# Exploring the surrounding area of schools in Stockholm, Sweden

Olga Krali

August 9, 2020

## 1. Introduction

### 1.1 Background

Sweden's capital, Stockholm, is the largest Swedish city with nearly 1 million population. However, more people who live outside Stockholm, tend to commute to the city for work, school, university etc. Thus, Stockholm hosts a lot of venues, universities as well as numerous schools in various neighborhoods. Taking a closer look on schools, it might be beneficiary to understand how the business works in the surrounding area to help stakeholders or parents to take serious decisions.

### 1.2 Business Problem

The purpose of this project was to find which are the most common venues next to schools around Stockholm region and if there was any correlation with the number of enrolled students. Furthermore, a final goal was to investigate if there was a specific pattern on the most common venues that was related to presence of schools.

### 1.3 Business Value

An area with schools can be appealing to many businesses that not only want to expand, but also increase their profits. After a long day at school, the students might decide to go for food with friends or family. A school with a lot of students for instance, could be a target for building a new bakery or a restaurant.

At the same time, getting information about what exists around a school can be signifant for parents, who want to choose a school for their kids that provides a lot of places where their kids and themselves can relax before or after school.

## 2. Data

### 2.1 Data access

The school data for Stockholm region were provided from an open source data portal: https://dataportalen.stockholm.se/. They were in a form of a dbf database, so the simpledbf python library was utilized so that to create a pandas DataFrame for further analysis. Simultaneously, the Foursquare API (https://developer.foursquare.com/places) was used to obtain all venues (up to 100) around the location of the schools in a radius of 500 m.

### 2.2 Data manipulation

Data cleaning and manipulation is a very important step before proceeding to any kind of modelling. First, the dataset was checked for missing values. The dataset contained a single missing value which was replaced by a simple string, since the information for the school year

feature column was not particularly used during modelling. The next step included the coordinate transformation from the SWEREF99 1800 to the standard WGS84 coordinate system which was used by the map library folium as well the Foursquare API. The data then were ready to be combined with the Foursquare dataset as it will be described during the next Section (Section 3).

### 2.3 Features

The school dataset contains 375 entries (372 different schools) and 17 features. From these 17 features, 7 were only kept and went through manipulation as explained in 2.2. More specifically: SDO- which is the region where the school is located, REGI- if the school is municipal or stand-alone, SKOLA-The name of the school, Årskurs- the school grade that the data came from, Elevanal19- the number of enrolled students for 2019-20, X_North & Y_East- the coordinates in SWEREF99 1800 coordinate system.

## 3. Methodology

### 3.1 Python Libraries

The Python libraries mostly used for this project included pandas for creating dataframes and manipulate the data, folium for data visualization in maps, pyproj for coordinate system transformation, matplotlib and seaborn for data visualization and scikit learn for the Machine Learning part.

### 3.2 Frequencies

After data manipulation, the venue data from Foursquare were converted to one hot encodings and the mean from each venue category per school were obtained. For each school the top 5 venues were selected and stored in a pandas DataFrame. This data was the input for the clustering algorithm explained during the next subsection (3.2).

### 3.2 Clustering

Since the venue data were unlabeled, a clustering algorithm could provide some useful insights. The selected clustering method was K-Means were a new sample belongs to the cluster with the nearest mean. However, to determine the number of clusters for this data, it was necessary to utilize a method to aid this process. This method is called the elbow method or within Cluster Sum of Squares (WCSS), which is based on how similar or dissimilar are the data within a cluster to choose the optimum number of clusters. The optimum number of clusters was determined by the less steep decrease of the WCSS, which created a formation which resembles an elbow. After running WCSS, the clustering algorithm was ready to run on the venue data.

### 3.3 Correlation

The labels from clustering were merged with the original dataset along with the data from the frequency analysis so that to have a dataset including the 5 most common venues per school and in which cluster it belongs to. Consequently, the most common venue was obtained for each and every of the top 5 venue categories for each cluster, along with the mean number of enrolled students and cluster size. This data was used for a correlation analysis provided from Pandas based on Pearson's correlation coefficient to investigate whether the number of students was affected by the cluster size or the venue category.

## 4. Results

According to the findings from the elbow method, a cluster size of 9 would provide the best possible outcome from the clustering algorithm.

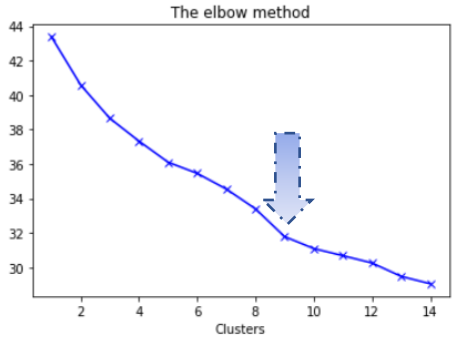



Figure 1: The elbow method (WCSS) showed a smaller decrease (less steep) on the WCSS values after 9 clusters. Thus, an optimal number of 9 clusters was selected.

Stockholm's map (Figure 2) shows that the most populated cluster (cluster 4-Table 1) included schools located in a very central region in Stockholm. The other clusters were more scattered around the city.

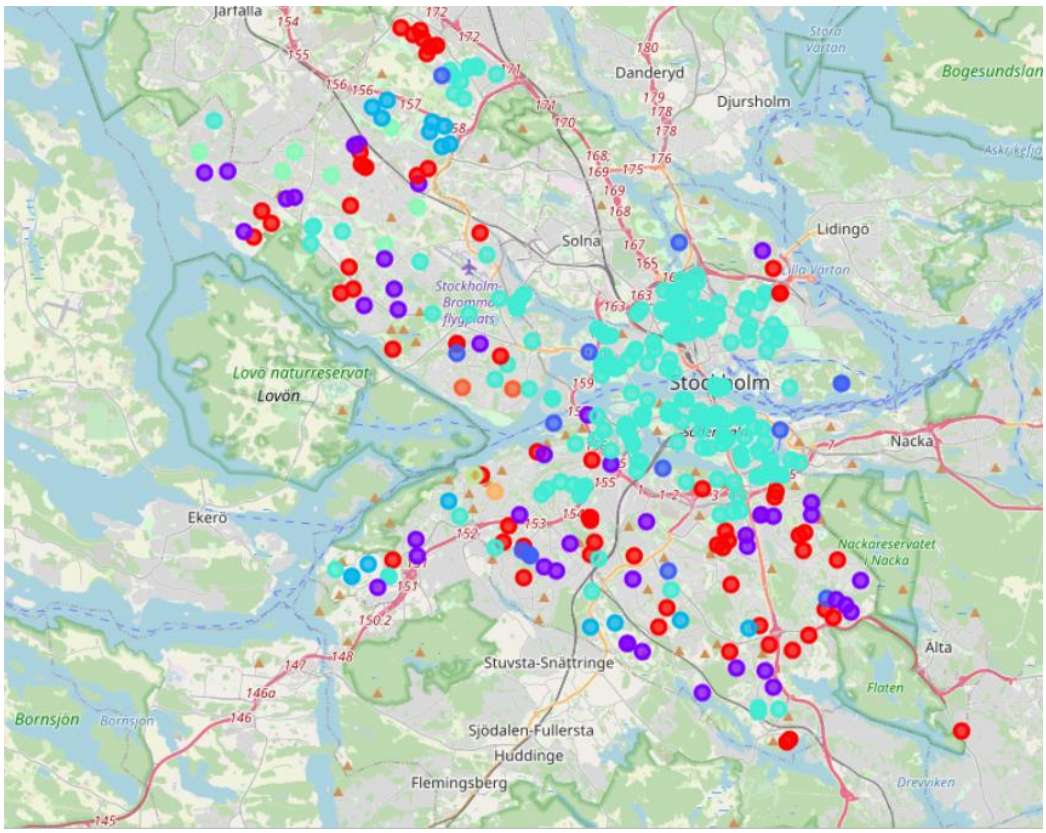



Figure 2: The schools in Stockholm region colored by cluster.

Table 1: Most common venues for each of the top 5 venue categories per cluster, average number of students and cluster size.

| Cluster | MostFreqVen1 | MostFreqVen2 | MostFreqVen3 | MostFreqVen4 | MostFreqVen5 | MeanNumStudents | ClusterSize |
|---|---|---|---|---|---|---|---|
| 0 | Bakery | Grocery Store | Gym / Fitness Center | Park | Accessories Store | 362.582090 | 67 |
| 1 | Pizza Place | Pizza Place | Pizza Place | Accessories Store | Accessories Store | 429.727273 | 44 |
| 2 | Bus Stop | Café | Park | Park | Scandinavian Restaurant | 238.733333 | 15 |
| 3 | Bus Stop | Convenience Store | Grocery Store | Metro Station | Accessories Store | 318.312500 | 16 |
| 4 | Café | Café | Bakery | Bakery | Gym / Fitness Center | 328.556561 | 221 |
| 5 | Bus Stop | Bus Stop | Accessories Store | Accessories Store | Advertising Agency | 497.714286 | 7 |
| 6 | Harbor / Marina | Accessories Store | Advertising Agency | American Restaurant | Amphitheater | 153.000000 | 1 |
| 7 | Deli / Bodega | Accessories Store | Advertising Agency | American Restaurant | Amphitheater | 1204.000000 | 1 |
| 8 | Tram Station | Accessories Store | Advertising Agency | Advertising Agency | Amphitheater | 398.500000 | 2 |

Each cluster tend to have a signature combination of venues that make them very distinct from each other. However, as appendix A shows, clusters 6 and 7 are very similar with 1 sample each and could be merged into one cluster. Looking through the correlation matrix in Table 2, it is

Table 2: Correlation analysis based on the data from Table 1. Only week correlation is observed between each venue category and the average number of enrolled students.

```
ven.corr()
```

| | MostFreqVen1 | MostFreqVen2 | MostFreqVen3 | MostFreqVen4 | MostFreqVen5 | MeanNumStudents | ClusterSize |
|---|---|---|---|---|---|---|---|
| MostFreqVen1 | 1.000000 | -0.278375 | -0.110390 | -0.659893 | 0.051873 | 0.085029 | -0.201720 |
| MostFreqVen2 | -0.278375 | 1.000000 | 0.842451 | 0.226066 | -0.547965 | -0.276911 | 0.296557 |
| MostFreqVen3 | -0.110390 | 0.842451 | 1.000000 | 0.366223 | -0.160335 | -0.288443 | 0.123720 |
| MostFreqVen4 | -0.659893 | 0.226066 | 0.366223 | 1.000000 | 0.215778 | -0.259541 | 0.207402 |
| MostFreqVen5 | 0.051873 | -0.547965 | -0.160335 | 0.215778 | 1.000000 | -0.047688 | 0.193308 |
| MeanNumStudents | 0.085029 | -0.276911 | -0.288443 | -0.259541 | -0.047688 | 1.000000 | -0.191926 |
| ClusterSize | -0.201720 | 0.296557 | 0.123720 | 0.207402 | 0.193308 | -0.191926 | 1.000000 |

clear that the mean number of students is weakly correlated with any of the most frequent venues. Thus, it is not easy to tell weather the presence of specific venues attracts more students in certain schools or not.

## 5. Discussion

The results of this project can create value for stakeholders and parents for various reasons. For each of them the results suggest the following:

### 5.1 Stakeholders

A suggestion could be to proceed with building cafeterias or a gym next to a school in Cluster 1, since Pizza place and accessories stores are the most common venues. Moreover, a restaurant could be a good addition in Cluster 4, since it is not included into the list of the common venues. As stakeholders their aim should be to pick an area, where the venue they desire to build is not present in a high extend.

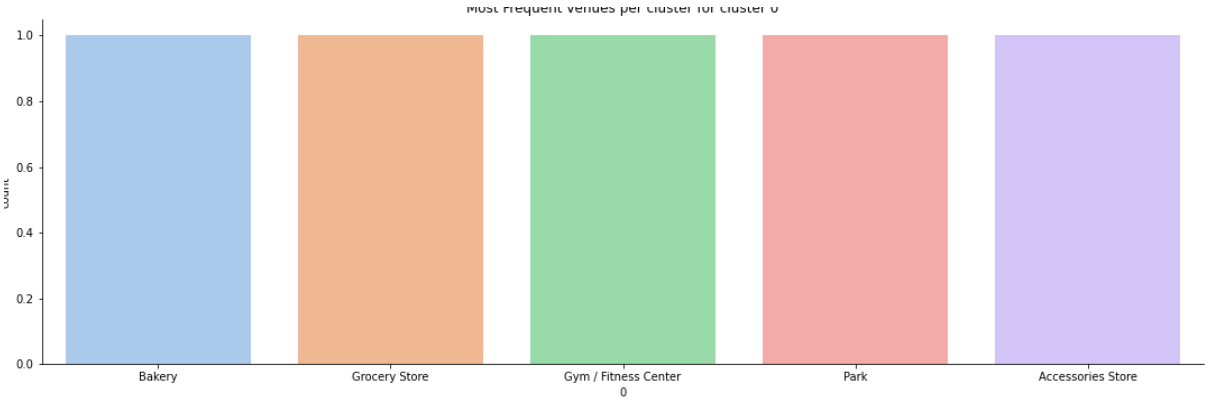
### 5.2 Parents

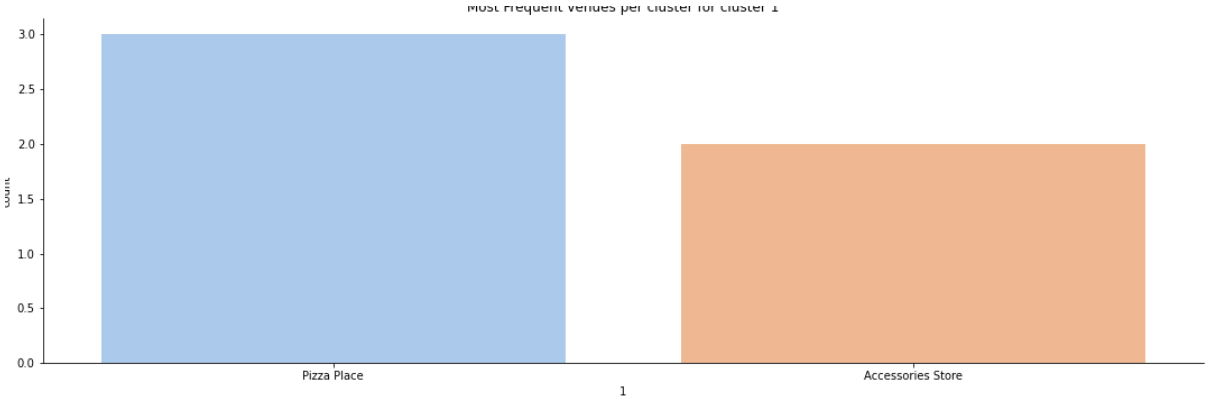
As a parent a good choice of school could be from Cluster 2, since it seems probable to have an available park to play, cafes and restaurants, as well as a bus stop in the proximity. A parent would prefer a school that is surrounded by many venues where someone or/and their kids can eat, relax, play or train. At the same time a bus stop, metro station etc. can help to get home easier if someone lives far away from the city center.

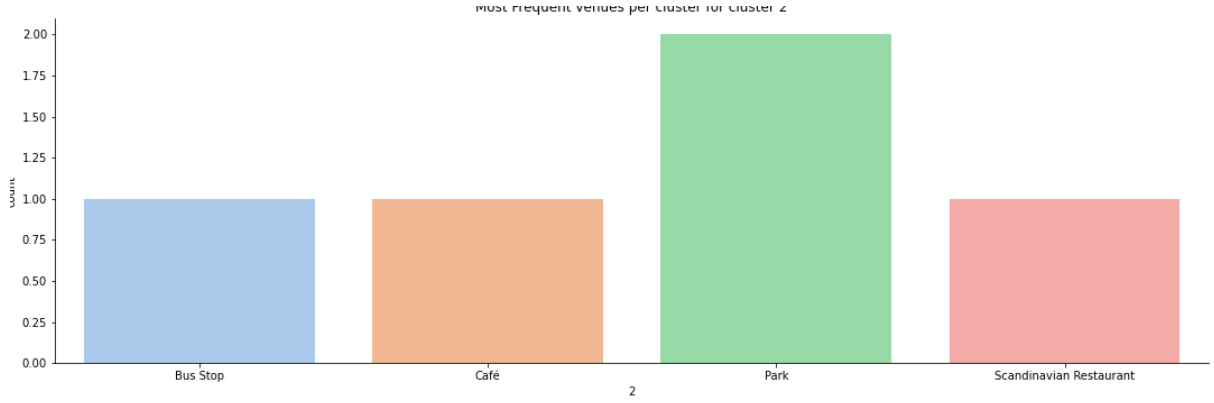
## 6. Conclusion

All in all, picking a location to expand business can be tricky, but open source data along with APIs, such as Foursquare's API can make the job easier. As a parent, it is important to pick a school after careful research, thus this kind of project can help to decide fast without having to visit all the candidate schools. Similar projects can be conducted in other fields such as: building a new park, a new block of flats, or even for conserving nature based on what exists in the surrounding area and so that people can help to preserve an ecosystem in the best possible manner.

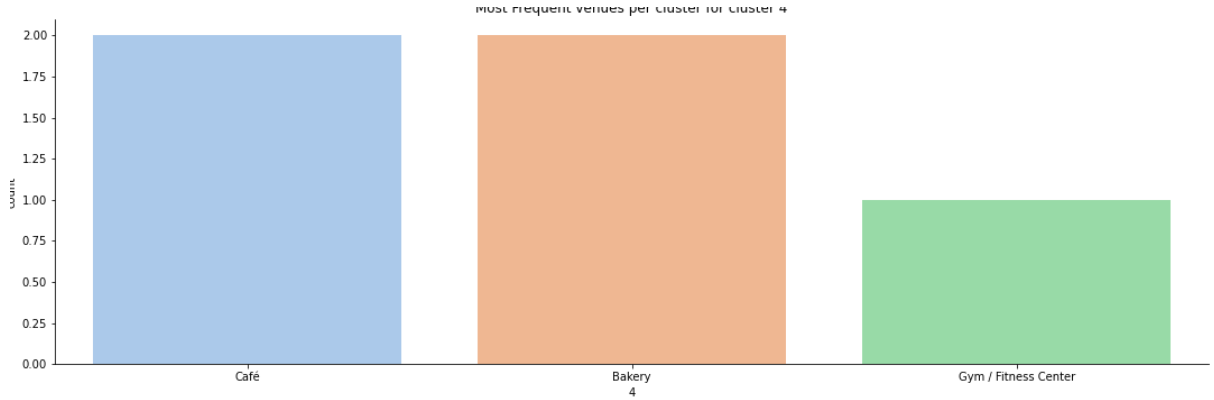
**Appendix A (0-8): Most common venues per cluster (0-8) based on frequency.**

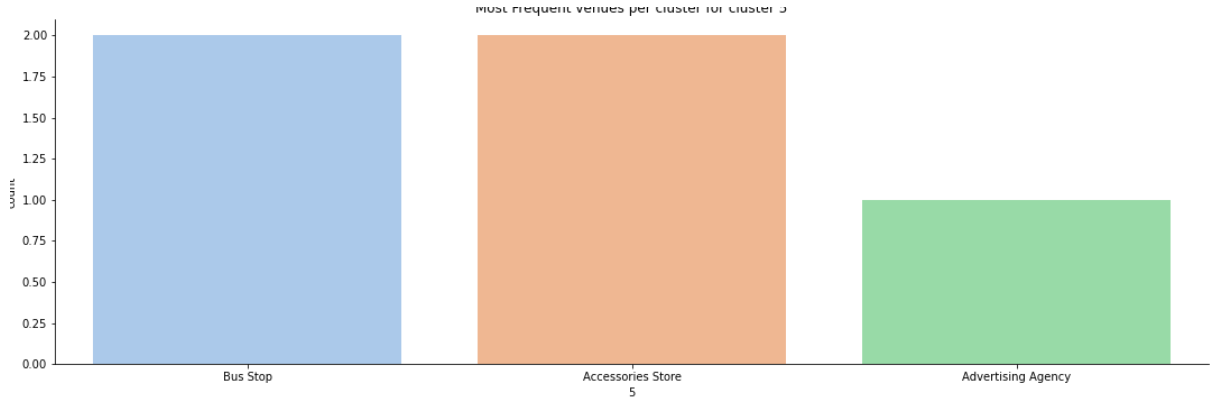


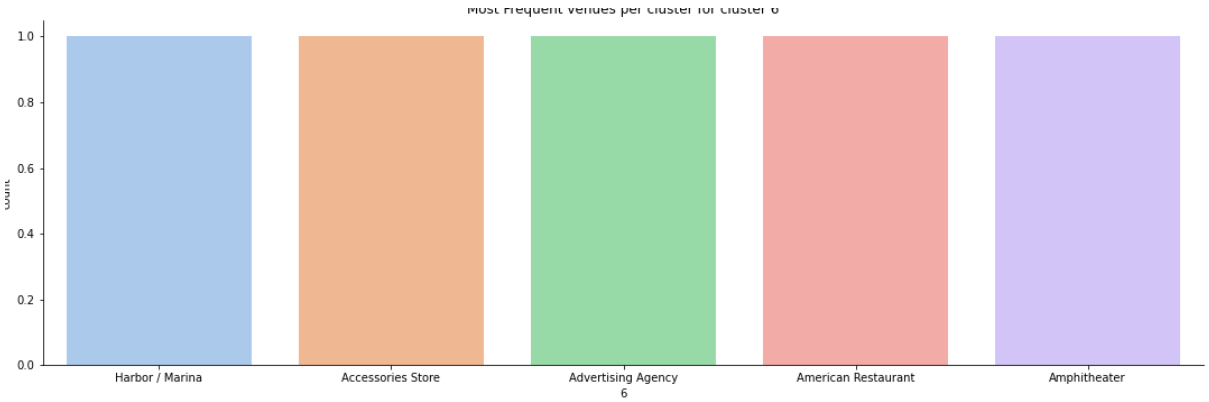



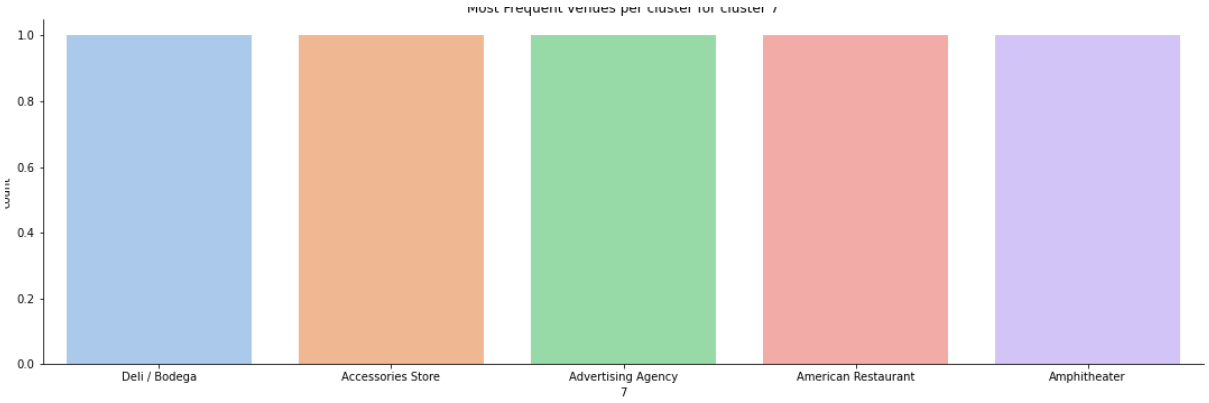

Most Frequent Venues per cluster for cluster 2
Count
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Bus Stop
Café
Park
Scandinavian Restaurant
2

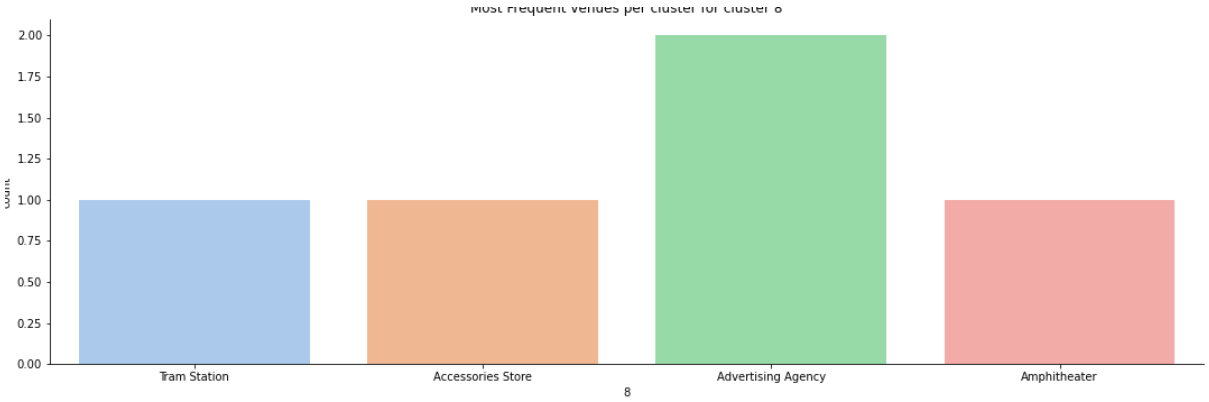

Most Frequent Venues per cluster for cluster 3
Count
1.0
0.8
0.6
0.4
0.2
0.0
Bus Stop
Convenience Store
Grocery Store
Metro Station
Accessories Store
3


Most Frequent Venues per cluster for cluster 4
Count
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Café
Bakery
Gym / Fitness Center
4


Most Frequent Venues per cluster for cluster 5
Count
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Bus Stop
Accessories Store
Advertising Agency
5


Most Frequent Venues per cluster for cluster 6
1.0
0.8
0.6
0.4
0.2
0.0
Harbor / Marina
Accessories Store
Advertising Agency
American Restaurant
Amphitheater
6


Most Frequent Venues per cluster for cluster 7
1.0
0.8
0.6
0.4
0.2
0.0
Deli / Bodega
Accessories Store
Advertising Agency
American Restaurant
Amphitheater
7


Most Frequent Venues per cluster for cluster 8
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Tram Station
Accessories Store
Advertising Agency
Amphitheater
8