Finding Where to Live in a New City:

A Case Study for the city of Hamburg, Germany

Mihael Machado de Souza

1. INTRODUCTION

Finding a place to live is always a challenge. When looking for a new home, to purchase or rent, you must consider whether the rent price is okay, the quality of the house/apartment, how much renovating it might need, whether it is already furnished or not. You also need to take into account the neighborhood the apartment is in and what type of amenities it offers. This can also highly depend on one's own family structure. If young children are involved, does it have good schools nearby? If you do not have your own vehicle, does it have good access to public transport? What about supermarkets? Or restaurants? How noisy is the neighborhood? These are just some of the questions we have to think about when choosing the next place to call home.

Now, on top of all of that, assume you are also moving to a city you now nothing about. You got this great opportunity, but it is in a whole different state/country. You have no family or connections there, so you cannot just ask for suggestions. Google is useful, but you first have to know what to search for. As an example, Hamburg has 103 neighborhoods. Where to start? General searches, such as "Where to live in Hamburg?", will find hundreds of pages worth of information. But without some firsthand knowledge, how to separate the good advice from the bad? Or to know if whoever wrote the advice has similar priorities to yours?

In those circumstances, a data driven approach can help you jumpstart your search the right way and can be done using a recommender system. On an individual level, if you can cross-reference what types of businesses or amenities are available in each neighborhood with their average rent prices, how safe they are and how far they are from your work place, for example, you can direct your searches towards a handful of preferred locations as opposed to the whole city. And you would know, in advance, that those neighborhoods/quarters are suited for your own personal preferences. On a business level, realtors or real state agencies could develop short questionnaires to probe a client's preferences and use that to get tailored recommendations for each potential client. City administration, itself, could add such questionnaires to their own websites and produce these recommendations for anyone looking to move into the city.

2. DATA

2.1 Data Sources

To build our content-based recommender system, we first have to gather the data that we will use to calculate the weights for our content matrix. For this proof of concept, we'll leverage data from three sources: Foursquare (https://pt.foursquare.com/), ImmobilienScout24 (http://immobilienscout24.de/) and BingMaps (https://www.bing.com/maps). Our neighborhoods from Hamburg, with their MultiPolygon bounding boxes, are obtained from the geojson file made available in https://github.com/blackmad/neighborhoods. We define the central coordinate for each neighborhood as the average latitude and longitude within each bounding polygon.

With Foursquare, we will gather data from all venues available within 500 meters of a neighborhood's central coordinate. To build our recommender system, we will then categorize these venues in some broad categories. For example, if we compared our application to a movie recommender system, these would be the genres associated with each film. For Foursquare, they already provide a top-level category system that encompass their set of unique venue labels (https://developer.foursquare.com/docs/build-with-foursquare/categories/). As a first step, we will use these 10 categories. They are:

- Arts & Entertainment → Group venues like Museums, Theaters and Stadiums.
- College & University → Group different university and higher education buildings.
- Event → Location of general seasonal events, e.g. Christmas Markets and Parades.
- Food → Groups Restaurants, Diners and Coffee Shops, for example.
- Nightlife Spot → Groups Bars, Pubs and Night Clubs.
- Outdoors & Recreation → Green areas (e.g. Parks) and Athletics venues (e.g. Gym).
- Professional & Other Places → Business Services and Government Buildings.
- Residence → Private Homes, Housing Developments and Assisted Living.
- Shop & Service → General service providers (e.g. ATM) and stores (e.g. Malls).
- Travel & Transport → General Travel (e.g. Hotels and Rental) and Public Transport.

Aside from venue information, we also need rent prices across all neighborhoods. On a more general manner, this could be obtained from state reports. We would like, however, this information to be as current as possible. In that case, we will search for all available homes for rent in the ImmobilienScout24 website. Their search result page contains the listing's address, price (in € per month) and size (in m^2). We will scrape this page and gather data from all listings, and then normalize the price based on the listing's size to get a price per square meter for each neighborhood.

Finally, we will use BingMaps' API to calculate the travel distance between each central coordinate and the address provided by the potential client. This address would be, for example, of his/her new work address. We will use this information to calculate a distance index between each quarter and the place of work and consider that into our recommender system as well. If no address is given, we can simply set the index to zero and not take it into account when creating our recommendation.

Together, information from the three data sources mentioned above will form the content matrix for our recommender system. Then, we will create a set of questions that addresses each of the categories defined to gather the client's perspective/preferences according to each category. This will establish our user profile that, in combination with the content matrix, will allow us to obtain neighborhood recommendations tailored for each client.

2.2 Data Wrangling

2.2.1 Rental Data

Rental data is scraped from the ImmobilienScout24 website using the BeautifulSoup4 package in Python. We gather three information points from every listing in a page, looping through all pages until we receive back a set of empty lists. Neighborhood location is scraped as plain text from the <buttom> tag, with the condition that buttom contains a <svg> tag and does not contain a tag. This combination ensures that only listing's addresses are returned, which are then saved into a list. The price and size of each listing is simpler, as we only need to find all <dd> tags within the page that does not contain an inner . Returning the contents of these tags will give us two list, with the respective prices and sizes.

The result of this scraping procedure will be a data frame containing all available listings within the city of Hamburg from the ImmobilienScout24 website, associated with the respective neighborhood. To establish a representative metric, we will normalize the prices by dividing them by the respective listing's sizes, yielding a price per square meter for each entry in our data frame. We can them average all normalized prices within the same neighborhoods to get the average square meter price. We will then apply a MinMaxScaler from the sci-kit learn package to these values to create a Price Index, varying from 0 to 1. 1 will represent the neighborhoods with the cheapest rents.

As a limitation from using a single data source, we can run into the situation where no information is obtained for a specific neighborhood. If that is the case, we will estimate the price index for that neighborhood as the average of the three closest neighborhoods with available data. Distances between each neighborhood central coordinate will be obtained using

the Haversine equation (https://en.wikipedia.org/wiki/Haversine_formula). Ideally, we would try to scrape multiple sources to ensure as diverse as possible of a data set, but for this proof of concept we will rely on interpolations to fill in any gaps.

2.2.2 Foursquare Data

A large portion of categories in our recommender system represents the different types of venues in each neighborhood, recovered from the Foursquare database. Access to the database is obtained through a free developer account in Foursquare (https://developer.foursquare.com/) and querying its API (https://api.foursquare.com/v2/). With a total of 10 categories (see section 2.1 Data Sources), we will loop through every neighborhood in Hamburg and request a search query for each specific category id. We will limit our search to a 500 meters radius around the neighborhood's central coordinates and request up to 20 venues in each search. Assuming all search queries, for all neighborhoods, returns the maximum amount of requested venues, we will have the name and category of 20600 venues.

We are not likely, however, to attain this maximum amount of venues for each category for all neighborhoods. This will depend heavily on whether the neighborhood is part of a busier, commercial district, or in a quieter, more residential section of the city. We will rely on this diversity of characteristics to create our index for each Foursquare category. After obtaining all entries available within our search parameters, we will one hot encode this information based on our categories. Summing up all entries for the same category in each neighborhood will form the basis of our category indexes, essentially representing their frequency of occurrence in every neighborhood. Any missing data, i.e. neighborhoods for which the search yielded no results, will be field with a zero. After the frequency of occurrence across every category is obtained, we will normalize the data with a MinMaxScaler again, to obtain values in the range 0 to 1. 1 will represent a neighborhood with ample offers in that category.

Additionally, the Event category will be dropped. It mostly consists of seasonal activities and preliminary analysis indicates only a few number of events within this category in Hamburg. This would like skewer results due to the strongly uneven distribution. Also, a Quiet Index will be computed, also in the 0 to 1 range. One of the given categories is Residences, which represents a limited set of venues and does not properly convey whether a neighborhood is truly residential or not. To remedy this limitation, we will establish our Quiet Index (see below) and drop the Residences category from our content matrix.

Quiet Index = (Residences + Outdoors & Recreation) –
(Professional & Other Places + Shops & Services)

2.2.3 Commute Distance with BingMaps

The coordinate of a client's given workplace address will be obtained from the Nominatim agent in the geopy package. This set of latitude and longitude will be used to compute the commute distance, in seconds, between each neighborhood central coordinate and the given address using the BingMaps API (http://dev.virtualearth.net/REST/v1/Routes/Driving). The set of distances will then be normalized using a MinMaxScaler to create a Distance Index varying from 0 to 1, in which 1 represents the closest neighborhood. The use of BingMaps limits our commute distance to reflect only the personal driving scenario. The use of public transport to compute this distance would be beneficial but unavailable within their current API.

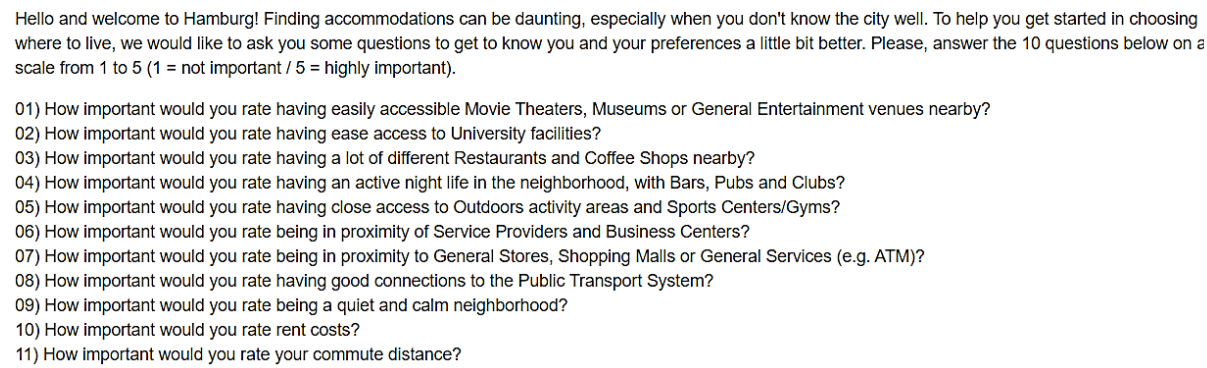
2.3 Final Deliverables

Our goal is to develop our Content Matrix. This matrix will have a set of meta data information, relating to the names of the available neighborhoods in Hamburg along with their central latitude and longitude. This data will be used primarily for visualization. Aside from this meta data, the Content Matrix will have 11 additional columns, indexed by neighborhood. The weights for all categories, aside from the Distance Index, will be pre-computed. The Distance Index, however, will have to be calculated in real-time due to the specificity to each client request. The defined categories are:

- Arts & Entertainment
- College & University
- Food
- Nightlife Spot
- Outdoors & Recreation
- Professional & Other Places
- Shop & Service
- Travel & Transport
- Quiet Index
- Price Index
- Distance Index

To complement our Content Matrix and create the neighborhood recommendations for each client, we will also need a set of client preferences to be applied for each category above. These could be in the form of a 1 to 5 scale, in which a 5 represents a category that is highly important to the client when choosing a neighborhood to live in. These values would be given

directly by the client, based on a previously established questionnaire. Below, we have a suggestion for such a questionnaire (fig. 1).


Hello and welcome to Hamburg! Finding accommodations can be daunting, especially when you don't know the city well. To help you get started in choosing where to live, we would like to ask you some questions to get to know you and your preferences a little bit better. Please, answer the 10 questions below on a scale from 1 to 5 (1 = not important / 5 = highly important).

01) How important would you rate having easily accessible Movie Theaters, Museums or General Entertainment venues nearby?
02) How important would you rate having ease access to University facilities?
03) How important would you rate having a lot of different Restaurants and Coffee Shops nearby?
04) How important would you rate having an active night life in the neighborhood, with Bars, Pubs and Clubs?
05) How important would you rate having close access to Outdoors activity areas and Sports Centers/Gyms?
06) How important would you rate being in proximity of Service Providers and Business Centers?
07) How important would you rate being in proximity to General Stores, Shopping Malls or General Services (e.g. ATM)?
08) How important would you rate having good connections to the Public Transport System?
09) How important would you rate being a quiet and calm neighborhood?
10) How important would you rate rent costs?
11) How important would you rate your commute distance?

Figure 1 – Example questionnaire to gather the user profile for our recommender system.

Our final recommendations would then be obtained by multiplying our Content Matrix with our User Profile and summing up the values across all categories for each neighborhood. This would yield an average neighborhood score. We could then provide a list of the top 5 or 10 recommendations to each client, along with a map of the city of Hamburg with these top choices highlighted. Additionally, a listing of all available apartments from our database for each of those top neighborhoods could be provided to jumpstart their search. These set of neighborhoods recommendation, visual map, and list of available rental properties would be our final deliverable.

3. EXPLORATORY DATA ANALYSIS

3.1 Rent Prices

We were able to scrape a total of 1009 listings from the ImmobilienScout24 website, for a total of 84 out of 103 neighborhoods. Amongst the scraped apartments, rental prices ranged from 205 €/month to 8500 €/month, with an average of 1279.71 €/month. Apartment sizes were between 14 m² for a single room location to 442 m², with an average apartment size of 83.3 m² (fig. 2, left). Based on these data, the normalized squared meter price per neighborhood in Hamburg was between 8 and 24 €/m², with an average 14 €/m². For 43 out of the 84 neighborhoods, these averages were composed based on a minimum of 10 listings, with up to 47 properties, while the rest had an average of four listings (fig. 2, right).

Overall, prices are higher in the central and western regions of Hamburg, with the two most expensive neighborhoods being Sternschanze and Hafen City (fig. 3). Within the city center, cheapest neighborhoods to live in were St. Pauli and Borgfelde, with prices around the general average of 14 €/month. In the west, Sulldorf and Finkenwerder had comparable cost to central

neighborhoods, and most other neighborhoods were within the average cost band. Prices in northern Hamburg were mostly in the general average, with neighborhoods such as Wellingsbüttel and Volksdorf just slightly above the 14 €/month mark. South and southeast Hamburg were the cheapest regions to live in. The exception being Harburg and Ochsenwerder, which were in the price range of Wellingsbüttel and Volksdorf.

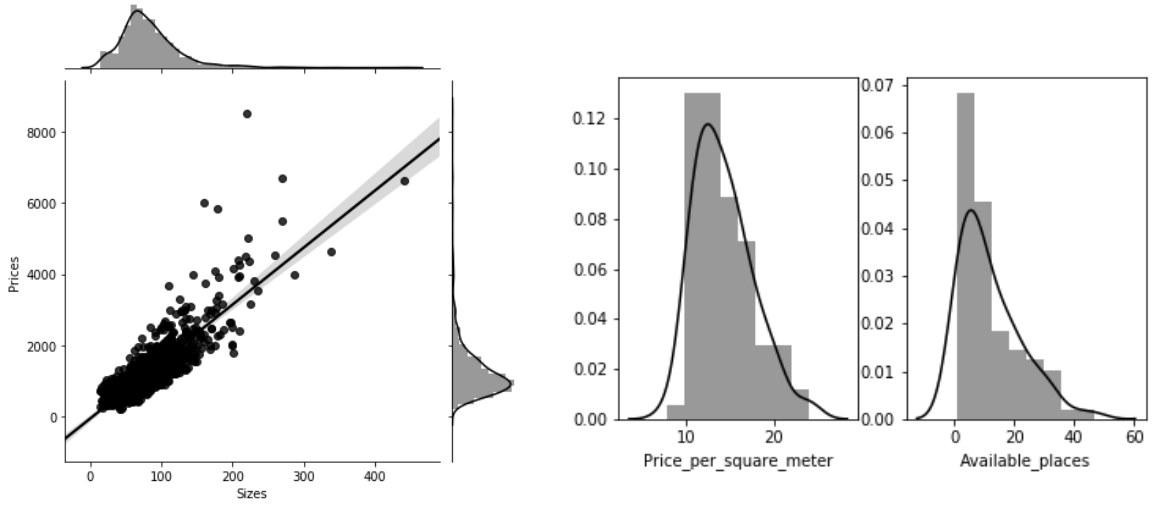


Figure 2 – On the left, scatterplot of all the prices and sizes for every listing scraped from ImmobilienScout24. The top and right figures show the kernel density estimation plot for each axis. On the right, histograms with the density distribution of the normalized price per square meter and the number of available listing used to compose these averages.

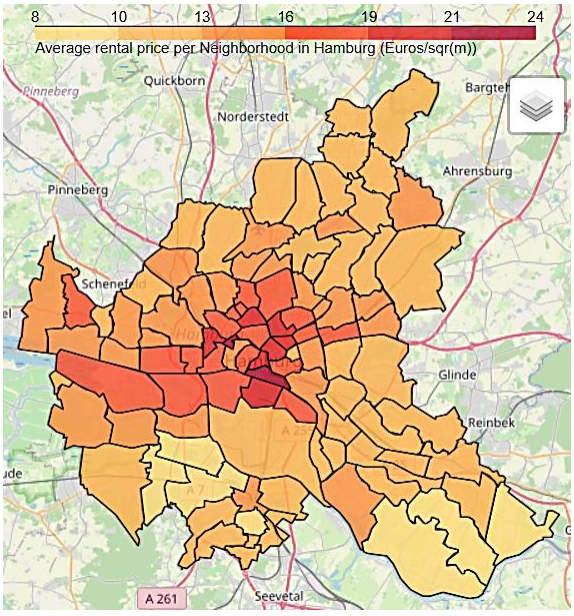


Figure 3 – Average price per square meter across all neighborhoods from Hamburg.

3.2 Foursquare Venues Data

We found a total of 6136 venues within our search parameters, of which only 7 were in the Event category (fig. 4). This reinforces our decision to remove the Event category from consideration when building our content matrix. Both the “Professional & Other Places” and the “Shop & Service” categories had the highest amount of returned venues, exceeding 1000, followed by “Food” with 919 and “Outdoors & Recreation” with 779. Aside from the Event category, all other had above 350 venues.

As for the distribution of returned venues for each neighborhood, the highest concentration was once again found in the city center, with over 140 venues per neighborhood (fig. 4). Harburg, on the south, was the only other neighborhood outside of central Hamburg to register a similar count. 6 neighborhoods, on the other hand, had no results from the Foursquare database. They are: Altengamme, Neuengamme, Kirchwerder and Reitbrook in the SE; Wohlsdorf-Ohlstedt in the north; and Francop in the southwest. For these neighborhoods, all values were set to zero.

Surprisingly, despite the high concentration of venues in the central areas of Hamburg, the Quiet Index classified most central neighborhoods to be on the average of our scale. This means that there is an overall balance between the amount of shop and businesses, and the amount of residences, parks and recreation spots. However, there is an overall higher number of busier/noisier neighborhoods (31) than truly quiet ones (13) in Hamburg.

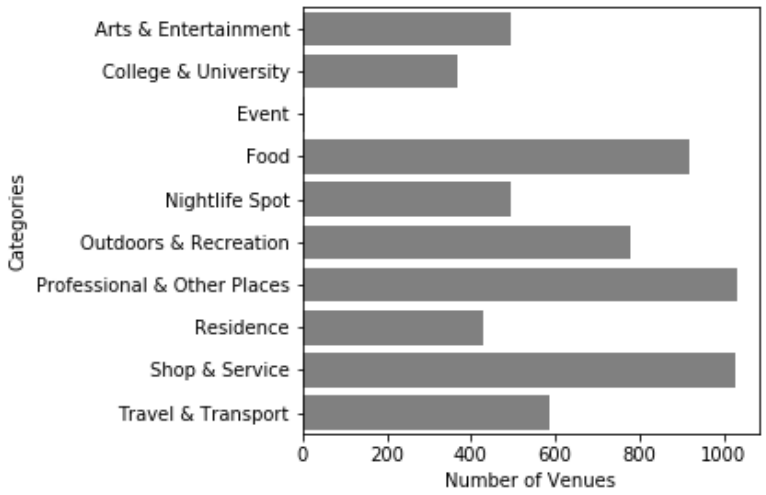


Figure 4 – Number of venues in each category obtained from our Foursquare search.

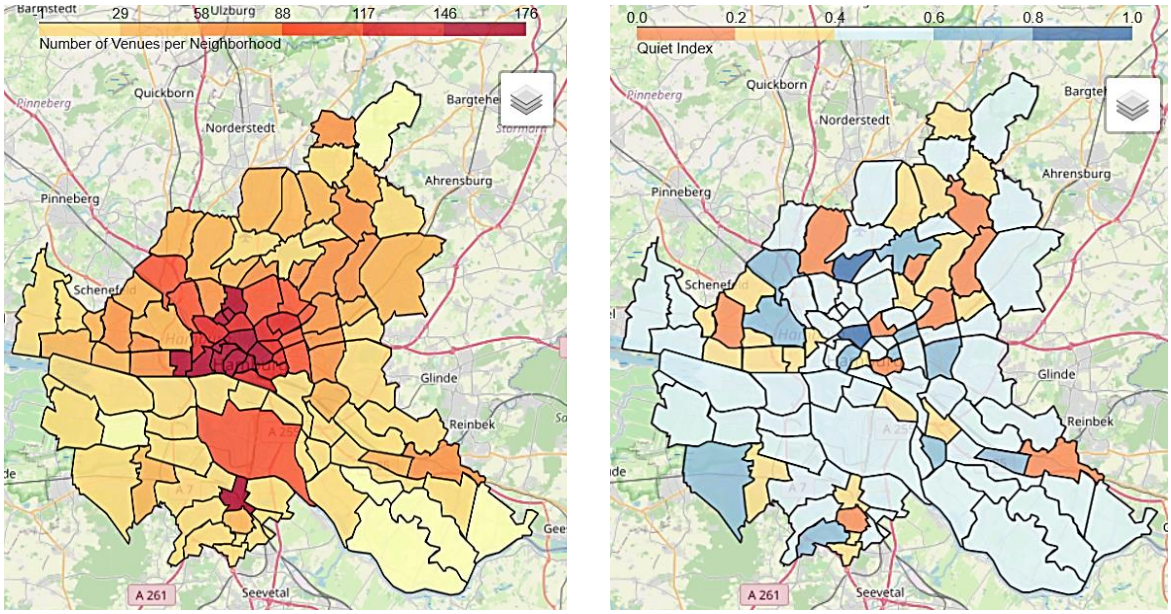


Figure 5 – On the left, total number of venues per neighborhood. On the right, the calculated Quiet Index.

3.3 Best Neighborhoods in Hamburg

In the end, our content matrix is composed of 11 categories. 8 of those are a direct reflection of number of venues in each neighborhood, found through our Foursquare search (see first 8 bullet points in section 2.3). Since the Foursquare data showed a clear bias towards the central neighborhoods due to the higher number of venues found (figure 5, left), this cascades down to affect our recommender system as all categories are given the same weight by default. The Quiet Index break this dominant pattern slightly, combined with the Price and Distance indices. However, mapping all neighborhoods in Hamburg based on our content matrix (fig. 6) still show a similar pattern as obtained from the mapping of the number of venues per neighborhood. According to this, the three best neighborhoods in Hamburg are Sternschanze, Hafen City and Harvestehude. Harburg is the first neighborhood outside the central cluster, scored at the 23rd position with an average score of 0.537.

Breaking down the top neighborhoods across each category individually (table 1), the central neighborhoods also dominate the top positions across all categories except for the Price Index. The cheapest neighborhoods to live, in this case, are Neuengamme and Altengamme in southeast Hamburg. Harburg, outside central Hamburg, was within the top five of 3 major categories (“Arts & Entertainment”, “Food” and “Outdoors & Recreation”), indicating that it might be a good option for people looking to live outside of central Hamburg but still have ample offers in terms of daily amenities, while having rental prices just slightly above average.

	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport	Quiet Index	Price Index
1	St. Pauli	Rotherbaum	Hamburg-Altstadt	Ottensen	HafenCity	Ottensen	Sternschanze	St. Georg	Groß Borstel	Neuengamme
2	HafenCity	Neustadt	Rotherbaum	Hammerbrook	Neustadt	Eppendorf	St. Georg	Hammerbrook	Rotherbaum	Altengamme
3	Altona-Altstadt	Sternschanze	Harburg	St. Georg	Hamburg-Altstadt	Neustadt	St. Pauli	Sternschanze	Bahrenfeld	Hausbruch
4	Harburg	St. Georg	HafenCity	Altona-Nord	Barmbek-Süd	Altona-Altstadt	Hoheluft-Ost	HafenCity	Neuallermöhe	Heimfeld
5	St. Georg	Hamburg-Altstadt	Dulsberg	Altona-Altstadt	Harburg	Harvestehude	Uhlenhorst	Neustadt	Ohlsdorf	Wilstorf

Table 1 – Top 5 neighborhoods in Hamburh for each category.

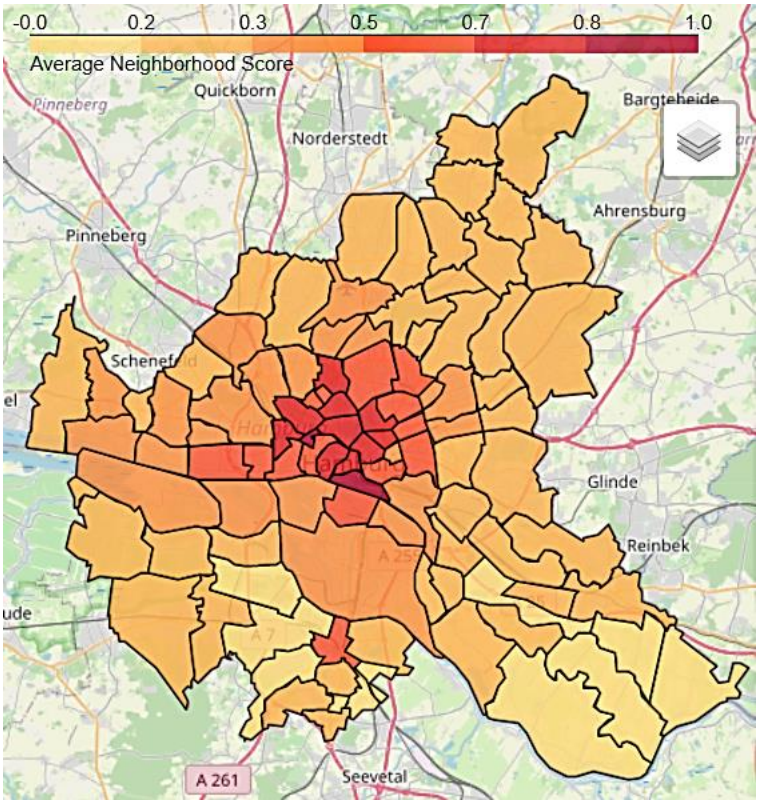


Figure 6 – Neighborhood score across Hamburg, defined based on the default state of our content matrix.

4. Our Recommender System

4.1 Parameter Boosting

As seen in section 3.3, our system has an overall bias towards the central neighborhoods due to the higher number of parameters dependent on the concentration of venues obtained through Foursquare. This implies that, by default, parameters such as Price Index or Distance Index will have only a 1/11 impact in the choice of neighborhoods to recommend. Considering that the cheapest regions are also found where the least amount of venues were gathered, this would further minimize the impact of rental price under default assumptions (see fig. 6). To ensure that these parameters had a stronger effect in the chosen neighborhoods, a simple boosting is applied. For both the Price and Distance indices, the user defined importance obtained through the questionnaire is multiplied by 5 before the user profile is applied to the content matrix. This ensures that those non-venue dependent parameters actively influence the

neighborhood choice but hopefully do not overpower the overall performance of the algorithm. This can be further optimized to ensure performance meets client's expectations.

4.2 Final Deliverables

The final deliverable from our recommender system is a set of neighborhood recommendations tailored towards each client's profile. These takes the form of a list, that could include the top 5 or 10 recommendations, and a map of the different neighborhoods in Hamburg, with all neighborhoods with a user score above 0.8 highlighted. Both the number of recommendations and the mapping threshold can be optimized if necessary. Along these two products, we can also provide a list of all available rental spaces from our rental dataset associated with each recommended neighborhood, along with the average rental's price and size. This last deliverable could be tailored to reflect only the offers of specific real state agencies if multiple databases are used to build the rental dataset. In our case, however, they would reflect only what is currently offered in ImmobilienScout24. To demonstrate the performance of our system, we will analyze the profile of three hypothetical clients. Their answers, on a scale of 1 to 5, for the questions shown in figure 1 can be found in table 2, below.

- Client_1 is a young woman, mid-20s, good job in a tech startup. Very social, enjoys the night club scene and loves going to theaters and watching movies. Self-described foodie. Would like a short commute, but it is not an absolute priority. Works in northeast Hamburg.
- Client_2 is a family of three. Both parents have good jobs and their son is just starting college in the city. Value the outdoors and would like to be close to good entertainment possibilities. Hates losing time in traffic, so would prefer a short commute. They both work in west Hamburg.
- Client_3 is retired. Values a quiet neighborhood and would like to be close to parks or the river.

	Client_1	Client_2	Client_3
Question 1: Arts & Entertainment	4	4	3
Question 2: College & University	1	5	1
Question 3: Food	5	2	1
Question 4: Nightlife Spot	5	1	1
Question 5: Outdoors & Recreation	2	5	5

Question 6: Professional & Other Places	1	1	1
Question 7: Shop & Service	2	2	3
Question 8: Travel & Transport	4	1	3
Question 9: Quiet Index	4	3	5
Question 10: Price Index	3	2	5
Question 11: Distance Index	3	5	1

Table 2 – Client importance rating for each category, on a 1 to 5 scale.

4.2.1 Client_1

Based on the profile of Client_1, the top 5 neighborhood recommendations are Borgfelde, St. Pauli, Harburg, Dulsberg and Hammerbrook (fig. 7). St. Pauli, Borgfelde and Hammerbrook are all central locations with a plethora of food and nightlife spots. Borgfelde is also a relatively calm neighborhood despite its central location. In comparison, Borgfelde, St. Pauli and Hammerbrook have 5, 4 and 11 available listings, with minimum rents of at least 480 €/month (St. Pauli). Dulsberg combines good offer of venues and easy access to the city center, while offering a closer commute. It also has relatively cheaper rents when compared to the previous three other recommendations, with a minimum cost of 477 €/month and 4 available listings. Finally, Harburg offers the cheapest rents among all recommendations with a diverse set of venues in the region, with the trade-off of a longer commute when compared to all other neighborhoods. It also has the highest number of available listings, with 32 apartments in offer with prices ranging between 380 to 2330 €/month.

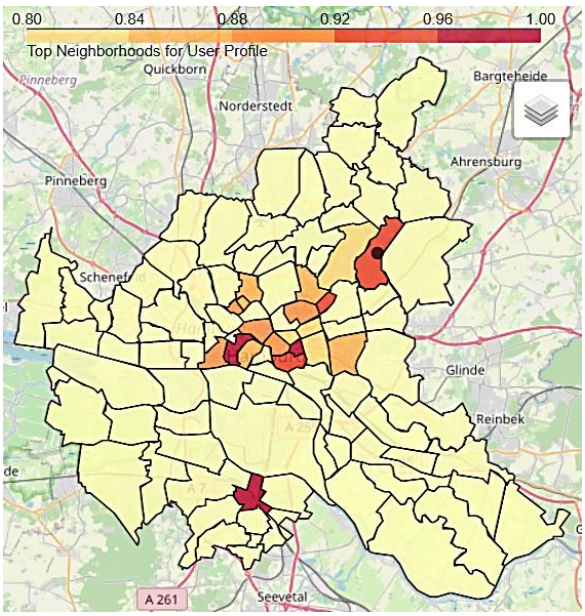


Figure 7 – Neighborhoods with a score higher than 0.8 for Client_1. The circle dot represents the workplace address informed by the client.

4.2.2 Client_2

Based on the profile of Client_2, the top 5 neighborhood recommendations are Altona-Altstadt, St. Pauli, Ottensen, Harburg and Neustadt (fig. 8). Since this client had a strong priority towards shorter commutes, we can see that the recommendations distribution is shifted westward, towards the workplace address indicated. Nevertheless, the central neighborhood bias is still significant, and all apart from Harburg are central. Harburg is still the cheapest neighborhood offered with a minimum rental price of 380 €/month, although St. Pauli is just 100 € more expensive at a minimum of 480 €/month with a significantly shorter commute. Altona-Altstadt's offers starts at 630 €/month and has 8 listings, Ottensen starts at 700 €/month with 4 listings and Neustadt is potentially the more expensive at a minimum of 890 €/month across 11 offers. Neustadt was also amongst the top 5 locations for the "College & University" category, which fits with the requested profile, as well as having ample offer of outdoors activities.

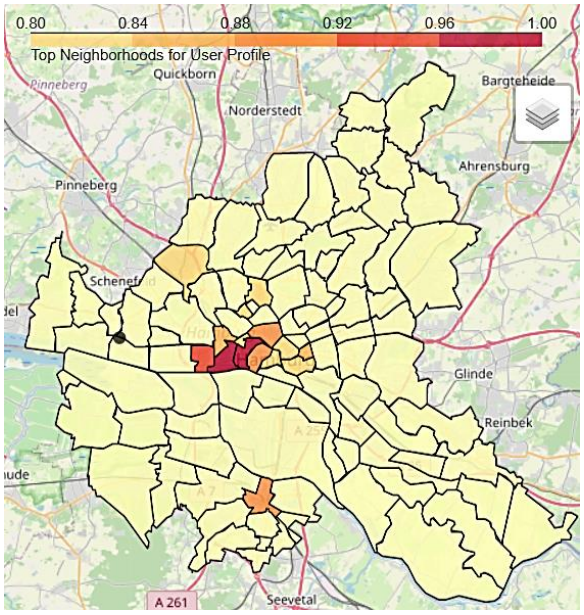


Figure 8 – Neighborhoods with a score higher than 0.8 for Client_2. The circle dot represents the workplace address informed by the client.

4.2.3 Client_3

Based on the profile of Client_3, the top 5 neighborhood recommendations are Borgfelde, St. Pauli, Harburg, Horn and Altona-Altstadt (fig. 8). Aside from Harburg, all these neighborhoods are considered quiet or on the average of our Quiet Index, which was one of the

absolute preferences of this client. Harburg, although slightly noisier/busier, ranked among the top five for the “Outdoors & Activities” and “Arts & Entertainment”, both categories were the client indicated stronger preferences. It is also the cheapest neighborhood at a minimum of 380 €/month, just slightly below Horn (422 €/month) and St. Pauli (480 €/month).

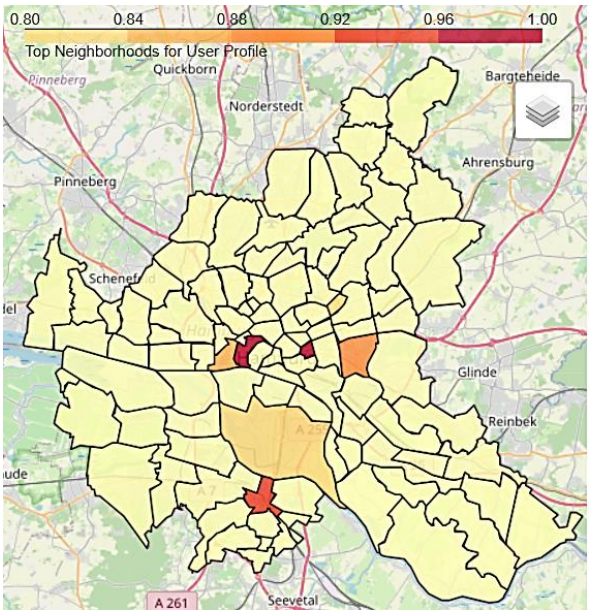


Figure 8 – Neighborhoods with a score higher than 0.8 for Client_2. The circle dot represents the workplace address informed by the client.

5. CONCLUSION AND RECOMMENDATIONS

Despite the limited amount of data sets utilized in building our recommender system, it is able to account for a set of client’s preferences when building a personalized recommended list of neighborhoods in the city of Hamburg. Our system recognizes 11 categories, accounting for different types of venues between commerce, education, services, and entertainment. Aside from venues, it also considers different rental prices across the city, distances between each neighborhood and a target (“workplace”) address and tries to reflect the level of noise in different regions. As deliverables, aside from a ranked list of neighborhoods, it provides a map of the city with the recommended locations highlighted that is fully interactable in a webpage. It can also be combined with a data set of rental properties available across the city, to give the prospective client a list of available homes in the neighborhoods chosen.

The current version is merely a proof of concept. The reliance on constructing our venues categories based on the returned number of venues introduces a clear bias towards central Hamburg. To counteract this limitation, we could gather rating scores for each venue alongside its name and category. The score of each neighborhood would then be calculated based on the

average score across each category, as opposed to simply number of venues. This is possible with the Foursquare API, although it relies on premium calls to the service which are extremely limited on personal accounts. This procedure should give a more unbiased approach to building the categories, as well as better curating the categories ids to reflect only specific types of business/services. For example, “Schools” should be with the “College & University” into a single “Education” category. Using additional services, such as Google’s API, could also improve the quality of our venue-based categories. For rental prices, the reliance on a single website is a bottleneck. Scraping additional websites would be beneficial. This would also be the easiest aspect of the system to tailor towards specific companies. Finally, BingMaps API allows only for the calculation of the drive distance between two address/locations. Utiliing an API that considers public transport and setting both as return options for the service would expand its capacities further. These modifications, however, can be easily adapted into the current coding of our recommender system.