Venues & House Price Data Analysis of Stockholm

Zhayida Simayijiang

2020-03-16

1. Introduction

1.1 Background

The entire Stockholm metropolitan area consisting of 26 municipalities and has a population of over 2.2 million making it the most populous city in the Nordic region. It has a population density of 5.200 people per square kilometer [1].

1.2 Problem

Stockholm is a city with a high population and population density in Sweden, and living cost is comparably higher than other cities of Sweden as well. Newcomers of the city may want to choose the area where housing price is lower but still not too far away from social places.

1.3 Interest

In this project, I aim to describe different municipalities of Stockholm by studying surrounding available venues. I create a map and information chart where the real estate index is placed on Stockholm and each municipality is clustered according to the venue density. The result may useful for people who are moving to Stockholm and wanted to find a place which is a match one's economy as well as lifestyle and hobbies.

2. Data acquisition and cleaning

What and where to collect the data are essential step in this project.

- I download the excel file from Svensk Mäklarstatistik [2]. The file has average house price of all cities of Sweden. I cleaned the data and reduced it to city of Stockholm.
- I used Google [4] to get the equivalent latitude and longitude values of each municipality in Stockholm.
- I used **Foursquare API** [3] to explore neighborhood venues in Stockholm.

After little manipulation the obtained dataframe looks as below (only the head of the data are shown here):

Municipality	Average price	Price Label	Latitude	Longitude
Upplands Väsby	4638	4	59.5196	17.9283
Vallentuna	4459	4	59.5357	18.078
Österåker	4421	4	59.4818	18.2979
Värmdö	4730	4	59.3164	18.4466
Järfälla	4883	4	59.4101	17.8368

Table 1. Data set after cleaning. Here "Average price" refers to average housing price; Price Label 1: when the price in the range [1000, 2000]; Price Label 2: when the price in the range [2000, 3000] and so on, Price Label 11: when the price in the range [11000, 12000].

The data set has information about longitude and latitude, then we can visualize each municipality geographical location on the map of Stockholm (see Figure 1) using python **folium** library. There are 26 municipalities in total and all the municipalities are used for data analysis.

Figure 1: Map of Stockholm with municipalities marked in dark blue circle

3. Explore Neighborhoods in Stockholm

I explore the surrounding neighborhood of each municipality using **Foursquare API**. Since there is a limit to Foursquare usage, I set the venue limit to **100 venues** that are within a 2km radius of each municipality. After using Foursquare lab, the result is then cleaned up from JSON to a structured pandas dataframe as shown below:

	name	categories	lat	lng
0	Science Fiction Bokhandeln	Bookstore	59.324047	18.070682
1	Tweed	Cocktail Bar	59.324471	18.067696
2	The Burgundy	Wine Bar	59.324434	18.068161
3	Stortorget	Plaza	59.324973	18.070727
4	Corner Club	Cocktail Bar	59.323342	18.069431

Table 2. Nearby venues

Overall, foursquare returned 917 venues and 165 unique categories curated from all the returned venues. Below Table 3 is the merged table of municipality and its venues:

	Municipality	Municipality Latitude	Municipality Longitude	Venue	Venue Latitude	Venue Longitude	Category
0	Upplands Väsby	59.5196	17.9283	Sibylla Mammas	59.513921	17.929729	Fast Food Restaurant
1	Upplands Väsby	59.5196	17.9283	Vilundabadet	59.517178	17.919924	Gym / Fitness Center
2	Upplands Väsby	59.5196	17.9283	Circle K	59.513968	17.919595	Convenience Store
3	Upplands Väsby	59.5196	17.9283	ICA Kvantum	59.518821	17.913883	Grocery Store
4	Upplands Väsby	59.5196	17.9283	Scandic Upplands Väsby	59.515610	17.922886	Hotel

Table 3. Merged dataframe

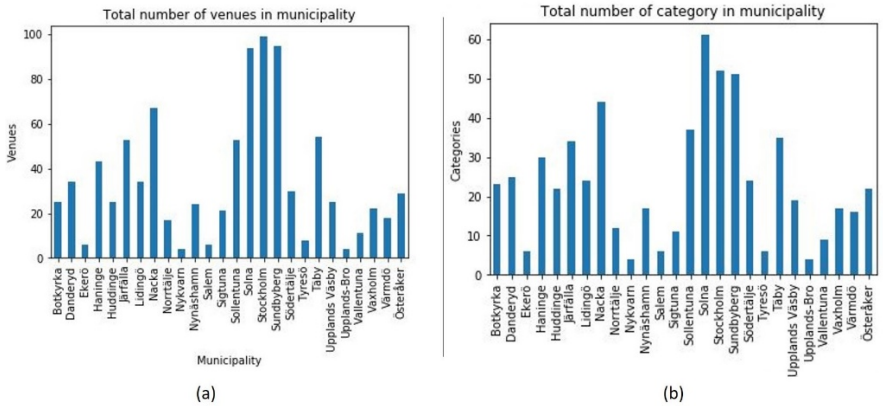

Figure 2. (a) the total number of venues and (b) the total number of categories in each municipality.

From Figure 2, we can see that the total number of venues returned for Stockholm municipality is the highest, but the total number of unique categories that venues belong to is highest in Solna municipality. One can understand that Solna municipality has a variety of venues.

	Category	Frequency
0	Grocery Store	50
1	Café	41
2	Pizza Place	37
3	Scandinavian Restaurant	37
4	Hotel	33
5	Gym / Fitness Center	32
6	Supermarket	32
7	Shopping Mall	23
8	Fast Food Restaurant	22
9	Park	22

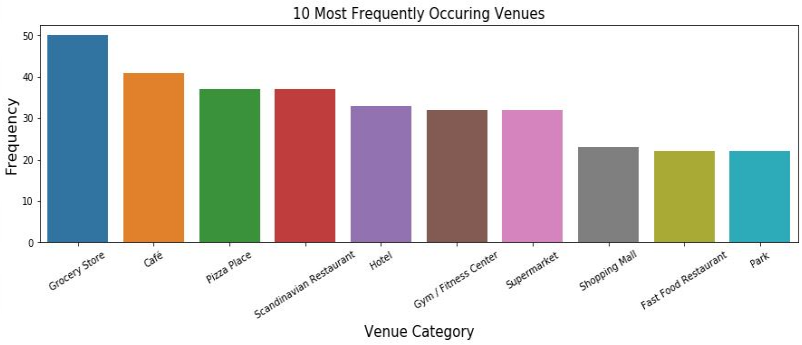

Figure 3. The 10 most frequent occurring venue category

We can see from Figure 3 that Grocery Store is one of the most frequent venues in Stockholm and on the other hand Fast Food Restaurant is a less occurred venue. This is good news for one who wants to have a healthy lifestyle. In the list, the park is also the last one, but no worry one can find the park-like area if one slightly out of the city center.

4. Analyze Each Neighborhood

As shown in Figure 2, some of the municipalities have too few venues. In order to get reliable classification result, I removed the municipalities which have less than 10 venues for further analysis. But I think they may someone who likes to live in such an area with much fewer venues, so I give cluster label for those 6 municipalities as the calm municipality. After removing venues of those calm municipalities from the data, there are 878 venues left.

One hot encoding was performed for preparing the dataset for clustering and deal with venue categories. New dataframe is created and in Table 3 top 10 venues for each municipality are displayed.

	Municipality	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	Botkyrka	Metro Station	Fast Food Restaurant	Department Store	Supermarket	Food	Café	Shopping Mall	Soccer Field	Electronics Store	Pizza Place
1	Danderyd	Bus Station	Hotel	Café	Burger Joint	Park	Golf Course	Light Rail Station	Pizza Place	Supermarket	Breakfast Spot
2	Haninge	Electronics Store	Bus Station	Supermarket	Coffee Shop	Bowling Alley	Convenience Store	Department Store	Shopping Mall	Gym / Fitness Center	Furniture / Home Store
3	Huddinge	Bakery	Gym / Fitness Center	Park	Fast Food Restaurant	Moving Target	Pizza Place	Furniture / Home Store	Farm	Stadium	Liquor Store
4	Järfälla	Sporting Goods Shop	Grocery Store	Supermarket	Fast Food Restaurant	Hotel	Furniture / Home Store	Electronics Store	Department Store	Pizza Place	Plaza

Table 3. The 10 most common venues of each municipality.

5. Clustering

5.1 Finding the Optimal Number of Clusters

For find the optimal number of clusters k, I tried two different method. The Figure 4 shows that, we know that the elbow method did not give us good answer since there are not clear bend. Then the silhouette method used to find the optimal number of clusters k. From Table 4, we learn that optimal k = 3 since it has the highest silhouette score.

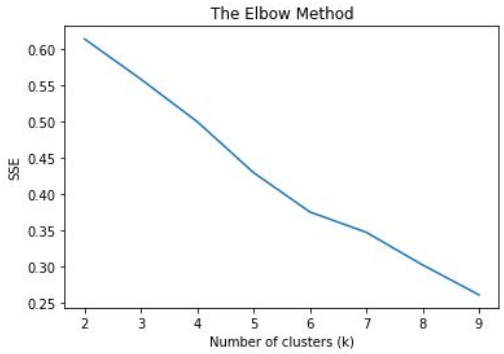

Figure 4. The Elbow Method

Number of clusters k	Silhouette score
2	0.1907
3	0.2317
4	0.1613
5	0.0301
6	0.1381
7	0.0095
8	0.0908
9	0.0277

Table 4. Silhouette score

5.2 K-Means Clustering

I used the K-means clustering algorithm to cluster the municipalities in the Stockholm metropolitan area. K-mean clustering is one of the most common clustering methods. The Stockholm municipalities divided into three clusters as shown in Figure 5 and in Table 5, one can see the number of venues and unique categories of each municipality.

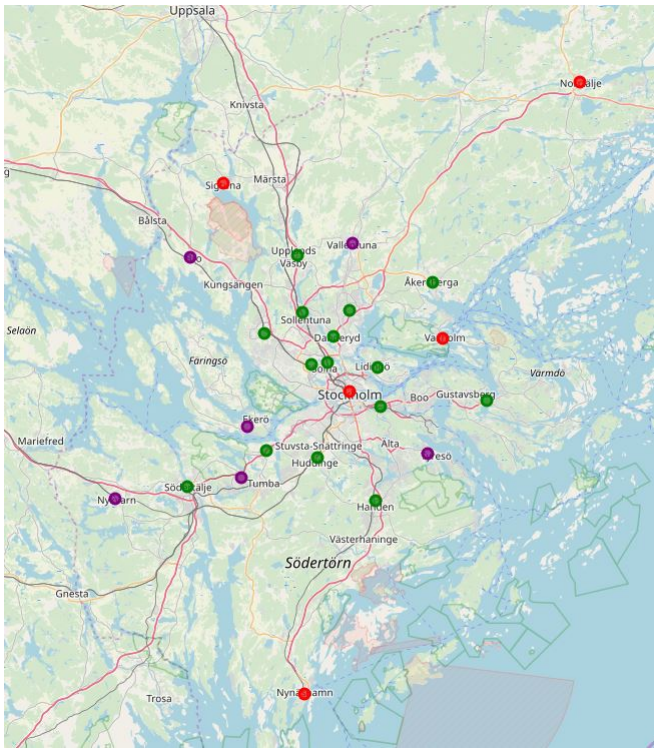

Figure 5. Map of Stockholm with municipality clustered to 3 labels.

Class Label	Municipality	Nr. of venues	Nr. of categories
Cluster 0	Stockholm	100	52
	Vaxholm	22	17
	Norrtälje	17	12
	Sigtuna	21	11
	Nynäshamn	24	17
Cluster 1	Upplands Väsby	25	19
	Österåker	29	22
	Värmdö	18	16
	Järfälla	55	34
	Huddinge	25	22

	Botkyrka	25	23
	Haninge	43	30
	Täby	56	35
	Danderyd	34	25
	Sollentuna	54	37
	Södertälje	30	24
	Nacka	69	44
	Sundbyberg	100	51
	Solna	96	61
	Lidingö	35	24
Cluster 2	Vallentuna	11	9
	Ekerö	6	6
	Salem	6	6
	Tryesö	8	6
	Upplands-Bro	4	4
	Nykvarn	4	4

Table 5. Number of venues and categories in each cluster label

6. Results and Discussion

If we group the **1st Most Common Venue** in Table 3, we get Figure 6. As I mentioned in above section, Cluster 2 is predefined according to number of venues which are few than 11 venues.

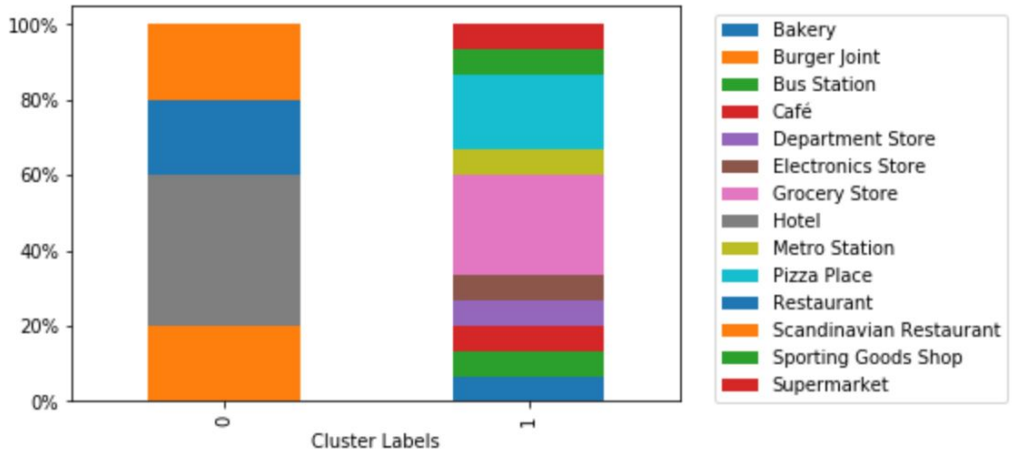

Figure 6. Stacked bar plot with normalized to 100%

Further we can label each cluster as follows:

- Cluster 0: Hotel & Restaurant
- Cluster 1: Multiple Social Venues
- Cluster 2: Calm area

From Figure 7, we can see that most common average price range is between 4000 to 5000 SEK.

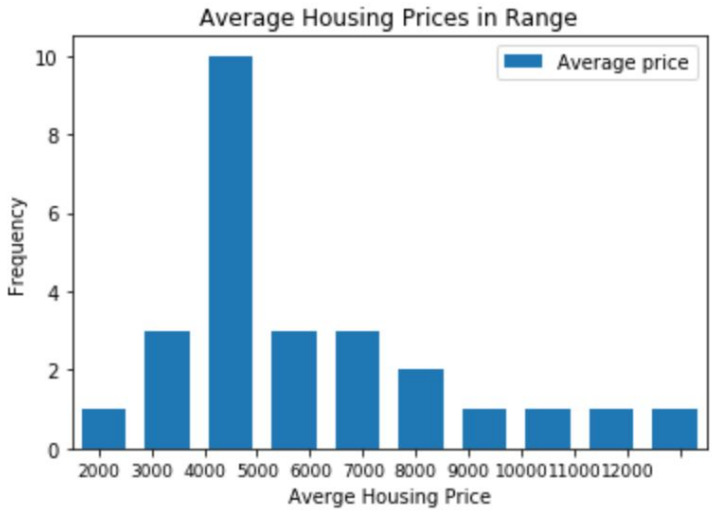

Figure 7. Histogram of average house price

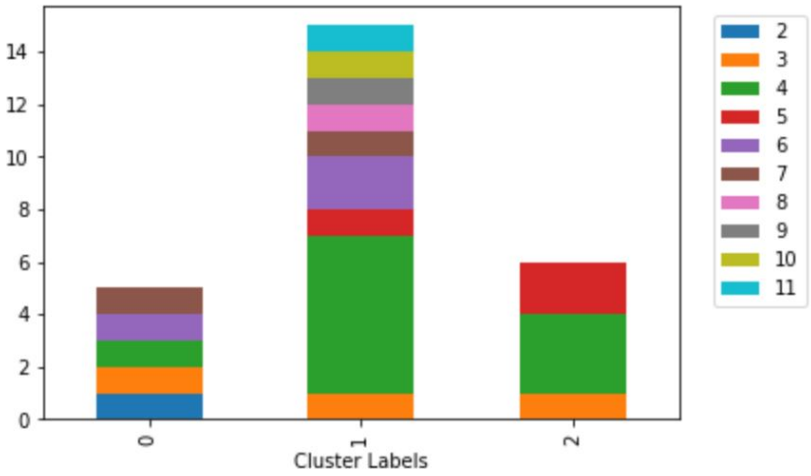

Figure 8. House price distribution in different clusters. Note: Legend “2” refers to price in the range [2000, 3000] and so on, legend “11” refers to price in the range [11000, 12000].

From Figure 8, we can see that Cluster 1 has the most expensive housing price includes price ranges from 8000 SEK to 12000 SEK (see legend 8,9,11), but not the other two clusters. And, according to neighborhood clustering result, we named Cluster 1 as "Multiple Social Venues". It aligns with our assumption that the municipalities surrounding by different social venues naturally have a higher housing price.

On the other hand, Cluster 2 has a price range from 3000 SEK to 5000 SEK. Those are the areas that surrounding venues are less than 11 in a 2-kilometer distance. We named this area a calm area.

Cluster 0 includes the average price range between 2000 SEK to 8000 SEK and has plenty of various venues.

Different clusters can be categorized in Stockholm area as following:

- Cluster 0 (k = 0, red) includes 5 municipalities. Cluster label name: "Hotel & Restaurant"
- Cluster 1 (k = 1, green) includes 15 municipalities. Cluster label name: "Multiple Social Venues"
- Cluster 2 (k=2, purple) includes 6 municipalities. Cluster label name: "Calm area"

7. Discussion

The aim of this project is to cluster different municipalities in Stockholm based on surrounding venues. For that, Foursquare venue data is used. The venues returned by Foursquare is mainly food, stores and lifestyle related categories, so the housing price information added for the further analysis. I believe it would be interesting if one can add information like metro station, daycare/school, university or hospital information. This could potentially valuable for who have kids or elder/sick people who needs to visit a hospital often.

I believe different classification methods can be tried for this project, but not guarantee that get best result.

8. Summary

What was my purpose?

The aim of the project is to cluster different municipalities of Stockholm so that newcomers can choose the municipality where they can live but also not too far away from social places.

What I did or learned?

In this project, I have studied how to convert addresses into their equivalent latitude and longitude values. Also, I have used the Foursquare API to explore neighborhoods in Stockholm City. I used the explore function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. In the end, I have used a k-means clustering algorithm to complete this task. Finally, I have used the Folium library to visualize the neighborhoods in Stockholm and their emerging clusters. Below flow chart shows how this project developed.

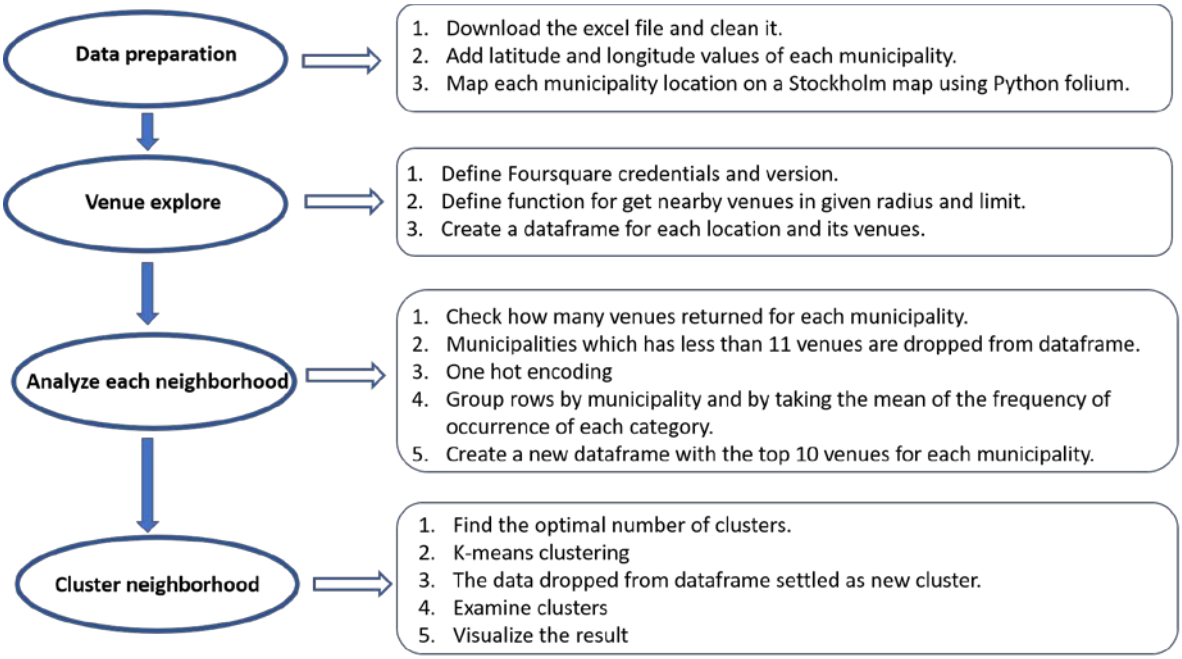

Table 6. Flow chart of the project

What is next?

One can do more advanced or accurate analysis with more information input as explained in the Discussion section. I would love to revisit this project in the future and do more advanced analysis! Until then!

9. References

[1] Stockholm --- Wikipedia (https://en.wikipedia.org/wiki/Stockholm)

[2] Svensk Mäklarstatistik (https://www.maklarstatistik.se/)

[3] Foursquare API (https://developer.foursquare.com/)

[4] Google Map (https://www.google.com)