Data Ninja

Secret places to open a Japanese restaurant

Fernanda Oliveira

December 28, 2019

Introduction

Background

The city of Berlin is well known to be a cosmopolitan city where you can find people from all around the world. Berlin offers a very wide commercial variety, especially in the area of gastronomy. The trend that comes to stay, are Asian restaurants, particularly Japanese restaurants. Although there are a lot of them spread in the city, there are new ones opening all the time. Therefore to analyze locations, types, and the number of these restaurants is a plus for those who want to open a new restaurant in the city.

Problem

Searching an optimal location to open a Japanese restaurant in the city of Berlin can be challenging. One could think that the better location for it should be at a place where there is no Japanese restaurant. But the problem is that perhaps most of the interested customers instead of going to an isolated neighborhood, prefer to go to a popular neighborhood, where there are more options and also there is movement of people. At the same time that the concurrence will be big in these regions, the flux of interested customers in this specific region will be relevant as well. Many people, for example, go on the weekends to a specific Japanese restaurant and when they arrive, there is a large line waiting for them. This usually happens because it is also a new trend in Berlin, in some popular restaurants, not to have an option to make a reservation. The good news is that perhaps some of the customers, those who do not want wait too long in line, might want to search for similar options in the neighborhood.

Interest

This project is ideal for a person or a branch that is interested in opening a Japanese restaurant.

Data acquisition

Data sources

The goal of this project is **to search for locations where the neighborhood is surrounded by Japanese restaurants.** Then the focus will be to find locations that have a distance of approximately 300 m from Japanese restaurants that already exists. After that, a research about the prices to rent a place for opening a restaurant will be made. In addition, the optimal location should be accessible by public transportation. Based on the goal of this project I describe below what is needed to perform the search and which data sources will be used:

1. number of existing Japanese restaurants in a neighborhood
2. segmentation of types of Japanese restaurants in a neighborhood
3. prices and locations of places in Berlin to open a restaurant

4. distance of the available places to rent to the Japanese restaurants that already exists and to the public transportation.

The data and tools that I will use are the following:

1. **Foursquare API** to select the number of restaurants and their location in some neighborhoods of Berlin
2. **Geocoder** to get the latitudes and longitudes of places to rent, together with information from this website
3. **k-means Clustering** to perform the segmentation of the categories of restaurants

Feature selection and data cleaning

The dataset that will be used in this project was obtained thought the Foursquare API, exploring several types of venues, such as, ID, name, category (Japanese restaurants), latitude, longitude and neighborhood. The Figs.(1) shows the five first lines of the dataset created.

	id	name	categories	lat	lng	neighborhood
0	55f9a48e498ee737a1893058	Heno Heno	Japanese Restaurant	52.503964	13.315578	Wielandstr. 37
1	4bbe353b9474c9b63e41d9b6	Kushinoya	Japanese Restaurant	52.505372	13.319982	Bleibtreustr. 6
2	570b97c4498e2c6e7c5eb991	Smart Deli	Japanese Restaurant	52.528094	13.389060	Novalisstr. 2
3	57c9e26a498ed1dcbbd0b461	Sticks'n'Sushi	Japanese Restaurant	52.502020	13.365064	Potsdamer Str. 85
4	4c0fde34ce57c928f7f580d2	Green Tea Café MAMECHA	Japanese Restaurant	52.527284	13.406305	Mulackstr. 33 (Rückerstr.)

Figure 1: Dataset created using Foursquare API exploring categories of Japanese restaurants.

The dataset 2 has information about avaiable places to rent in Berlin. First, it was select the postal codes and prices of these places and then with the help of Geocoder was possible to get the latitude, longitude features.

	Postcode	Price	Latitude	Longitude
0	12683	2900.00	52.503731	13.559540
1	10247	2400.00	52.516340	13.463990
2	10777	1142.36	52.497685	13.342285
3	10713	3269.00	52.485240	13.311870
4	10719	5900.00	52.498245	13.327140

Figure 2: Dataset created using Geocoder.

Using again Foursquare API, I searched for categories of public transportation in Berlin (S-Bahn and U-Bahn) and then, I selected the following features: ID, name, category, latitude and longitude locations. The five first lines of the third dataset is shown in the Fig.(3) and (4).

	id	name	categories	lat	lng	neighborhood
0	4a1c8506f964a520457b1fe3	Berlin Hauptbahnhof	Light Rail Station	52.525220	13.369369	Europaplatz 1 (Washingtonplatz)
1	4af5f0c7f964a52020ff21e3	Bahnhof Berlin Friedrichstraße	Light Rail Station	52.520284	13.387063	Georgenstr. 14/17
2	4b05bf38f964a5204ce222e3	Bahnhof Berlin Potsdamer Platz	Light Rail Station	52.509723	13.376597	Potsdamer Platz (Potsdamer Str.)
3	4adcda91f964a520ba4b21e3	Bahnhof Berlin Zoologischer Garten	Light Rail Station	52.506642	13.332513	Hardenbergplatz 13
4	4b01859ef964a520174322e3	S Savignyplatz	Light Rail Station	52.505093	13.319847	Bleibtreustr. 49

Figure 3: Dataset created using Foursquare API exploring categories of public transportation (S-Bahn).

	id	name	categories	lat	lng	neighborhood
0	4bfb2cf765fbc9b66f23916c	U Rehberge	Metro Station	52.555570	13.343412	Müllerstr. (Dubliner Str.)
1	4b538a1af964a52043a127e3	U Wilmersdorfer Straße	Metro Station	52.506312	13.306770	Wilmersdorfer Str. (Kantstr.)
2	4b5de986f964a520387329e3	U Adenauerplatz	Metro Station	52.499950	13.307203	Adenauerplatz (Kurfürstendamm)
3	4b47845cf964a5209e3426e3	U Güntzelstraße	Metro Station	52.490989	13.330868	Bundesallee (Güntzelstr.)
4	4b2a3edbf964a52076a624e3	U Deutsche Oper	Metro Station	52.511193	13.311905	Bismarckstr. (Krumme Str./Weimarer Str.)

Figure 4: Dataset created using Foursquare API exploring categories of public transportation (U-Bahn).

Exploratory Data Analysis

Here we will understand more our data collection and we will apply some descriptive statistics and visualization to answer the following questions:

- How many restaurants exist in each dataset?
- How many avaiable places to rent there are?
- How many categories exist in each dataset?

Using the **describe** method in Python, we can already have some results. The dataset (1) contains 100 restaurants, which 63 are Japanese restaurants.

To see all the categories collected using the Fousquare API, I will plot the category feature. See the result in Fig. (5). Observe that in the Fig. (5) we can already see all categories that we collected using Fousquare API and the number of restaurants of each category.

Predictive Modeling

I will apply the machine learning algorithms called **K-means Clustering** to perform a segmentation in the Japanese restaurants dataset. K-means Clustering is a simple and popular unsupervised algorithms that can be used to make segmentations. Segmentation is a practice of divide a feature into groups with similar characteristics. Therefore one can get some insights about the characteristics of the data.

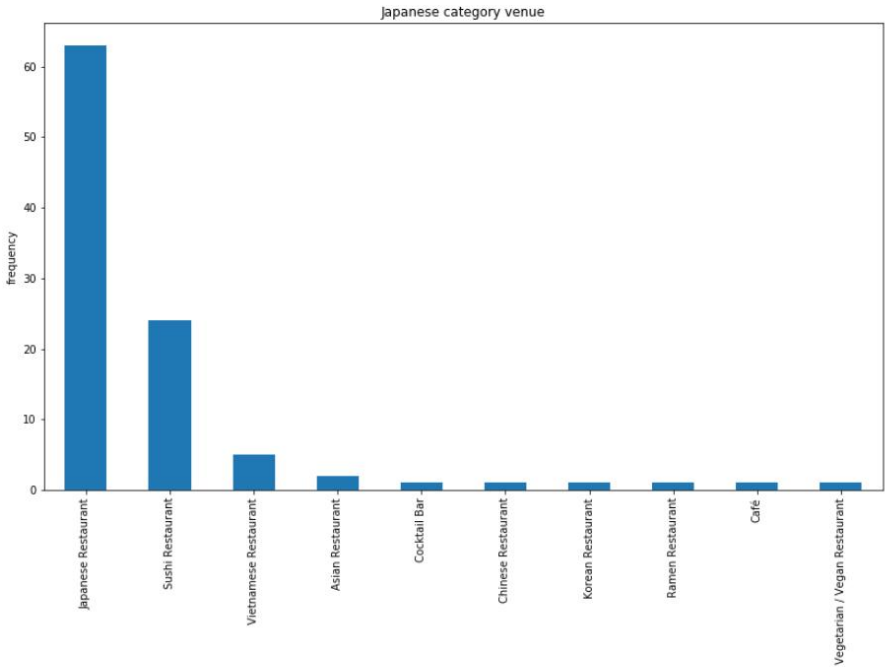

Figure 5: Category of restaurants using the Japanese category venue from Foursquare API.

First, I will start applying the **One-hot Encoding** function to convert categorical variations to numerical ones. This facilitated for Machine Learning algorithms to perform a better prediction. The results 0 indicates non existent while 1 indicates existent, see Table (6).

	neighborhood	Asian Restaurant	Café	Chinese Restaurant	Cocktail Bar	Japanese Restaurant	Korean Restaurant	Ramen Restaurant	Sushi Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant
0	Wielandstr. 37	0	0	0	0	1	0	0	0	0	0
1	Bleibtreustr. 6	0	0	0	0	1	0	0	0	0	0
2	Novalisstr. 2	0	0	0	0	1	0	0	0	0	0
3	Potsdamer Str. 85	0	0	0	0	1	0	0	0	0	0
4	Mulackstr. 33 (Rückerstr.)	0	0	0	0	1	0	0	0	0	0

Figure 6: Dataset after applying one-hot Encoding

Now I will create a dataframe in **Pandas**. For this I will use a function to sort the venues in descending order and then I will create a new dataframe and display the top 7 venues for each neighborhood. The result is in Fig. (7).

I will run **k-means** to cluster the neighborhood into 5 clusters using the K-means Clustering function. The dataset for the **cluster 0** is showed in the Fig. (8).

	neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Ahornstr. 32	Sushi Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Ramen Restaurant	Korean Restaurant	Japanese Restaurant	Cocktail Bar
1	Albrechtstr. 131	Sushi Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Ramen Restaurant	Korean Restaurant	Japanese Restaurant	Cocktail Bar
2	Alte Schönhauser Str. 13 (Mulackstr.)	Japanese Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Sushi Restaurant	Ramen Restaurant	Korean Restaurant	Cocktail Bar
3	Alte Schönhauser Str. 7-8	Japanese Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Sushi Restaurant	Ramen Restaurant	Korean Restaurant	Cocktail Bar
4	Bergmannstraße 93	Japanese Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Sushi Restaurant	Ramen Restaurant	Korean Restaurant	Cocktail Bar

Figure 7: Dataset with the neighborhood in the index

	id	neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
28	4b3ba703f964a5200a7825e3	Kottbusser Damm 102	0	Sushi Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Ramen Restaurant	Korean Restaurant	Japanese Restaurant	Cocktail Bar
48	4d358f7f2c76a1438bd18fc7	Goethestr. 37-38	0	Sushi Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Ramen Restaurant	Korean Restaurant	Japanese Restaurant	Cocktail Bar
53	4adcda88f964a520724921e3	Albrechtstr. 131	0	Sushi Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Ramen Restaurant	Korean Restaurant	Japanese Restaurant	Cocktail Bar
54	5571f00e498e055a7d77700d	Dahlmannstr. 14	0	Sushi Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Ramen Restaurant	Korean Restaurant	Japanese Restaurant	Cocktail Bar
55	4c84d8ab51ada1cd472a3210	Wilmersdorfer Str. 22 (Thrasoltstr.)	0	Sushi Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Ramen Restaurant	Korean Restaurant	Japanese Restaurant	Cocktail Bar

Figure 8: Cluster 0

Score calculations

Here I will calculate a score of the available places to rent to the Japanese restaurants that already exists and to the public transportation. For this I used **pyproj** library. First, I created 4 lists with the latitudes and longitudes using the following datasets Figs. 1, 2, 3 and 4. Then I transformed the latitudes and longitudes to Euclidian coordenates (X, Y). Then I calculated the distance from the avaiable places to rent to the Japanese restaurants and the public transportations and I took the minimum value for each indices. Finally, I created a list called **optimal list** and I transformed it to dataframe and I added it to the dataframe of avaiable places to rent a restaurant (2). The result is showed in the Fig. (9).

	index	Postcode	Price	Latitude	Longitude	Score
Address						
Zinnowitzer Straße 2, Mitte	16	10115	0.0	52.531570	13.383444	748.154426
Zehdenicker Straße 21, Mitte	12	10119	3000.0	52.530505	13.405483	1310.302883
Barstraße, Wilmersdorf	3	10713	3269.0	52.485240	13.311870	1353.984241
Ebertstraße, Mitte	20	16727	0.0	52.516040	13.376910	1382.880829
Fasanenplatz, Wilmersdorf	4	10719	5900.0	52.498245	13.327140	1390.663694

Figure 9: Dataset created post calculations of the score.

Results and Discussions

In this section I will show some of the results obtained. We segmented the category features into five Clusters and I can see the **1st Most Common Venue** in each of these clusters:

1. **Cluster 0**: 24 Sushi restaurants
2. **Cluster 1**: 2 Ramen, 1 Chinese, 1 vegetarian/vegan restaurants, 1 Cafe and 1 cooktail bar
3. **Cluster 2**: 62 Japanese restaurants
4. **Cluster 3**: 5 Vietnamese restaurants
5. **Cluster 4**: 2 Asian restaurants

The Fig. (10) shows the result of applying the **K-mean Clustering**.

I calculated the score of the available places to rent to the Japanese restaurants that already exists and to the public transportation. The lower score more optimal is the place. The results are showed in the Fig. (11).

I will now explore the results using the **Folium map**. In the Fig. (12) is showing the locations of each cluster and each type of the restaurant. In the

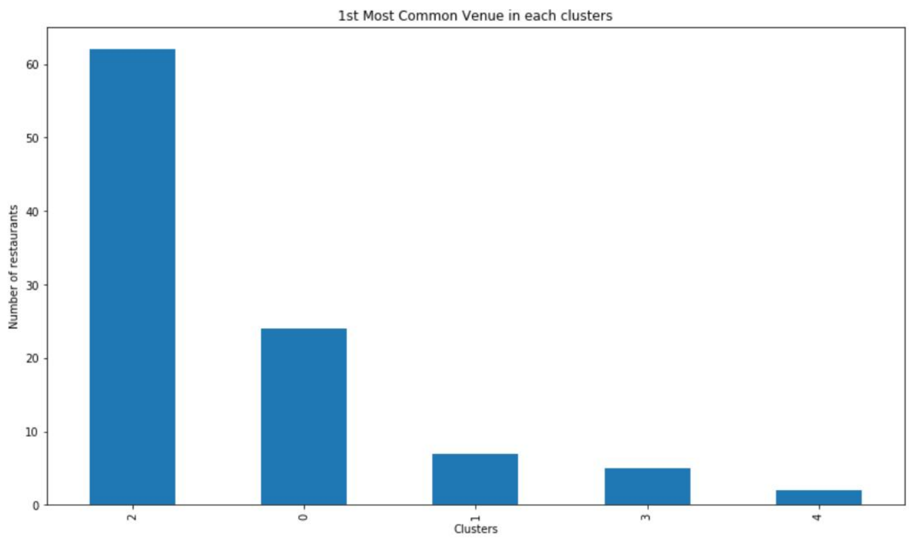

Figure 10: Number of restaurants in each cluster.

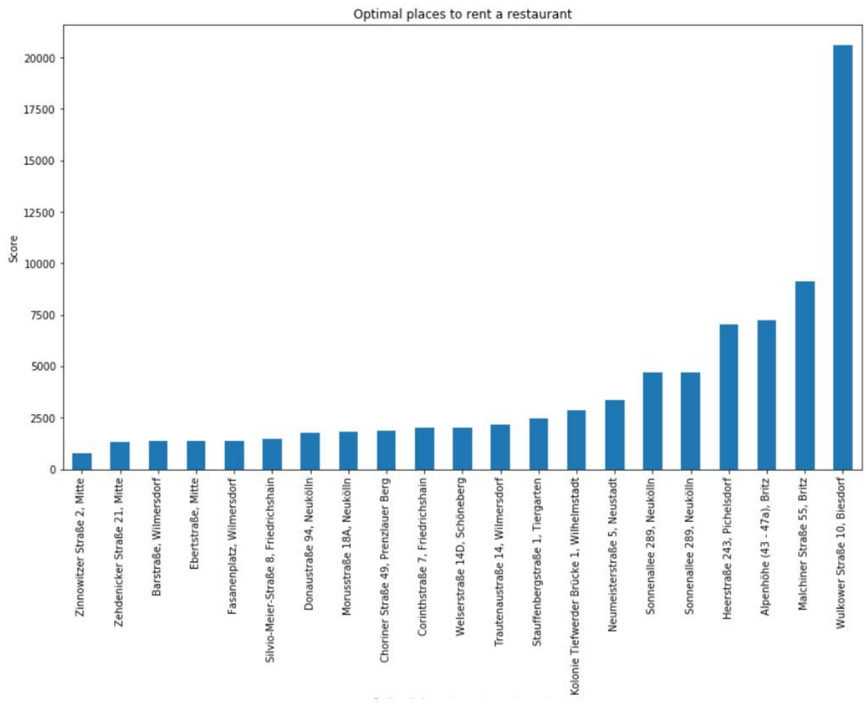

Figure 11: Optimal places to open a restaurant in Berlin.

same figure, is also the optimal places (yellow points) with the score label in each point and includding the public transportation, the city train - S-bahn (green points) and the metro - U-bahn (blue points).

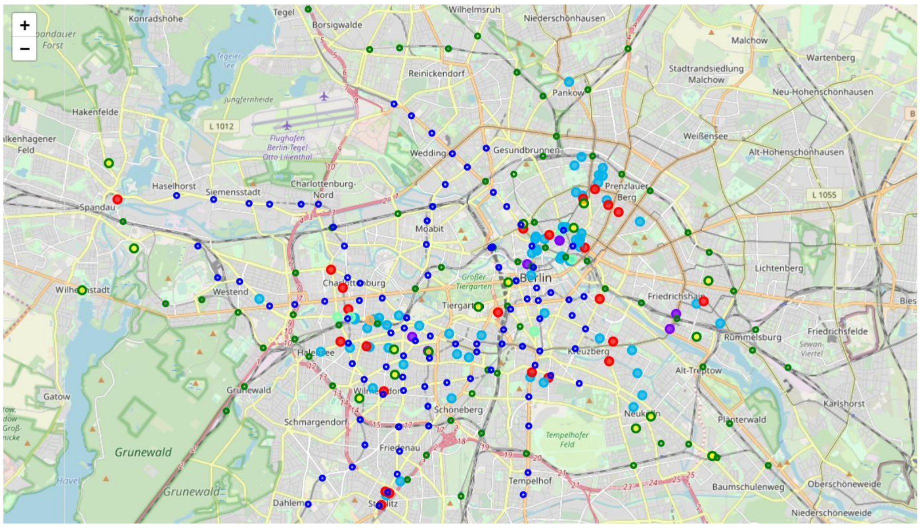

Figure 12: In the map is the five clusters: **Cluster 0 (red)**: 24 Sushi restaurants, **Cluster 1 (purple)**: 2 Ramen, 1 Chinese, 1 vegetarian/vegan restaurants, 1 Cafe and 1 cooktail bar, **Cluster 2 (light blue)**: 62 Japanese restaurants, **Cluster 3 (light green)**: 5 Vietnamese restaurants and **Cluster 4 (orange)**: 2 Asian restaurants. The public transportation are the U-bahn (dark blue) and S-Bahn (green). The avaiable places are showed in yellow points.

Could be interesting in the future, to increase the dataset with avaiable places. Here it was used only one agency website to collect the data of avaiable places in Berlin.

Conclusions

In this data science project, I showed how to explore venues using Fousquare API and how to get latitudes and longitudes using Geocoder. I chose the Japanese restaurant category to explore Foursquare venues in the city of Berlin.

I applied the Machine Learning algorithm K-means Clustering and I made segmentations of the types of Japanese restaurants. Therefore, It was possible to observe in the 'Folium map' the locations of the restaurants in each of the cluster created.

I collected prices of avaiable places for opening a restaurant in Berlin and created a dataset.

I calculated the score for locations that have a distance of approximately 300 m from Japanese restaurants that already exists and from public transportations, such as, the city train and the metro of Berlin. At the end, I obtained the results of the *secret* places to open a restaurant in Berlin.