A new restaurant in Copenhagen

In which region in Copenhagen would a new restaurant mostly survive?

1. Introduction

1.1 Background

A client wants to open a new restaurant in big Copenhagen area in Denmark. There are 51 regions in this area based on the postal codes. Before the client starts to think about the costs in different areas, he would like to firstly know which regions are more suitable for a new restaurant considering the competitions and economic conditions in a specific region. There are too many things that could have an impact on the survival of a restaurant. For example, local policy, tax rate, rental, transport, public infrastructure, shops, schools, etc. However, with the help of Foursquare data, we could easily get a dataset of surrounding venues which basically covers all kinds of assets around a region center.

1.2 Problem

The relations of all the venues around a region may not be explicit, but the structure of these venues could largely reflect the economy conditions of the region. How can we utilize all these venue data to get a view of how similar, from economic conditions point of view, these different regions can be? And based on the conclusion, can we decide whether or not should we open a new restaurant? Or in which regions should we suggest the client to open a restaurant? In another word, would a new restaurant generally survive the competition in a specific region?

1.3 Interest

Based on the venue dataset, we could group the 51 regions into clusters with similar economy conditions. Assume that the economy is stable, and all existing restaurants in each region are running in a good condition, we could then reasonably tell that a region with relatively less restaurants (compared with another region in the same cluster) could have more new restaurants. For example, assume that region A and B are in the same cluster, meaning these two regions have similar economic conditions (similar venue structures). If the restaurants in Region A takes 5% of all venues, and the number for region B is 2 %, we can say that the economy or competition in region B would allow more restaurants, as region A actually has proved that restaurants as 5% of all venues could survive in this region cluster.

There can be two phases to get the final conclusion. Phase 1: group the 51 regions into clusters and investigate which ones to choose for further selection. Phase 2: analyze each cluster and find the regions which is more suitable for opening new restaurant.

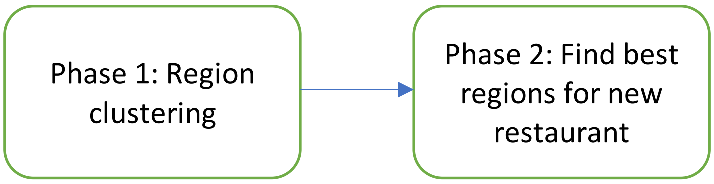

Phase 1: Region clustering
Phase 2: Find best regions for new restaurant

Chart1. phase chart

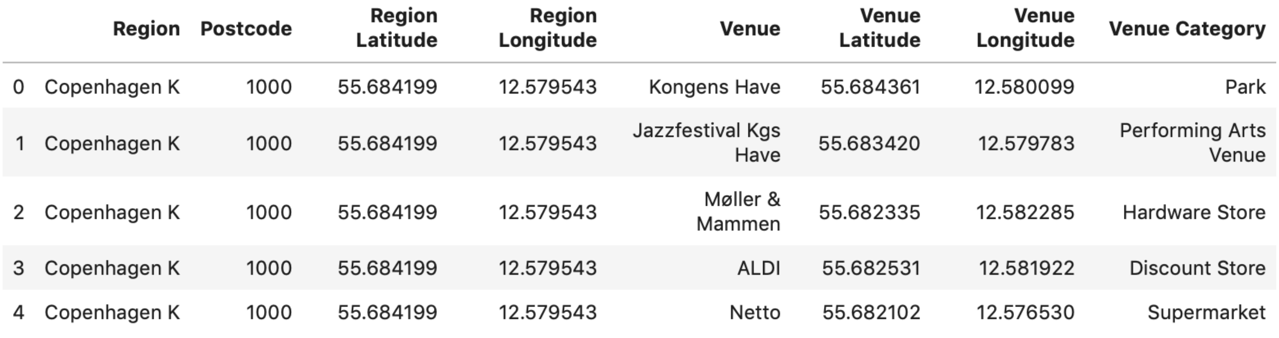
2. Data

2.1 Location data acquisition

The postal code data can be easily found from Wikipedia, together with the region name, the coordinates of the region centers can be fetched using Google map geocoding API. The python library *geopy.geocoder* can also help to get the coordinates, but some addresses are missing in this case, thus geocoding API is used here.

2.2 Venue data acquisition

As the coordinates data for all 51 regions in Copenhagen is available, the Foursquare API can be used in order to get the surrounding venues, with a radius of 500 meters and venue limits of 200. The venue categories are also extracted and stored for further analysis. The venue dataset should essentially contain the following fields: *region name, postal code, region coordinate, venue name, venue category, etc*.



	Region	Postcode	Region Latitude	Region Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Copenhagen K	1000	55.684199	12.579543	Kongens Have	55.684361	12.580099	Park
1	Copenhagen K	1000	55.684199	12.579543	Jazzfestival Kgs Have	55.683420	12.579783	Performing Arts Venue
2	Copenhagen K	1000	55.684199	12.579543	Møller & Mammen	55.682335	12.582285	Hardware Store
3	Copenhagen K	1000	55.684199	12.579543	ALDI	55.682531	12.581922	Discount Store
4	Copenhagen K	1000	55.684199	12.579543	Netto	55.682102	12.576530	Supermarket

Pic1, an example of the venue information

2.3 Feature selection

As described in the introduction part, there can be two phases to solve the problem. There could be different features that need consideration in these two phazes.

For phase 1, the venue structures in all regions need to be compared with one another, and the rate of occurrence of each venue category in each region could be used as the feature to compare, in order to calculate the similarity of the regions. For phase 2, as the purpose is to analyze and find the best regions for new restaurant, the venues with category of restaurant or food related should be all considered. The occurrence rate of the restaurants in each region of the same cluster need to be calculated and used as the feature for comparison.