Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods

Introduction

For the Battle of Neighborhoods assignment, I've chosen the provided prompt of

In a city of your choice, if someone is looking to open a restaurant, where would you recommend that they open it?

A person looking to open a restaurant in Berlin will have to make many decisions while scouting a location. One of the factors to consider when choosing a location of a restaurant is the profile of the neighborhood. A hip fusion cuisine migh not fit well into a neighborhood populated mostly by retirees. A pizza place might struggle in a neighborhood with a high concentration of other pizza places. A coffee shop could do well in a popular tourist spot.

Data

Let's say the restaurant owner is interested in answering these questions:

- What is the age profile of the residents?
- What competition exists in the vicinity?
- Are there other venues people might go to in the vicinity and then opt to grab a bite nearby?

To try to answer these questions, I'll be working with demographic data of Berlin's residents along with the existing venue data from Forsquare to analyze the neighborhoods and create different neighborhood profiles to assess the locations. I aim to categorize the neighborhoods based on their profile into clusters with similar profiles, to help narrow down the selection of neighborhoods for an owner of a restaurant to choose from.

Demographic data

Data I'll be working with is from the www.statistik-berlin-brandenburg.de web page. I'm going to be working with the data from this data set, from sheet "T14": https://www.statistik-berlin-brandenburg.de/publikationen/stat_berichte/2021/SB_A01-05-00_2020h02_BE.xlsx

This data set contains data for **Berlin** from **December 31st 2021** and includes the following fields:

- **Postleitzahl** - postal code
- **Bezirk** - borough
- **Insgesamt** - total number of residents

- **unter 6** - residents under the age of 6
- **6-15** - residents between the ages 6 and 15
- **15-18** - residents between the ages 15 and 18
- **18-27** - residents between the ages 18 and 27
- **27-45** - residents between the ages 27 and 45
- **45-55** - residents between the ages 45 and 55
- **55-65** - residents between the ages 55 and 65
- **65** und mehr = residents over the age of 65
- **Darunter Weiblich** - number of female residents

The age categries do not include the upper limit. Meaning for example the age group of 45-55 includes people who are 45 years old, or 54 years old, but not people who are 55 years old. The 55-year-olds are in the next category of 55-65.

Since this data set doesn't have all the column labels in one row, and they are in German, I've opted not to use the existing column names, and instead rename the default names.

The new column names are:

- Postal Code
- Borough
- Residents
- Under 6
- 6-15
- 15-18
- 18-27
- 27-45
- 45-55
- 55-65
- Over 65
- Female

I'll need to convert the demographic data into percentages of population, and to make the data easier to read, I'll group the age ranges into few bigger groups:

- Kids - ages 0 - 18
- Young Adults - ages 18 - 27
- Adults - ages 27 - 45
- Older Adults - ages 45 - 65
- Retirees - ages over 65

	Female	Kids	Young Adults	Adults	Older Adults	Retirees
Postal Code						
10115	0.492448	0.168237	0.092082	0.421340	0.252281	0.066061
10117	0.482862	0.127617	0.095116	0.356776	0.260742	0.159750
10119	0.491692	0.165244	0.083632	0.405081	0.277511	0.068532
10178	0.484443	0.127110	0.100670	0.354263	0.225617	0.192340
10179	0.495657	0.126057	0.113029	0.346819	0.220343	0.193752

Coordinates

I'll be using the pgeocode library to query the latitude and longitude of the neighborhoods. I can use this library to extract the latitudes and longitudes based on the postal code of each neighborhood. This'll help with extracting nearby vanue data, as well as drawing the results on a map.

Foursquare

Next there will be the data from Foursquare. I'll query foursquare by postal code to extract the venues in the vicinity of the neighborhoods. From experience I know this data includes the category of venues, that'll help us profile them.

Looking at the high -level categories of venues, I've chosen 6 categories of interest:

- Arts and Entertainment
- Food
- Nightlife Spot
- Outdoors and Recreation
- Shop and Service
- Travel and Transport

Totaling these will give me a venue profile of the neighborhood.

	Culture	Food	Nightlife	Recreation	Shopping	Travel
Neighborhood						
10115	5	13	6	4	6	6
10117	8	23	6	3	18	6
10119	5	48	14	6	21	5
10178	5	36	2	8	38	10
10179	4	9	9	0	1	2

I have considered adding the rent prices for these neighborhoods; however, I wasn't able to find something not behind a paywall. Therefore, for the purpose of this exercise, we will assume money is no object, or that the restaurant owner will be choosing the neighborhoods by profile first.

Methodology

I'm aiming to cluster neighborhoods based on the demographic data and the Venue type data. For that I'll need to transform the the data into usable data sets.

With this data, I'll be using th K-MEANS clustering method to create the different neighborhood clusters and I'll be using the average values for each of the above age-ranges and venue categories to describe the clusters.

Lastly, I'll draw the neighborhood points on a map and label them with the postal code of the neighborhood and the neighborhood profile based on their cluster assignment.

Results

Running the k-means algorithm several times with different k values quickly showed that 5 is the ideal number of clusters, as they remain distinct enough from one another an it eliminates 1-item clusters.

The created clusters were labeled 0-4. However the numbers aren’t really descriptive, so I needed to analyze the composition of each cluster.

	Female	Kids	Young Adults	Adults	Older Adults	Retirees	Culture	Food	Nightlife	Recreation	Shopping	Travel
Cluster Labels												
0	0.511118	0.165760	0.090057	0.247539	0.270032	0.226612	0.271845	1.699029	0.116505	1.203883	2.067961	1.000000
1	0.490344	0.149513	0.098121	0.388262	0.257453	0.106653	1.933333	46.466667	9.600000	3.000000	13.266667	1.933333
2	0.485346	0.127782	0.098617	0.336357	0.256839	0.180404	6.600000	29.600000	6.400000	4.600000	19.400000	7.600000
3	0.499760	0.139919	0.095536	0.317828	0.271832	0.174885	2.526316	23.842105	3.105263	2.947368	7.000000	3.368421
4	0.498277	0.152255	0.110557	0.324728	0.246626	0.165835	1.394737	10.157895	1.736842	2.552632	4.921053	1.921053

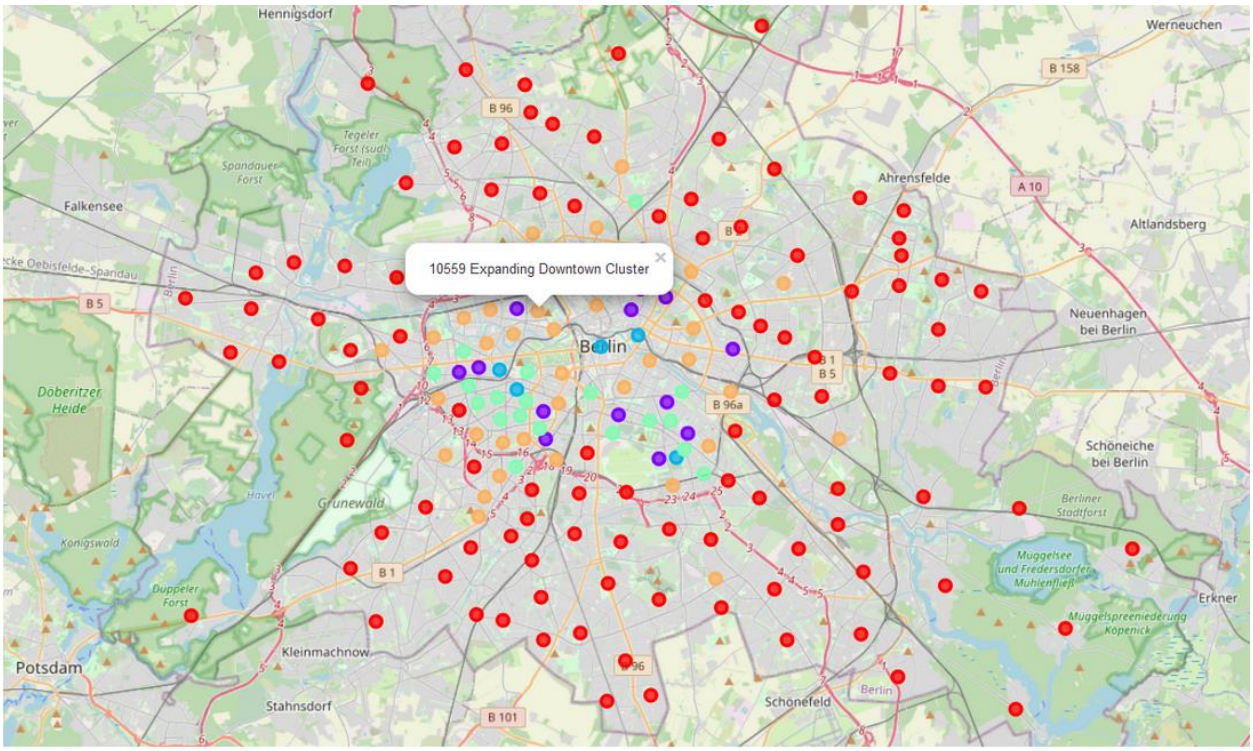
Using this table of average values I've arrived at these descriptive labels:

- Tourist Traps - high in culture, recreation and shopping, mid-level in competition, not many children living in the area.
- Wider Downtown - second highest concentration of places of culture, these places are just off the main tourist traps and could be a great place for a restaurant aimed at locals.
- Food Central - largest concentration of competition, also highest in nightlife venues
- Expanding Downtown - highest concentration of young adults, not too much competition, possible areas for downtown expansion
- Residential - largest cluster with low sores in all venue types, highest concentration of retired residents and kids

Adding the data back to the original data set results in the reference table below.

Female	Kids	Young Adults	Adults	Older Adults	Retirees	Culture	Food	Nightlife	Recreation	Shopping	Travel	Cluster Labels	Category	postal_code
0.492448	0.168237	0.092082	0.421340	0.252281	0.066061	5	13	6	4	6	6	4	Expanding Downtown	10115
0.495657	0.126057	0.113029	0.346819	0.220343	0.193752	4	9	9	0	1	2	4	Expanding Downtown	10179
0.488476	0.130933	0.118787	0.395478	0.205338	0.149464	2	8	7	4	8	3	4	Expanding Downtown	10243
0.476699	0.145493	0.095713	0.511121	0.194660	0.053013	3	8	6	6	2	0	4	Expanding Downtown	10245
0.495305	0.149195	0.084943	0.426693	0.205265	0.133903	2	8	0	3	3	2	4	Expanding Downtown	10249

Using coordinates of each neighborhood we can create an interactive map that displays these neighborhoods, heir postal codes and their cluster type.



Discussion

We've created 5 categories of neighborhoods based on the averages of the data.

The Tourist Trap cluster is great spot for a restaurant aimed at tourists. Naturally, it comes with a high concentration of places of entertainment and culture, and a higher density of other restaurants. These places may be a good fit for a franchise-type restaurant, as it's familiar brand may draw in those who are only visiting and look for something familiar. On the other hand it could also be a good spor for a restaurant serving local food, that aims to sell to tourists and visitors.

The Wider Downtown cluster consists of places that seem to be just off the most tourist-y areas. Competition is still fairly high, but these could be a good spot for a higher-end, or more specialty restaurant, that people don't mind to make a little detour for. These places could also market themselves to locals, thanks to their proximty to downtown but also being not directly in the tourst zones.

Food Central are place with high concentration of restaurants and other food places. The competition is high, the nightlife is booming. The new restaurant may try to cater to the party crowd, but unless the owner looks for strong competition, these places might not be good fit.

Expanding Downtown has the highest concetration of younger people living in the area. These spots could be a good fit for more experimental/ fusion cuisine, artisanal restaurants and other novelty restaurants. Anything that caters to a younger crowd. On the other hand the young adults

are likely to start families soon, so a restaurant aimed at families with small kids could do well in time.

Residential areas are the largest cluster. Restaurants in these areas will likely only see quests living or working in the vicinity. Price of the meal may play the largest role in these areas. Also thanks to the high concentration of retired residents, more traditional restaurants will likely have better chance of survival.

The detailed neighborhood data is available in the df_berlin dataframe. If the person looking for a location alrady has a few listings in mind, they can compare their postal codes with the df_berlin data set and see if the area matches their vision.

Conclusion

In this project I analyzed the age composition and venue composition of the neighborhoods and clustered them using the k-means method into 5 clusters. These clusters share similar characteristics, and the resulting data set can be used to more quickly sift through for-sale listings to rule out areas by their profile using the postal code of the address on the listing.

Alternatively, one may use the map to pick locations of interest, based on the preferred neighborhood profile, then use the postal codes to filter the listings to the areas of interest.