Capstone Report - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Introduction and Business Problem - Looking for restaurant locations in Berlin

Introduction and Business Problem Berlin is Germany is a big city with twelve boroughs that are made up of a total of 96 officially recognized localities. There are some upcoming neighborhood which are packed with restaurants, night life and hip people. For people that are new to Berlin it can be hard to figure out what restaurants are worth going to and where they are. For people that used to live in Berlin or are visiting, how do they know what kind of food they get in the different neighborhoods of Berlin?

For this project a simple guide is created on where to eat based on Foursquare ratings, price, category and geographic location data for restaurants in Berlin. These restaurants will be clustered based on their similarities so that users can easily check what kind of food they are looking for and in what price range.

Data

For this assignment the Foursquare API will be utilize to pull the following location data on restaurants in the different neighborhoods in Berlin, Germany:

- Venue Name
- Venue ID
- Venue Location
- Venue Category
- Rating
- Price

To acquire the data mentioned above first the names of the 96 neighborhoods in Berlin are required. These data is used to get the coorinates (latitue & longitude) of each neighborhood with help of the geolocator library. Afterwards the list of all venues for each neighborhood is gathered using the Foursquare API. With the venue list an the containing "Venue ID" the location, category, rating, price is gathered. Before using k-means clustering the date is checked, filtered and prepared.

Methodology

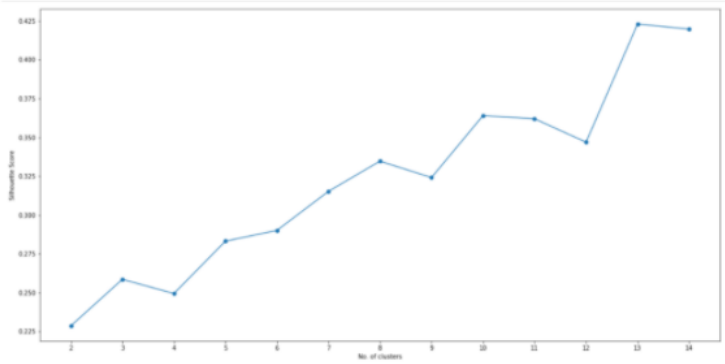
First the a list with all veneus is gathered. Following the creation of a dataset of neighbourhoods and their locations, we used the Foursquare API to retrieve data on venues using the location of each neighborhood. Then we filter the 'Venue Category' for restaurants and summarize some categories.

	Neighborhoods	Neighborhoods Latitude	Neighborhoods Longitude	Venue ID	Venue	Venue Latitude	Venue Longitude	Venue Category
15	Mitte	52.517885	13.404060	5a958ec0e4c459472938359f	Wilde Matilde	52.517475	13.405384	German Restaurant
17	Mitte	52.517885	13.404060	584c882dd702824c51e2be9a	Balthazar	52.515913	13.406160	Restaurant
27	Mitte	52.517885	13.404060	59eb78736bdee6069ed97d77	EL COLMADO	52.519412	13.409681	Spanish Restaurant
33	Tiergarten	52.509778	13.357260	4cc6c1d5c844721ea24ef601	Kantine im Felleshus	52.508555	13.350813	Scandinavian Restaurant
38	Tiergarten	52.509778	13.357260	4cdae7f2930af04dfbb08797	Eventlocation Alte Pumpe	52.505481	13.358203	German Restaurant

Using one hot encoding, for each neighbourhood the most common categories of restaurant are calculated.

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adlershof	Mediterranean Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant	Seafood Restaurant	Oriental Restaurant
1	Alt-Hohenschönhausen	Mediterranean Restaurant	Asian Restaurant	Vegetarian / Vegan Restaurant	Seafood Restaurant	Oriental Restaurant
2	Alt-Treptow	Seafood Restaurant	Mediterranean Restaurant	Italian Restaurant	Central and South American Food	Asian Restaurant
3	Blankenburg	Mediterranean Restaurant	Vegetarian / Vegan Restaurant	Seafood Restaurant	Oriental Restaurant	Italian Restaurant
4	Bohnsdorf	Italian Restaurant	Vegetarian / Vegan Restaurant	Seafood Restaurant	Oriental Restaurant	Mediterranean Restaurant

Using this data the best number of clusters is determined. Using the "elbow method" the number of clusters is set to k=4.



To visualize the data, we used k-means clustering to group together the neighbourhoods based on the venues categories. The clusters are:

- First cluster mainly Asian Food (red marks)
- Second cluster mainly Oriantal Resaurants or Doner Places (purple marks)
- Third cluster mainly Italian and Mediterranean Restaurants (turquoise marks)
- Fourth cluster mainly German cuisine or vegetarian/vegan food (green marks)

Results

The whole process generated four clusters, where you can find specific kinds of food per neighborhood. So it's easy to get an overview. The price and ratings of the restaurants had to be excluded because there was not enough data for every venue of the list. Only 42 got a price rating and 47 a rating of the food quality out of 265 restaurants found. So for clustering there were just the restaurant categories be used. But in the end the result is still useful

Discussion

So the approach was working so far. The results would be even better if we could use the price and ratings of the restaurant. Mainly the problem is, that Foursquare is not very common in Germany. Probably it would be way more efficient to gather the data from Google in this case.