Similar to Odenplan?

Introduction

Different cities in the world have perfect venues that you cannot live without. In this project we will help people who are living in Odenplan, a central city part of Stockholm, to find a new home with similar features as Odenplan in Stockholm city. Despite of having dissimilarities of taste, we will look for city parts in Stockholm that looks like Odenplan. By clustering different neighborhood according to venue category may give insights on what defines the Norrmalm, the district of Odenplan. Furthermore, measure quaintly by venues that show how similar city parts are to Odenplan might serve as a variable that can help people to make a decision when they are trying to explore new residence in Stockholm.

Problem

Identify neighborhoods in Stockholm city that gives a good perception of similar neighborhoods to Odenplan.

Data gathering

The data on the most common Stockholm areas are scraped from Stockholm stads website. Google maps geocoder API was used to obtain the latitude and longitude values for each city part of interest, which then could be used with Foursquare API to collect the closest venues (supermarket, restaurant, park, etc.). The data was pre-processed and inserted into a dataframe, see table 1.

	City parts	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bandhagen	59.267846	18.051589	Bandängens parklek	59.266979	18.048587	Playground
1	Bandhagen	59.267846	18.051589	Café Chateau	59.270393	18.048338	Café
2	Bandhagen	59.267846	18.051589	Anjas Vedugn	59.266567	18.052061	Pizza Place
3	Bandhagen	59.267846	18.051589	ICA Supermarket	59.263242	18.043206	Grocery Store
4	Bandhagen	59.267846	18.051589	Fitness24Seven	59.263750	18.041561	Gym / Fitness Center
...	...	...	...	...	...	...	...

Tablel 1. The table shows the pre-processed dataframe which was obtained from the callbacks of the APIs

The dataframe size is 5468 rows × 7 columns and was further cleaned by removing NaN-values and outliers.

Feature selection

Foursquare API gives very rich information of venues, for instance we can see from the dataframe that each Venue Category seems pretty accurate. The venue Category for each city part was obtained One-Hot-encoded and grouped by city part. By taking the mean value on the the transformed dataframe a densification-factor was obtained per venue category, see table 2.

	City parts	Accessories Store	Advertising Agency	Airport Gate	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Antique Shop	Aquarium	...	Train Station	Tram Station	Turkish Restaurant
0	Abrahamsberg	0.0	0.0	0.000000	0.000000	0.000000	0.00	0.0	0.0	0.0	...	0.0	0.000000	0.0
1	Akalla	0.0	0.0	0.000000	0.000000	0.000000	0.00	0.0	0.0	0.0	...	0.0	0.000000	0.0
2	Alvik	0.0	0.0	0.000000	0.000000	0.000000	0.00	0.0	0.0	0.0	...	0.0	0.032258	0.0
3	Annedal	0.0	0.0	0.015873	0.047619	0.031746	0.00	0.0	0.0	0.0	...	0.0	0.015873	0.0
4	Aspudden	0.0	0.0	0.000000	0.000000	0.000000	0.00	0.0	0.0	0.0	...	0.0	0.000000	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
162	Örby	0.0	0.0	0.000000	0.000000	0.000000	0.00	0.0	0.0	0.0	...	0.0	0.000000	0.0
163	Örby Slott	0.0	0.0	0.000000	0.000000	0.000000	0.00	0.0	0.0	0.0	...	0.0	0.000000	0.0
164	Östberga	0.0	0.0	0.000000	0.000000	0.000000	0.00	0.0	0.0	0.0	...	0.0	0.000000	0.0
165	Östermalm	0.0	0.0	0.000000	0.000000	0.000000	0.01	0.0	0.0	0.0	...	0.0	0.000000	0.0
166	Östermalmstorg	0.0	0.0	0.000000	0.000000	0.000000	0.01	0.0	0.0	0.0	...	0.0	0.000000	0.0

167 rows × 288 columns

Table 2. The table shows a density-factor for each venue category

After the data-preperation a K-means clustering was performed to cluster the city parts into areas data gives insights of Odenplan.

Methodolgy

K-means is unsupervised cluster analysis which used to find groups from data which have not been explicitly labeled. The K-means algorithm tries cluster data by separating samples in n groups by the density features from the dataframe, see table 2. The elbow method was used to determine the number of clusters for the data set. Basically, the “elbow” is when the gradient change from the distortion score (from different Ks) is small, i.e the line chart will resemble an elbow of an arm.

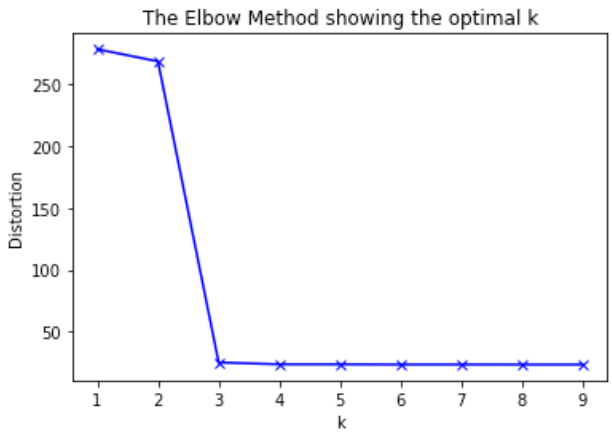

Chart 1. The graph shows that the optimal value for k = 3

This gives a good indication of an inflection point on the curve which shows that the underlying model fits best at that point. Chart 1 shows that K=3 was optimal for this case which means that there will be 3 clusters in total.

A new dataframe is created to get an overview of the 10 most common venues at each city part, see table 3

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
Gym / Fitness Center	Grocery Store	Convenience Store	Pizza Place	Playground	Burger Joint	Mediterranean Restaurant	Food & Drink Shop	Metro Station	Pharmacy
Stadium	Thai Restaurant	Bus Station	Light Rail Station	Park	Nightclub	Bakery	Gym / Fitness Center	Scenic Lookout	Restaurant
Bus Station	Supermarket	Metro Station	Bakery	Pizza Place	Thai Restaurant	Furniture / Home Store	Stadium	Stables	Athletics & Sports
Pizza Place	Metro Station	Bakery	Scandinavian Restaurant	Cemetery	Grocery Store	Indian Restaurant	Sushi Restaurant	Italian Restaurant	Park
Soccer Field	Tennis Stadium	Gym / Fitness Center	Business Service	Supermarket	Restaurant	Shopping Mall	Food & Drink Shop	Grocery Store	Gym Pool

Table 3. The table show common venues in different city parts

Table 4 show the difference between most common venues between the different clusters

Norra Ängby	0	Pizza Place	Metro Station	Soccer Field	Plaza	Bus Stop	Food & Drink Shop	Bakery	Asian Restaurant	Convenience Store	Chinese Restaurant
Olovslund	0	Pizza Place	Tram Station	Soccer Field	Sushi Restaurant	Grocery Store	Thai Restaurant	Plaza	Concert Hall	Escape Room	Dog Run
Vasastan	1	Italian Restaurant	Sushi Restaurant	Pizza Place	Café	Bakery	Scandinavian Restaurant	Indian Restaurant	French Restaurant	Ice Cream Shop	Coffee Shop
Hötorget	1	Hotel	Scandinavian Restaurant	Café	Burger Joint	Gym / Fitness Center	Salad Place	Coffee Shop	Clothing Store	Bakery	Pizza Place
Odenplan	1	Scandinavian Restaurant	Café	Bakery	Middle Eastern Restaurant	Coffee Shop	Pizza Place	Indian Restaurant	Sushi Restaurant	Italian Restaurant	Burger Joint
Kista	2	Convenience Store	Metro Station	Bus Stop	Shopping Mall	Mattress Store	Zoo Exhibit	Event Space	Eastern European Restaurant	Electronics Store	Escape Room
Rinkeby	2	Convenience Store	Metro Station	Bus Stop	Shopping Mall	Mattress Store	Zoo Exhibit	Event Space	Eastern European Restaurant	Electronics Store	Escape Room

Table 4. The table show common venues in different city parts

Folium maps is used to obtain a visual perception of how the different clusters look on a map of Stockholm, see figure 1. The most notable is that central parts of Stockholm, which has cluster label 1 (blue) are pretty accurate clustered with other central city parts.

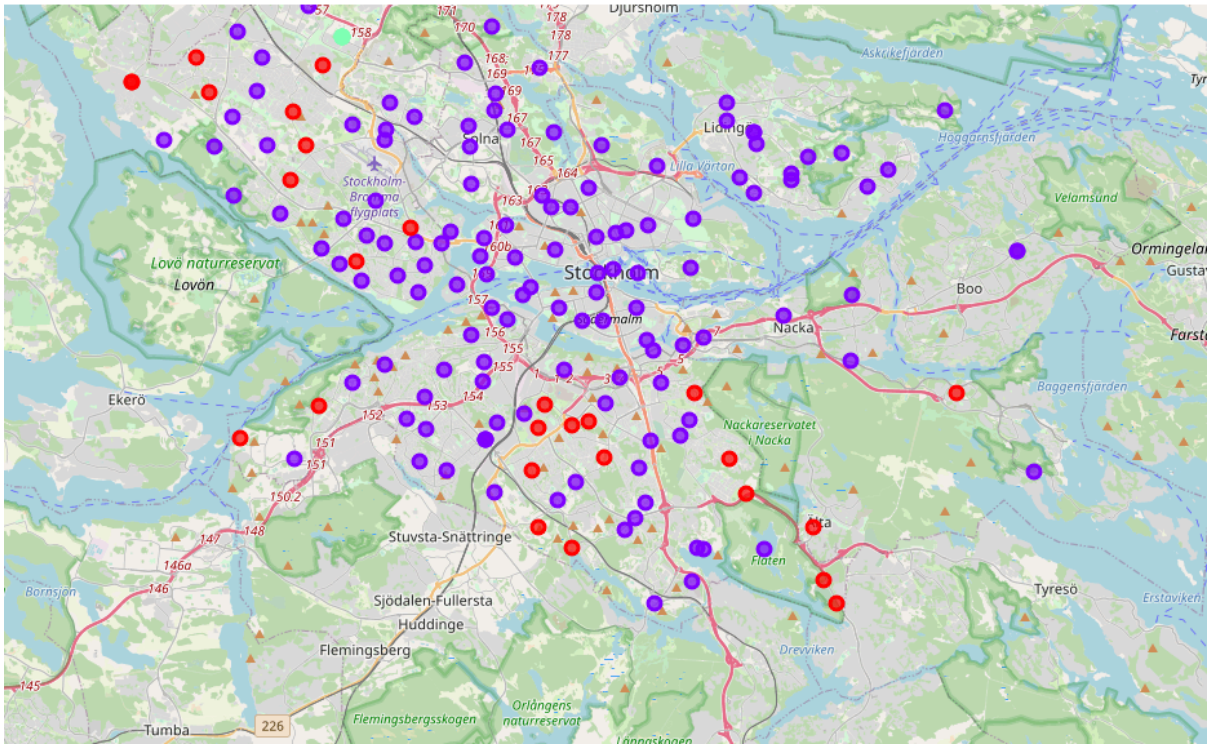

Figure 1. The table show common venues in different city parts

There are various algorithms to help to evaluate the performance of the clustering algorithm and in this case was adjusted rand index used. Rand index computes a similarity measure between two clusters, in this case → all city parts are compared with Odenplan. Perfect similarity would give a score of 1 and bad labelling or independent labelling Is scored 0 or a negative score to -1.

	City parts	score_rand		City parts	score_rand
1	Akalla	-0.028129	**112**	Sibirien	0.695401
126	Spängadalen	-0.028129	**146**	Vasastan	0.599132
71	Kista	-0.028129	**43**	Hagastaden	0.574900
107	Rinkeby	-0.028129	**73**	Kungsholmen	0.490694
103	Orhem	-0.022882	**72**	Kristineberg	0.398062
163	Örby Slott	-0.010294	**88**	Marieberg	0.397491
44	Hagsätra	-0.002479	**47**	Hamnvakten	0.386930
111	Saltsjö-Duvnäs	-0.002332	**137**	Södermalm	0.368120
28	Fisksätra	0.000642	**70**	Katarina-sofia	0.349404
76	Käppala	0.001611	**127**	Stadshagen	0.347872

Table 4. The table adjusted rand index.

The results from table 4 show that city parts in the suburbs likely, e.g. Rinkeby, are very low scored and city parts that are very central, e.g Kungsholmen, are scored higher.

Results

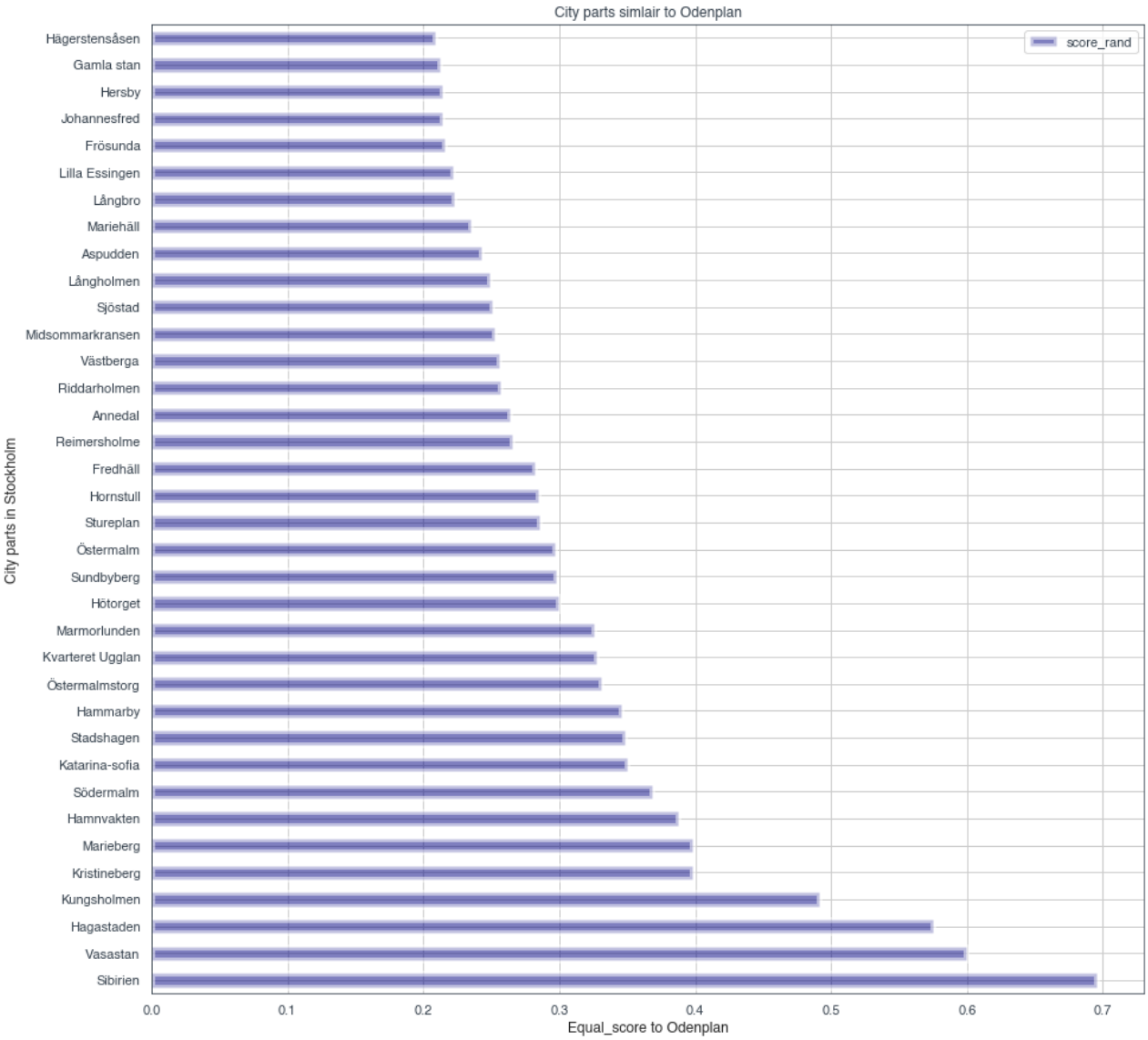

Figure 2. The figure shows city parts that are simlair to Odenplan based on adjusted rang index

Discussion

K-means is very simple clustering algorithm that necessary dosent converge to the most optimal result and there are many other different unsupervised clustering algorithms that probably would achieve better results. The project was performed with a selected data set of different popular city parts with 286 different features. However, have more data point of the city and different features could lead to totally different result.

Conclusion

The results from figure 2 are very promising and if you ask me who lives in Stockholm can say that the adjusted rand index is pretty accurate when it comes to similarity of city parts.