6/3/2021

Capstone Project

The battle Of Neighborhoods

Evert Johannes Woudstra
CONTACT INFORMATION

Table of Contents

1 Introduction

This project is the final assignment of the IBM Professional Data science Certificate. This certificate consists of a series of 9 courses in which data science skills, including Data Science Methodology, data mining and analysis with python has been studied and applied. The main objectives for this project are the following:

- leverage location data provided by Foursquare and the Data of Amsterdam website
- applying data science skills in machine learning and data visualization
- Segmentation and clustering of location data in order to explore neighborhoods to help addressing a business problem

The results will be presented in this report as well as in a blogpost online and a Jupyter notebook published on Github.

1.1 Context of the project

Amsterdam is the capital city of the Netherlands and is famous for its multicultural identity. With 872.922 inhabitants from 177 different nationalities, it belongs to one of the most divers cities of the world.
The city has the largest number of bars and pubs of any city in the Netherlands, although the number is declining over the last decade. Small traditional pubs are transitioning to more fancy venues targeting the market of “hipsters” and the modern millennial customer seeking for experiences and high-quality food.
The opposite trend is visible in the category of restaurants. Since 2010 the number of restaurants is increasing not only limited to the old city center but also surrounding neighborhoods. New forms like food markets as “Rollende Keukens” and concepts restaurants are being developed over the last few years. The growing popularity of Amsterdam worldwide and the annually increasing number of tourists indicates good prospects for opening a new restaurant concept.

1.2 Business problem

Opening a restaurant in a city like Amsterdam is a complex project. Although a good food concept can be a profitable and fulfilling business, reality is that many restaurants fail during their first year, mainly due to a lack of planning and having a sound business plan.
In the business plan many topics like funding, choosing a concept, advertising and logistics and choosing an optimal location has to be carefully considered.
The scope of this project is addressing the question what the optimal location could be for opening an Italian restaurant in the city of Amsterdam with a concept of choice.
This project could be potentially interesting to:

- entrepreneurs who want to start a food business in Amsterdam
- brokers to advise people on buying properties based on their business plan

2 Data description

To find an answer to the business problem we will make use of the following two data sources:

- Geographical data provided by the Amsterdam Data Website
- Location data provided by Foursquare

In this chapter these sources are further introduced and explained.

2.1 Data sources

The Amsterdam data website (https://maps.amsterdam.nl) has all the geographical data this project needs. The geographical data of the neighborhoods is downloadable in different formats like .csv and Geo Json.
The advantage of the Geo Json format, which is used in this project, is its precise definition of each neighborhood where the borders are defined by a polygon (fig 1) consisting of al the latitudes and longitudes.

```
In [15]: amsterdam_data.geometry.head(1)

Out[15]: 0    POLYGON ((4.90326 52.37658, 4.90298 52.37668, ...
         Name: geometry, dtype: geometry
```

Fig 1. Example of a Geo Json polygon object.

This list of location coordinates describes the borders of each neighborhood as shown in figure 2.

```
In [8]: p_list = amsterdam_data.geometry[0]
        p_list

Out[8]:
```

Fig 2. Plot of a polygon entry in the geojson dataset amsterdam_data.

The second source which is used in this project is the Foursquare database.
Foursquare (https://foursquare.com) is a company which build and maintains a massive dataset of accurate location data. This data is freely available using the RESTful API.
After registration on their website on can do a limited number of queries with different search parameters so called “endpoints” and collect a dataset containing all the interesting venues a location has to offer.
Figure 3 shows what a RESTfull API url would looks like:

```
# create the API request URL
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'
```

Fig 3. Example of quiry on Foursquare database.

As show in figure 3, one of the limitations of this API is the search method “radius” around a location of interest.

Since neighborhoods in old cities like Amsterdam tends to have irregular forms and sizes there is a need for a more specific definition of the search area.
If the radius has been defined to small, interesting venues could be missed in the neighborhood. On the other side, choosing a to large radius will result in duplicate venues and venues assigned to the wrong neighborhood.

Therefore, a method has to be developed finding most of the important venues in each neighborhood while preventing duplicates or venues wrongly assigned.

2.2 Data cleaning and preprocessing

To prepare and use the data for the analysis it has be free of errors and be in a format which can easy be used and manipulated.

As introduced in the previous chapter. Searching neighborhoods varying in shape and size by an API request with a fixed radius can result in either missing many venues or having duplicated values spread over multiple neighborhoods. To overcome this problem an optimal radius has to be chosen and the dataset will be cleaned of duplicates. Finally, each venue assignment will be checked based on its location and the definition of the neighborhood polygon.

Once the foursquare data is retrieved, cleaned and preprocessed, the JSON format has to converted to a data frame on which the analysis can be performed. Many attributes in the retrieved data consist of categorical data which have to be converted to numerical values.

Finally, before starting the analysis the data has to be normalized to guarantee all features will have an equal and predictable impact on the final analysis and data visualizations.