# Open source data analysis for Stockholm neighborhoods classification

***Yevgen Pogoryelov,*** *November 2019*

**Summary**

This project aims to analyze the neighbourhoods within and around Stockholm municipality by exploring various venues in the vicinity of subway and commuter train stations. Unsupervised machine learning method of K-means clustering is utilized to conduct the analysis. Three main clusters of locations were identified and studied. The main difference between these clusters is either the type of most common venues or the number of venues surrounding the stations, with the cluster of areas situated predominantly close to city center being most different. Results of this study will be of great interest to people who are deciding to move to Stockholm.

## 1. Introduction

### 1.1. Background

Stockholm is the capital of Sweden and the most populous urban area in the Nordic countries. Stockholm has topped a list of Europe's fastest growing places, with the city's population expected to overcome one million in 2020. The Stockholm Region houses 2.5 million residents, just over one-quarter of Sweden's population, and generates over 30 percent of national economic output.

Stockholm is one of the most successful entrepreneurial environments in Europe, receiving $2.5 billion in venture capital investments since 2005.

Stockholm was recognized as the most innovative region in EU according to the 2017 Regional Innovation Scoreboard, presented by the European Commission. Dubbed the “unicorn” factory by the Financial Times in 2015, the Stockholm Region continues to attract more companies, investments and talents. As a result of a strong innovation infrastructure, based on academic and scientific research, Stockholm has a higher number of start-ups per capita than any other European city and one of the largest life science clusters in Europe.

### 1.2. Problem

A small group of talented minds is planning to move to Stockholm and start an innovative IT company. Some of the members of this group already have families with small kids. They are looking for a nice neighborhood to settle down. Considering sky high housing prices in central Stockholm they do not mind settling at some distance from the city.

### 1.3. Interest

Results of this study will be interested to anyone who is planning to move to Stockholm and settle down in a place that would fit their lifestyle.

## 2. Data

### 2.1. Data requirements

In order to successfully solve the problem, we would need next data:

- List of municipalities within Stockholm Region including geographical data
- List of districts within Stockholm municipality including geographical data
- List of subway and railway stations and their coordinates. In case some of team members decide to live outside central Stockholm, subway and railway represent the fastest means of commuting.
- Demographical data for municipalities and districts
- Venue information that can be obtained from Foursquare. This would help to select a place which would match the lifestyle of the person
- List and location of preschools and daycare facilities for those with small kids

### 2.2. Data sources

The list of municipalities and districts can be obtained from the Wikipedia. Information is retrieved by web-scraping corresponding pages using BeautifulSoup library.

The geo-information on the administrative division of Stockholm Region in the form of GeoJson file was found here: https://matsw.carto.com/tables/stockholms_l_n/public/map

The geo-information on the administrative division of Stockholm municipality is loaded from GeoJson file openly provided by AirBnB (http://insideairbnb.com/get-the-data.html).

The list of Stockholm's subway (Tunnelbana) stations including their coordinates was extracted from Wikipedia: https://sv.wikipedia.org/wiki/Lista_%C3%B6ver_tunnelbanestationer_i_Stockholm

The list of railway (Järnvägsstation.csv) and commuter train (Pendeltåg.csv) stations was downloaded from https://www.koordinater.se/

List and location of preschools and daycare facilities was acquired from the City of Stockholm open data portal (https://dataportalen.stockholm.se/dataportalen/) utilizing their API.

House sales prices for Stockholm Region and separate districts of Stockholm municipality were obtained from Svensk Mäklarstatistik (Swedish Broker Statistics) website: https://www.maklarstatistik.se/

Demographic information for Stockholm Municipality and Stockholm Region were web-scraped from https://ugeo.urbistat.com

### 2.3. Data preparation

For the subway station data:

- Actual coordinates (latitude and longitude) from the 'Position' column
- Names of unopened stations were removed
- Removed duplicate station names. Some stations are listed several times in the original table since they belong to different lines.
- Removed the old names of stations

List of commuter train (Pendeltåg) stations was reduced to those located within 20 km distance from Stockholm center.

Finally, subway and commuter train station datasets were combined.

From the Foursquare venue dataset the category “Metro station” was removed.

For the list of preschools in Stockholm municipality it appeared that their coordinates were in the X/Y coordinate system. After short investigation I could find out that this format corresponds to a Swedish coordinate system RT90 2.5 gon V.

| | name | x | y |
|---|---|---|---|
| 0 | Kavat, Filipstadsbacken 18 | 6570864 | 1631350 |
| 1 | Brostugan, Lofotengatan 30 | 6589738 | 1620526 |
| 2 | Förskolan Björnlandet | 6583935 | 1629809 |
| 3 | Blomsterkungen | 6586569 | 1613337 |
| 4 | Regnbågen, Rågsveds skolgränd 5 | 6572835 | 1626518 |

| | Name | Latitude | Longitude |
|---|---|---|---|
| 0 | Kavat, Filipstadsbacken 18 | 59.238431 | 18.106699 |
| 1 | Brostugan, Lofotengatan 30 | 59.410952 | 17.927724 |
| 2 | Förskolan Björnlandet | 59.356146 | 18.087560 |
| 3 | Blomsterkungen | 59.384518 | 17.799540 |
| 4 | Regnbågen, Rågsveds skolgränd 5 | 59.257580 | 18.023255 |

Coordinates had to be converted into latitude and longitude of the WGS84 coordinate system (GPS coordinates).

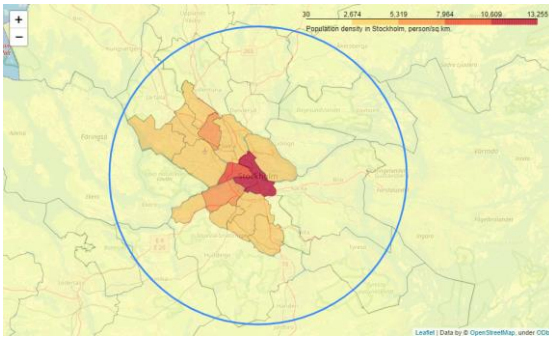
Using the API calls to the website Lantmäteriet (https://www.lantmateriet.se/en/) belonging to Swedish Ministry of Finance the coordinates of preschools were converted into GPS coordinate system

## 3. Methodology

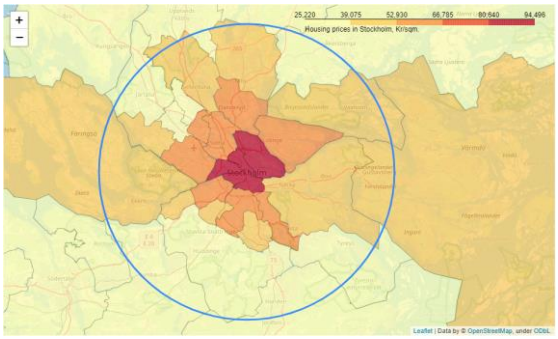
### 3.1. Exploratory data analysis

#### 3.1.1. Map visualization

Choropleth maps representing population density and housing prices.



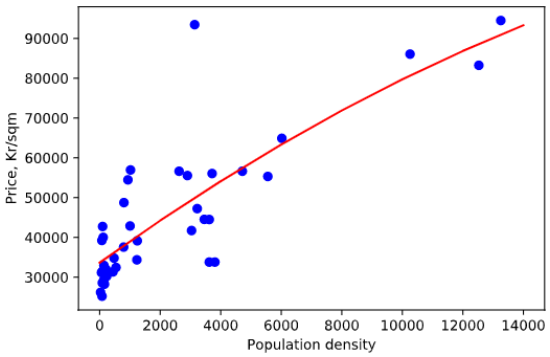

*Density of population. Person / square kilometer*




*Housing prices. Kr / square meter*

As we can see there seem to be a correlation between population density and the housing prices in Stockholm Region, which is actualy not that surprising.

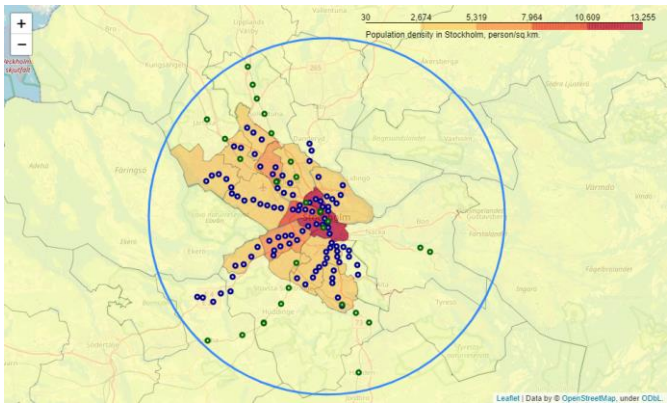
But let's look at this dependency a bit deeper and use the method of polynomial regression, utilizing the sklearn package to model our data.




*Correlation between population density and housing prices*

Although we will not try to do any predictions at this point as we simply do not have enough data, and there might be other factors affecting housing prices.

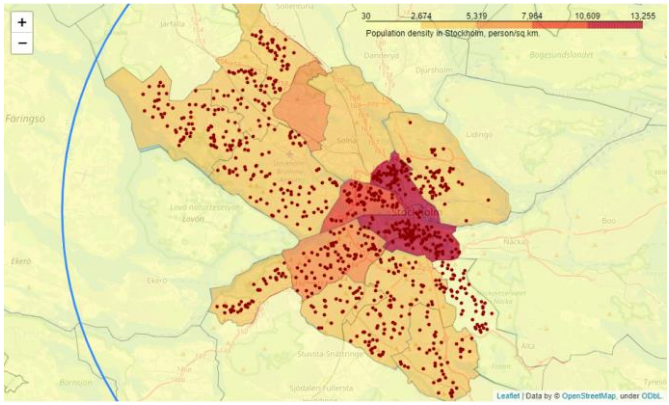
Display subway and commuter train stations on the map




*Population density with the location of subway (blue) and commuter train (green) stations*

As we can see from the map most subway and some commuter train stations are located within areas with higher population density. Only few are located in less populated areas.

Now let's how preschools are distributed within Stockholm municipality.




*Red dots on the map are locations of preschools in Stockholm municipality*

The distribution looks quite uniform. Unfortunately there was no data available on the position of preschools in the neighboring municipalities. Special interest would be Solna.

### 3.1.2. Foursquare data

The venue data from Foursquare with the limit of 100 venues 500 meters from each station resulted in the dataframe containing 2965 venues.

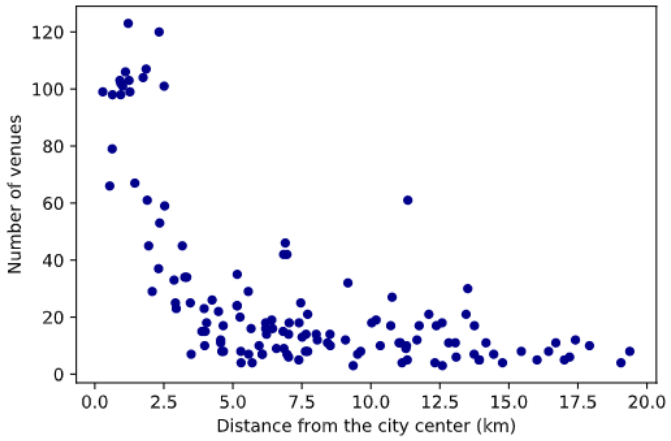
However Foursquare dataset does not contain any locations of preschools.

After merging Foursquare venue dataset with the preschool dataset we have a more complete picture of what are the top venue categories. As a result the complete dataset now contains 3598 venues.

| | Main Category | Venue Category | Percent |
|---|---|---|---|
| 2 | Food | 1473 | 41.0 |
| 5 | Professional & Other Places | 669 | 18.6 |
| 6 | Shop & Service | 575 | 16.0 |
| 4 | Outdoors & Recreation | 309 | 8.6 |
| 7 | Travel & Transport | 234 | 6.5 |
| 3 | Nightlife Spot | 199 | 5.5 |
| 0 | Arts & Entertainment | 133 | 3.7 |
| 1 | College & University | 3 | 0.1 |

*Top main categories of the venues in Stockholm*

As is expected, the number of venues decrease with the distance from the city center




*Number of the venues with the distance from Stockholm center*

### 3.2. Data preparation

Since there are 14 stations that have 5 and less venues around them, we will exclude them from clustering analysis. However, we will keep information related to them in a separate dataframe.

We are left with 3538 venues in 261 categories located around 116 stations.

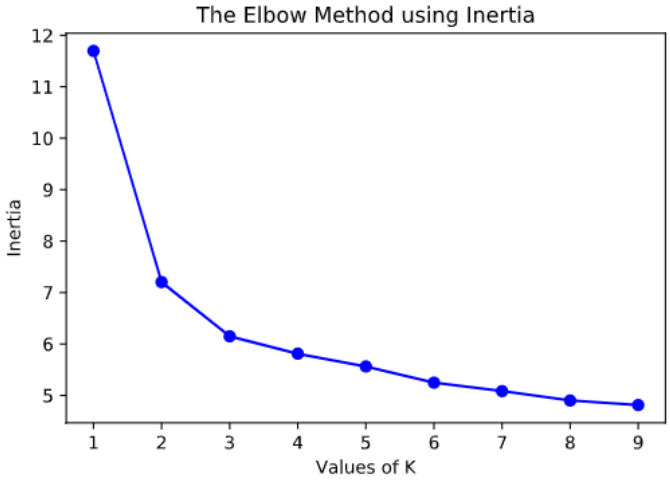
To prepare the dataset for k-means clustering and deal with categorical values, such as venue categories, we apply one-hot encoding. After data normalization the new dataframe contains 116 rows for every subway and commuter train station and 261 column for every category.

### 3.3. Clustering

#### 3.3.1. Finding the optimal number of clusters

In order to determine the optimal number of clusters we us so-called Elbow method.

If the line chart resembles an arm, then the “elbow” (the point of inflection on the curve) is a good indication that the underlying model fits best at that point.

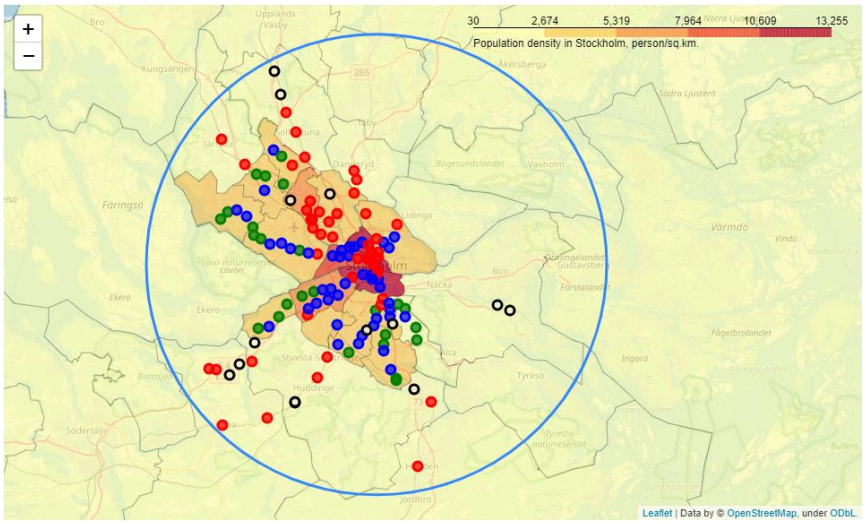



*Elbow plot to identify optimal number of clusters*

According to the plot the optimal number of clusters is ***k* = 3**.

#### 3.3.2. K-Means clustering

K-Means clustering was carried out with the number of clusters ***k* = 3**. Below is the visualization of the cluster distribution on the choropleth map of Stockholm showing the population density.




*Distribution of the clusters*

Red markers correspond to Cluster 0, green – to Cluster 1, and blue – to Cluster 2. Finally, open black markers represent stations, which have 5 or less venues around them.

To describe clusters, I have calculated the number and percentage of venues belonging to main categories, as well as 10 most common venues on the lower level of classification in each of the clusters.

## 4. Results

As a result of our analysis we have identified 4 clusters of locations centered around subway and commuter train stations.

- **Cluster 0 (Red)**: has the highest number of venues, large portion of which is located close to the city center. The most common categories of venues beginning with the largest category are Food, Shop & Service, Travel & Transport, Outdoors & Recreation, Nightlife Spot and Arts & Entertainment. Within category Food, the most common venues are cafés, Scandinavian restaurants, pizza places, Italian restaurants and bakeries. This cluster would be of interest to those who either not yet planning to fill up their life with family/kids, or whose kids already grew up.
- **Cluster 1 (Green)**: has the lowest number of venues among the clusters. This cluster is located outside the central part of Stockholm. The most common categories of venues are Professional & Other Places, Food, Shop & Service, Outdoors & Recreation and Travel & Transport. Over half of the Professional & Other Places consist of preschools. Within category Food, the most common venues are pizza places. This cluster might be interesting to young families with small kids. Considerable number of preschools will satisfy the requirements for child day care. And small number of other venues mean rather quiet atmosphere around.
- **Cluster 2 (Blue)**: is located mainly outside of the city center. The most common categories of venues are Food, Professional & Other Places, Shop & Service, Outdoors & Recreation and Nightlife Spot. Although Food category is the larges for this cluster,main venues are preschools. Within category Food, the most common venues are cafés, bakeries, pizza places, Thai and Sushi restaurants. This cluster will be of interest to young families who either already got small kids, or have such plans for near future. Preschools are at almost every corner ready to take care of small ones. And the rest of infrastructure seem to be well thought of with many restaurants, shops and gyms.
- And the remaining category includes areas with very small number of venues. Might be a good place for those who either want to have some quietness in their lives.

## 5. Discussion

The purpose of this study was to analyse neighbourhoods within and around Stockholm municipality and identify promising locations that would fit requirements and lifestyles of some of the members of newborn IT startup company. One of the main requirements was easy access to subway and commuter trains, which represent rather reliable and fast means of transportation.

Analysis was done using some openly available data, both from swedish government sources and Foursquare. Since Foursquare is mainly focused on food related venues, the venue data from Foursquare was supplemented by a complete list of preschools within Stockholm municipality. Added information about population density and housing prices may help better identify the place where a person would want to settle down after deciding to move to Stockholm.

Further development of this study might be to include information on the venues not covered by Foursquare:

- Preschools outside Stockholm municipality (e.g., in Solna, which looks like an attractive area to live)
- Schools (for families with school-age kids)

- Healthcare infrastructure
- Housing prices that include not only apartment prices, but also houses (e.g., villas)

Analysis can be also extended to areas beyond 500 m distance from subway and commuter train stations.

## 6. Conclusion

In this study I have tried to provide some recommendations to those who decided to move to Stockholm in order to find a good place to settle down, which would fit their lifestyle. Subway and commuter train stations were chosen as a sort of "anchor" points and were clustered into three main clusters based on the type of venues that surround them. As it is the case in many modern cities, the central areas of Stockholm are more suitable for people with active lifestyle, while more distant areas would better satisfy needs of families with younger kids.