	Taastrup	55.6522	12.292
11	Copenhagen	Ishøj	55.6094	12.3583
12	Copenhagen	Hedehusene	55.652	12.1971
13	Copenhagen	Hvidovre	55.6442	12.4772
14	Copenhagen	Brøndby Strand	55.621	12.4219
15	Copenhagen	Vallensbæk Strand	55.6209	12.3861
16	Copenhagen	Greve	55.5834	12.2999
17	Copenhagen	Solrød Strand	55.5333	12.2181
18	Copenhagen	Karlslunde	55.5663	12.2697
19	Copenhagen	Brønshøj	55.7041	12.4985

Once we have the geographical coordinates, the neighbourhoods' can be displayed on a map with folium, a map rendering library. These neighbourhoods' are going to be grouped together based on their similarities.

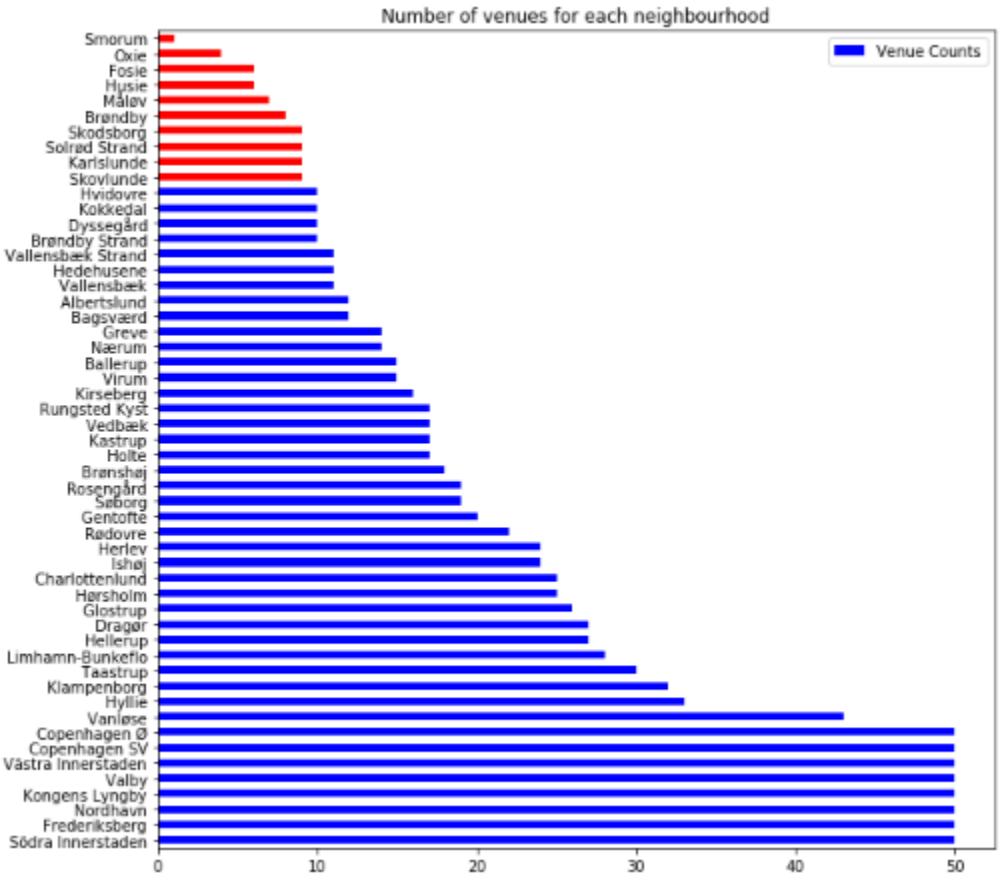


With the Foursquare credentials for a developer and a user-defined function, we can get a list of places to go for each neighbourhood. Few rows of that resultant table are shown.

```
Size is (1071, 5)
```

	Neighbourhood	Venue	Venue Category	Venue Latitude	Venue Longitude
0	Frederiksberg	Café Intime	Lounge	55.677221	12.532585
1	Frederiksberg	Frederiksberg Hovedbibliotek	Library	55.680724	12.530827
2	Frederiksberg	Creative Space	Art Gallery	55.677572	12.536766
3	Frederiksberg	Halifax	Burger Joint	55.680438	12.530850
4	Frederiksberg	Edison Teatret	Theater	55.676647	12.536828
5	Frederiksberg	Hi-Fi Klubben	Electronics Store	55.678346	12.533251
6	Frederiksberg	Restaurant Frederiks have	French Restaurant	55.679178	12.525047
7	Frederiksberg	Sokkelund Cafe & Brasserie	Café	55.679298	12.526044
8	Frederiksberg	Smalle Vine	Wine Shop	55.678974	12.528744
9	Frederiksberg	Forno a Legna	Pizza Place	55.682382	12.535324

A count for the venues listed for each neighbourhood is made into a bar chart. The maximum number of venues to be found was set to 50. One can observe that for some neighbourhoods', the count is low (less than 10). This is most likely due to the unavailability of data in Foursquare. People are yet to find good places and post them on Foursquare.



Since we will be comparing all neighbourhoods' with one another, it's better to remove neighbourhoods' with low count on venues to get a fair comparison. The ones marked in red are removed from the table.

Next, a method known as the one-hot encoding is done. It basically groups the rows by neighbourhood and inserts all the found venue categories as columns. The table formed after this procedure happens to have 190 columns. That seems to be a lot of venues categories. Having a look at the first table below, it is easy to spot that the column names are similar. Such similar venue categories can be combined, e.g., Art Gallery and Art Museum or all restaurants into one. Later, we use the k-means technique to find out similar neighbourhoods' based on the venues they have in them.

```
Size is (42, 190)
```

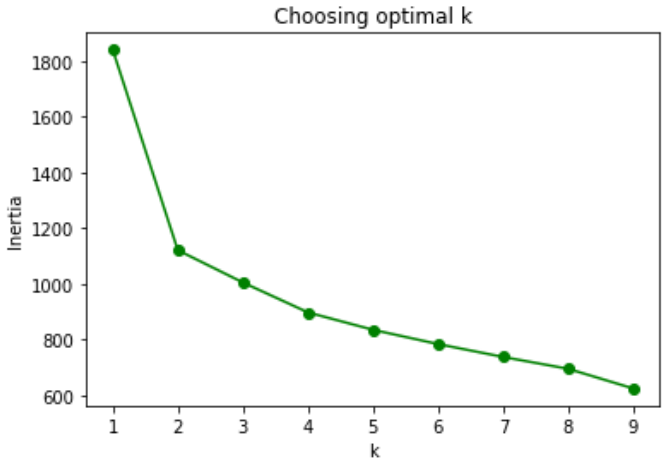
	Neighbourhood	American Restaurant	Apres Ski Bar	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bar
0	Albertslund	0	0	0	0	0	0	0	0	0	0	0	0
1	Bagsværd	0	0	0	0	0	0	0	0	1	0	1	0
2	Ballerup	0	0	0	0	0	0	0	0	0	0	0	0
3	Brønshøj	0	0	0	0	0	0	0	0	0	0	2	1
4	Charlottenlund	0	0	0	0	0	0	1	0	0	0	3	0

```
Size is (42, 130)
```

	Neighbourhood	Art	Auto	BBQ Joint	Bagel Shop	Bakery	Bar	Bathing Area	Beach	Beer	Boat or Ferry	Book	Breakfast Spot	Brewery	Burger Joint
0	Albertslund	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Bagsværd	0	0	1	0	1	0	0	0	0	0	0	0	0	0
2	Ballerup	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Brønshøj	0	0	0	0	2	1	0	0	0	0	0	0	0	0
4	Charlottenlund	0	0	0	0	3	2	0	0	0	0	0	0	0	0

K-means clustering

This is a simple unsupervised machine learning technique that can cluster similar data points together. What're the criteria for the rows (or neighbourhoods') to be similar? Well, it is the Euclidean distance from the cluster centers. Elbow method helps to choose the optimal number of clusters. From the below plot, we pick the point (value of 'k') from which the graph is linear. It is more linear from k = 5 and this is the chosen value of 'k'.



Inertia is the sum of squared distances of samples from the nearest cluster center. Choosing k = 2 would result in only two types of neighbourhoods'. The group of just two will not be helpful to the reader in providing insights to situations that were mentioned in the problem description.

3) Results

After applying k-means method, we look at the different clusters formed along with their count. Most of the neighbourhoods' are grouped within two clusters (label 1 and label 2).

Cluster Label	Counts
0	3
1	17
2	15
3	5
4	2

These cluster labels are inserted into the dataframe corresponding to its neighbourhood.

```
Size is (42, 133)
```

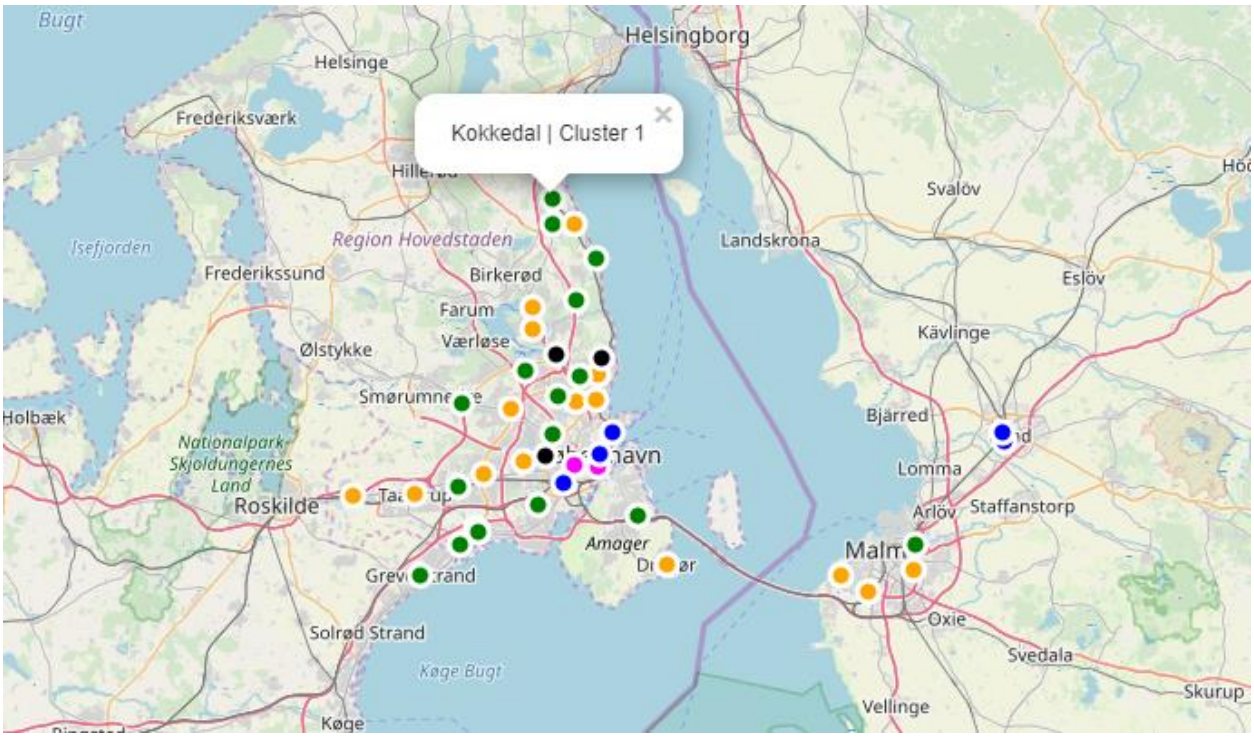
:

	Neighbourhood	Cluster Label	Art	Auto	BBQ Joint	Bagel Shop	Bakery	Bar	Bathing Area	Beach	Beer	Boat or Ferry	Book	Breakfast Spot	Brewery
0	Albertslund	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Bagsværd	1	0	0	1	0	1	0	0	0	0	0	0	0	0
2	Ballerup	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Brønshøj	1	0	0	0	0	2	1	0	0	0	0	0	0	0
4	Charlottenlund	2	0	0	0	0	3	2	0	0	0	0	0	0	0

Once we have this table, we append the latitude and longitude values for the neighbourhoods' so that we can plot them on a map. Assigning different colors for the cluster labels, **['black', 'green', 'orange', 'blue', 'magenta']** for ['label 0', 'label 1', 'label 2', 'label 3', 'label 4'], gives us the representation seen below.

Popup labels are defined in a way that they display the neighbourhood followed by its equivalent cluster label. An example of a popup is shown on the map.

With this map, we can find out which neighbourhoods' are similar and/or dissimilar. Notice there are fewer neighbourhoods' in the city of Malmö. It is because we had to drop out neighbourhoods' that had a small number of venue categories (from the bar chart).



Let us club all the similar cluster labels into one and sum up all its column values. That would end up with this table below. The number seen in each column is the total number of venue categories.

	Cluster Label	Art	Auto	BBQ Joint	Bagel Shop	Bakery	Bar	Bathing Area	Beach	Beer	Boat or Ferry	Book	Breakfast Spot	Brewery	Burger Joint
0	0	1	0	0	0	10	3	0	0	1	1	0	1	0	3
1	1	0	0	1	0	11	3	0	5	0	0	1	0	0	2
2	2	2	1	0	1	10	7	1	2	0	0	2	0	0	3
3	3	3	0	1	0	12	17	0	0	4	0	0	1	1	6
4	4	3	0	0	0	8	4	0	0	4	0	2	1	1	4

Upon further analysis, we can list out the top venues seen in each cluster type neighbourhood. This is done by sorting each row individually and getting the first ‘n’ column names (here ‘n’ is set to 14).

The most frequent venues seen in cluster 0 are:
Restaurant, Bakery, Supermarket, Park, Gym, Café, Grocery Store, Pizza, Train, Burger Joint, Bar, Plaza, Diner, Shopping Mall **(Set 0)**

The most frequent venues seen in cluster 1 are:
Grocery Store, Restaurant, Pizza, Park, Discount Store, Gym, Bakery, Supermarket, Café, Bus, Convenience Store, Train, Stadium, Shopping Mall **(Set 1)**
The most frequent venues seen in cluster 2 are:

Restaurant, Grocery Store, Gym, Supermarket, Café, Pizza, Train, Bakery, Electronics Store, Bus, Pharmacy, Hotel, Bar, Ice Cream Shop **(Set 2)**

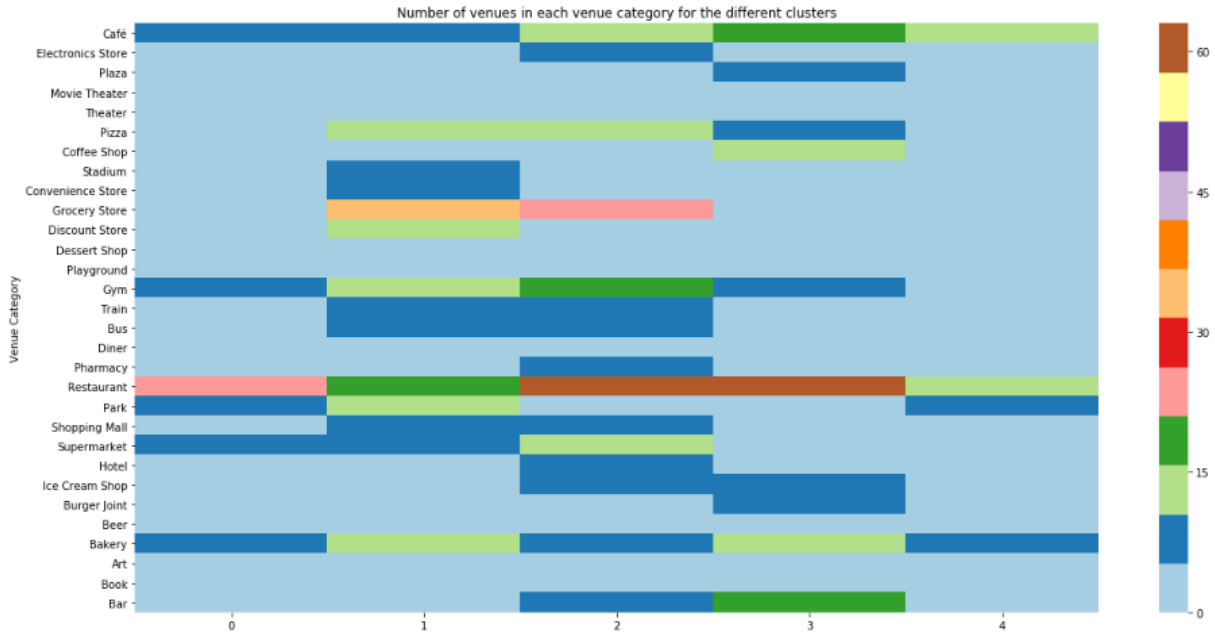
The most frequent venues seen in cluster 3 are:
Restaurant, Café, Bar, Coffee Shop, Bakery, Pizza, Gym, Ice Cream Shop, Plaza, Burger Joint, Park, Movie Theater, Beer, Supermarket **(Set 3)**

The most frequent venues seen in cluster 4 are:
Restaurant, Café, Bakery, Park, Bar, Burger Joint, Beer, Dessert Shop, Coffee Shop, Art, Book, Plaza, Playground, Theater **(Set 4)**

Lists like these are a good start to differentiate the clusters but it will be easier if it can be somehow visualized. Below, we show three ways to compare and visualize the 5 clusters.

Heatmap

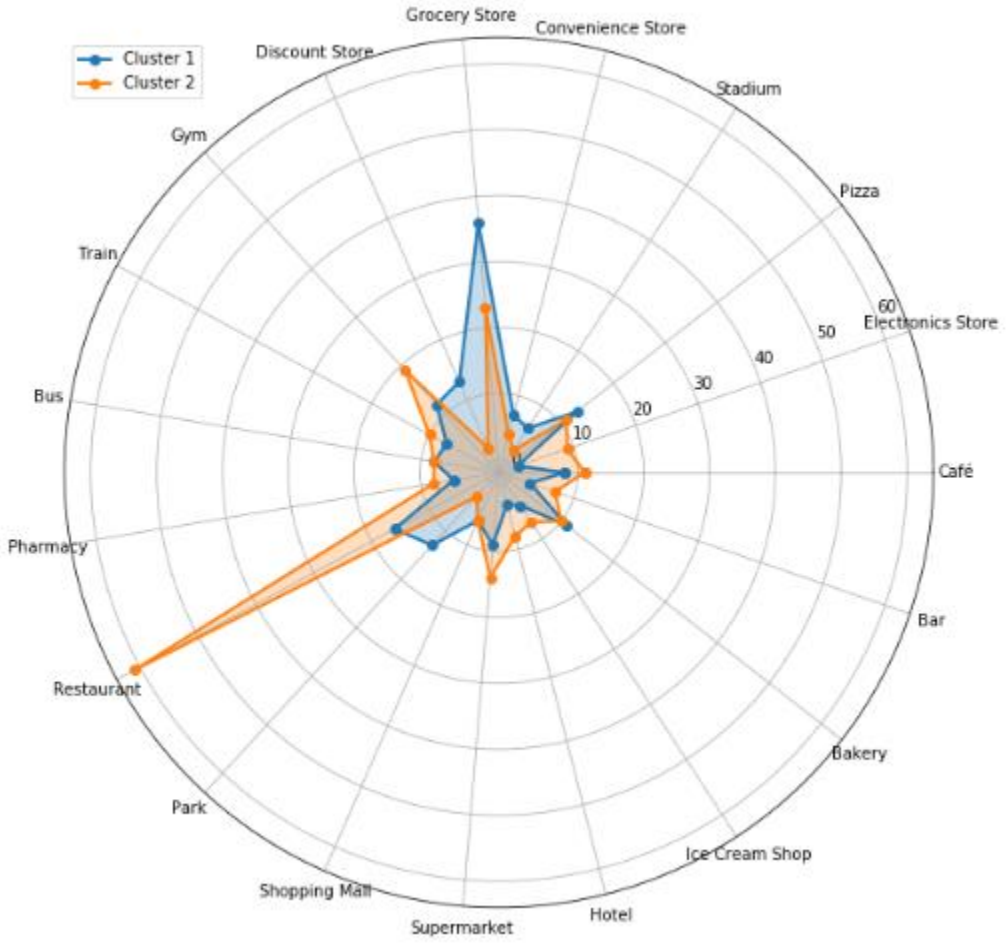
This is the only plot in this report that accounts for all the clusters. The colors help the reader to quickly find a specific venue count in any cluster. The union of sets 0 to 4 is found and used as the y-axis labels. Light blue color corresponds to the count between 0 and 5.



Instantly, it can be inferred that there are plenty of restaurants in cluster type 2 and 3. This is beneficial to the people as they have lots of choices to choose from. It could be a good reason to move into neighbourhoods' belonging to cluster 2 or 3. Anyone who wants to open a new restaurant should know better and not open one in these neighbourhoods' as it is already competitive enough.

Radar Chart

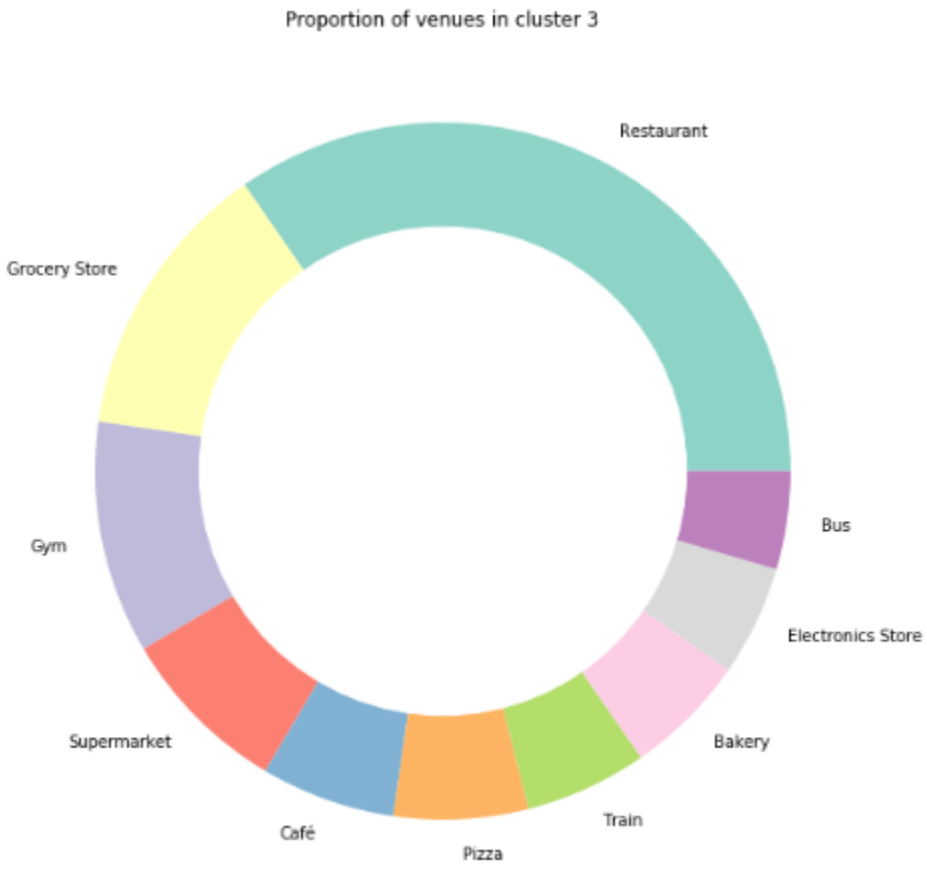
From the cluster count, we know the top two cluster labels are 1 and 2. These two clusters are differentiated in the below chart. Plotting more than two clusters would be a mess and might confuse the reader. The union of sets 1 and 2 is found and used as the labels on the circumference. The inner radial circles represent the venue category counts.



Apart from the obvious (restaurants in cluster 2), we see that there is a good amount of grocery stores in cluster 1. This is an attractive feature of cluster 1 as people will find it convenient to have many grocery stores around their house. The number of electronic stores, bars, ice cream shops, hotels, pharmacies in cluster 1 are all low when compared to cluster 2. So maybe starting a business in one of these mentioned categories in neighbourhoods' belonging to cluster 1 wouldn't be a bad idea.

Donut Plot

This plot is useful if we must analyze one particular cluster type separately.



4) Discussion

From the observations made from the results, I strongly recommend people to choose neighbourhoods' from cluster 2 if they are looking to move into a new place. These neighbourhoods' have the greatest number of different venue categories. I also recommend neighbourhoods' from cluster 4 for aspiring entrepreneurs to start a profitable business. Cluster 4 has the least number of venues.

5) Conclusion

This assignment's objective was to provide meaningful insights to help make a person or business on their decision. I thereby conclude that it has indeed done that by comparing the advantages/disadvantages of different neighbourhoods'.