# REPORT

IBM Applied Data Science Capstone

# Segmenting and Clustering Neighbourhood venue types in Helsinki

By: Mikael Björkqvist

January 2020

# Introduction

All cities on earth are full of different kind of venues. Would it be nice to segment and rank venues in a city. More specifically, which venue types people use the most in in each city neighbourhoods.

I.E. What are the top 10 used venue types in Konala, Helsinki, Finland

To understand what types of venues people of Helsinki like and use the most, and where they are located, is valuable information.

## Business problem

Create system that can search and rank venue types in chosen city neighbourhoods by their usage. This report and methods build for it, will also act as a solid base for further processing of similar ideas. I.E. What type of food venue types are most visited in each neighbourhood.

## Target audience

Investors, loan givers, entrepreneurs, city planners/management etc.

# Data

To solve the problem, we will need the following data:

- List of all neighbourhoods in Helsinki (postal code). We ruled out Espoo and Vantaa.
- Coordinates for all those areas. We used OpenCage GeoCoder API for that.
- Venue data near those neighbourhood coordinates. Foursquare API was used for this.
  API is limiting results from these queries. Max 100 venues per area was gained.

## Sources of data and methods to extract them

We use google for Helsinki coordinates, Helsinki database service for postal areas table and Python Geocoder for coordinates. Opencage Geocoder is API to convert coordinates to and from places. Venue location data we got from Foursquare service using their API queries per area. Foursquare has one of the largest databases of 105+ million places and is used by over 150,000 developers. These venues are updated with data by over 13 billion venue check ins starting 2009. Foursquare API will provide many categories of the venue data which is good when we want to expand our research. We used only free versions of these API:s so some data queries were limited. Upgrading these services gives lot more API calls and results. Which in return will increase datasets and improve report accuracy. Tools for this project are highly customizable.

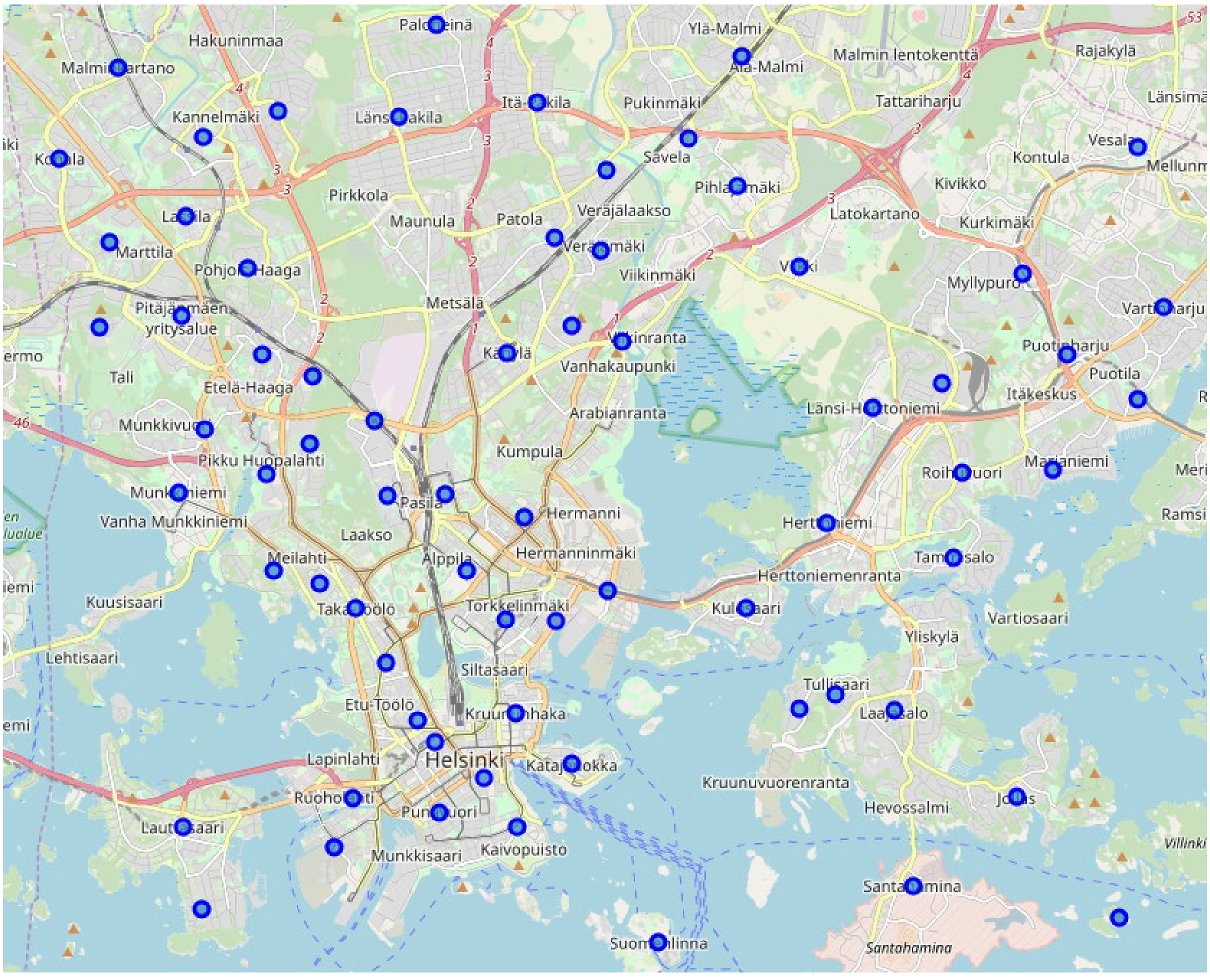
All data cleaning, wrangling, machine learning (K-means clustering) and map visualization (Folium) is done in Python 3 using Jupyter notebook, that is also shared for review.

# Methodology

First we got coordinates of Helsinki with google search. Next we downloaded official excel file for 2019 Helsinki area codes and corresponding names from [https://www.hsy.fi/fi/asiantuntijalle/avoindata/Sivut/AvoinData.aspx?dataID=35](https://www.hsy.fi/fi/asiantuntijalle/avoindata/Sivut/AvoinData.aspx?dataID=35). Table was very simple, so we manually excluded Espoo and Vantaa area data from table and saved table as CSV file. These postal codes formed the scope of this project.

Dataframe was made from gained postal area table, and Geocoder was used to gain coordinates for postal areas in that dataframe.

Map of Helsinki with marker on every neighbourhood center was created with Python Folium. This showed the real coverage on map and gave good visual aid for determining average venue search radius. See picture below.



Using new coordinates, we made Foursquare API query calls, to search venues near those postal area centers (neighbourhoods) with a customizable radius. We chose 300 meters from center point because of the result limitations, that Foursquare free personal account dictates. 100 venue results per API call is possible. So we made this query call for every neighbourhood coordinates. Increasing max venue results and possibly radius, we could get better accuracy for this research.

We are interested in type of category that every venue falls in to, so we made custom table from all gained venue data. See example picture below.

| | Neighbourhood | Neighbourhood Latitude | Neighbourhood Longitude | Venue Name | Venue Category | Venue Latitude | Venue Longitude |
|---|---|---|---|---|---|---|---|
| 0 | Helsinki Keskusta - Etu-Töölö | 60.17207 | 24.931243 | Arkadia Oy International Bookshop | Bookstore | 60.173369 | 24.929330 |
| 1 | Helsinki Keskusta - Etu-Töölö | 60.17207 | 24.931243 | Taidehalli | Art Gallery | 60.172127 | 24.931014 |
| 2 | Helsinki Keskusta - Etu-Töölö | 60.17207 | 24.931243 | Ateljé Finne | Scandinavian Restaurant | 60.171198 | 24.928515 |
| 3 | Helsinki Keskusta - Etu-Töölö | 60.17207 | 24.931243 | Luonnontieteellinen museo | Science Museum | 60.171350 | 24.931549 |
| 4 | Helsinki Keskusta - Etu-Töölö | 60.17207 | 24.931243 | Terrace | Beer Garden | 60.173639 | 24.932516 |

Amount of unique categories was solved from this. 155 Unique categories were found.

At this point we summarized all found unique categories, and made dataframe from that so we could make bar graph to visualize most common venue types (Check “results” section).

Now we performed “One-hot” encoding and aggregated venues by neighbourhoods.

Then returned most common venues with a function and created dataframe of that data to get the top 10 venues in neighbourhoods. Example picture below.

| | Neighbourhood | 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 | Aurinkolahti | Beach | Restaurant | Ice Cream Shop | Yoga Studio | Food Court | Gastropub | Garden Center | Garden | Furniture / Home Store | French Restaurant |
| 1 | Etelä-Haaga | Bus Stop | Convenience Store | Cafeteria | Falafel Restaurant | French Restaurant | Grocery Store | Gift Shop | Gastropub | Garden Center | Garden |
| 2 | Etelä-Vuosaari | Recreation Center | Pizza Place | Taxi Stand | Gym / Fitness Center | Cafeteria | Café | Fast Food Restaurant | Filipino Restaurant | Flea Market | Flower Shop |
| 3 | Etu-Vallila - Alppila | Theme Park Ride / Attraction | Gym | Event Space | Chinese Restaurant | Pizza Place | Bar | Park | History Museum | Forest | Garden Center |
| 4 | Helsinki Keskusta - Etu-Töölö | Scandinavian Restaurant | Bookstore | Art Gallery | Jazz Club | Beer Garden | Science Museum | Gastropub | Garden Center | Garden | Furniture / Home Store |

After this, neighbourhoods with top 10 venue similarities were clustered using K-Means clustering. Then visualized on map by color codes. (Check “results” section).

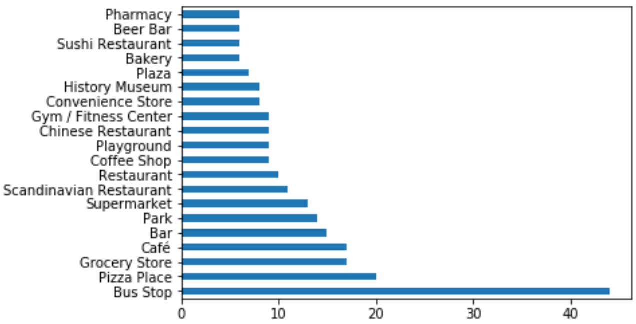
# Results

To understand our results better. We needed to visualize our data. This can be done in many ways and to many questions, but here we concentrate to the following

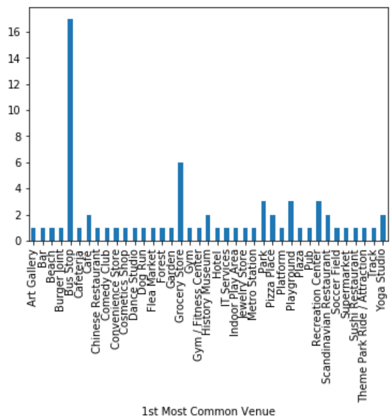
We can now answer three basic questions.

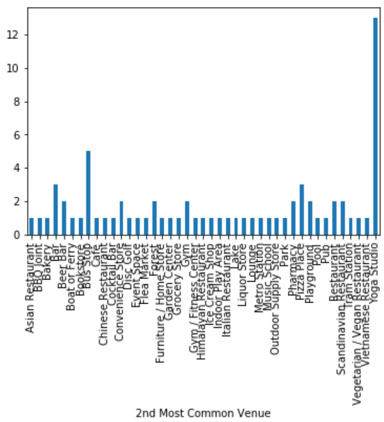
1. What are most used venue types in whole city combined.
2. What are 10 most visited venue types in each neighbourhood.
3. Which neighbourhoods have similarities in venue usage.

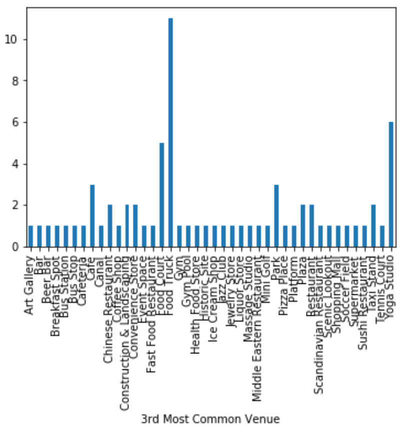
Graph below shows top 20 used venue types in whole city of Helsinki combined.



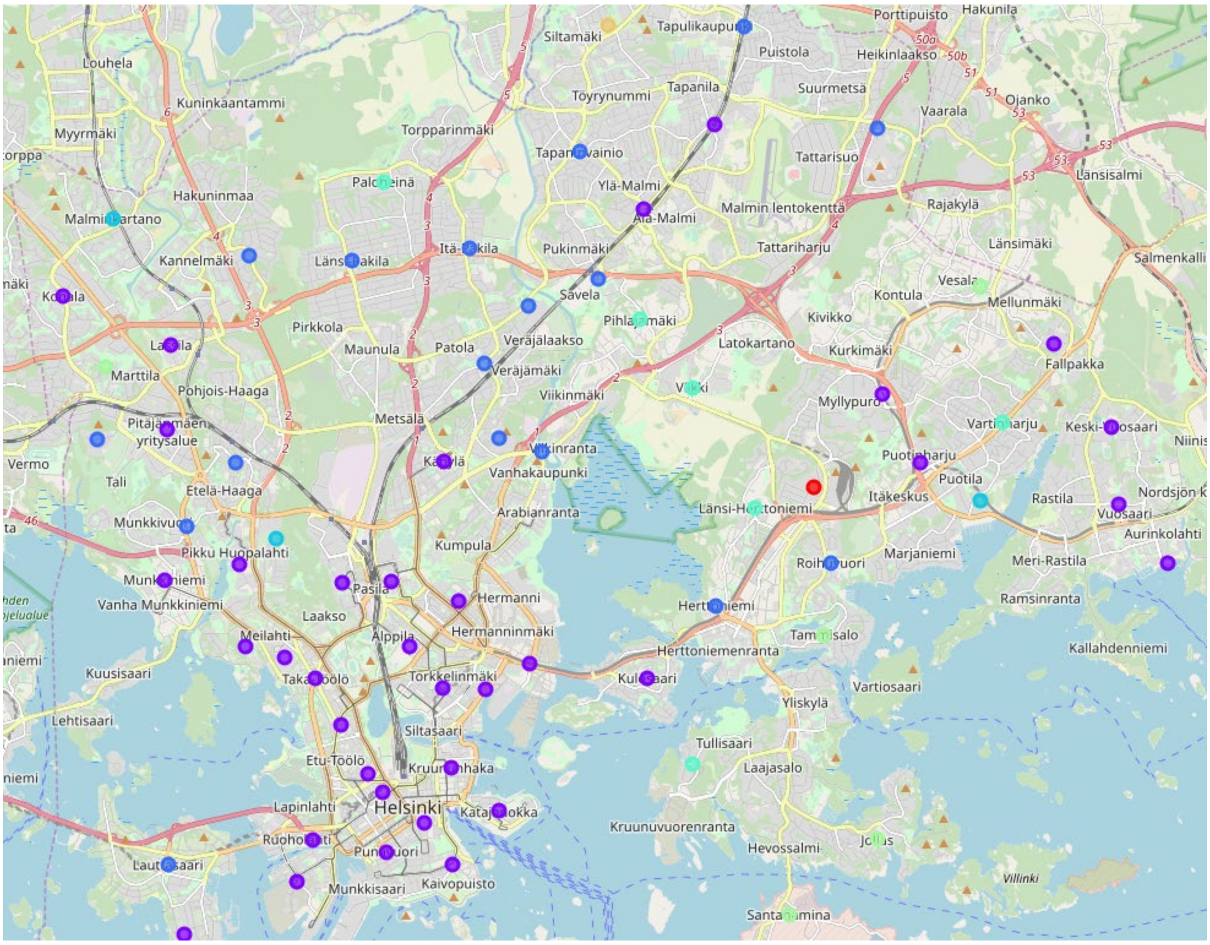

Next three graphs show what are top 1, 2 and 3 most commonly used venue types in neighborhoods. I.E. Some areas most commonly used venue is a bus stop. But, some other neighborhoods most commonly used venue is yoga studio.







Map below shows clustering of neighbourhoods. We have defined 7 clusters, each with it’s own color. “K-Means” clustering Machine Learning algorithm, finds venue usage similarities in neigbourhoods and puts similar neighbourhoods in a group(cluster).




We can also inspect clusters closely with table formed by cluster data. See example table of cluster 4 data below (color turquoise on map above).

| | Neighbourhood | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 43 | Kaitalahti | Bus Stop | Yoga Studio | Food Truck | Gift Shop | Gastropub | Garden Center | Garden | Furniture / Home Store | French Restaurant | Forest |
| 51 | Paloheinä | Bus Stop | Pool | Yoga Studio | Forest | Gift Shop | Gastropub | Garden Center | Garden | Furniture / Home Store | French Restaurant |
| 55 | Pihlajamäki | Bus Stop | Yoga Studio | Food Truck | Gift Shop | Gastropub | Garden Center | Garden | Furniture / Home Store | French Restaurant | Forest |
| 63 | Viikki | Bus Stop | Lake | Yoga Studio | Forest | Grocery Store | Gift Shop | Gastropub | Garden Center | Garden | Furniture / Home Store |
| 64 | Länsi-Herttoniemi | Bus Stop | Yoga Studio | Food Truck | Gift Shop | Gastropub | Garden Center | Garden | Furniture / Home Store | French Restaurant | Forest |
| 79 | Vartioharju | Bus Stop | Garden Center | Yoga Studio | Food Truck | Gift Shop | Gastropub | Garden | Furniture / Home Store | French Restaurant | Forest |

# Discussion

Using only cluster map and cluster 4 table in “results” section we can draw conclusion that people in neighborhoods, that are not in dead center of Helsinki, needs of course bus stops. But the fact that yoga studios are in high demand is surprising. Also food truck usage is high. Most used restaurants are French and only “Viikki” neighbourhood has grocery store in top 10.

Caution must be used using this data. Thinking that French restaurants are in high demand in cluster 4, could be misleading. Or it could give us a hint about type of people and income levels in those areas. Notice that top 20 most used venue types in Helsinki, does not include French restaurants. But it does include sushi and Scandinavian restaurants. Which are more commonly used in other and more bigger neighbourhood clusters.

Not many certain answers can be given with this broad search of venue types. But when user of these tools narrows venue searches to I.E. Only food venues. And filtering only á la carte restaurants, user gets much better and accurate data to questions like, what areas miss some kind of restaurant, or have too many of them. What type of restaurants are a hit in city of Helsinki, and where they are located. Where should you drive with your food truck in example.

Upgrading your Foursquare service, can make these tools very powerful and gives much more accuracy through venue results that could be almost unlimited. With more venue results, you can also adjust venue query radius more accurately

# Conclusion

We successfully created data science tools and methods to solve our business problem.
Reader should also notice that making these tools with such large venue category query scope, was to make customizing of these tools easier. Changing few parameters here and there, and using more powerful visualizing plots, these methods can be made to answer many questions of people behavior in different areas. These tools can be configured to handle much larger geographical areas and much bigger venue datasets.  Only thing missing methods in this project, are data scraping and cleaning. I.E. If user need to get data from much more complicated tables of data, or to scrape websites for data.