IBM CAPSTONE PROJECT REPORT

Where would you open a Turkish Restaurant?

October 2020

Introduction

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. The city is also one of Germany's 16 federal states. It is surrounded by the state of Brandenburg, and contiguous with Potsdam, Brandenburg's capital. The two cities are at the center of the Berlin-Brandenburg capital region, which is, with about six million inhabitants and an area of more than 30,000 km2, Germany's third-largest metropolitan region after the Rhine-Ruhr and Rhine-Main regions.

Problem

Searching an optimal location to open a Turkish restaurant in the city of Berlin can be difficult. One could think that the better location for it should be at a place where there is no restaurants. 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.

Interest

This project is to find an optimal place to open Turkish Restaurant.

Data acquisition

Data sources

The goal of this project is to search for locations where the neighborhood is surrounded by Turkish restaurants. Then the focus will be to find optimal locations that have a distance of approximately 400 m from Turkish restaurants that already exist. 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. Data sources will be used:

1. number of existing Turkish restaurants in a neighborhood

2. segmentation of types of Turkish 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 Turkish restaurants that already exist 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
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 (Turkish restaurants), latitude, longitude, and neighborhood.

	id	name	categories	lat	lng	neighborhood
0	4af95d40f964a520771122e3	Restaurant Tuğra	Turkish Restaurant	52.498673	13.298555	Kurfürstendamm 96 (Markgraf-Albrecht-Str.)
1	504db200e4b05828d1ff6d4d	Bey Simit Haus	Turkish Restaurant	52.511215	13.298407	Kaiserdamm 66
2	4cf95af334c1a093ae95390e	Mercan	Turkish Restaurant	52.498589	13.427954	Wiener Str. 10
3	4b464c82f964a520e81c26e3	Doyum Grillhaus	Turkish Restaurant	52.498259	13.417375	Admiralstr. 38
4	4c45d1fbf0bdd13a317acbcc	Adana Grillhaus	Turkish Restaurant	52.499641	13.426887	Manteuffelstr. 86

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

The rent dataset has information about available 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: Avaliable for renting

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.

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

Using the describe method in Python, we can already have some results. To see all the categories collected using the Fousquare API, I will plot the category feature. See the result in Fig. (5).

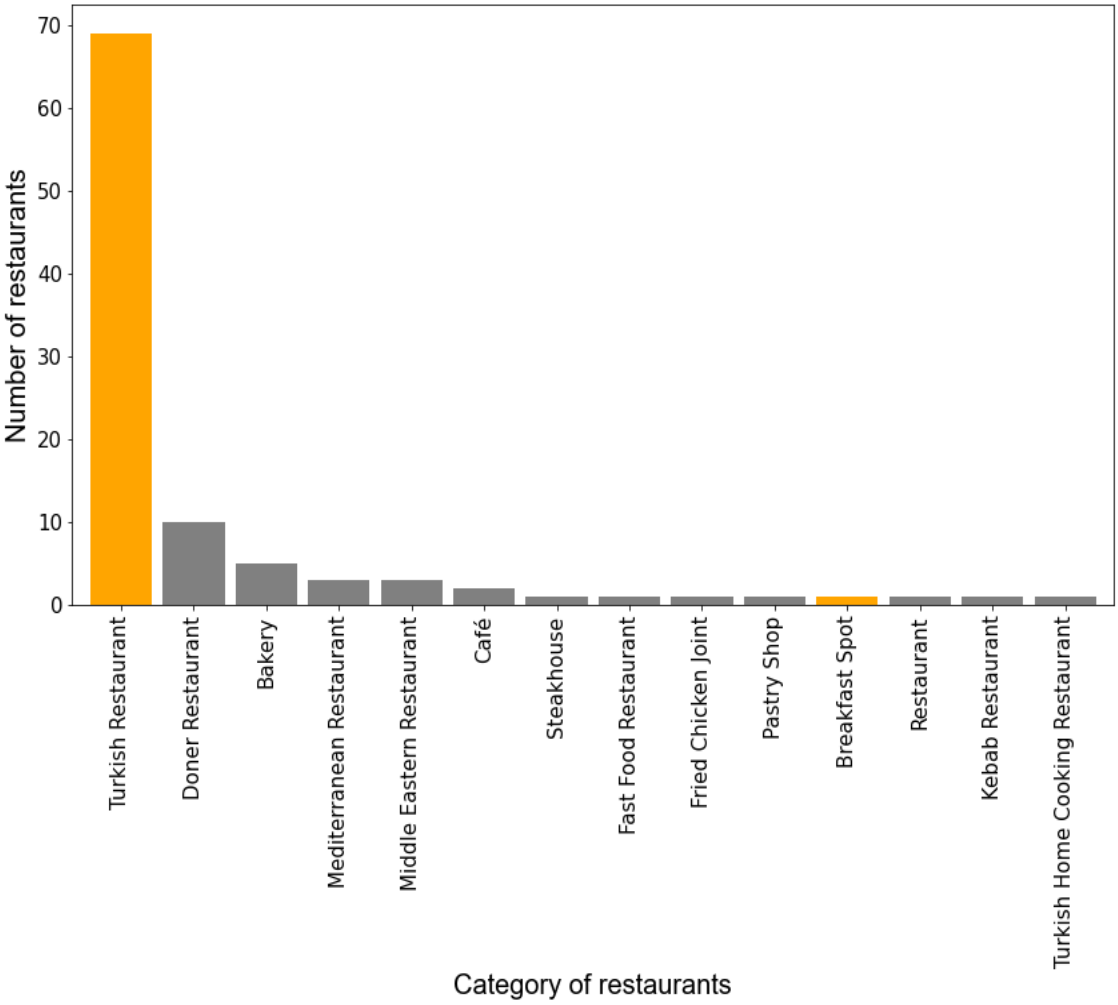

Figure 5

It can be already seen 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 Turkish restaurants dataset. K-means Cluster- ing 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.

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.

	neighborhood	Bakery	Breakfast Spot	Café	Doner Restaurant	Fast Food Restaurant	Fried Chicken Joint	Kebab Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Pastry Shop	Restaurant	Steakhouse	Turk Ho Cooki Restaur
0	Kurfürstendamm 96 (Markgraf-Albrecht-Str.)	0	0	0	0	0	0	0	0	0	0	0	0	
1	Kaiserdamm 66	0	0	0	0	0	0	0	0	0	0	0	0	
2	Wiener Str. 10	0	0	0	0	0	0	0	0	0	0	0	0	
3	Admiralstr. 38	0	0	0	0	0	0	0	0	0	0	0	0	
4	Manteuffelstr. 86	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 6: Dataset after applying one-hot Encoding

Now I will create a dataset in Pandas. For this I will use a function to sort the venues in descending order and then I will create a new dataset and display the top 7 venues for each neighborhood.

	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	Adalbertstr. 10	Turkish Restaurant	Turkish Home Cooking Restaurant	Steakhouse	Restaurant	Pastry Shop	Middle Eastern Restaurant	Mediterranean Restaurant
1	Adalbertstr. 12	Turkish Restaurant	Turkish Home Cooking Restaurant	Steakhouse	Restaurant	Pastry Shop	Middle Eastern Restaurant	Mediterranean Restaurant
2	Adalbertstr. 97	Bakery	Turkish Restaurant	Turkish Home Cooking Restaurant	Steakhouse	Restaurant	Pastry Shop	Middle Eastern Restaurant
3	Adalbertstraße 98	Turkish Restaurant	Turkish Home Cooking Restaurant	Steakhouse	Restaurant	Pastry Shop	Middle Eastern Restaurant	Mediterranean Restaurant
4	Admiralstr. 38	Turkish Restaurant	Turkish Home Cooking Restaurant	Steakhouse	Restaurant	Pastry Shop	Middle Eastern Restaurant	Mediterranean Restaurant

Figure 7: Dataset with the neighborhood in the index

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

	Cluster Labels	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	1	Adalbertstr. 10	Turkish Restaurant	Turkish Home Cooking Restaurant	Steakhouse	Restaurant	Pastry Shop	Middle Eastern Restaurant	Mediterranean Restaurant
1	1	Adalbertstr. 12	Turkish Restaurant	Turkish Home Cooking Restaurant	Steakhouse	Restaurant	Pastry Shop	Middle Eastern Restaurant	Mediterranean Restaurant

Fig. (8). Cluster 0 head(2)

Score calculations

Here I will calculate a score of the available places to rent to the Turkish restaurants that already exist and to public transportation Then I calculated the distance from the available places to rent to the Turkish restaurants and the public transportations and I took the minimum value for each index.

	Postcode	Price	Latitude	Longitude	Score
0	12683	2900.00	52.503731	13.559540	16757.044480
1	10247	2400.00	52.516340	13.463990	1487.424275
2	10777	1142.36	52.497685	13.342285	1992.547554
3	10713	3269.00	52.485240	13.311870	2602.664191
4	10719	5900.00	52.498245	13.327140	1525.615759

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 we can see the 1st Most Common Venue in each of these clusters:

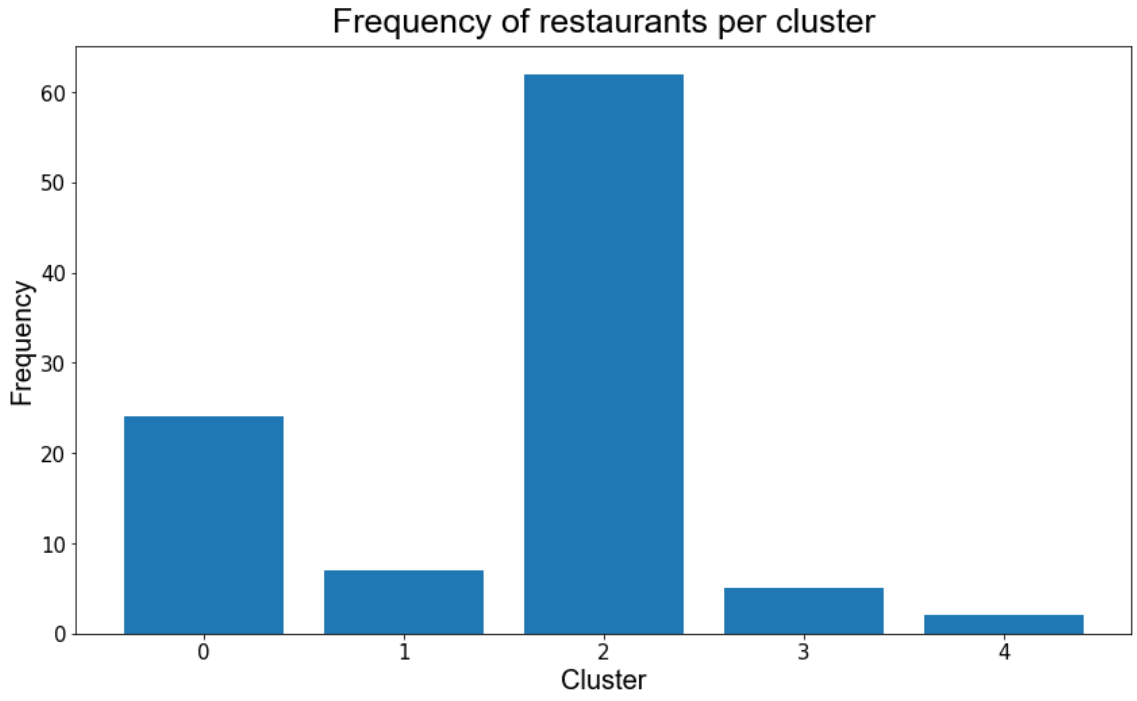

Figure 10: Number of restaurants in each cluster.

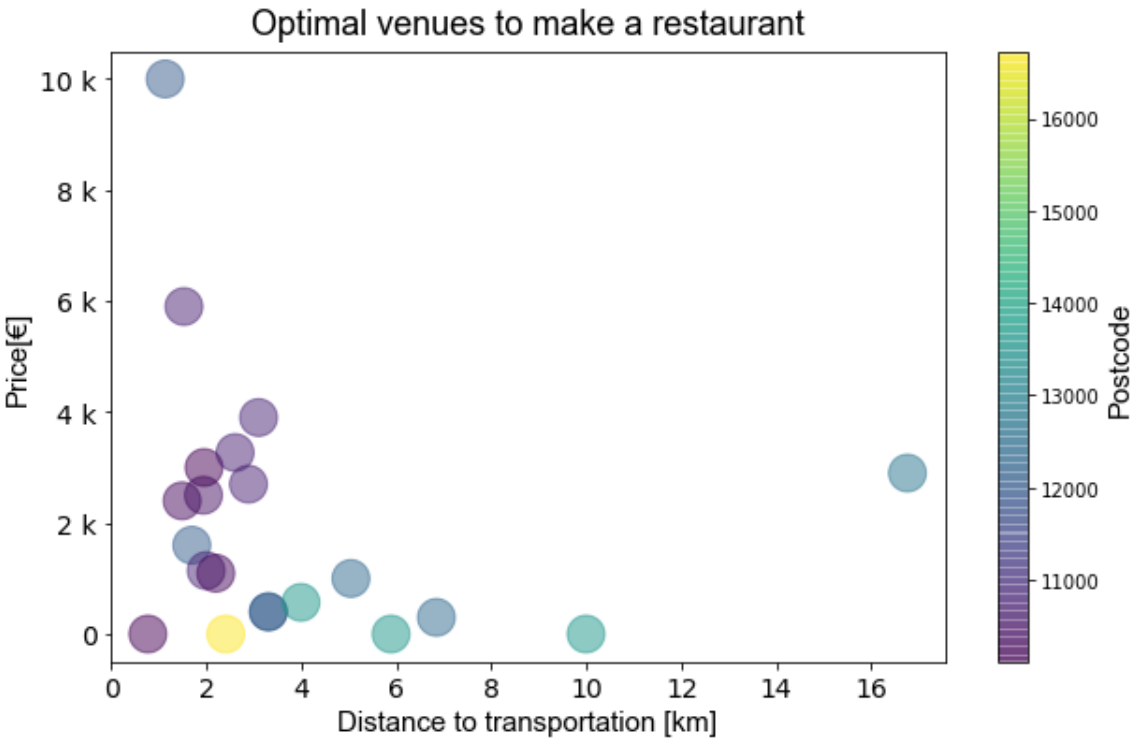

Figure 11: Optimal places to open a restaurant in Berlin.

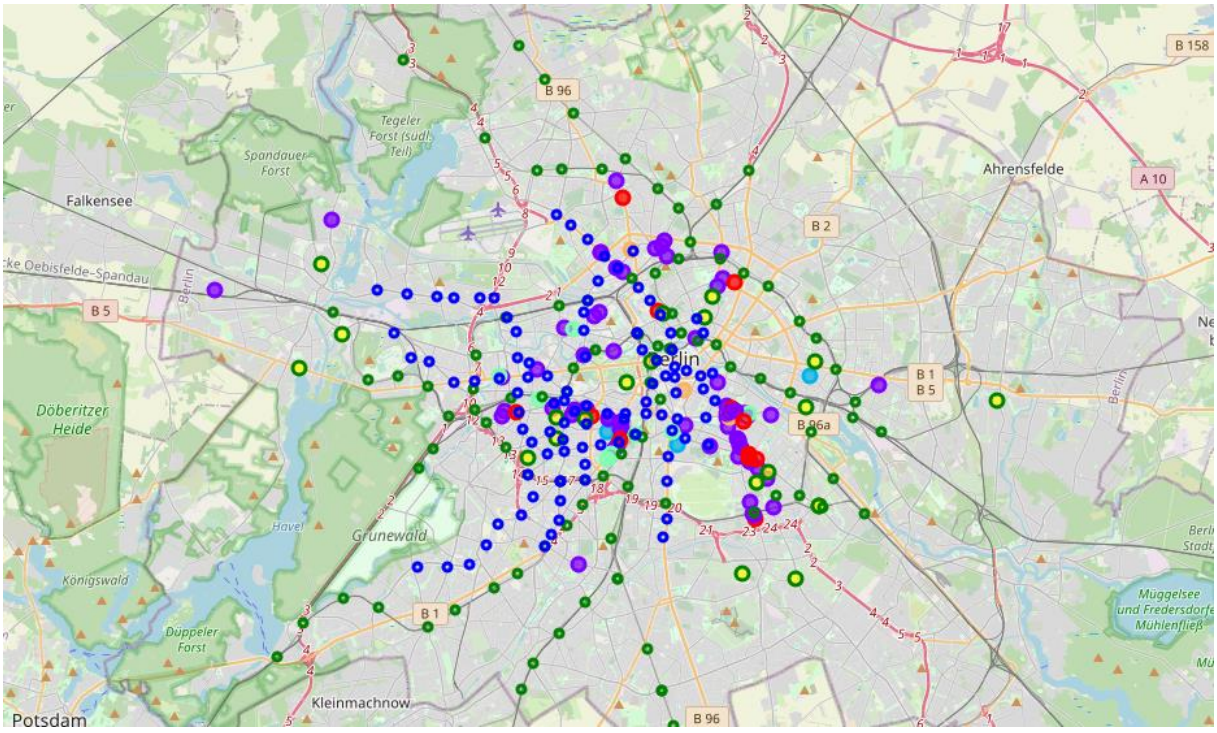

Figure 12: In the map is showing all the optimal locations where a Turkish restaurant can be opened.

The available places are showed in yellow points.

Conclusions

In this data science project, showed how to explore venues using Foursquare API and how to get latitudes and longitudes using Geocoder. I chose the Turkish 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 Turkish restaurants. Therefore, It was possible to observe in the `Folium map` the locations of the restaurants in each of the clusters created. I collected prices of available places for opening a restaurant in Berlin and created a dataset. I calculated the score for locations that have a distance of approximately 400 m from Turkish restaurants that already exist and from public transportations, such as, the city train and the metro of Berlin. In the end, I obtained the results of the secret places to open a restaurant in Berlin.