Neighborhood Segmentation to Find the Best Location for a Sustainable Meat Restaurant

A submission for the final project from the Coursera Course in Applied Data Science Capstone from Johanna Schulte-Drüggelte

1. Introduction

About 6.1 million people in Germany at least largely forgo meat and meat products for many reasons. Between 2016 and 2018 market revenues made with the sale of vegetarian products increased by about 400 million euros in Germany and approximately 13 percent of all new food products launched in 2018 were vegan products in nature. How can we turn this trend into a business? Is it really sustainable to cut all the meat sources we have? Or can we also rarely rely on sustainably sourced meat products instead of highly processed vegan alternatives?

While many young people are looking for healthy and sustainable alternatives to conventionally sourced meat there is still a source of meat that is actually underrated. Game often is found in regions where their natural predators are extinct. Forest owners are looking for ways to protect areas of reforestation which are as important as ever, regarding their role in reducing CO_2 emissions. Also certain species contribute to the spread of dangerous viruses such as African Swine Fever and Influenza. Therefore it is evident that certain game populations need to be regulated. Hunting may have a bad reputation with some people in the society. Yet there is an increase in demand of game, because of the so to speak negative CO_2 balance. While game is mostly found in rural areas, many cities suffer from increasing populations of e.g. raccoons and wild boars. In order to help hunters and land owners with the sales of game, and in order to meet the demands of an increasingly ecologically thinking urban population the intent is to open a restaurant that provides meals using responsible meat sources, including game but also with meat from cattle from extensive farming when no other land use is possible.

1.1 Problem

How can we find the best location for the restaurant in order to have the right kind of people who care about the environment so that the restaurant can be run efficiently?

This project is dedicated to possible restaurant entrepreneurs from outside of Berlin who want to bring this kind of restaurant to the people and want to run their business as profitable as possible with the right kind of people frequenting this neighborhood to go out to eat.

1.2 Approach

We can use Neighborhood Segmentation and the Machine Learning Algorithm k-means to come up with the kind of neighborhood suited to the opening of this restaurant. We are looking for venues attracting younger and alternative people such as coffee shops, event locations, Yoga studios among others.

2. Methodology

In this section the whereabouts of the methodology will be explained. Google colaboratory was used to create the notebook. Programming language was Python.

2.1 Use of Data

For our restaurant we choose the city of Berlin. The Wikipedia page for Berlin's district will give us the names of boroughs and districts using web scraping techniques such as beautifulsoup. In order to find latitude and longitude we will use a Geocoder and create a dataframe that combines borough, neighborhood and the location. With these data the first Folium map of Berlin can be created. Then, we add the Foursquare API derived data and can find venues for each neighborhood, e.g. the ten most common in each neighborhood. After using one-hot-encoding, the k-means algorithm will be applied and we can find the right cluster of neighborhoods for our restaurant.

2.2.1 Web scraping

A simple pandas function was used to scrape the Wikipedia site in German. pd.read_html which worked fine.

The result of the cleaning of the dataframe is this:

Table 1 Head of Neighborhood dataframe

	Neighborhood	borough
0	Mitte	Mitte
1	Moabit	Mitte
2	Hansaviertel	Mitte
3	Tiergarten	Mitte
4	Wedding	Mitte

2.2.2 Geocoding API

Nominatim was used as a geolocator. Empty lists were created to be filled with the geocode. These lists were than inserted into the Berlin_neighbourhoods dataframe, shown in Table 2.

Table 2 Head of dataframe with geocode

	Neighborhood	borough	Latitude	Longitude
0	Mitte	Mitte	52.517690	13.402376
1	Moabit	Mitte	52.530102	13.342542
2	Hansaviertel	Mitte	52.519123	13.341872
3	Tiergarten	Mitte	52.509778	13.357260
4	Wedding	Mitte	52.550123	13.341970

2.3 Create a map with Folium

Now that the geocode of all neighborhoods is available, a Folium map is created with blue pop up labels for each neighborhood as seen in figure 1.

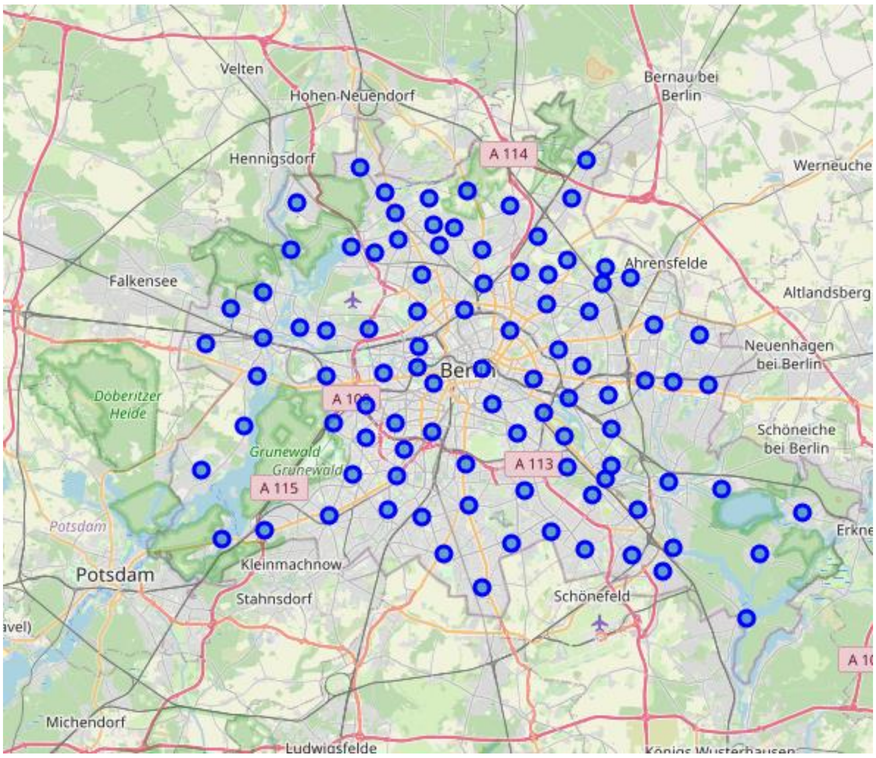

Figure 1 Map of Berlin with pop up markers for each neighborhood

2.4 Retrieving the venue data from the Foursquare API

The venue data for Berlin were obtained using the Foursquare API. A URL was created using the credentials and geocode. As an example the Top 5 venues for the first neighborhood Mitte were obtained. Then the categories of those venues were extracted.

Afterwards a dataframe was created providing every venue in Berlin and by creating a new URL.

Table 3 Head of the dataframe with all Berlin venues from Foursquare API

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Mitte	52.51769	13.402376	Lustgarten	52.518469	13.399454	Garden
1	Mitte	52.51769	13.402376	Kuppelumgang Berliner Dom	52.518966	13.400981	Scenic Lookout
2	Mitte	52.51769	13.402376	Radisson Blu	52.519561	13.402857	Hotel
3	Mitte	52.51769	13.402376	Bronzestatue "Heiliger St. Georg im Kampf mit ...	52.516290	13.405558	Outdoor Sculpture
4	Mitte	52.51769	13.402376	Designpanoptikum - surreales Museum für indust...	52.516941	13.406072	Museum

The venues of Berlin were grouped by neighborhood and counted. Afterwards the method one-hot-encoding was used, and the results were grouped by neighborhood and the mean of the frequency was calculated.

The ten most common venues per neighbourhood were then identified in order to make clusters.

2.5 Forming clusters of all neighborhoods in Berlin

The default number of clusters was 5. It already turned out early that one label was more abundant than others. The berlin_grouped dataframe was used for clustering.

The cluster label was inserted into the Berlin_neighbourhoods dataframe.

Table 4 Head of neighborhoods with Cluster label

	Neighborhood	borough	Latitude	Longitude	Cluster Labels 36	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	Mitte	Mitte	52.517690	13.402376	4	German Restaurant	Museum	History Museum	Art Gallery	Café	Hotel	Fountain	Art Museum	Concert Hall	Brewery
1	Moabit	Mitte	52.530102	13.342542	4	Café	Hostel	Doner Restaurant	Hotel	German Restaurant	Bar	Gym / Fitness Center	Burger Joint	Drugstore	Bakery
2	Hansaviertel	Mitte	52.519123	13.341872	4	Café	Art Museum	Bakery	Plaza	Bus Stop	Metro Station	Farmers Market	Sporting Goods Shop	Mediterranean Restaurant	Boat or Ferry
3	Tiergarten	Mitte	52.509778	13.357260	4	Lounge	Hotel Bar	Hotel	Memorial Site	Sculpture Garden	German Restaurant	Garden	Historic Site	Scandinavian Restaurant	Café
4	Wedding	Mitte	52.550123	13.341970	4	Bus Stop	Park	Café	Tram Station	Bakery	Tennis Court	Big Box Store	Gas Station	Organic Grocery	Supermarket

2.6 Visualisation with Folium

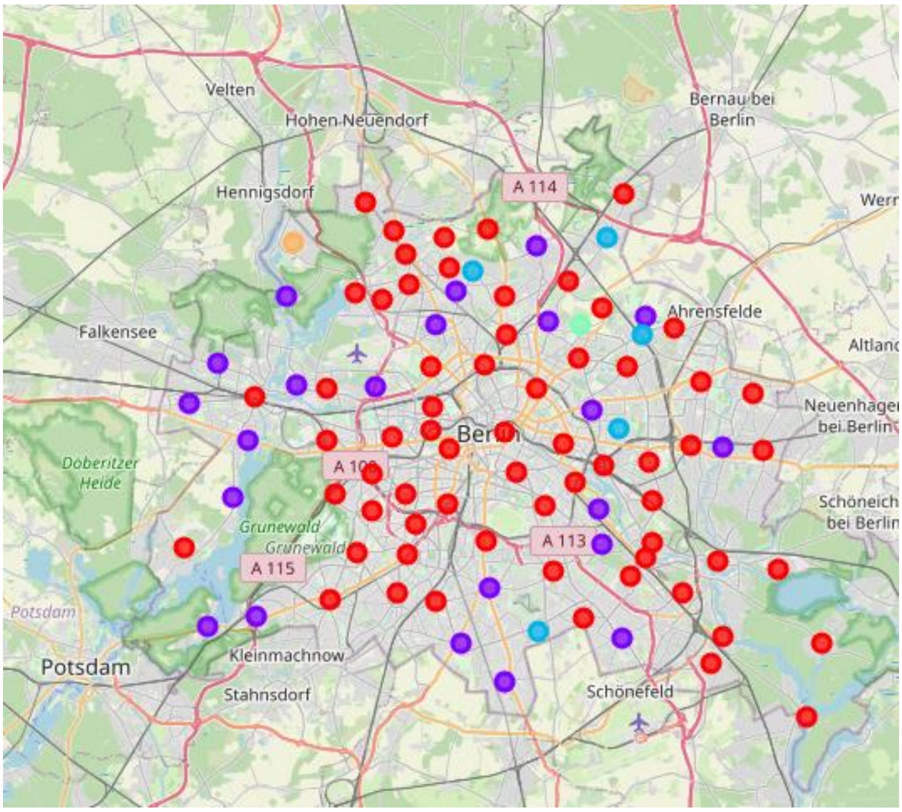

Figure 2 Berlin map with colored labels representing the different cluster labels

3. Results

An overview of the different clusters is given in this section.

3.1 Cluster 1

The outskirts of Berlin have a very basic set of venues and are more practical with supermarkets very often as the most common venue. So it is not recommended for our restaurant.

3.2 Cluster 2

This cluster is very similar to the first one, but has a slightly higher choice of restaurants.

Table 5 Cluster 2 of the Berlin neighborhoods

	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
12	Karow	Supermarket	Restaurant	Bus Stop	Zoo Exhibit	Ethiopian Restaurant	Fountain	Food Court	Food & Drink Shop	Flower Shop	Fishing Store
19	Rosenthal	Tram Station	Supermarket	German Restaurant	Automotive Shop	Currywurst Joint	Event Space	French Restaurant	Fountain	Food Court	Food & Drink Shop
51	Buckow	Supermarket	Women's Store	Pizza Place	Zoo Exhibit	Fish Market	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Flower Shop
74	Lichtenberg	Tram Station	Automotive Shop	Gym / Fitness Center	Bowling Alley	Park	Supermarket	Diner	Discount Store	French Restaurant	Fountain
78	Neu-Hohenschönhausen	Supermarket	Movie Theater	Tram Station	Fried Chicken Joint	French Restaurant	Fountain	Food Court	Food & Drink Shop	Flower Shop	Fishing Store

3.3 Cluster 3 and 4

Only one neighborhood was found for each of these neighborhoods.

3.4 Cluster 5

This cluster was very big and was cleaned of unwanted venues in the area, like supermarkets, fast food restaurants and German restaurants. There we have 16 suitable neighborhoods to open a modern sustainable restaurant. They all provide an infrastructure with different kinds of places to eat. Yet certain venues indicate to attract younger people such as museums and galleries, pubs and cafés, but also rare kinds of restaurants with e.g. Ethiopian cuisine.

Table 6 Results of the most favorable cluster for the restaurant

	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
2	Hansaviertel	Café	Art Museum	Bakery	Plaza	Bus Stop	Metro Station	Farmers Market	Sporting Goods Shop	Mediterranean Restaurant	Boat or Ferry
6	Friedrichshain	Coffee Shop	Café	Pub	Bar	Vegetarian / Vegan Restaurant	Middle Eastern Restaurant	Bookstore	Bagel Shop	Ice Cream Shop	Pizza Place
8	Prenzlauer Berg	Café	Bakery	Cocktail Bar	Beer Bar	Vietnamese Restaurant	Falafel Restaurant	Park	Organic Grocery	Coffee Shop	Donut Shop
10	Blankenburg	Playground	Bus Stop	Café	Greek Restaurant	Zoo Exhibit	Falafel Restaurant	Fountain	Food Court	Food & Drink Shop	Flower Shop
16	Buch	Italian Restaurant	Art Gallery	Bakery	Asian Restaurant	Big Box Store	Drugstore	Zoo Exhibit	Flower Shop	Fast Food Restaurant	Fish Market
23	Schmargendorf	Italian Restaurant	Ice Cream Shop	Drugstore	Gym	Chinese Restaurant	Restaurant	Café	Trattoria/Osteria	Deli / Bodega	Coffee Shop
25	Westend	Café	Bar	Art Museum	Gourmet Shop	Drugstore	Liquor Store	Italian Restaurant	Plaza	Bus Stop	Ice Cream Shop
37	Lichterfelde	Bakery	Italian Restaurant	Vietnamese Restaurant	Café	Liquor Store	Chinese Restaurant	Sculpture Garden	Eastern European Restaurant	Pool	Dive Bar
49	Neukölln	Bar	Café	Coffee Shop	Middle Eastern Restaurant	Dive Bar	Cocktail Bar	Bistro	Italian Restaurant	Vegetarian / Vegan Restaurant	Nightclub
57	Johannisthal	Tram Station	Pub	Park	Taverna	Dessert Shop	Burger Joint	Sushi Restaurant	Movie Theater	Pizza Place	Food & Drink Shop
58	Niederschöneweide	Indian Restaurant	Restaurant	Greek Restaurant	Harbor / Marina	Fast Food Restaurant	Gas Station	Motel	History Museum	Convenience Store	Italian Restaurant
60	Bohnsdorf	Park	Italian Restaurant	Flower Shop	Ethiopian Restaurant	French Restaurant	Fountain	Food Court	Food & Drink Shop	Fishing Store	Fish Market
64	Grünau	Tram Station	Restaurant	Historic Site	Dessert Shop	Boat or Ferry	Flower Shop	Food & Drink Shop	Food Court	Ethiopian Restaurant	Fish Market
66	Schmöckwitz	Harbor / Marina	Accessories Store	Italian Restaurant	Gas Station	Tram Station	Ice Cream Shop	Dance Studio	Ethiopian Restaurant	Food Court	Food & Drink Shop
68	Biesdorf	Big Box Store	Liquor Store	Plaza	Bakery	Palace	Outdoor Event Space	Park	Light Rail Station	Farmers Market	Farm
75	Falkenberg	Nature Preserve	Coffee Shop	Café	Zoo Exhibit	Falafel Restaurant	French Restaurant	Fountain	Food Court	Food & Drink Shop	Flower Shop

4. Discussion

The clusters found were distinct. Yet we have to argue that another approach could have been chosen.

It could be argued that the algorithm used and the number of k could be improved and more accurate.

'Clustering algorithms seek to learn, from the properties of the data, an optimal division or discrete labeling of group points' (Jake Vanderplas). In k-means clustering the definition is that each point is closer to its own cluster center than to other cluster centers. Is this algorithm even appropriate to compare the venue categories in different neighborhoods? The advantage of k-means is that it is very easy to understand. The biggest disadvantage is that it cannot learn the number of k by itself and it has to be determined beforehand. An alternative algorithm would be DBSCAN but its use for this project is out of the scope here. Having applied some critical thinking the number of neighborhoods of the suitable cluster could be reduced. It could also be argued that the data for clustering could be simplified beforehand, this was not done because we wanted an idea of what the entire neighborhood looks like. The mention of many supermarkets in the first cluster already gave us the idea, that it might not be a popular area for restaurant visits.

Also we don’t know how complete the data from Foursquare and Wikipedia are, so there could be other data providers useful. However all of the data were free to be used.

5. Conclusion

In the notebook provided we found a number of neighborhoods suitable for the opening of a restaurant providing sustainably sources of meat for its dishes such as venison. This restaurant is supposed to attract younger alternative people that want to make more sustainable choices and to reduce their CO_2 footprint. The found cluster contains neighborhoods such as Prenzlauer Berg and Friedrichshain which are obvious choices for people who know Berlin. However, other neighborhoods rise in popularity which is shown with the neighborhoods in this cluster and might have other advantages such as lower rents and might be easier to reach. But these parameters are out of the scope of this project.