Opening a new Burger Restaurant in Hamburg, Germany

By: Simon Schuster

July 2021

Coursera Capstone Project

IBM Data Science

1. Introduction

A thick, juicy beef patty, cheese losing its hard edges as it melts overtop, a slathering of sauce, a spoonful of tangy relish and a couple of slices of smoky, salty bacon, a bit of onion, some tomato and crisp lettuce all tucked between two halves of a bun: it is all the things you could want.[1]
The fast-food trend is growing in Germany. As a result, burger restaurants become more and more popular in big cities like Hamburg. Opening a new restaurant requires serious consideration and is a lot more complicated than it seems. The location of a burger restaurant is one of the most important decisions that will determine whether it will be a success or a failure.

1.1 Business Problem

The objective of this capstone project is to find the best neighborhood in Hamburg to open a Burger Restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide a solution to answer the business question: In the city of Hamburg, Germany, where would you recommend opening a new burger restaurant?

2. Data

To solve the problem, we need following data:

- List of neighborhoods in Hamburg
- The coordinates of the neighborhoods. This is required to plot the map and also get the venue data from the Foursquare API
- Venue data, particularly data related to food and restaurants. We will use this data to cluster the different neighborhoods. We will also be using population density data for clustering.

Source of the data and how to extract them:

This Wikipedia page (https://de.wikipedia.org/wiki/Liste_der_Bezirke_und_Stadtteile_Hamburgs) contains a list of neighborhoods in Hamburg. With the help of Python requests and beautifulsoup packages, we will use web scrapping techniques to get the data from the page. We store the data in a pandas dataframe. To get the coordinates of each neighborhood, we use a Python Geocoder package. This will give use the latitude and longitude. After this the dataframe looks like follows:

	Neighborhood	Latitude	Longitude	Borough	Population density
0	Hamburg-Altstadt, Hamburg	53.550468	9.994640	Hamburg-Mitte	979
1	HafenCity, Hamburg	53.542913	9.995835	Hamburg-Mitte	2239
2	Neustadt, Hamburg	53.549881	9.979048	Hamburg-Mitte	5549
3	St. Pauli, Hamburg	53.553935	9.959432	Hamburg-Mitte	8839
4	St. Georg, Hamburg	53.557149	10.014256	Hamburg-Mitte	4733

Then we will use the Foursquare API to get the venue data of the neighborhoods. Foursquare has one of the largest databases. It will provide many categories of the venue data, we are particularly interested in the food and restaurants.

This is a project that will make use many data science skills, from web scrapping, working with Foursquare API, data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization.

3. Methodology

Firstly, we need to get the list of neighborhoods in the city of Hamburg. Fortunately, the list is available in the Wikipedia page (https://de.wikipedia.org/wiki/Liste_der_Bezirke_und_Stadtteile_Hamburgs). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Kuala Lumpur.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 1000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Burger Joint” data, we will filter the “Burger Joint” as venue category for the neighborhoods. [2]

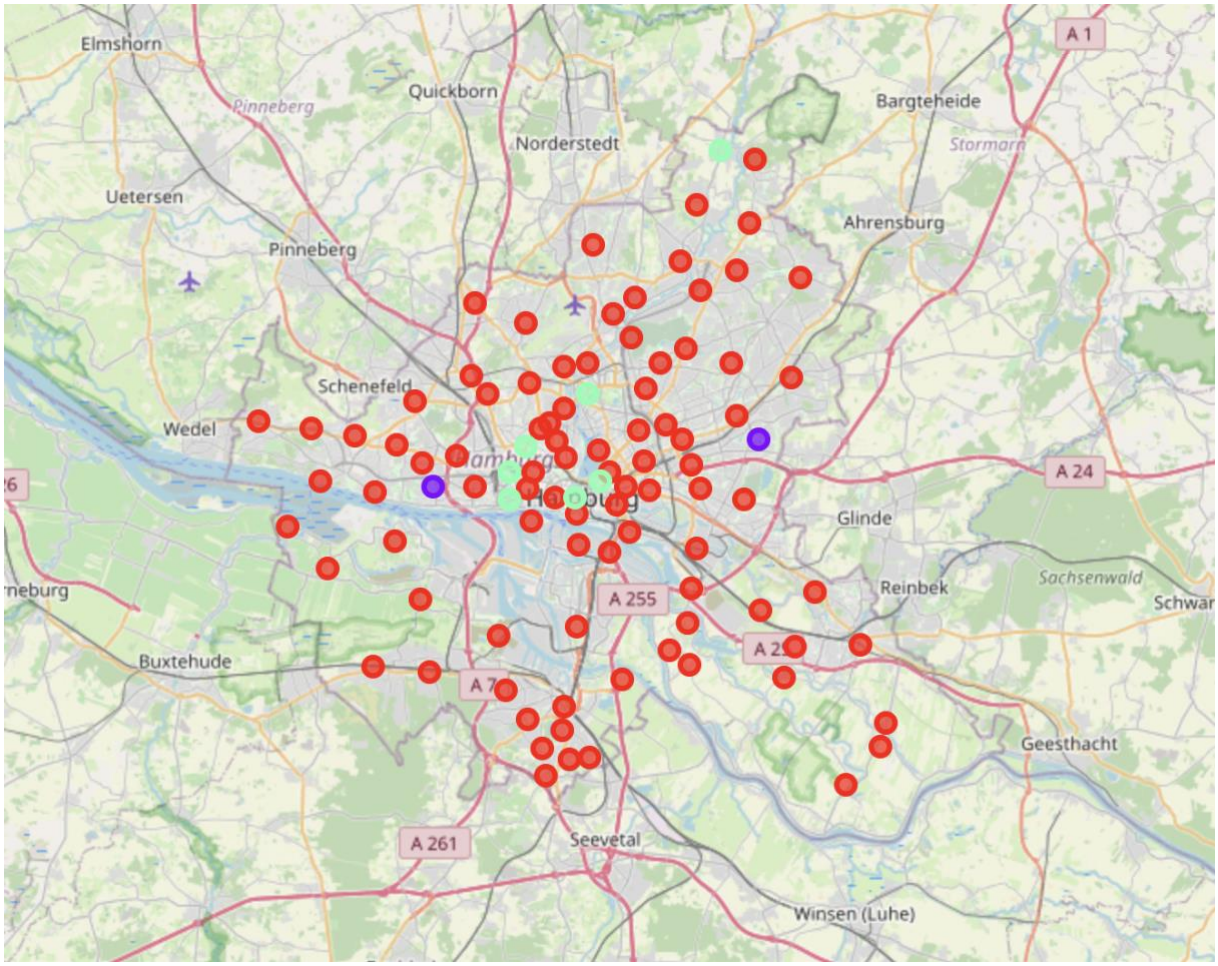
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Burger Joint”. The results will allow us to identify which neighborhoods have higher concentration of burger restaurants while which neighborhoods have fewer number of burger restaurants. Based on the occurrence of burger restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new burger bar.

4. Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Burger Joint”:

- Cluster 0: Neighborhoods with low number to no existence of burger restaurants
- Cluster 1: Neighborhoods with moderate number of burger restaurants
- Cluster 2: Neighborhoods with high concentration of burger restaurants

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.



5. Discussion

As observations noted from the map in the Results section, most of the burger restaurants are concentrated in the central area of Hamburg, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number of burger restaurants in the neighborhood. This represents a great opportunity and high potential areas to open new restaurant as there is very little to no competition from existing bars. Meanwhile, burger restaurants in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of burger restaurants. From another perspective, the results also show that the oversupply of burger restaurants mostly happened in the central area of the city, with the suburb area still have very few

bars. Therefore, this project recommends to open new burger restaurant in the neighborhood “Sternschanze” in cluster 0 with little to no competition. This neighborhood has a small number of burger bars and a high population density. Lastly, this project recommends avoiding neighborhoods in cluster 2 which already have high concentration of burger restaurants and suffering from intense competition.

6. Limitations and Suggestions for future research

In this project, we only consider one factor i.e., frequency of occurrence of burger restaurants, there are other factors such as tourists and income of residents that could influence the location decision of a new restaurant. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new burger restaurant. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

7. Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhood “Sternschanze” in cluster 0 is the most preferred location to open a new burger restaurant.

8. References:

[1]: https://calgaryherald.com/life/food/the-battle-of-the-burgers-an-introduction

[2]: https://developer.foursquare.com/docs