Coursera Capstone

IBM Data Science Capstone Project

Strategic Locations for Supermarket-Chain in Berlin, Germany

By: Helal Chowdhury

April 2020

1. Background

Almost every single adult living in the world needs to visit supermarkets or marketplaces at least few times in a month. Supermarket is a place where you can buy most the things necessary for everyday-life. A traditional supermarket offers groceries, utensils, toiletries cosmetics etc. In addition, few supermarkets are multipurpose offering furniture, clothing etc. and many more. Considering demands from every part of the city, often we can see the appearance of one or more supermarkets. Berlin is a well-developed and capital city of Germany, with lots of business opportunities and business friendly environment attracting many different players into the market. To be specific, there are some hundred supermarkets in the city of Berlin. Since the overall population in Berlin is on the rising trend, new supermarkets or new branches of an existing chain are appearing on a regular basis. This is an opportunity for any new company or an existing one offering new facilities in different parts of Berlin. However, the market is highly competitive, for any business decision, opening a number of new locations needs serious consideration and is a lot more complicated than usual. Any new business decision or expansion endeavor from a venture needs to be reviewed carefully and strategically so that the return on investment will be sustainably reasonable, and more importantly can be considerably less risker. Particularly, strategic locations are of paramount importance in this case. Selecting few random locations or only central locations might not lead to success always.

2. Business Problem

Berlin is one of the densely populated cities in Germany, with more than 3.6 million residents and an average of almost 1.2 million visitors each month [1]. Let's say a large supermarket chain is interested in opening some new branches in Berlin and therefore looking for some strategic locations. The objective of this project is to analyze and select such locations, say five for example, in the capital city of Berlin. Using data analysis and machine learning techniques like clustering, this project aims to provide answer to the business question: In the city of Berlin, if a local or foreign supermarket giant looks for opening few new spots, which strategic locations should be preferred considering business potential?

3. Target Audience

This project is particularly useful for local or foreign supermarket chains or for a new venture willing to do similar business in the capital city of Berlin. The objective is to locate and recommend to the management of interested parties which set of neighborhoods will be the best choice to start their services. The management also expects to understand the rationale of the recommendations in the report. This project is timely as the city is currently shows an increasing trend of population, especially due to the flood of new people from war-hit zones. The success criteria of this project will be a good recommendation of the neighborhood choice to the management based on some key factors: higher population, less competition, higher density and neighborhood similarities.

4. Working with data

To find the answer of the above business problem, we will use the following data:

- List of neighborhoods including borough in Berlin from Wikipedia [2].
- Latitude and longitude of all neighborhoods. This is required for getting venue data and for the visualization of neighborhoods. Related data will be obtained using Python geocoder package.
- Population and density of each neighborhood with which we will shortlist the candidate locations, from Wikipedia [2].
- Venue data with the density of supermarkets with which we will perform clustering on the neighborhoods in order to identify a strategically similar location.

The Wikipedia page in [2] contains all basic information of Berlin including borough, neighborhood, population, area, density etc. for each locality. The page shows that there are 12 borough and 96 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python Pandas. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us latitudes and longitudes. After that, we will use APIs from one of the popular sources Foursquare to get the venue data from all neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers [3]. One of the features of Foursquare APIs is to provide a list of venues within a specific location, based on latitude/longitude coordinates and a radius. By passing the proper parameters via an HTTP request, we get the required data. The *location* object contains the coordinates of each venue, which will be used to associate it with its respective neighborhood. The *categories* array will be used to categorize the neighborhood. Basically, we will count how many venues from all available categories are found on each neighborhood, and then use that information to compare neighborhoods in Berlin. Foursquare API provides many categories of the venue data; we are particularly interested in similar neighborhoods considering supermarket category in order to solve the business problem.

This project requires many data science skills, from web scraping (from Wikipedia), working with API (of Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (using Folium).

5. References

[1] https://www.statista.com/statistics/568463/tourism-arrivals-berlin-germany-by-origin/
[2] https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin
[3] https://developer.foursquare.com