20 september 2019

Finding optimal salad bar location

IBM Applied Data Science course Capstone Project

Xander Mol

Table of Contents

Cover page photo credit:

Jakub Kapusnak

https://www.foodiesfeed.com/free-food-photo/healthy-vegetarian-salad-buffet/

Executive Summary

Introduction

Background

This exercise has its origin in the IBM Applied Data Science Capstone course on Coursera and the Peer Graded Assignment concluding this course.

Assignment: ‘The Battle of Neighborhoods’
‘Come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve’.

The chosen problem is a fictional research assignment to research the optimal place to start a specified type of new business given the specified requirements.

Problem

What is the optimal neighborhood to start a new salad bar, based on a comparison between neighborhoods in the four cities considered using demographic and income data combined with data on other venues in the area, based on the following requirements and background:

- Only locations in the four biggest cities in the Netherlands are considered, being Amsterdam, Rotterdam, Utrecht and The Hague;
- From market research a projected ideal target group is defined: single household females;
- Areas with abundant availability of other restaurants are preferred as experience indicates that this attracts more potential customers to new restaurants;
- Preferably none or few competing salad bars in the area.

Target group for research results

Results of this research is targeted on the aspiring new founders of this salad bar to be established. There interest is to choose a location for their new businesses that has the highest chance of being successful given market research on target group they already did perform.

Their interest in using a data science approach for this problem is to leverage data to make the most objective decision possible.

Data

Demographic data

Given the target group requirement of single household females, demographic data is necessary for the different areas we want to compare.

As basis for the area boundaries, in the Netherlands the postal code system is a very good proxy of the different neighborhoods as it is more consistent and also much more granular than for example borough ('wijken') and neighbourhood ('buurten') names.

Dutch postal code have a four digit two letter format (1234 AB). Together with the house number it uniquely identifies an individual postal address. Good proxy for neighbourhoods are the four digits, the full postal code is on street level so very low level.

For demographic data, the Dutch Central Bureau of Statistics (CBS) is by far the most logical source. It derives it's data from the government own systems and from data delivery that companies are legally obliged to deliver to the government. Therefore, CBS data is considered to be the most accurate, complete and comprehensive.
Additionally, the CBS provides almost all data for free and open for public use in several formats and views.

After browsing available datasets, this set was selected as best fitting the requirements for this project, being demographic data on at least sex and household size per four digit postal code, where the dataset as of 1 January 2016 was the most recent to be found:

Population by four digit postal code as of 1 January 2016

https://www.cbs.nl/nl-nl/maatwerk/2016/51/bevolking-per-viercijferige-postcode-op-1-januari-2016

Format is an Excel file, fields in this file include:

- Four digit Dutch postal codes;
- Total population;
- Population split per sex and five-year age cohort;
- Migration background;
- Household size and nature of households (kids/no kids)

Geolocation data

To be able to map our postal codes, and also to be able to retrieve venue data for this postal code, geolocation data (latitude/longitude) is required for all four digit postal codes.

Amongst others on this location a file can be found exactly doing this:

https://git.tuxm.nl/tuxmachine/postcodes/src/4329c858db24b79523fd3fbbaf2df138ccaf16cd

Format is a Comma Separated Values (CSV) file.

Income data

To assess how affluent our prospected customers are, income data per postal code is handy as well to enrich the data.

For this, another dataset of the CBS is used:

Spendable income per postal code, 2004-2014

https://www.cbs.nl/nl-nl/maatwerk/2017/15/besteedbaar-inkomen-per-postcodegebied-2004-2014

Format is an Excel file, fields in this file include:

- Four digit Dutch postal codes;
- Number of households;
- Average spendable income;
- Spendable income standardised for householdsize.

Only the 2014, the most recent year available in this format, will be used as we will not research changes in time.

Names of areas and neighbourhoods based on postal code

As names are more easy to relate to as numbers, it is handy to enrich the four digit postal code to the area/neighbourhood/city names it relates to.

For this, again a CBS dataset is used:

Neighborhood, district and municipality 2018 for postcode house number

https://www.cbs.nl/nl-nl/maatwerk/2018/36/buurt-wijk-en-gemeente-2018-voor-postcode-huisnummer

Format is a zip file with four Comma Separated Values (CSV) files:

- pc6hnr20180801_gwb-vs2.csv with link per postal code/housenumber to neighborhoodm district and municpality codes;
- buurtnaam2018.csv with all names of the neighborhoods per neighborhood code;
- gemeentenaam2018.csv with all names of the municipality per municipality;
- wijknaam2018code.cvs with all names of the districts per district code.

Foursquare location data

To obtain information of surrounding venues for a given longitude/latitude, the Foursquare API is used.

See https://developer.foursquare.com/docs for API documentation.

This API will be used to obtain a list of the (max 100) venues surrounding the center of each postal code area and to obtain specific data on presence of salad bars.

Methodology

Results

Discussion

Conclusion