Helsingborg City Data

Data published by the city of Helsingborg is used for this project. It is hard to get a sense of raw data. The data consist of geospatial locations of different business point in different categories like restaurants, cafe, etc from Helsingborg. Preprocessing of data is done using Python libraries. K-Means clustering is performed on this data to see the patterns and the results are visualized using Folium map interface. optimal k value for k-means is chosen using the Elbow method. The data is so complex that it is really hard to see the optimal value from the Elbow method. Later in this project foursquare API is used to fetch data for the same postal codes and clustered data using the above methodology and the results are shown below.

```
import numpy as np
import pandas as pd
import json
import matplotlib as mlt
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import matplotlib.colors as colors
from sklearn.cluster import KMeans
```

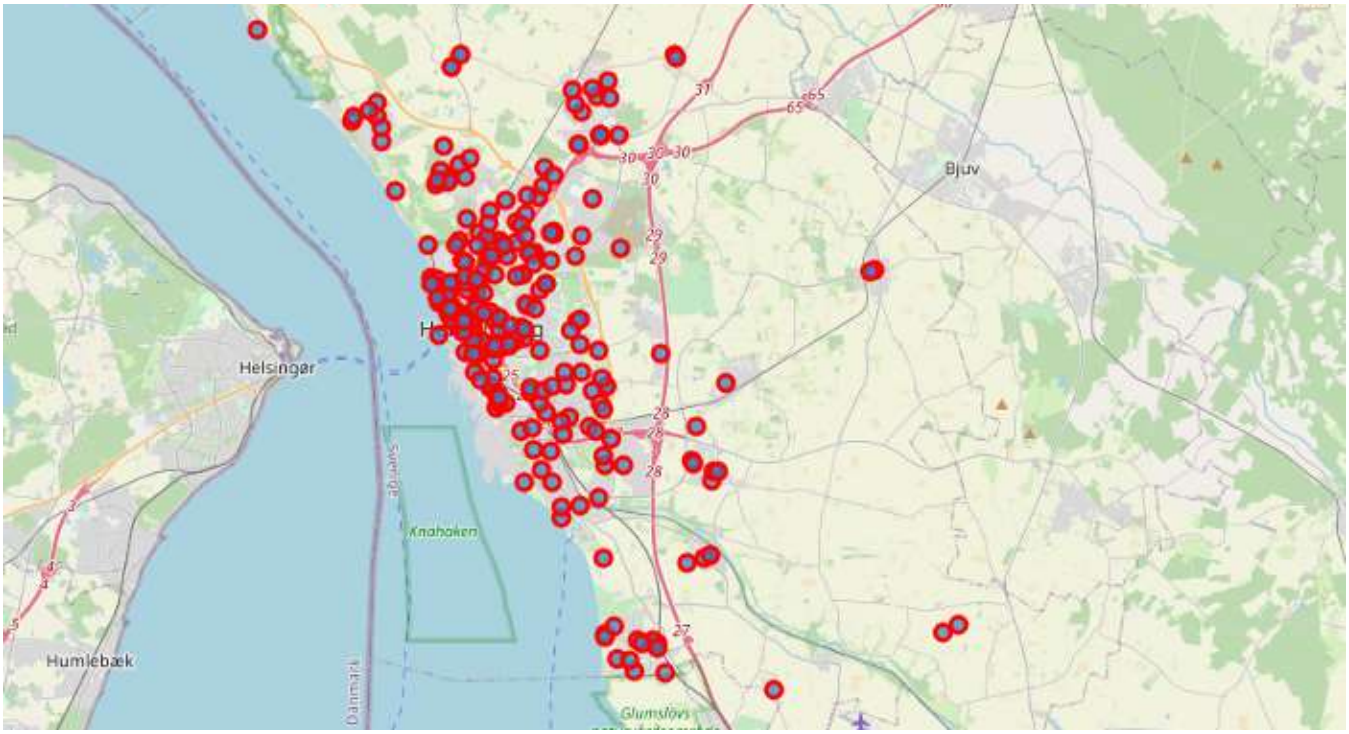
Python libraries used in this project

Sample data

	CITY	POSTCODE	geo_shape	geo_point_2d	area
0	Helsingborg	25482	{"type": "Polygon", "coordinates": [[[12.65947...	56.0894565738,12.6568019271	Hittarp - Laröd
1	Helsingborg	25361	{"type": "Polygon", "coordinates": [[[12.74727...	56.0132736825,12.7530666401	Ättekulla

KATEGORI	NAMN	ADRESS	HEMSIDA	Kategori webb	Exportdatum	geo_shape	geo_point_2d	area	postalcode
Butiker	Borgmans Fisk Eftr	Kopparmöllegatan 17	http://www.borgmansfisk.net/	Butiker	2020-07-16	{"type": "Point", "coordinates": [12.697555275...	56.0534616304,12.6975552755	Slottshöjden	25435
Butiker	Polshop	Södergatan 16	http://www.polshop.se/	Butiker	2020-07-16	{"type": "Point", "coordinates": [12.701065141...	56.0421044698,12.7010651416	Söder	25225

Data that we are interested in is Geospatial data, Category of each entry, name.



Data before clustering

One hot encoding is performed on categorical variables to feed it to the clustering algorithm

Butiker	Caféer	Personalrestauranger	Restauranger	Skolrestauranger	Snabbmat	Vårdverksamheter
1	0	0	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0

Data is grouped over postal codes and mean is performed at the grouped level

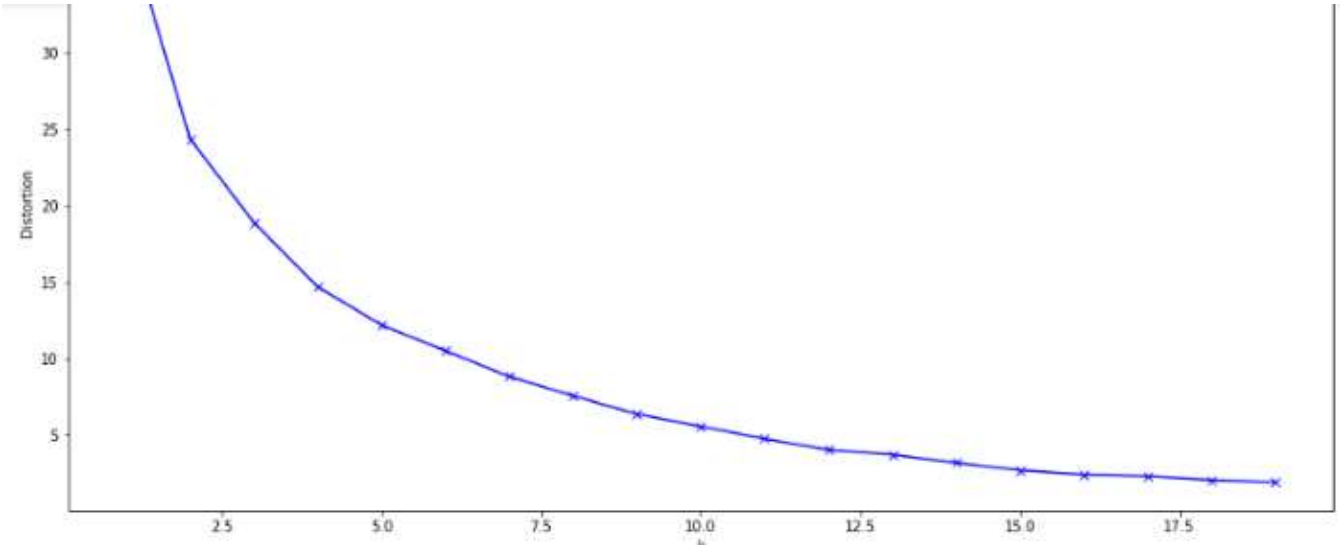
postalcode	Butiker	Caféer	Personalrestauranger	Restauranger	Skolrestauranger	Snabbmat	Vårdverksamheter
25220	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25221	0.000000	0.250000	0.000000	0.583333	0.083333	0.083333	0.000000
25222	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25223	0.090909	0.090909	0.090909	0.272727	0.272727	0.181818	0.000000
25224	0.000000	0.000000	0.000000	0.777778	0.000000	0.111111	0.111111

the top places for each postal code are performed

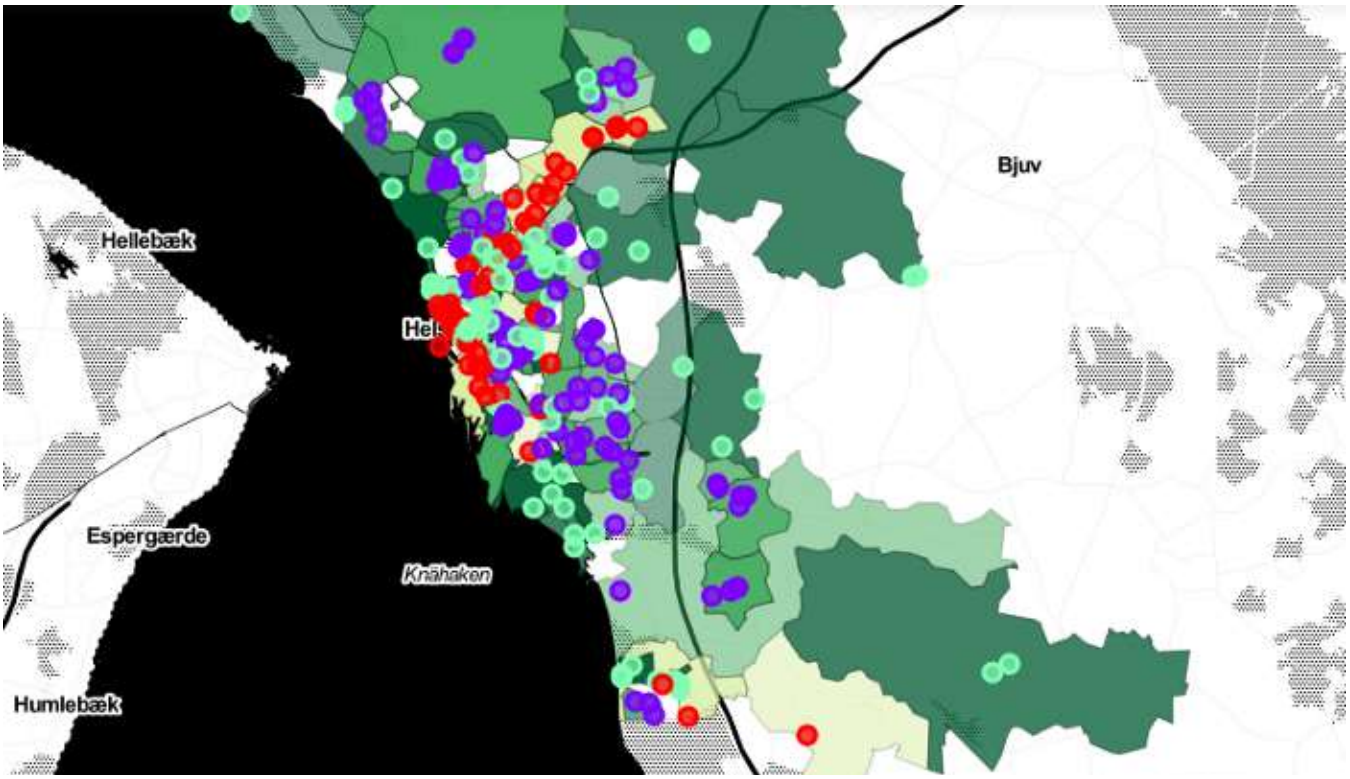
postalcode	1st	2nd	3rd	4th	5th
25220	Restauranger	Vårdverksamheter	Snabbmat	Skolrestauranger	Personalrestauranger
25221	Restauranger	Caféer	Snabbmat	Skolrestauranger	Vårdverksamheter
25222	Caféer	Vårdverksamheter	Snabbmat	Skolrestauranger	Restauranger
25223	Skolrestauranger	Restauranger	Snabbmat	Personalrestauranger	Caféer
25224	Restauranger	Vårdverksamheter	Snabbmat	Skolrestauranger	Personalrestauranger

Find optimal k value by The Elbow Method

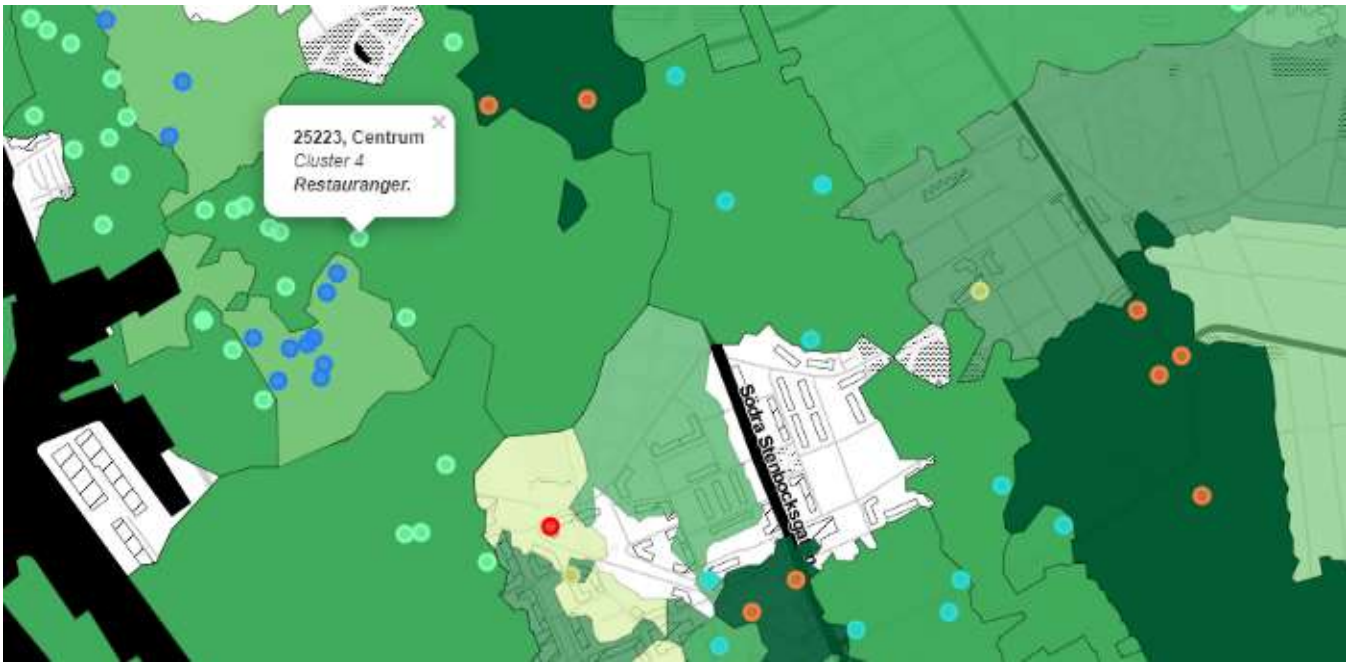
It was difficult to choose an optimal value for this dataset



3 Clusters data is visualized by Folium maps



. . .



. . .

Foursquare API is used to fetch venues for each of the above-used postal codes. Data is processed in the same way as the above and the k-means clustering algorithm is used to get patterns from the data.

POSTCODE	PostalCode Latitude	PostalCode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
25482	56.089457	12.656802	Sofierokiosken	56.087651	12.660768	Candy Store
25482	56.089457	12.656802	La Pizza	56.091222	12.654753	Pizza Place
25482	56.089457	12.656802	Hemköp Laröd	56.091335	12.654943	Grocery Store
25482	56.089457	12.656802	Chateau de Samsailles	56.087559	12.652621	Playground
25361	56.013274	12.753067	City Gross Ättekulla	56.016197	12.754077	Supermarket

Onehot encoding categorical data

POSTCODE	American Restaurant	Art Museum	Asian Restaurant	Athletics & Sports	Auto Dealership	Auto Garage	Bagel Shop	Bakery	Bar	Beach	Beach Bar	Beer Bar	Bistro	Bookstore	Bowling Alley	Brewe
25482	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25482	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25482	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25482	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25361	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

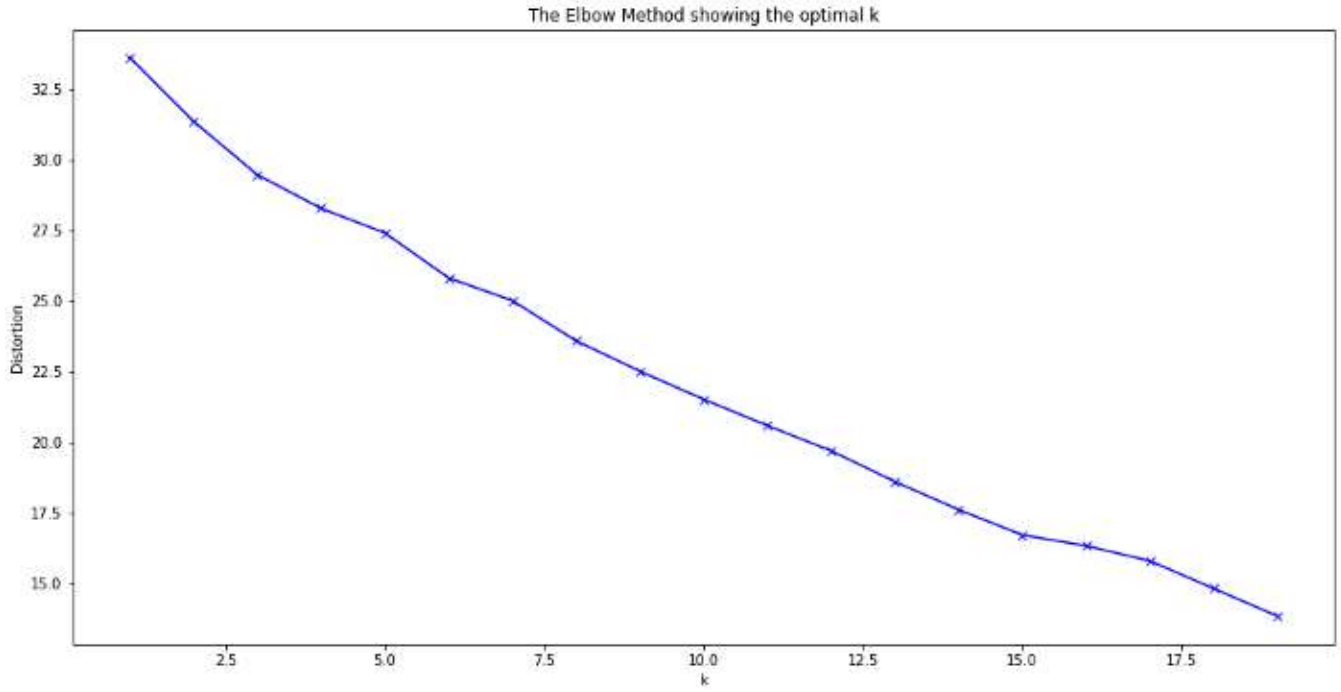
Data grouped on postal code and mean is performed on group-level data.

POSTCODE	American Restaurant	Art Museum	Asian Restaurant	Athletics & Sports	Auto Dealership	Auto Garage	Bagel Shop	Bakery	Bar	Beach	Beach Bar	Beer Bar	Bistro	Bookstore
25220	0.0	0.020833	0.0	0.0	0.0	0.0	0.0	0.020833	0.041667	0.0	0.0	0.020833	0.020833	0.020833
25221	0.0	0.025641	0.0	0.0	0.0	0.0	0.0	0.000000	0.051282	0.0	0.0	0.025641	0.025641	0.025641

Most common business places based on the data from Foursquare API for each postal code of Helsingborg

POSTCODE	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
25220	Hotel	Scandinavian Restaurant	Restaurant	Café	Coffee Shop	Burger Joint	Ice Cream Shop	Salad Place	Bar	Italian Restaurant
25221	Scandinavian Restaurant	Restaurant	Pizza Place	Bar	Burger Joint	Hotel	Ice Cream Shop	Salad Place	Nightclub	Café
25222	Park	Scandinavian Restaurant	Italian Restaurant	Grocery Store	Playground	Gym	Deli / Bodega	Pizza Place	Plaza	Coffee Shop
25223	Coffee Shop	Hotel	Café	Restaurant	Italian Restaurant	Park	Asian Restaurant	Pub	Scandinavian Restaurant	Thai Restaurant
25224	Hotel	Restaurant	Coffee Shop	Bar	Café	Thai Restaurant	Scandinavian Restaurant	Asian Restaurant	Pub	Italian Restaurant

picking an optimal k-value was not easy for this dataset

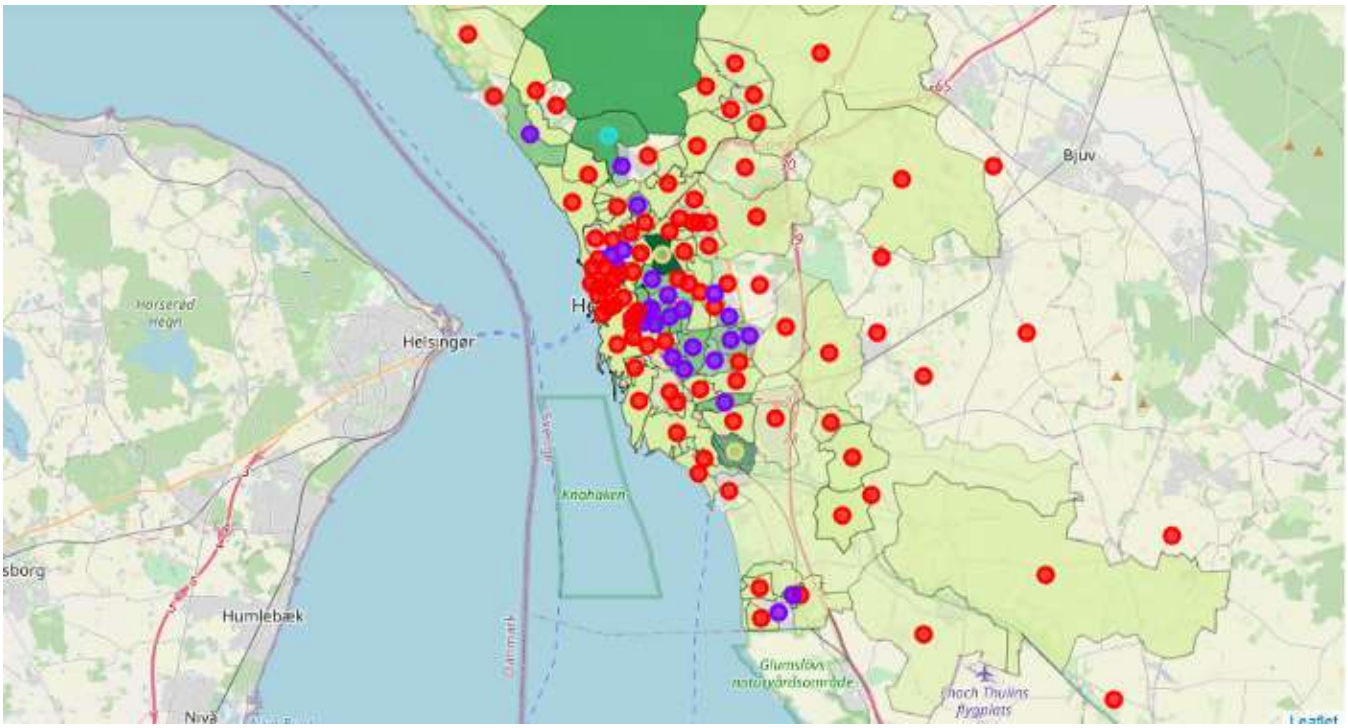


Finally, data with clusters are merged

POSTCODE	geo_shape	geo_point_2d	area	type	coordinates	polygon_coordinates	Cluster Labels	1st Most Common Venue	2nd Most Common Venue

25482	{"type": "Polygon", "coordinates": [[[12.65947...	56.0894565738,12.6568019271	Hittarp - Larod	Polygon	[[[12.65947745564023, 56.08065244077402], [12....	[(12.65947745564023, 56.08065244077402), (12.6...	1	Playground	Pizza Place
25361	{"type": "Polygon", "coordinates": [[[12.74727...	56.0132736825,12.7530666401	Ättekulla	Polygon	[[[12.747272091715816, 56.01611001498235], [12...	[(12.747272091715816, 56.01611001498235), (12....	0	Brewery	Supermarket
25465	{"type": "Polygon", "coordinates": [[[12.72806...	56.0721772539,12.7345182361	Dalhem	Polygon	[[[12.728065859762449, 56.07223268630626], [12...	[(12.728065859762449, 56.07223268630626), (12....	0	Playground	Shopping Mall
25665	{"type": "Polygon", "coordinates": [[[12.74776...	56.0499416462,12.7499930315	Adolfsberg	Polygon	[[[12.747764472346708, 56.04706467772815], [12...	[(12.747764472346708, 56.04706467772815), (12....	0	Business Service	Athletics & Sports
25244	{"type": "MultiPolygon", "coordinates": [[[[12...	56.0413191382,12.7085138805	Eneborg	MultiPolygon	[[[[12.709012002277971, 56.04312887626643], [1...	[[(12.709012002277971, 56.04312887626643), (12...	0	Bus Stop	Pizza Place

4 clusters with the Foursquare data



Conclusions:

Data preprocessing is done in Python and prepared data for k-means clustering to discover patterns from complex unlabeled data and the results are visualized using the folium maps. If we see the clustered data from the first experiment, red circle clusters are the areas where the main city centre is located and most of the popular restaurants, supermarkets are located and also most populated areas. Foursquare API data clearly lacks distinguishable clusters, require more data and good features to perform better in discovering patterns.