Coursera.org

IBM Applied Data Science Capstone

Proposal for a new vegetarian / vegan restaurant in Berlin, Germany

By: Tatiane MP

January 2020

Introduction

In recent years, the German vegetarian/vegan population has rapidly grown. As of 2015, the vegetarian/vegan market was worth $520 million and saw a increase rate of 17%. This is the result of a population that is more and more concerned about animal welfare, the environment, and, especially, their health.

According to the European Vegetarian Union, 10% of German consumers (7.8 million individuals) are vegetarians, and 1.1% are vegans (900,000 individuals). These numbers have doubled since 2006 and many believe that these numbers will only continue to increase in the coming years.

For this reason, the opening of a vegetarian / vegan restaurant is a great business opportunity in a so eco – friendly city like Berlin. But of course, as with any business decision, opening a new venue requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the vegetarian / vegan restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best location in the city of Berlin, Germany to open a new vegetarian / vegan restaurant. For this purpose, it will be used data science methodology and machine learning techniques like clustering to find out which part of the city would be the ideal place to open a new vegetarian / vegan restaurant. This project aims to provide an answer to a very important question: If a property developer is looking to open a new vegetarian / vegan restaurant in Berlin, Germany, where would be the best recommend place to do it?

Target Audience

As with any restaurant, vegetarian and vegan businesses need a location where they have a strong chance of building positive buzz and developing a base of repeat customers.

Entrepreneurs pointed out that new owners must consider questions like how accessible a spot is to likely patrons, the foot traffic in the area, plans for future development nearby and the affordability of the lease. For a restaurant specializing in vegetarian cuisine, the best option is usually an relatively upscale area frequented by health-conscious people.

Therefore, this project is particularly useful to entrepreneurs, property developers and investors looking to open or invest in vegetarian / vegan restaurants in Berlin (DE). This project is timely as the city has a great potential to such a kind of business, because like the Culture Trip Internet Page[1] writes: “But the ever increasing number of vegetarians in Berlin has given rise to a vibrant meat-free dining scene: vegan currywurst, all-you-can-eat buffets, and Himalayan dumplings are just a few of the vegetarian offerings in the German capital.”

A survey from "Forsa" also revealed that approximately 42 million people in Germany identify as flexitarians aka "part time vegetarians." Professionals at the German Official Agencies estimate that by 2020 over 20% of Germans will eat mostly vegetarian and this fact creates an excellent opportunity for those ones who would like to start a vegetarian / vegan restaurant. Besides that, many vegetarian / vegan restaurants also sell food products as well as vegetarian / vegan cookbooks, and this can be a terrific way to increase profits.

[1] https://theculturetrip.com/europe/germany/articles/the-best-vegetarian-restaurants-in-berlin/

Data

To solve the problem, we will need the following data:

- List of Berlin's districts. This defines the scope of this project which is confined to the city of Berlin, the capital of the Germany.
- Latitude and longitude coordinates of those districts. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to vegetarian / vegan restaurants. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them:

- Wikipedia[2], that contains a list of all districts of Berlin, with a total of 12 districts. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and BeautifulSoup packages. Then we will get the geographical coordinates of the districts using a GitHub CSV file that's already there.
- Foursquare API: we will use Foursquare API to get the venue data for those districts. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Vegetarian / Vegan Restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

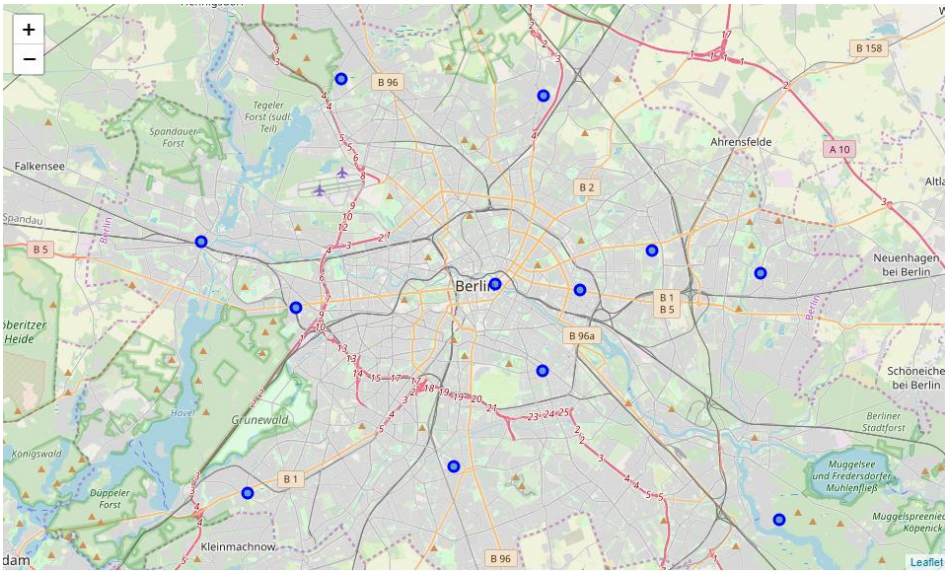
[2] https://de.wikipedia.org/wiki/Kategorie:Bezirk_von_Berlin

Methodology

Firstly, we need to get the list of districts in the city of Berlin. Fortunately, the list is available in the Wikipedia page (https://de.wikipedia.org/wiki/Kategorie:Bezirk_von_Berlin).

We will do web scraping using Python requests and BeautifulSoup packages to extract the list of districts data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use a CSV file hosted on GitHub, that already has all informations we will need. An alternative is to use the Geocoder package or the Geopandas package, which allow us to convert address into geographical coordinates in the form of latitude and longitude, but this option was too slow on my laptop, so I decided to use the CSV file instead, also because Berlin has only 12 districts. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the districts in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Berlin.

The generate map of Berlin's districts looks like this (for a better visualization, see the Jupyter Notebook uploaded at Github):



Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the district in a Python loop.

Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each district and examine how many unique categories can be curated from all the returned venues.

Then, we will analyse each district by grouping the rows by district and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering.

Since we are analyzing the "Vegetarian / Vegan Restaurant" data, we will filter the "Vegetarian / Vegan Restaurant" as venue category for the districts.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the districts into 3 clusters based on their frequency of occurrence for "Vegetarian / Vegan Restaurant". The results will allow us to identify which district has higher concentration of vegetarian / vegan restaurants while which district have fewer number of them.

Based on the occurrence of vegetarian / vegan restaurants in different districts, it will help us to answer the business question we proposed above, that means which district is most suitable to open new vegetarian / vegan restaurant.

Results

After cleaning and separating the data related to the presence of vegetarian / vegan restaurants in Berlin, we obtained the following table, which confirms the great business potential that this type of venture can bring, since there is not a wide variety of restaurants of this type listed by Foursquare.

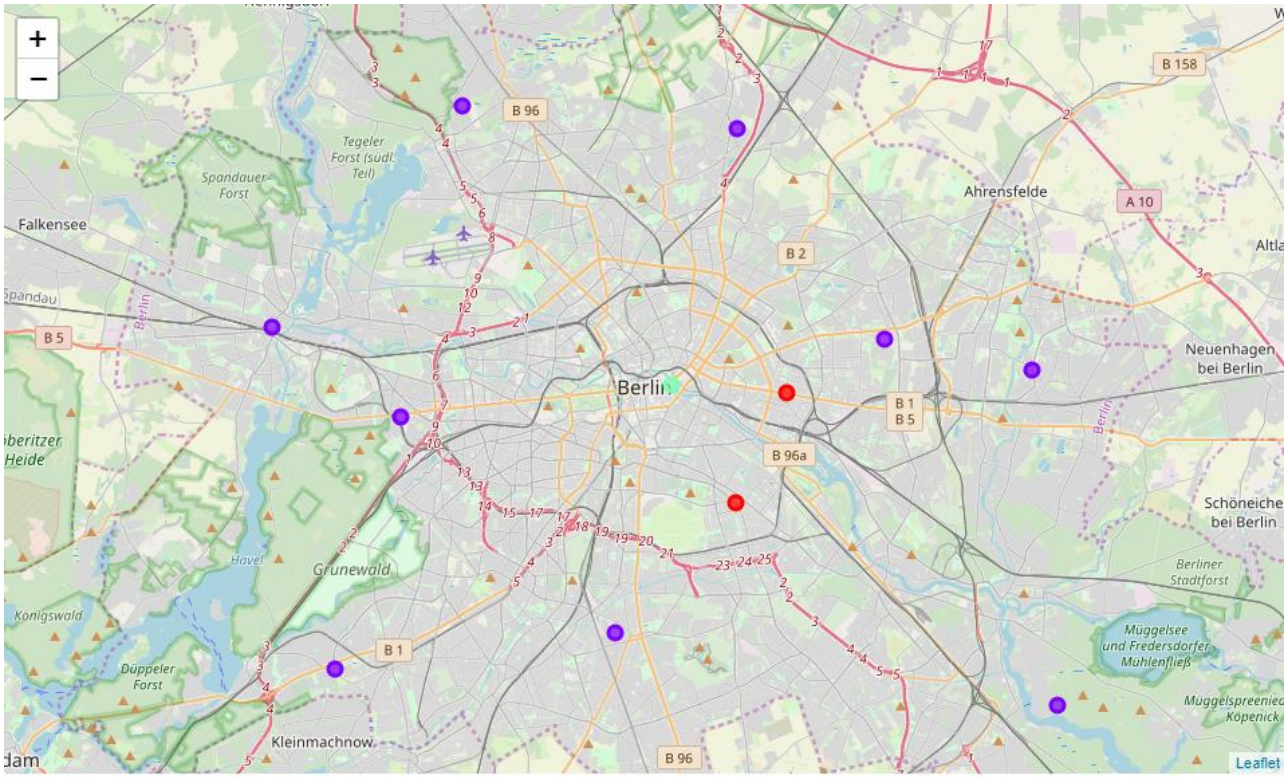
	Districts	Vegetarian / Vegan Restaurant
0	Charlottenburg-Wilmersdorf	0.00
1	Friedrichshain-Kreuzberg	0.04
2	Lichtenberg	0.00
3	Marzahn-Hellersdorf	0.00
4	Mitte	0.01
5	Neukölln	0.04
6	Pankow	0.00
7	Reinickendorf	0.00
8	Spandau	0.00
9	Steglitz-Zehlendorf	0.00
10	Tempelhof-Schöneberg	0.00
11	Treptow-Köpenick	0.00

The results from the k-means clustering, where we divided the districts of Berlin into 3 clusters based on the frequency of occurrence for "Vegetarian / Vegan Restaurant" is showed below.

We can see that the map is not very populated, and the most of the vegetarian / vegan restaurants are located around Berlin's downtown. This result matchs with the table (above) we've obtained after separate this specific data into a *pandas* DataFrame.

The clusters are divided in this way:

- Cluster 0: Districts with a few number of vegetarian / vegan restaurants, represented in purple color.
- Cluster 1: Districts with no existence of vegetarian / vegan restaurants. No representation.
- Cluster 2: Districts with very low concentration of vegetarian / vegan restaurants. Represented in red color.



Discussion

As observations noted from the map in the Results section, most of the vegetarian / vegan restaurants are concentrated in the central area of Berlin, with the highest number in cluster 0 and a few number in cluster 2. On the other hand, cluster 1 has so good like no vegetarian / vegan restaurant in their districts. This represents a great opportunity and high potential areas to open a new vegetarian / vegan restaurants as there is very little to no competition from existing restaurants of this kind.

Meanwhile, the results also show that vegetarian / vegan restaurants mostly happened in the central area of the city, with the suburb area still have very few vegetarian / vegan restaurants. Therefore, this project recommends property developers to capitalize on these findings to open new vegetarian / vegan restaurants in districts in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new vegetarian / vegan restaurants in districts in cluster 2 with a few competition. Lastly, property developers are advised to avoid districts in cluster 0 which already have a moderate concentration of vegetarian / vegan restaurants and suffering from a more intense competition.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new vegetarian / vegan restaurant.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: The districts in cluster 1 are the most preferred locations to open a new vegetarian / vegan restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new venue.

In this project, we only consider one factor i.e. frequency of occurrence of vegetarian / vegan restaurants, but there are other important factors such as gender, education level and income of residents that helps the stakeholders' decision making and is better tool to find an optimal place to open a new vegetarian / vegan restaurant.