Location Suggestion for a new shopping center in Hamburg

Shahab Shadan

May 2020

1. Introduction

1.1 Background

Hamburg is the second largest metropole in Germany with approximately 1,7 millions population after the capital, Berlin. Growing industry, tourism attractions, fish market, ports and etc. are the features of this city over more than a century on the first glance. Hamburg has seven boroughs representing 104 quarters with almost 100 postal codes.

1.2 Problem

Europa Passage is one of the most famous shopping centers in Hamburg. To this date, it has the highest number of reviews on Google Maps among shopping centers in Hamburg. Its owner has decided to open a new shopping center, but he is unsure about its location, since he wants it to succeed and make profits approximately as much as Europa Passage does.

1.3 Interest

This interests not only the owner of the Europa Passage but also other start-ups who want to build their first shopping center and they want it to succeed as much as Europa Passage did.

2. Data (Pre-)Processing

2.1 Data Sources

The data regarding the population and area of regions based on their postal codes was gathered from https://postal-codes.cybo.com and the one regarding the location coordinates was loaded from https://www.geonames.org. Moreover, the location data about the venues was obtained from https://www.foursquare.com through its API.

2.2 Data Cleaning and preparation

Invalid values were removed from the data frame row-wise. Inconsistency issues resulting from obtaining data from various sources with not suitably overlapping data were solved. Only postal codes were kept which did not have any flaw or anomaly. The number of removed postal codes is in percentage neglectable. The following head of a data frame shows the not preprocessed data acquired from “geonames”:

	Unnamed: 0	Place	Code	Country	Admin1	Admin2	Admin3	Admin4
0	1.0	Hamburg	20095	Germany	Hamburg	NaN	Hamburg, Freie und Hansestadt	Hamburg, Freie und Hansestadt
1	NaN	53.552/10	53.552/10	53.552/10	53.552/10	53.552/10	53.552/10	53.552/10
2	2.0	Hamburg	20097	Germany	Hamburg	NaN	Hamburg, Freie und Hansestadt	Hamburg, Freie und Hansestadt
3	NaN	53.548/10.019	53.548/10.019	53.548/10.019	53.548/10.019	53.548/10.019	53.548/10.019	53.548/10.019
4	3.0	Hamburg	20099	Germany	Hamburg	NaN	Hamburg, Freie und Hansestadt	Hamburg, Freie und Hansestadt

Figure 1 Not cleaned data of postal codes' coordinates

Afterwards, the datasets from different sources were reworked, so that they are compatible and can be represented in a single dataset whose first rows are as follows:

	Postal Code	Population	Area	Latitude	Longitude
0	20095	3574	0.855	53.552	10.000
1	20097	12023	2.179	53.548	10.019
2	20099	4612	1.490	53.558	10.011
3	20144	5893	1.196	53.574	9.975
4	20146	4402	1.000	53.567	9.980

Figure 2 First rows of the cleaned and prepared dataset

Figure 3 represents the locations of the postal codes on the map (Europa Passage is in 20095):

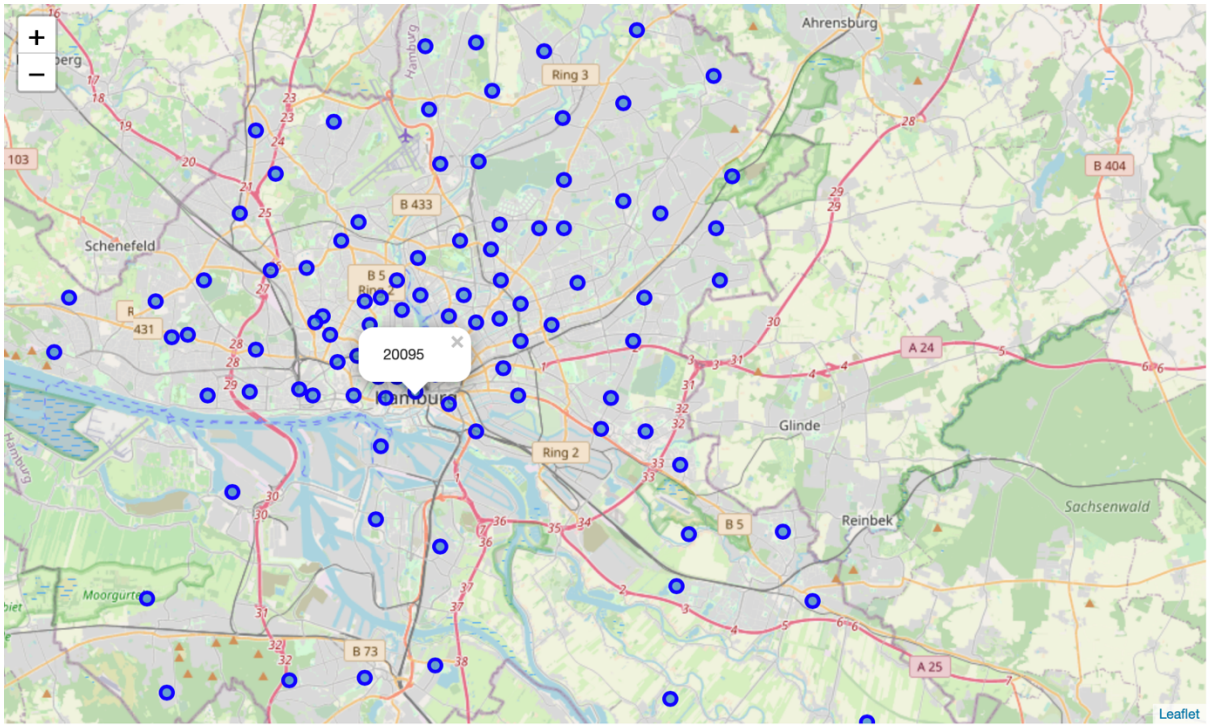

Figure 3 Hamburg postal codes

Furthermore, venues information was obtained from Foursquare through the API. The prepared and preprocessed dataset which is the input of our model is as follows:

	Postal Code	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Aquarium	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports
0	20095	0.000000	0.0	0.0	0.00000	0.000000	0.011111	0.0	0.00000	0.0	0.011111	0.0	0.011111	0.00000
1	20097	0.058824	0.0	0.0	0.00000	0.058824	0.000000	0.0	0.00000	0.0	0.000000	0.0	0.058824	0.00000
2	20099	0.000000	0.0	0.0	0.00000	0.000000	0.000000	0.0	0.01087	0.0	0.000000	0.0	0.010870	0.01087
3	20144	0.000000	0.0	0.0	0.00000	0.000000	0.000000	0.0	0.00000	0.0	0.000000	0.0	0.000000	0.00000
4	20146	0.000000	0.0	0.0	0.02439	0.000000	0.000000	0.0	0.00000	0.0	0.000000	0.0	0.024390	0.00000

Figure 4 The input dataset

The first two selected features are the population and the area of each postal code. In addition, venues information in each postal code was obtained from Foursquare through the API. There are about 270 venue categories in Hamburg. The next main derived feature was the average frequency of categories per postal code

3. The predictive unsupervised model

Due to the fact that we do not know the ground truth, which could tell how many groups or classes of postal codes exist, an unsupervised model shall be built and trained. In this case, the K-Means Clustering algorithm is used. The number of clusters (n=5) was chosen via the elbow method based on the fit time and distance score (Figure 5).

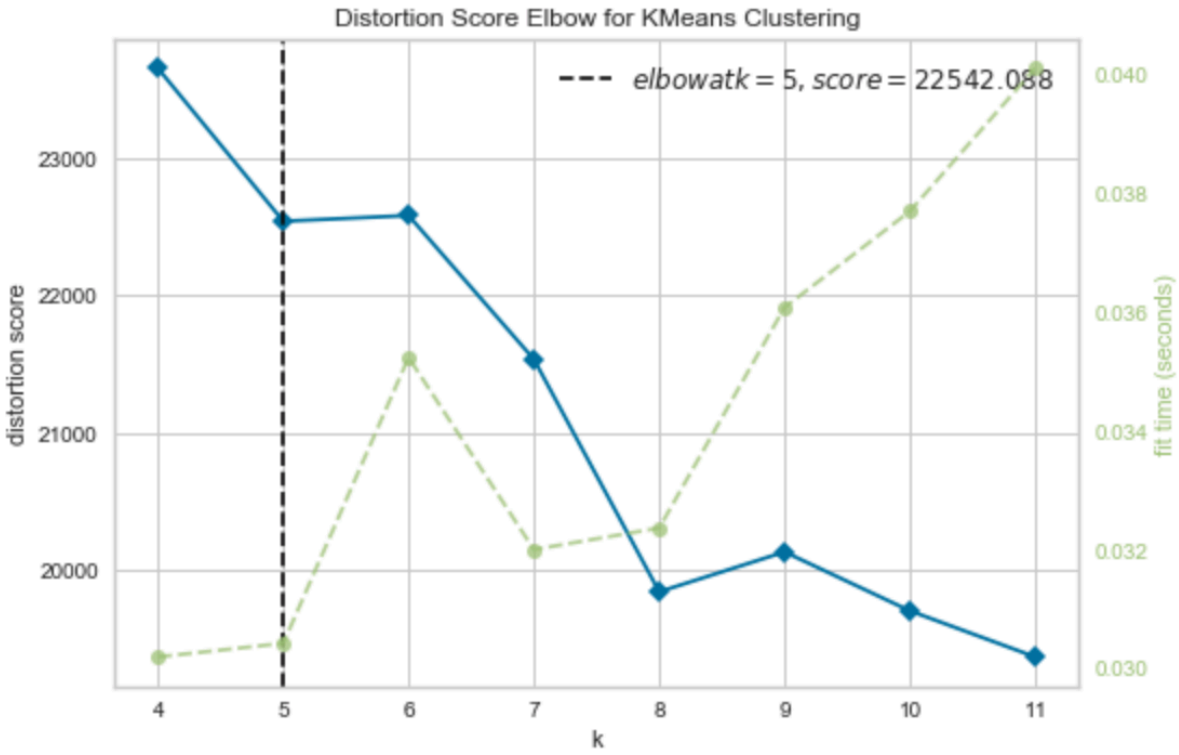

Figure 5 K-Means Elbow

After fitting the model, Europa Passage, meaning the postal code 20095, is put in cluster number 0. Other members of this cluster are two postal codes, namely 20355 and 22299. These two are the location suggestion of the model based on the population, area, the number and categories of the venues inside each postal code. The suggestions are marked blue and the postal code 20095 (Europa Passage) is marked red:

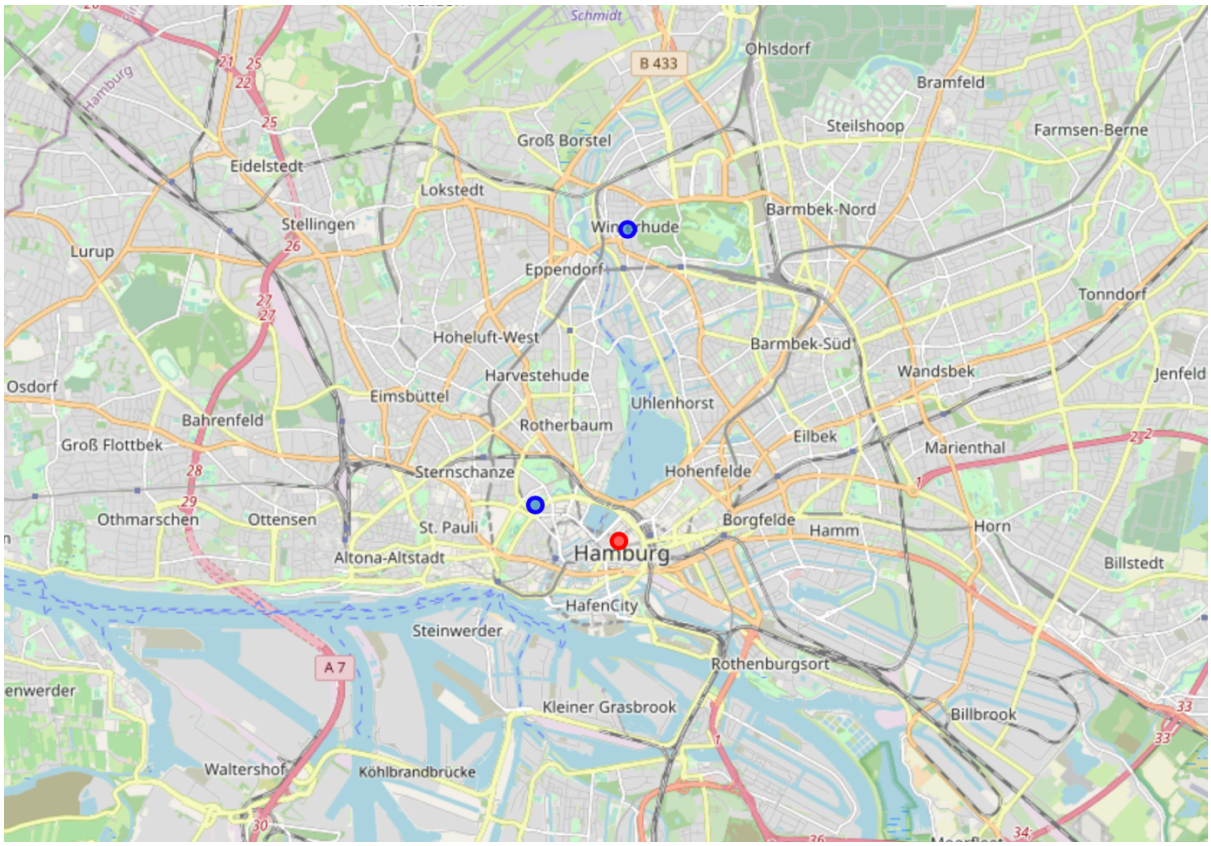

Figure 6 Final location suggestions

In the next dataset you can see the most common venue categories in each of these postal codes:

Postal Code	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	Population	Area
20095	Vietnamese Restaurant	Cosmetics Shop	Café	Bookstore	Drugstore	Burger Joint	Bakery	Hotel	Plaza	Electronics Store	3574	0.855
20355	Hotel	Restaurant	Gym / Fitness Center	Botanical Garden	Café	Burger Joint	Garden	Ramen Restaurant	Plaza	Playground	7564	1.423
22299	Café	Italian Restaurant	Drugstore	Trattoria/Osteria	Bus Stop	Vietnamese Restaurant	Bakery	Plaza	Bank	Bar	6460	1.708

Figure 7 Properties of Cluster 0 members

Both of these suggestions seem reasonable due to their similarities with the postal code 20095, located in different sight of the city, so that it does affect or get affected through its short distance to Europa Passage. For a more detailed observation and

evaluation of the result, both the absolute and squared error were calculated relative to the centroid of cluster 0 and to the postal code 20095. These values represent the similarity of these postal codes to the centroid and to 20095, respectively:

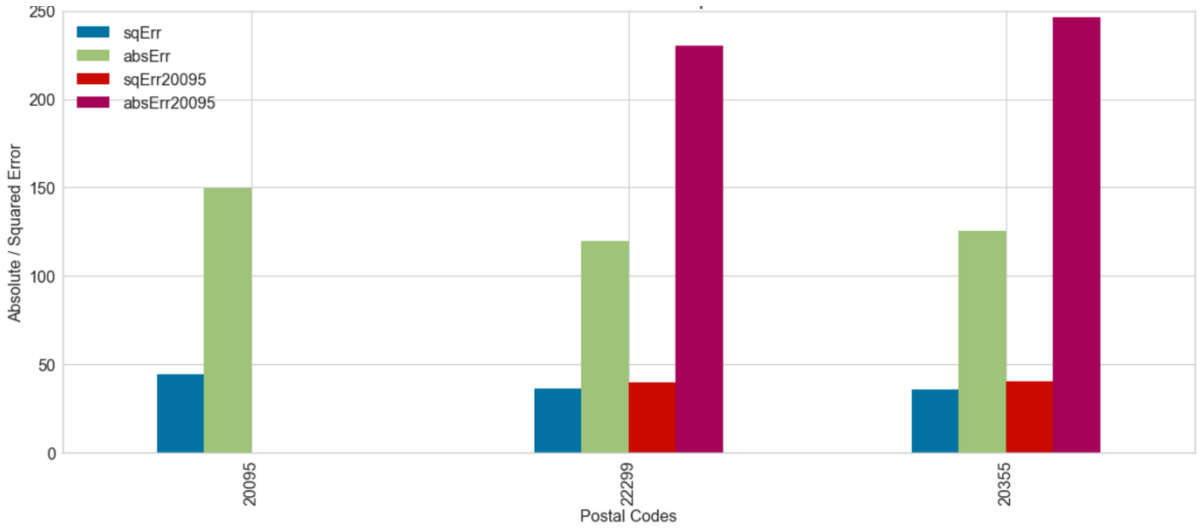

Figure 8 Absolute and squared error diagram of cluster 0

4. Conclusion

The aim of this project was to suggest locations for a new shopping center in Hamburg, which could be as successful as Europa Passage in postal code 20095. As you see in Figure 8, all the three postal codes in cluster 0 have almost the same absolute and squared error relative to the centroid ensuring us that the suggested locations are similar to 20095 and share the same properties and opportunities for a new successful shopping center.