Title: Finding suitable areas for a new restaurant in greater Helsinki area

Introduction

The purpose of this project was to explore neighbourhoods in greater Helsinki area (Finland) to find the best places for a new restaurant. The target audience for this report is anyone looking to setup a certain type new restaurant in the area. The analysis will shed light into questions like: how many restaurants of each type already exists in different areas, how many people live there and how many restaurants there is per capita.

This project used multiple data sources that will be described more in detail in the data section of this report. The types of data include : basic information about neighbourhoods in greater Helsinki area, such as: neighbourhood names, coordinates and population. Information about current restaurants will be fetched from Foursquare.

The analysis regarding current restaurant offering in the neighbourhoods includes finding out what kind of restaurants are most prominent within each area and per capita. K-means clustering method will be used to cluster the neighbourhoods based on their similarities.

The data analysis was carried out using Jupyter notebook with Python programming language and was published in Github. Python's pandas library was used for the analysis and findings will be visualised using Matplotlib and Folium libraries.

Data

In this project I used the following data sources from Helsinki region data share (https://hri.fi/en_gb/):

* Greater Helsinki area neighbourhood list and their coordinates
* Population per neighbourhood

This data is available in XML format. This is an example of one Neighbourhood and its' coordinate data:

```
<Placemark>
<name>Tullisaari</name>
<Point>
<coordinates>25.0277633723003,60.1778929296579,0</coordinates>
</Point>
</Placemark>
<Placemark>
```

Restaurant related information was fetched from Foursquare: how many restaurants of each type exists within given range of neighbourhood centre. Foursquare data comes in json format, out of which I will use the following information;

* name of restaurant
* category
* location coordinates

Methodology

The first task was to find out the different neighbourhoods, their location coordinates and merge that with population information from another data source. The neighbourhoods were illustrated on a graphical area map.

The restaurants for each neighbourhood were gathered from Foursquare. A fixed 500m perimeter circular area was used to figure out how many restaurants there were within each neighbourhood. Is it to be noted that not all neighbourhoods had a single restaurant within the given radius. I decided to leave these neighbourhoods out of further analysis although for someone looking to open a new restaurant they could be of interest to look into.

The next exploratory method used was to figure out how many restaurants there were within each neighbourhood and compare that with the population number. The result of this part was figure depicting number of restaurants per 1000 capita for each area. This was also illustrated on a map with colour coded markers.

Then the neighbourhoods were clustered based on their similarities for most common restaurant types. First the most common types of restaurants were explored for each area. K-means clustering, a typical non-supervised machine learning algorithm, was used to cluster the neighbourhoods into five distinct clusters. I used Python's popular Scikit-learn package to do the heavy lifting. Finally the clusters were illustrated on a map and neighbourhoods belonging the different clusters listed.

Results

The main finding of the exploratory part of the analysis was that the "number of restaurants per capita" figure varies massively between different neighbourhoods although this was to be expected.

Five clusters were for the K-means clustering analysis. Cluster number zero turned out to be the largest with most central Helsinki neighbourhoods belonging to it. In addition many suburban neighbourhoods, such as Kivenlahti in Espoo belong to cluster zero. Cluster zero is also the most diverse what comes to what types of restaurants are most common.

Cluster one is a lot more evenly spread-out cluster with only or two central neighbourhoods belonging to it. The restaurant split is also surprising even in cluster one. The most common category is "restaurant" which does not tell us much about the place. This would warrant a more thorough analysis of the Foursquare data. It is evident that there are many restaurants missing and also quite a few might be placed in slightly misleading category.

Cluster two is the "Indian restaurant" cluster with an Indian restaurant being the most common in all but one occasion. Vietnamese and German restaurants are also prominent. The neighbourhoods in the cluster are typically suburban.

Cluster three is the "fast food cluster" with places like Pukinmäki along the main ring highway as well as Otaniemi which is known for its high concentration of university students.

The final cluster, cluster four, is a small cluster with just five neighbourhoods in it. Scandinavian is the most common restaurant type within this cluster.

Discussion

The original tasks was to examine and analyse restaurants in Greater Helsinki area to provide help for anyone looking to setup a new restaurant. The analysis carried out in this project gives valuable pieces of data based insight that can certainly help, for example the locations and types of existing restaurants and neighbourhoods' "restaurant to population" ratio. The clustering provides intriguing view into neighborhoods' similarities by type of most common restaurant types.

However, it must be noted that incorporating more data into the study would be beneficial: for example the movements of people during the day and night could reveal more lucrative areas for a new restaurant. Information about large concentration of workplaces would be helpful to have. Also, many touristic areas with large hotels could shift the result as these were not available for this study.

The foursquare data as only source for restaurants in Greater Helsinki is also somewhat unreliable: many restaurants are missing altogether and categories can be misleading causing the analysis to be distorted or skewed. Only the decision to use fixed 500 meter radius where to look for restaurants is not very accurate. It would have been better to use the exact coordinates of neighborhoods.

Keeping the original business question in mind it would also make sense to study demographics and disposable income levels of people living in the different neighbourhoods.

Conclusion

The results demonstrate and visualise what types of restaurants are most common in each neighbourhood and how different neighbourhoods' relate with each other in terms of restaurant types. This type of information is hard to come by without a study like this and can, in my opinion, surely help in solving the original business question.

However, my recommendation is to use the insights provided as a starting point only keeping in mind the limitations and potential sources of error as described in the discussion section of this report. As an example of highly recommended further study would be to carry out a market research about people's preferred restaurant types.