# 1 Introduction/Business problem

The idea of this study is to help people planning to move to Sweden to find the best municipality to reside in. This will be done by providing data regarding average income, crime rate and municipal tax rates rate as a feature set for clustering. A selected cluster will be displayed as dots on the Swedish map. After this the foursquare API will be used to pull popular venues around the municipalities.

# 2 Data

## 2.1 Data source

To provide the stakeholders the necessary information I'll be combining data from several sources. I will be using the ‘Sweden statistics’ (SCB) API to obtain data regarding tax rates. I will be creating a dataset of data from SCB regarding the average income. This will be picked up as a excel from their website. The crime rates will be obtained from BRÅ´s website as an excel.

The geographical data will be taken from ‘Geocode.xyz’, where I will first extract the names of the municipalties into a csv. Then run batch geocoding from their website to get the coordinates.

Last but not least I will use the foursquare API to collect the top venues around the municipalities.

SCB = <https://scb.se/en/>

SCB = https://www.scb.se/hitta-statistik/statistik-efter-amne/hushallens-ekonomi/inkomster-och-inkomstfordelning/inkomster-och-skatter/pong/tabell-och-diagram/inkomster--individer-lankommun/sammanraknad-forvarvsinkomst-2018--per-kommun-efter-percentiler/

Foursquare = <https://foursquare.com/city-guide>

BRÅ = http://statistik.bra.se/solwebb/action/index

Geocoder = https//www.geocoder.xyz

Foursquare = https://www.foursquare.com

# 3 Methodology

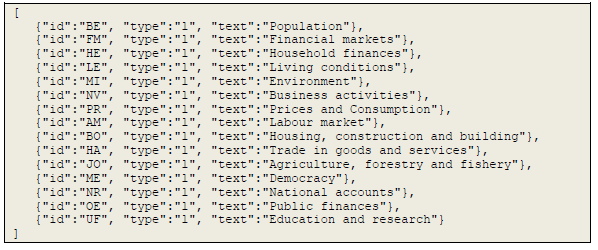
## 3.1 Data collection

The data collection from all sourced is from 2018. The primary key for the dataset(s) are the municipal.

So all data collection is targeted around the swedish municipalities.

## 3.1.2 SCB API

Connection to the SCB API is done according to the specification on their website.



The database consists of several layers which means that through the notebook you are able to go deeper into the database structure. Once the desired level is reached, a query can be made which collects the data in a json file. This is later placed in a data frame named ‘df\_taxes’.

## 3.1.2 SCB website

The average income is collected using the online service on the SCB website. The data is collected in an excel and stored locally on the machine. The data from the excel is stored in a dataframe named ‘df\_income’

## 3.1.3 BRÅ website

The crime rate is collected using the online service on the BRÅ website. The data is collected in an excel and stored locally on the machine. The data will be stored in the dataframe ‘df\_crime’

## 3.1.4 Geocoder.xyz website

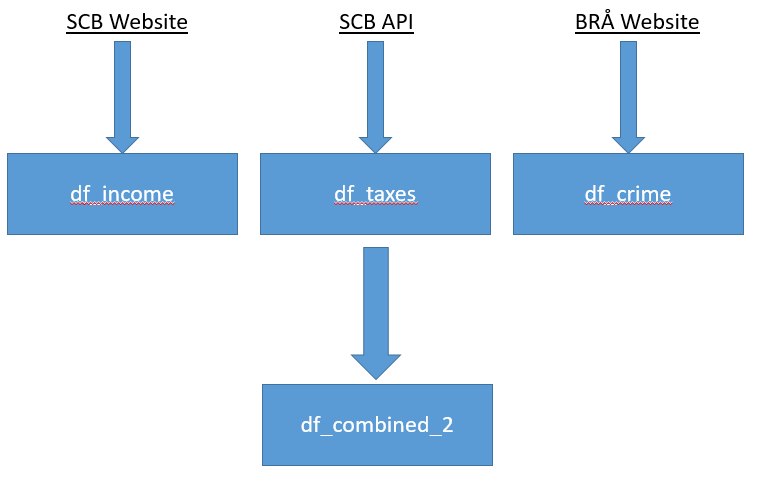
The coordinates for each municipal is collected using batch geocoding from the Geocoder.xyz website. This is selected since it is one of the few free geocoding sites available. The municipal names are extracted from the main dataset and stored as a CSV file. The CSV file in executed through the Geocoder.xyz website and returns a json file with the coordinates. The json file is read in as a dictionary and later added to a data frame named ‘df\_coordinates’.

## 3.1.5 Foursquare API

The municipal venue data is collected through the foursquare API. The request is made through the API and it returns a json file which is entered into a data frame named ‘Sweden\_venues’.

## 3.2 Data cleaning and wrangling

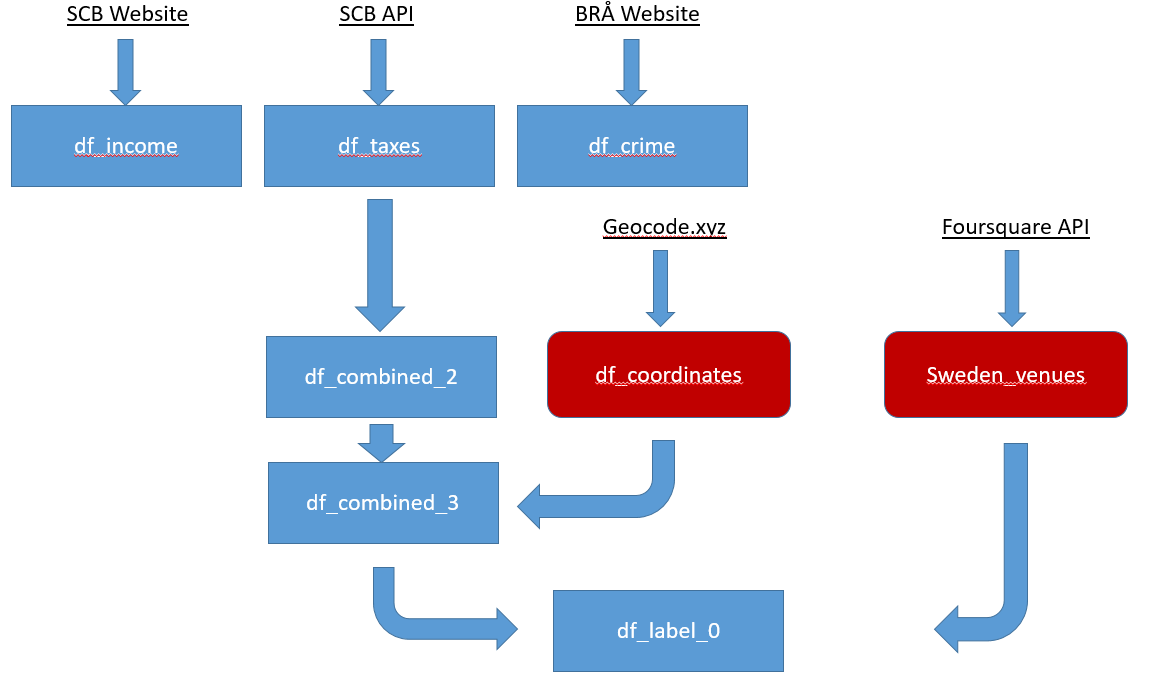
The datasets are reviewed. Not a number values are removed. Any unused columns are dropped from the acquired data. Some column names were renamed to ensure understanding. It was also required to change data types for some columns. As the picture below describes the three collected dataframes were merged into one.



The dataframes where merged in two steps. First combining df\_income and df\_taxes into df\_combined. Then combining df\_combined and df\_crime into the new dataframe df\_combined\_2. All merging was done using the ‘Area’ as the primary key. Also using an inner joint. This was the ensure that in case there was a mismatch we would only store the full data. Some mergers was a bit complicated as the area names(String) contained spaces. This was solved by running the following code before merging.



After running the first machine learning clustering, the coordinate data and venue data is added to df\_label\_0.



## 3.3 Machine learning

In the project there are two steps of machine learning. The first is taking all the municipals and clustering them based on income, taxes and crime rate. The second machine learning step is clustering based on the venues of a given cluster from the first machine learning outcome.

## 3.3.1 Clustering based on income, taxes and crime

For this clustering I selected to use k-means clustering. Mainly based on it´s simplicity as a unsupervised machine learning algorithm. There are 290 municipals in Sweden. We select to have these in 10 clusters. Prior to modelling, the dataset df\_combined\_2 is preprocessed by removing categorical values as well as normalizing the data.

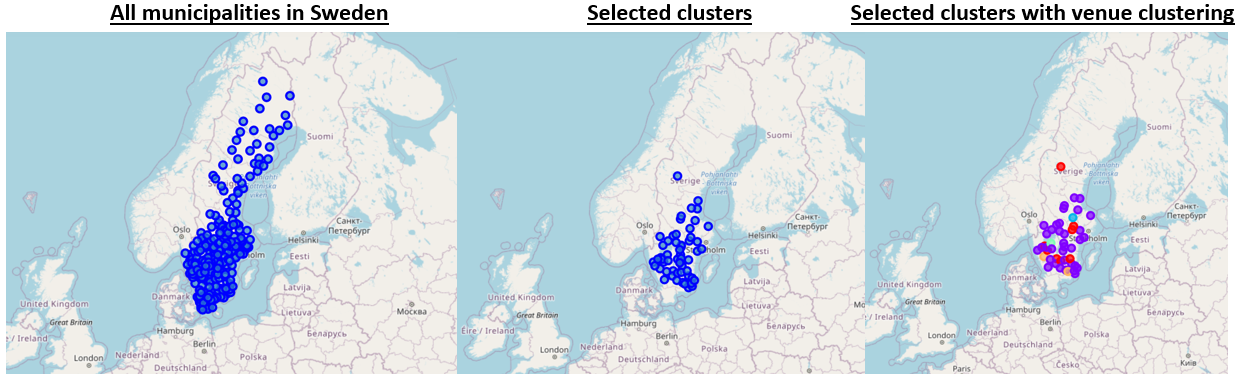


## 3.3.2 Clustering based venue

For the second machine learning step. We have honed in on one specific cluster from the first model. Data regarding surrounding venues are obtained using the foursquare API. Prior to running the model, the data is processed using ‘one hot encoding’. We run the model to cluster the most popular venues around the municipalities. The target is to use this clustering to visualize in which areas certain venues are most popular.

## 3.3.4 Visualization

The visualization is created using the Folium library. Example below:



# 4 Results

The outcome of the project provides a good overview of the different municipalities in Sweden. The visualization is a good way to show potential users which areas provide the basic needs as well as the details regarding the surrounding venues.

# 5 Discussion

The number of clusters in both K-means algorithms could probably be optimized. The current K factor in the Jupyter notebook was a good guess of a reasonable number. One can further dive into this and define this in a more scientific way. I feel that in the context of the main purpose it was good enough for a potential user.

During the Geocoding sequence of the project there was a mismatch when the batch geocoding was running on the geocoding.xyz website and the entry from the csv. Some of the Swedish names was difficult to process and about 16 areas were misplaced into other countries. This could have been solved on a more sophisticated platform. But since I only had access to a free website I was restricted to a fixed number of requests. I decided to go ahead with this loss since I would have required to pay for additional requests.

# 6 Conclusion

In this project I obtained data from several different sources. The data was cleaned and wrangled. There were two different machine learning unsupervised cluster algorithm used. Prior to modelling the clusters the data was pre-processed. These models are helpful for people thinking of moving to Sweden to get a good overview of the different municipals and what they have to offer.