IBM DATA SCIENCE CAPSTONE PROJECT

Location Selection for A New Restaurant in Berlin

F CAN ÖZER

DECEMBER 25, 2020

The Battle of Neighborhoods

1. Introduction/Business Problem

Berlin is one of Europe's most cosmopolitan cities. Its best restaurants are spanning the globe in their offering – from Turkish Döner Kebap to Japanese Sushi. The city's food culture is an insight into its multi-cultural nature, and nothing shows this better than its street food markets. The town offers foods from all corners of the globe; these markets are a food lover's dream and a must-experience in Berlin.

For this project, I decided to solve a business problem in Berlin. A businessman who has fast-food restaurants in the US wants to open his new fast-food branch in Berlin, but he is unsure where he should open his new restaurant branch in the town. So, I will help him make a great decision to make his fast-food restaurant one of the tops visited venues in its borough.

Firstly, we need to know that customers go for eating at which type of restaurants mostly in each Berlin district. Districts with fast-food restaurants in their top-visited venue lists should be considered to open a new restaurant because it clarifies customer behaviors in selecting restaurants and the city's food culture based on location. Also, we need to check the number of fast-food restaurants in that district to avoid a competitive food market and powerful rivals, making some trouble for a new restaurant.

2. Data

We need a couple of Datasets to solve this problem.

1. Boroughs and neighborhoods of Berlin

2. Geo-coordinates of boroughs in Berlin

3. Top venues in each borough from the Foursquare API

Berlin is both a city and one of Germany's federated states. As of 2012, the 12 boroughs are made up of a total of 96 officially recognized localities. The dataset will be scraped from the Wikipedia Page: https://en.wikipedia.org/wiki/Category:Localities_of_Berlin

Then, Geo-coordinates of Boroughs will be obtained with the Google Maps Geocoding API. After gathering the Boroughs/Neighborhoods and the Boroughs coordinates datasets, we can display boroughs on a map combining these two datasets.

Finally, we will use the Foursquare API to obtain the most visited restaurant categories and the number of restaurants belonging to different types in each borough of Berlin.

3.Methdology

After scraping the data and analyzing it, we will K- Means Clustering algorithm used as an unsupervised machine learning method to cluster boroughs based on the most visited venues. K value of the K means algorithm will be selected to plot the sum squared error graph and find its elbow point. According to clustering results, the most promising clusters for a new fast-food restaurant will be detected, and neighborhoods located in those clusters would be considered potential restaurant regions.

3.1. Data Exploration and Preprocessing

We conducted several preprocessing steps on the dataset, which is obtained by scraping from the Wikipedia webpage. Table 1. shows the raw dataset contains boroughs and neighborhoods in Berlin.

	Borough	Neighborhood
0	Charlottenburg-Wilmersdorf	Charlottenburg,Charlottenburg-Nord,Grunewald,H...
1	Friedrichshain-Kreuzberg	Friedrichshain,Kreuzberg
2	Lichtenberg	Alt-Hohenschönhausen,Falkenberg,Fennpfuhl,Frie...
3	Marzahn-Hellersdorf	Biesdorf,Hellersdorf,Kaulsdorf,Mahlsdorf,Marzahn
4	Mitte	Gesundbrunnen,Hansaviertel,Mitte,Moabit,Tierga...
5	Neukölln	Britz,Buckow,Gropiusstadt,Neukölln,Rudow
6	Pankow	Blankenburg,Blankenfelde,Buch,Französisch,Buch...
7	Reinickendorf	Borsigwalde,Frohnau,Heiligensee,Hermsdorf,Konr...
8	Spandau	Falkenhagener,Feld,Gatow,Hakenfelde,Haselhorst...
9	Steglitz-Zehlendorf	Dahlem,Lankwitz,Lichterfelde,Nikolassee,Stegli...
10	Tempelhof-Schöneberg	Friedenau,Lichtenrade,Mariendorf,Marienfelde,S...
11	Treptow-Köpenick	Adlershof,Alt-Treptow,Altglienicke,Baumschulen...

Table 1. Raw Dataset (Boroughs and Neighborhoods)

There 12 central districts in Berlin, and these districts cover 96 neighborhoods. This table was our starting point, and then a geocoding process was applied using 12 central districts. For the geocoding process, we used the geocoding tool of python geopandas packages. It allows you to choose one of the geocoding providers such as Google, Bing, Yahoo, and Openstreetmap. This service provides limited usage for free users, limited to 250 calls per day and four requests per second.

We did not have any trouble while making the geocoding request for 12 locations. After the geocoding process, we obtained Table 2. which shows the latitude and longitude of all boroughs in Berlin.

	Borough	Neighborhood	Latitude	Longitude
0	Charlottenburg-Wilmersdorf	Charlottenburg,Charlottenburg-Nord,Grunewald,H...	52.586016	13.283000
1	Friedrichshain-Kreuzberg	Friedrichshain,Kreuzberg	52.586016	13.450000
2	Lichtenberg	Alt-Hohenschönhausen,Falkenberg,Fennpfuhl,Frie...	52.514557	13.498307
3	Marzahn-Hellersdorf	Biesdorf,Hellersdorf,Kaulsdorf,Mahlsdorf,Marzahn	52.533001	13.583000
4	Mitte	Gesundbrunnen,Hansaviertel,Mitte,Moabit,Tierga...	52.516998	13.367000
5	Neukölln	Britz,Buckow,Gropiusstadt,Neukölln,Rudow	52.483002	13.450000
6	Pankow	Blankenburg,Blankenfelde,Buch,Französisch,Buch...	52.569595	13.403235
7	Reinickendorf	Borsigwalde,Frohnau,Heiligensee,Hermsdorf,Konr...	52.574093	13.345394
8	Spandau	Falkenhagener,Feld,Gatow,Hakenfelde,Haselhorst...	52.534073	13.181689
9	Steglitz-Zehlendorf	Dahlem,Lankwitz,Lichterfelde,Nikolassee,Stegli...	52.432999	13.252171
10	Tempelhof-Schöneberg	Friedenau,Lichtenrade,Mariendorf,Marienfelde,S...	52.466999	13.383000
11	Treptow-Köpenick	Adlershof,Alt-Treptow,Altglienicke,Baumschulen...	52.450001	13.567000

Table 2. Geocoded Dataset (Boroughs, Neighborhoods, and coordinates)

The coordinates of boroughs were used to display regions on a map with the Python package Folium's help, below (Figure 1).

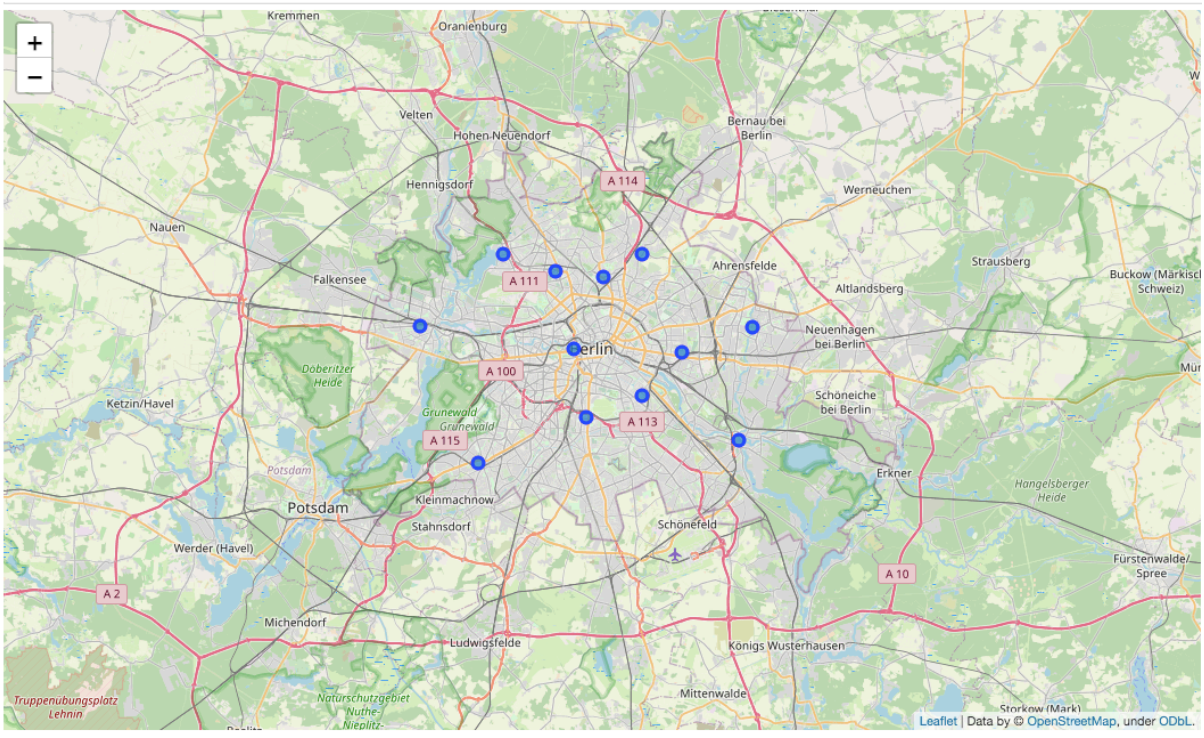

Figure 1. Berlin map with borough locations

In the next step, these locations were analyzed using venue data gathered from the Foursquare API in detail. We applied 1000 meters radius for each borough while gathering the venue information from the Foursquare API. Also, there was a limitation that allows us to get the top 100 venues for each borough. The obtained data was GeoJson format, and it was transformed into the pandas DataFrame format, as can be seen in Table 3.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Charlottenburg-Wilmersdorf	52.586016	13.283	Rüan Thai Restaurant	52.586872	13.283095	Thai Restaurant
1	Charlottenburg-Wilmersdorf	52.586016	13.283	SuperFit	52.585236	13.285668	Gym / Fitness Center
2	Charlottenburg-Wilmersdorf	52.586016	13.283	Hax'nhaus	52.589790	13.282210	German Restaurant
3	Charlottenburg-Wilmersdorf	52.586016	13.283	Cafe Wetterstein	52.589049	13.279866	Café
4	Charlottenburg-Wilmersdorf	52.586016	13.283	Creme de la Creme - Florida Eis	52.588646	13.278311	Ice Cream Shop

Table 3. Foursquare Dataset (Venues)

Table 3 shows the first five venues in Charlottenburg-Willmersdorf with their names, coordinates, and venue categories.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Charlottenburg-Wilmersdorf	55	55	55	55	55	55
Friedrichshain-Kreuzberg	6	6	6	6	6	6
Lichtenberg	38	38	38	38	38	38
Marzahn-Hellersdorf	18	18	18	18	18	18
Mitte	100	100	100	100	100	100
Neukölln	100	100	100	100	100	100
Pankow	69	69	69	69	69	69
Reinickendorf	17	17	17	17	17	17
Spandau	10	10	10	10	10	10
Steglitz-Zehlendorf	35	35	35	35	35	35
Tempelhof-Schöneberg	57	57	57	57	57	57
Treptow-Köpenick	60	60	60	60	60	60

Table 4. Venue distribution based on boroughs

Also, we checked how many venues were returned for each borough in Table 4. It shows there are top 100 venues, but most of the regions have less than 100 venues even the request was made within 1000 meters radius.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
43	Charlottenburg-Wilmersdorf	52.586016	13.283000	McDonald's	52.583974	13.286770	Fast Food Restaurant
359	Pankow	52.569595	13.403235	Ari's Dinner	52.564356	13.392959	Fast Food Restaurant
433	Steglitz-Zehlendorf	52.432999	13.252171	Sixteen Bites	52.431895	13.259409	Fast Food Restaurant
486	Tempelhof-Schöneberg	52.466999	13.383000	McDonald's	52.469660	13.371427	Fast Food Restaurant
544	Treptow-Köpenick	52.450001	13.567000	Mecklemburger Dorf	52.451654	13.575024	Fast Food Restaurant
554	Treptow-Köpenick	52.450001	13.567000	Rathaus Bistro	52.445068	13.574251	Fast Food Restaurant

Table 5. Top Fast – Food Restaurants

When we filter the venues data set by the venue category = Fast Food Restaurant, we can see the Fast-Food restaurants with their locations in the town (Table 5). Besides, the Fast – Food restaurant ratio in each region could be calculated, as shown in Table 6. According to the results, Treptow-Köpenick has the highest number of Fast-Food Restaurants among all boroughs in Berlin.

	Neighborhood	Fast Food Restaurant
0	Treptow-Köpenick	0.033333
1	Steglitz-Zehlendorf	0.028571
2	Charlottenburg-Wilmersdorf	0.018182
3	Tempelhof-Schöneberg	0.017544
4	Pankow	0.014493

Table 6. Neighborhoods and Fast-Food Restaurant Ratio

Then, we sorted the venue category dataset and created a dataset containing the ten most common venue categories for each region in Table 7. This table gives an idea about the region's settlement type.

	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Charlottenburg-Wilmersdorf	Café	Clothing Store	Drugstore	Supermarket	Restaurant	Italian Restaurant	Shopping Mall	Ice Cream Shop	Indian Restaurant	Sandwich Place
1	Friedrichshain-Kreuzberg	Bakery	Light Rail Station	Greek Restaurant	Bus Stop	Electronics Store	Café	Farm	Forest	Food & Drink Shop	Flower Shop
2	Lichtenberg	Bakery	Supermarket	Park	Coffee Shop	ATM	Soccer Stadium	Bookstore	Mexican Restaurant	Pharmacy	Breakfast Spot
3	Marzahn-Hellersdorf	Garden	Supermarket	Scenic Lookout	Drugstore	Tea Room	Mountain	Theme Park Ride / Attraction	Yoga Studio	Falafel Restaurant	Food & Drink Shop
4	Mitte	Hotel	Plaza	Hotel Bar	Museum	Spa	Concert Hall	Coffee Shop	Italian Restaurant	Monument / Landmark	Park
5	Neukölln	Café	Bar	Supermarket	Italian Restaurant	Coffee Shop	Park	Vegetarian / Vegan Restaurant	Bistro	German Restaurant	Cocktail Bar
6	Pankow	Café	Supermarket	Bakery	Drugstore	Park	Dance Studio	Organic Grocery	Sushi Restaurant	Burger Joint	Gym / Fitness Center
7	Reinickendorf	Soccer Field	Supermarket	Metro Station	Plaza	Light Rail Station	Pool	Bus Stop	German Restaurant	Dry Cleaner	Park

Table 7. The most common venues

For instance, in Reinickendorf, there is a soccer field which is mostly visited, and the regions have some workplaces with commonly used public transportation facilities.

3.2. K-Means Clustering

Our dataset is ready for clustering the regions based on the venues they have after data preprocessing and exploration steps. As is mentioned before, the K-Means Clustering algorithm was used in this step. However, the K- Means Clustering model needs a vital input to detect clusters, which indicates the number of groups. It is called the k value and can be identified by several methods.

In this project, we used the sum squared error (SSE) and elbow point technique. Firstly, we define a range for K-value from 2 to 10 and calculated SSEs for each k value. Then, errors were plotted to catch the elbow point of the graph (Figure 2).

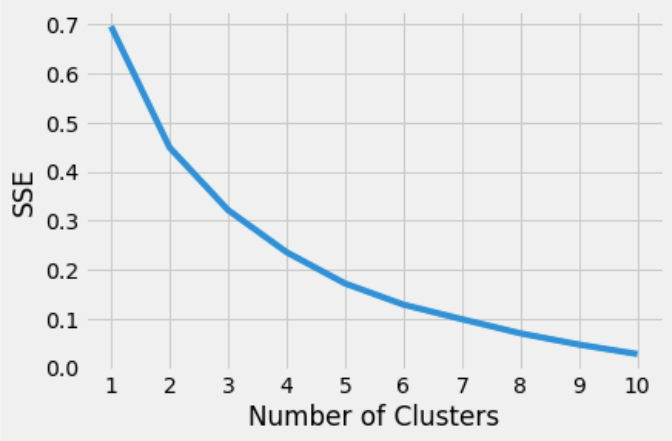

Figure 2. SSE – Number of Clusters Graph

Nevertheless, determining the elbow point in the SSE curve isn't always straightforward. In case choosing the elbow point of the curve is difficult, there is a Python package named 'kneed,' can be used to identify the elbow point programmatically:

```
from kneed import KneeLocator
```

```
kl = KneeLocator(range(1, 11), sse, curve="convex", direction="decreasing")
kl.elbow
```

```
4
```

Figure 3. Elbow Point Detection

In Figure 3, KneeLocator is used to identify the elbow point in our SSE curve, and it is detected as 4. The function is needed some parameters such as curve type; it is convex in our example and direction of the curve; we set it to decreasing based on our graph. Finally, our k value input was set to 4 for K-Means Clustering Algorithm.

4. Results

After building the K-Means model, we run it on the prepared venues data frame. As a result, we obtained 4 clusters, which were already defined in the beginning.

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels
0	Charlottenburg-Wilmersdorf	Charlottenburg,Charlottenburg-Nord,Grunewald,H...	52.586016	13.283000	1
1	Friedrichshain-Kreuzberg	Friedrichshain,Kreuzberg	52.586016	13.450000	0
2	Lichtenberg	Alt-Hohenschönhausen,Falkenberg,Fennpfuhl,Frie...	52.514557	13.498307	1
3	Marzahn-Hellersdorf	Biesdorf,Hellersdorf,Kaulsdorf,Mahlsdorf,Marzahn	52.533001	13.583000	2
4	Mitte	Gesundbrunnen,Hansaviertel,Mitte,Moabit,Tierga...	52.516998	13.367000	1
5	Neukölln	Britz,Buckow,Gropiusstadt,Neukölln,Rudow	52.483002	13.450000	1
6	Pankow	Blankenburg,Blankenfelde,Buch,Französisch,Buch...	52.569595	13.403235	1
7	Reinickendorf	Borsigwalde,Frohnau,Heiligensee,Hermsdorf,Konr...	52.574093	13.345394	1
8	Spandau	Falkenhagener,Feld,Gatow,Hakenfelde,Haselhorst...	52.534073	13.181689	3
9	Steglitz-Zehlendorf	Dahlem,Lankwitz,Lichterfelde,Nikolassee,Stegli...	52.432999	13.252171	1
10	Tempelhof-Schöneberg	Friedenau,Lichtenrade,Mariendorf,Marienfelde,S...	52.466999	13.383000	1
11	Treptow-Köpenick	Adlershof,Alt-Treptow,Altglienicke,Baumschulen...	52.450001	13.567000	1

Table 8. K-Means Clustering results

As it is seen in Table 8, cluster 1 has the highest number of entities (region) compared to the other groups. When we display the clusters on a map, cluster members' locations can be seen (Figure 5).

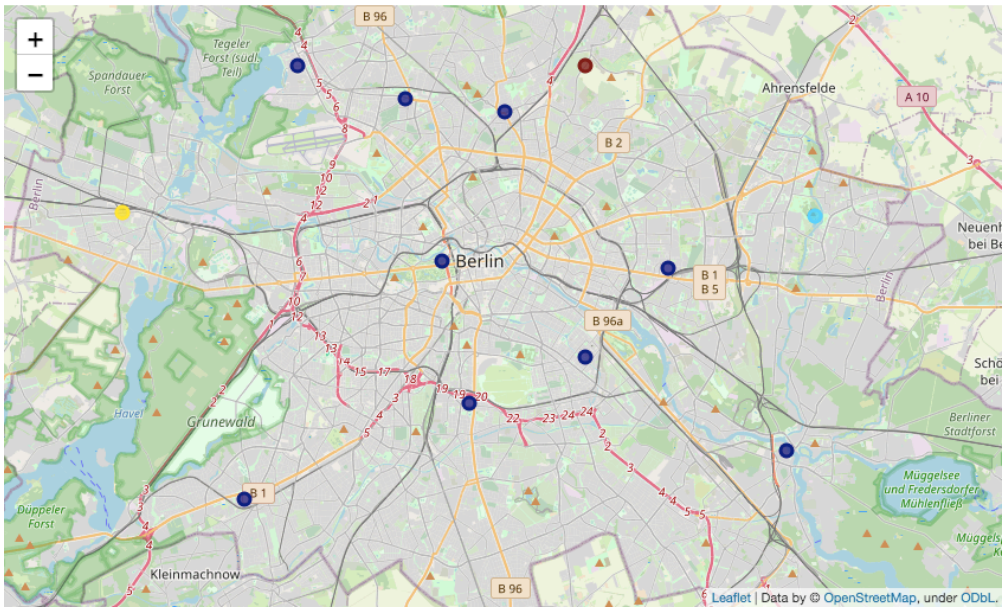

Figure 4. K-Means Clustering results on a map

In Figure 4, Dark blue color clusters indicating cluster 1 are outnumbered, and they have mainly cafes and recreational venues. The locations also have a high number of fast-food restaurants located in cluster 1 (Table 9).

	Borough	Longitude	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Charlottenburg-Wilmersdorf	13.283000	1	Café	Clothing Store	Drugstore	Supermarket	Restaurant	Italian Restaurant	Shopping Mall	Ice Cream Shop	Indian Restaurant	Sandwich Place
2	Lichtenberg	13.498307	1	Bakery	Supermarket	Park	Coffee Shop	ATM	Soccer Stadium	Bookstore	Mexican Restaurant	Pharmacy	Breakfast Spot
4	Mitte	13.367000	1	Hotel	Plaza	Hotel Bar	Museum	Spa	Concert Hall	Coffee Shop	Italian Restaurant	Monument / Landmark	Park
5	Neukölln	13.450000	1	Café	Bar	Supermarket	Italian Restaurant	Coffee Shop	Park	Vegetarian / Vegan Restaurant	Bistro	German Restaurant	Cocktail Bar
6	Pankow	13.403235	1	Café	Supermarket	Bakery	Drugstore	Park	Dance Studio	Organic Grocery	Sushi Restaurant	Burger Joint	Gym / Fitness Center
7	Reinickendorf	13.345394	1	Soccer Field	Supermarket	Metro Station	Plaza	Light Rail Station	Pool	Bus Stop	German Restaurant	Dry Cleaner	Park
9	Steglitz-Zehlendorf	13.252171	1	Café	Italian Restaurant	Supermarket	Doner Restaurant	Drugstore	Pizza Place	Yoga Studio	Fast Food Restaurant	Mobile Phone Shop	Gas Station
10	Tempelhof-Schöneberg	13.383000	1	Supermarket	Café	Bakery	Park	Doner Restaurant	Bus Stop	Italian Restaurant	Fried Chicken Joint	Drugstore	Plaza
11	Treptow-Köpenick	13.567000	1	Café	German Restaurant	Tram Station	River	Hotel	Soccer Stadium	Gas Station	Fast Food Restaurant	Boat or Ferry	Plaza

Table 9. Cluster 1

In Table 10, there are other clusters with their most common venues. For example, the region of Friedrichshain - Kreuzberg was labeled as cluster 0, and its most common venue is bakeries. For cluster 2, the most common venues are the garden, supermarket, and scenic lookout. On the other hand, cluster 3 has a supermarket, restaurant, and burger joint as the most visited venues.

	Borough	Longitude	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Friedrichshain-Kreuzberg	13.45	0	Bakery	Light Rail Station	Greek Restaurant	Bus Stop	Electronics Store	Café	Farm	Forest	Food & Drink Shop	Flower Shop

	Borough	Longitude	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Marzahn-Hellersdorf	13.583	2	Garden	Supermarket	Scenic Lookout	Drugstore	Tea Room	Mountain	Theme Park Ride / Attraction	Yoga Studio	Falafel Restaurant	Food & Drink Shop

	Borough	Longitude	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	Spandau	13.181689	3	Supermarket	Restaurant	Burger Joint	Mobile Phone Shop	Bus Stop	Automotive Shop	Clothing Store	Bakery	Big Box Store	Farmers Market

Table 10. The Cluster 0,2 and 3

5. Discussion

Based on the clustering results, we recommend that cluster 3, which included the Spandau region, might be a suitable region for the new fast-food restaurant. Its most common venue types are supermarket, restaurant, and burger joint. It shows that fast-food costumers already visit this region to have a meal. Additionally, it is far from the city center, and this can allow becoming an unrivaled fast-food restaurant around the area.

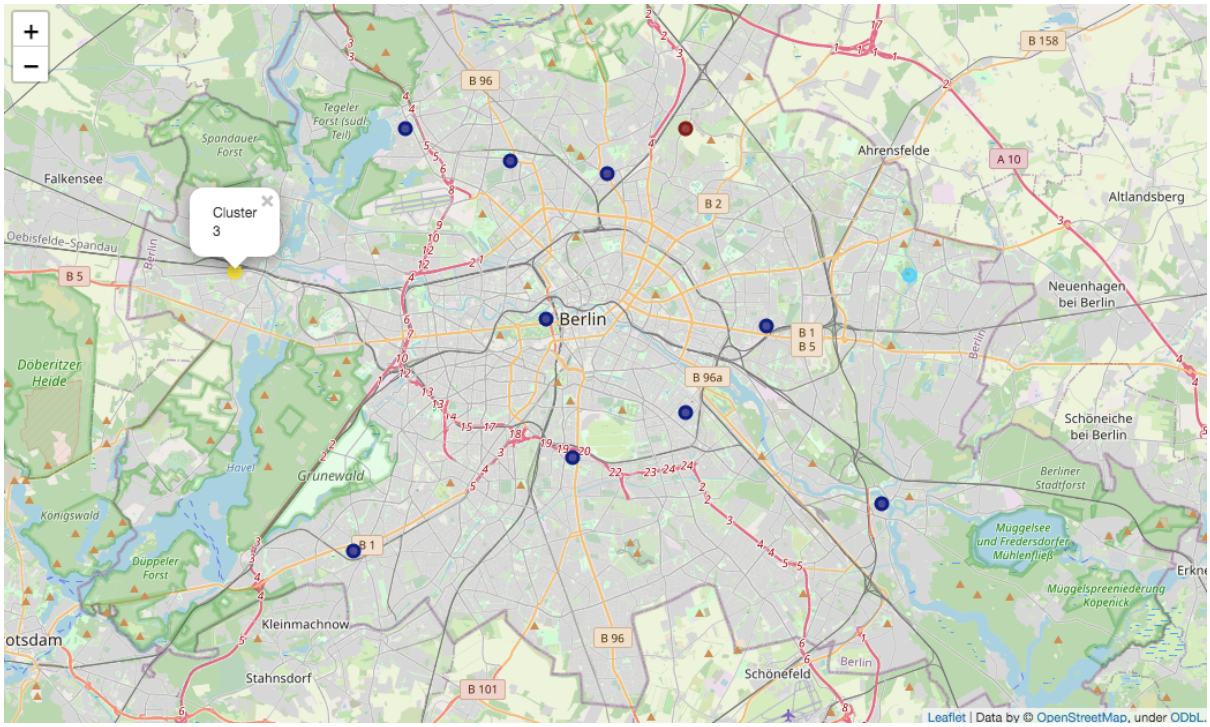

Figure 5. Cluster 3

Even cluster 1 has some boroughs mostly visited by fast-food lovers; these locations may lead to a competitive market for a new restaurant. Thus, we think that a detailed study should be conducted using additional information such as restaurant properties if you will select a region located in cluster 1.

6. Conclusion

This paper developed a solution to pick the most suitable location for a new fast-food restaurant. We obtained datasets from different resources and processed them to become input for an unsupervised machine learning model. We used Python package sci-kit learn while building K- means clustering model after selecting the most appropriate cluster number using SSE and elbow point method.

Based on the clustering results, we made recommendations regarding the best region for a new fast-food restaurant in Berlin.

6. References

https://en.wikipedia.org/wiki/Category:Localities_of_Berlin

https://foursquare.com/developers

https://geopandas.org/geocoding.html