Coursera Capstone

IBM Applied Data Science Capstone

Opening a Falafel fast food restaurant in Berlin, Germany

1. Introduction

Vegetarian food becomes a habit for plenty of people, oriental cuisine is also a favorite one sense it serves many vegetarian dishes. For many people, Falafel sandwich or snack is one of the favorite oriental vegetarian food in European countries like Germany.

In recent years, the country has embraced vegetarian and vegan cuisine, and today it is almost always possible to find several delicious vegetarian options on every restaurant menu.

se, as with any business decision, opening a new fast-food restaurant – even as small business - requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurant is one of the most important decisions that will determine whether it will be a success or a failure.(source: https://fearlessfemaletravels.com/eating-vegetarian-in-germany/)

"It's certainly fashionable. If you really want to be cool in Berlin, you've got to be vegan," Sebastian Joy of Vebu, Germany's largest vegetarian and vegan organization, told DW. (source: https://www.dw.com/en/berlin-vegan-capital-of-the-world/a-35951064-0)

So in Germany and especially in the multicultural Metropol Berlin such restaurants will be a good and profitable low cost business.

- ## *Business Problem*

The objective of this capstone project is to analyze and select the best locations in the city of Berlin, Germany to open a new Falafel restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question:

In the city of Berlin, if an investor is looking to open a new Falafel fast-food restaurant or even a series of restaurant with a special brand, where would you recommend that they open it?

- ***Target Audience of this project***

This project is particularly useful to middle and small investors looking to open or invest in new Falafel fast-food restaurant in the capital city of Germany i.e. Berlin. This project is timely as the city is currently suffering from oversupply of restaurants but not vegan ones!

According to "organic-market.info": The vegan trend has now reached the gastronomy sector. According to the GV-Barometer 2017, a rich variety of vegetarian food is one of the most important trends in gastronomy of the future. More and more guests want to see healthy vegetarian dishes on the menu, and 58 percent of entrepreneurs in gastronomy assume that the importance of vegetarian food will continue to grow. In big and medium-size cities in Germany there are currently at least 616 purely vegetarian eateries, which is nearly twice as many as last year. Berlin is clearly in the lead with its 193 wholly vegetarian restaurants and cafés but in other cities too you find plenty of places where you can try out the vegetarian offer. (source: https://organic-market.info/news-in-brief-and-reports-article/germany-9-3-million-vegetarians-and-vegans.html)

2. Data

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

• _List of neighborhoods in Berlin. This defines the scope of this project which is confined to the city of Berlin, the capital city of the country of Germany.

• _Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

• _Venue data, particularly data related to restaurants. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://de.wikipedia.org/wiki/Verwaltungsgliederung_Berlins) contains a list of neighborhoods in Berlin, with a total of 12 boroughs. 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 neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

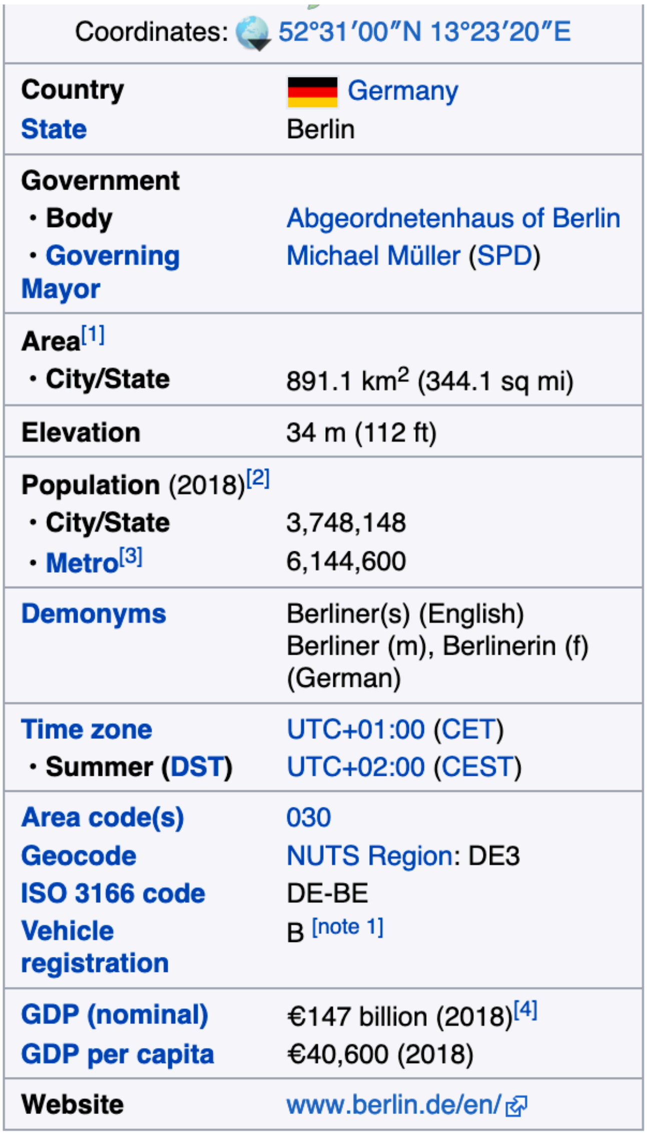
After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 150,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the restaurants 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). Next, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

3. Mythology:

Firstly, we need to get the list of boroughs in the city of Kuala Lumpur. Fortunately, the list is available in the Wikipedia page:

https://de.wikipedia.org/wiki/Verwaltungsgliederung_Berlins

Berlin is the capital and largest city of Germany by both area and population



Coordinates: 52°31′00″N 13°23′20″E

Country	Germany
State	Berlin
Government	
• **Body**	Abgeordnetenhaus of Berlin
• **Governing Mayor**	Michael Müller (SPD)
Area[1]	
• **City/State**	891.1 km^2 (344.1 sq mi)
Elevation	34 m (112 ft)
Population (2018)[2]	
• **City/State**	3,748,148
• **Metro**[3]	6,144,600
Demonyms	Berliner(s) (English) Berliner (m), Berlinerin (f) (German)
Time zone	UTC+01:00 (CET)
• **Summer (DST)**	UTC+02:00 (CEST)
Area code(s)	030
Geocode	NUTS Region: DE3
ISO 3166 code	DE-BE
Vehicle registration	B [note 1]
GDP (nominal)	€147 billion (2018)[4]
GDP per capita	€40,600 (2018)
Website	www.berlin.de/en/

As of 2012, the twelve boroughs are made up of a total of 96 officially recognized localities (Ortsteile). Almost all of them are further subdivided into several other zones (defined in German as Ortslagen, Teile, Stadtviertel, Orte etc.). The largest Ortsteil is Köpenick (34.9 km2 or 13.5 sq mi), the smallest one is Hansaviertel (53 ha or 130 acres). The most populated is Neukölln (154,127 inhabitants in 2009), the least populated is

Malchow (450 inhabitants in 2008).

We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods and boroughs 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 Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After getting the data of Brandenburg Gate which is the main place in Berlin, we will populate the data into a pandas DataFrame and then visualize the neighborhoods 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.

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 neighborhoods 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. By doing so, we are also preparing the data for use in clustering. So we check How Far are Restaurants from the core location, and explore the other venues around it. We will extract venues using search queries and will collate venues provided by foursquare and the ones extracted through hitting search query

API. We will get the categories of venues, then get the rating and tips for all venues to get the final list.

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 be collating the location of centroid of all clusters and midpoint of all venues to get more accurate location. The results will allow us to identify the final location suitable for the restaurant.

4. Discussion

most common categories of venues are temple, church, palace, museum in Berlin.

Average distance between Restaurants and core location is 2711 meters, which is normal sence i choose 10km radius in such a big city like Berlin.

there is no Falafel Restaurant close to the location, the only possible concurrent is 'Samadhi Vegetarian Restaurant' but its a Vietnamese one, so deferent cuisine.

From above report ,we could get an idea why the predicted one is pointed/clustered on the given spot. First most thing could be the center of attraction for the place.

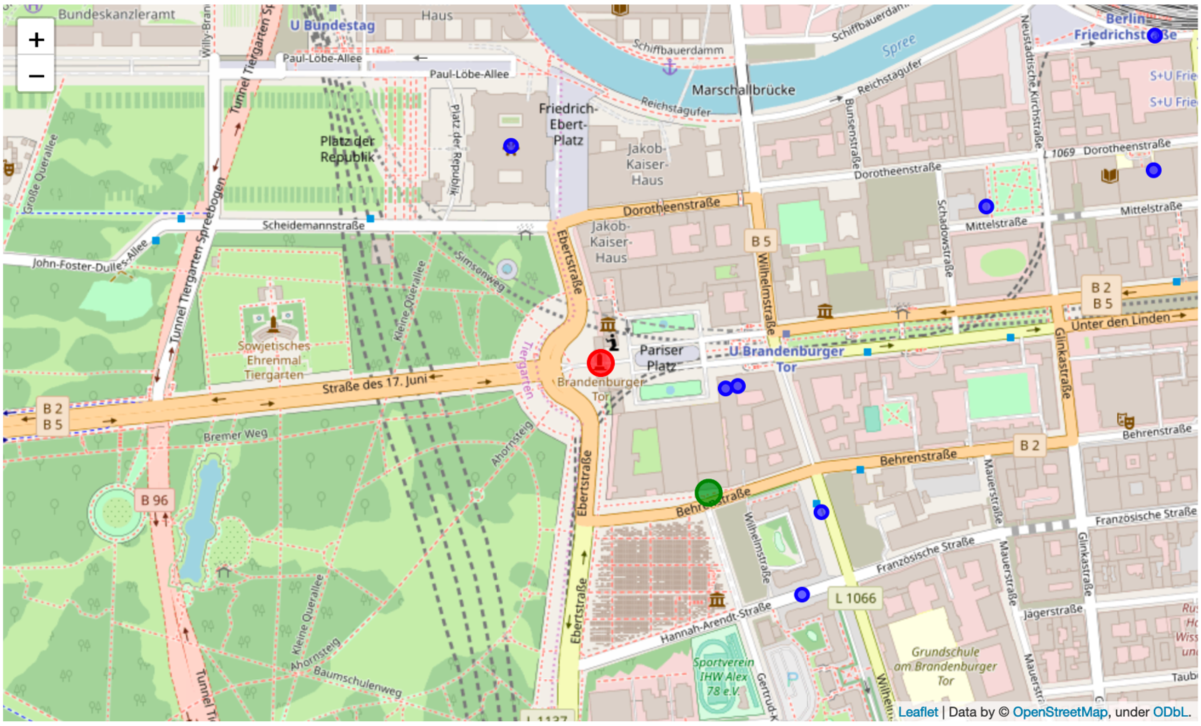
KMeans have figured out the most common place for all the venues. This output was very adjacent to the core location. This proves the accurate spotting of our predicted algorithm.

Despite of the findings, there were some lack in data. Tips and ratings were missing for most of the venues. Also when I compared foursquare data with google map ,i could see there were many restaurants and venues found missing in foursquare.

foursquare also limits the free use of it, it was very difficult to extract ratings & tips, but the results I've got are reasonable and make sense.

5. *Conclusion*

As a business person, one would be able to set up a restaurant on given spot. This will bring revenue automatically as we have located in very near to core one. We proved this with Kmeans.



The green circle shows the predicted location

Picture from google maps shows a view of Behrenstrasse (the chosen location)

6. *Future Expectation:*

As mentioned earlier, most of data needs to be extracted from googlemaps. Even though we got somewhat accurate prediction. To be very confident on concluding our output, we may need more data to analyze. More accurate data could be found in German language.

Research based on restaurant reviews and menus could be used for future purpose.

7. *My Experience:*

It was wonderful journey for me in IBM capstone and other courses. I learned a lot about Data science, it attracts me to go further and try to expand my knowledge and combine my experience in telecommunication with it in a professional way. Thanks to Coursera for keeping Skillful instructors with their awesome materials.