A Venue Placer

Michael Korbman

Final assignment for the IBM Data Science Professional Certificate

Business Problem

Until a few years ago there was a clear distinction between family shops, operated in a single physical place in one town (for example a pizzeria), and shops owned by big franchise (for example McDonald). This difference is becoming blurrier while more and more small companies are opening more than one sample of their business in different towns or even in different neighborhoods of the same town. While the big franchise will have branches in each possible promising spot, for the smaller companies it is crucial to identify the optimal neighborhood to open the next shop of their small chain. They will also benefit from a forecast of the expected demand (quality and size) as a stronger basis for their strategic evaluation.

In this report we will describe a tool performing a statistical analysis of the neighborhoods in a city. Leveraging the data of existing venues, the tool will be able to forecast - within a certain margin of error – which neighborhood is "missing" a shop of a certain type and, on the other side, which shops are "inflated" in the same area. According to the scope of the project, the result is no more than a Proof-of-Concept (POC) in a specific area (Manhattan, NY) but it will reveal what the potentials of the tool are, as well as which areas would require a more detailed development.

The human analysis is still fundamental at this stage, for example in order to identify whether an observed large deviation from the average can be explained from concurrent reasons. Given the scale of the problem (the number of cities multiplied by the number of neighborhoods in each city) however, an automated screening will be necessary in most cases.

The proposed tool can then be offered to the small chains evaluating their specific kind of shops, as a support in the process of localizing the area to open a new branch.

Data

For the POC we will focus on the neighborhoods of Manhattan, New York. We want to take advantage of correlations between the neighborhoods, which we will evaluate based on the relative density of venues. To retrieve data from the venues, the Foursquare API is called providing as input parameters the coordinates of the neighborhoods and the radius. We are using a radius of 500 meters (thus retrieving venues in a 0.8 Km^2 circle around the center).

In order to increase the statistical significance of the results, we discard all venues having less than 20 shops in Manhattan. This reduces the 339 types of venues returned to just 46. In each Manhattan neighborhood the absolute number of a certain type of venues is normalized over the total number of venues in that neighborhood (i.e. averaged). The resulting table (`df_nfreq` in the notebook) contains the average density of venues in the neighborhoods sorted according to their type; we will perform the analysis described in the following sections using this data.