Coursera Capstone Project for IBM Data Science Professional Certificate

Branching out in the Restaurant Business:

2. Data

Three types of data have been important in this Capstone project:

Statistics from Wikipedia: The name of the districts and general information about their number of inhabitants and geographical size was scraped directly from a Norwegian Wikipedia page about the districts in Oslo (https://no.wikipedia.org/wiki/Liste_over_Oslos_bydeler).

Geolocation data from Latitude: Geolocation data was used to find the Decimal degrees (DD) coordinates for the geographical centres of the 15 different districts in Oslo. The coordinates were gathered from Latitude (https://latitude.to)

Location data from Foursquare: Location data on venues was accessed through API and downloaded from Foursquare (https://foursquare.com). The venues were selected by specifying an area in each district that was the size of a circle with a radius of 2000 meters and with a centre in the middle of the district.

The three data sources were combined to create two datasets. The first dataset included information about all restaurant venues in each of the 15 districts (limited to 100) and the second dataset included information about all venues in each of the 15 districts (limited to 100). Each of the datasets included information about venue characteristics, such as name, location and type of venue. The first dataset was used to measure both competition and demand, where competition was understood as the presence of Chinese and other Asian restaurants in a district and demand was understood as similarities in food venue types between the district of Nordre Aker and the other districts in Oslo. The second data set was used only to measure demand, defined as similarities in (all) venues types between the district of Nordre Aker and other districts in Oslo (similarities in lifestyle).