IBM Data Science Capstone Project

The Battle of Neighborhoods

Finding the best place to open Falafel and Middle east vegan restaurant in Berlin

By

Nedal Shami

Problem

In this project we will explore Berlin city to find out where the most appropriate place to open a Flafel Middle east vegan food.

Background

Berlin is the capital and largest city of Germany by both area and population. Its 3,769,495 inhabitants as of 31 December 2019 make it the most populous city of the European Union, according to population within city limits.

Berlin is well known for its offerings of vegetarian and vegan cuisine and is home to an innovative entrepreneurial food scene promoting cosmopolitan flavors, local and sustainable ingredients, pop-up street food markets, supper clubs, as well as food festivals, such as Berlin Food Week.

Therefore, it is one of the potential places for the success of this type of restaurant (Falafel and Middle east vegan restaurant).

Methodology

- Data was scraped and cleaned, and features selection was made.
- Map created for neighborhood of Berlin
- We use Foursquare API was used to figure out all neighborhood venues in Berlin.
- Neighborhoods are classified using K-means unsupervised ML algorithm.

Data

The main data sources are:

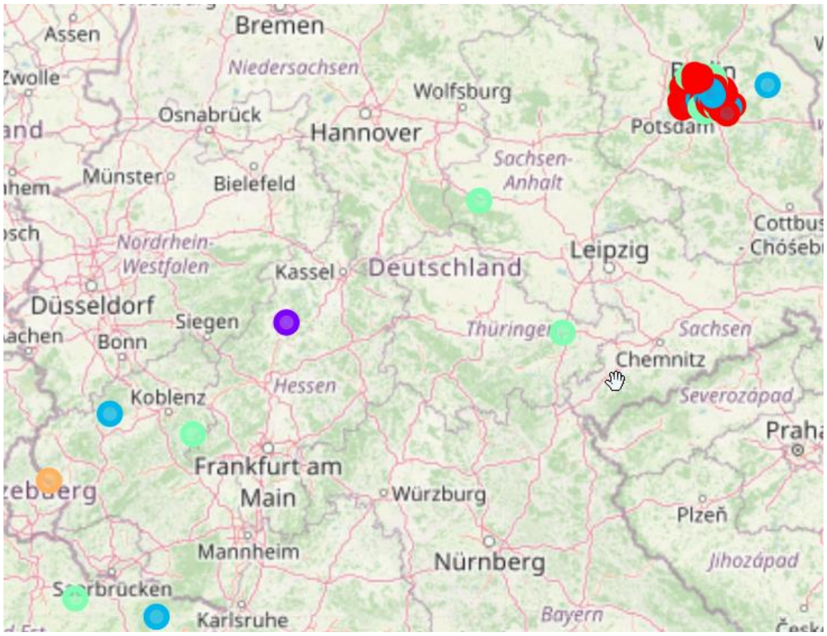
- Wikipedia, which provide a list of districts and neighborhoods of Berlin at: https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin
- Venues data from Foursquare API

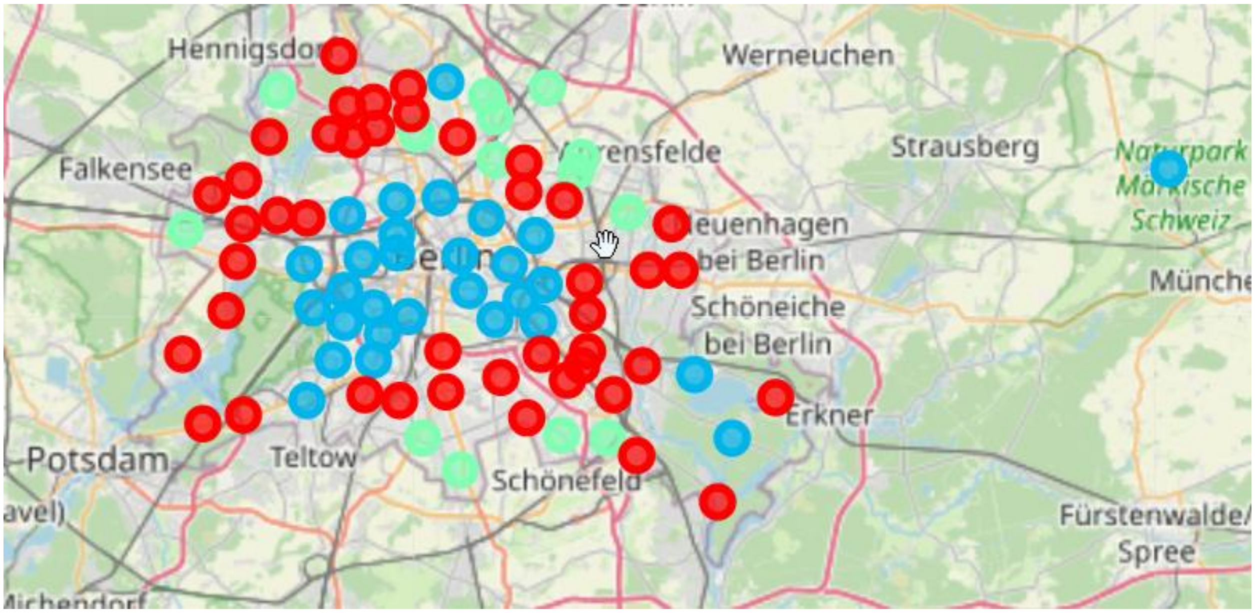
Data acquisition and cleaning:

- We use web scraping techniques to get data from the web (Wikipedia)
- Convert data to Pandas dataframe, and structure all data in one table.
- Clean data (remove unwanted columns and remove the unneeded characters from some columns).
- Using geopy library to add latitude and longitude to each neighborhood.
- Use Foursquare API to get venues for each neighborhood. To make it simpler we set a limit to 100 venues with radius of 1000 meters of every neighborhood.

Using unsupervised learning (K-means) to classify neighborhoods

- Create one-hot encoding to the venue's category for each neighborhood.
- Calculate the frequency of categories for each neighborhood and get the top 5 most common venues in each neighborhood.
- We use K-means clustering algorithm to classify neighborhood, we choose k to be 5 clusters.
- We plot the clustered neighborhood to visually illustrate it.


Bremen
Assen
Zwolle
Niedersachsen
Wolfsburg
Osnabrück
Hannover
Potsdam
Münster
Bielefeld
Sachsen-Anhalt
Cottbus - Chóśebu
Nordrhein-Westfalen
Kassel
Deutschland
Leipzig
Düsseldorf
Siegen
Thüringen
Sachsen
Bonn
Chemnitz
Koblenz
Hessen
Severozápad
Praha
Frankfurt am Main
Würzburg
Plzeň
Mannheim
Nürnberg
Jihozápad
Saarbrücken
Karlsruhe
Bayern


Hennigsdo
Werneuchen
Strausberg
Naturpark Märkische Schweiz
Falkensee
Ahrensfelde
Neuenhagen bei Berlin
Berlin
Schöneiche bei Berlin
Münche
Erkner
Potsdam
Teltow
Schönefeld
Fürstenwalde/ Spree
Michendorf

Results and conclusion:

Obviously, there are three main clusters:

- Cluster 0 with Red color on the map.
- cluster 2 with blue color.
- cluster 3 with green color

the other two clusters seem located in remote areas, and they are - definitely- not in our target.

From examining the mentioned three cluster we can note that clusters 0 and 3 are located in an industrial – companies areas, and cluster 2 is located at residential areas.

Neighborhoods that are classified as cluster 2 are seems good candidates of our target, especially that we see from our examining to this cluster that there are a few restaurants that offers Falafel and Middle east food.