Introduction

Background

The experience of moving home is often a stressful one. Whether you are relocating to a new city or just a new neighbourhood, there is plenty to manage and organise. Arguably the most draining activity of all is the search for your new dream home. I have experienced this myself more than once, most recently when I have decided to relocate from the small, peaceful town of Basel, Switzerland to busy, bustling, gigantic London. Where does one start to look for a home in a place like this?

Problem

There are the obvious constraints of the workplace commute time and price range, but that still leaves more options than one can ever hope to find the time to go through. Despite having a fairly clear idea of what I was looking for, not only in my ideal home but also in the area I wanted to live in, I could only check how well each property matched by going over them one by one. The number of choices was overwhelming and exhausting, all the while I could not shake the feeling that I was missing out on the perfect property.

Proposed approach

What if we could make this process easier by narrowing down our search and only focusing on a subset of neighbourhoods that fulfil criteria of our choice beyond location?

My project will use London as a proof of concept for how this could work: use venue information to cluster similar neighbourhoods, then apply additional filters, such as distance from a given location and other random desired features (e.g. green spaces, restaurants, ethnic shops), in order to narrow down the selection to only the few most suitable areas that fulfil these criteria. This tool can benefit anyone looking to move to a new house, making the process of finding a new home easier and more enjoyable.

Data

To identify the required data, some preparatory research is needed. We first need to identify the optimal level of territorial subdivision within the city of London. A quick search reveals that the borough compartmentalisation corresponding to London postcodes is at just the right level of granularity for our scope. The list of boroughs is easily available for scraping on multiple websites. Next, we will need to retrieve the coordinates of each district using a geolocation tool before we can employ the Foursquare API to complete our dataset with venue information for each district. This will provide us with all the necessary information in order to cluster similar neighbourhoods based on their venues, then apply additional filters to identify the most suitable areas according to our criteria.
On a side note, London is a dangerous city. Information on crime rates in each borough is available on several websites. We can use this information to assign a safety score to each neighbourhood and include this feature in either the clustering criteria or post-clustering, when filtering for most suitable areas to live in.

Methodology

Data collection

In the first step, the desired data was retrieved by scraping the borough list and corresponding crime numbers from *'https://www.finder.com/uk/london-crime-statistics'* into a data frame. This was an already clean and easy to use dataset. The crime numbers for each borough were transformed to express the crime rate as a percentage of total. Additionally, the latitude and longitude of each borough were retrieved using Python's *Nominatim* geocoder. Lastly, based on the coordinates, a distance metric was computed between each borough and a target location. This data frame constitutes the backbone of all subsequent analysis.

Exploratory analysis

Next, venue information was retrieved for each borough using the Foursquare API. At this step, exploratory analysis was performed to compare the different boroughs in terms of their total number of venues (Fig 2), as well as venue diversity (i.e. the number of unique venue categories in each borough, Fig 3).

K-means clustering

The venue data was grouped and transformed in order to obtain a dataset suitable for clustering the boroughs based on their venue information and crime rate. These transformations included one-hot encoding the venue categories retrieved from Foursquare and calculating the frequency at which each venue category occurs in each of the boroughs. The k-means clustering algorithm was selected to allow the grouping of boroughs sharing similar venue features, in order to aid the process of identifying suitable boroughs. The algorithm was initially run multiple times for a range of cluster numbers (*k*) to allow identifying and settling on an optimal number. The within-cluster sum of squares (*wcss*) was plotted for each *k* value and the optimal value was identified as the elbow point of the curve.

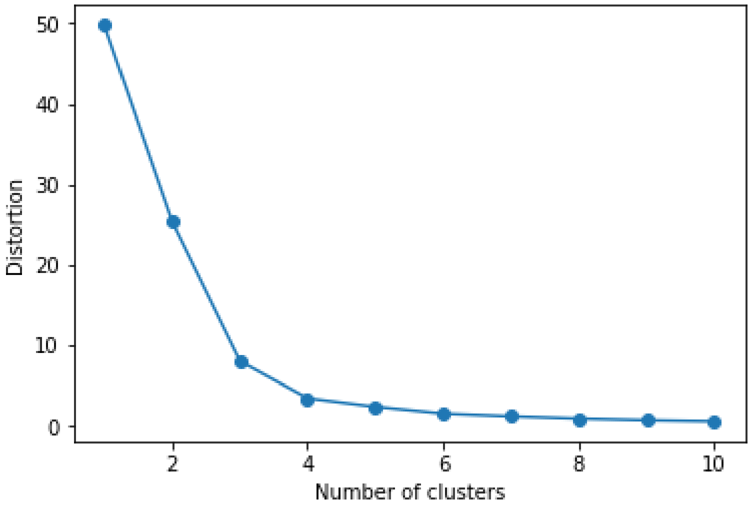

Figure 1. Relationship between the number of clusters and the wcss (distortion). The curve shows a clear elbow point corresponding to an optimal number of clusters equal to 4.

Additional analysis

Once the clusters were obtained, the cluster labels were then used to further explore the defining features of each neighbourhood. For each cluster, a word cloud was generated to help visualise these features. This provided a an easy-to-parse overview of each cluster and allowed identifying the cluster with the most overall appeal, narrowing down the options. Additionally, it was assessed how well the crime rate discriminates the clusters and which clusters are preferable based on this parameter.

To further filter the remaining boroughs, a set of keywords was chosen using specific requirements that the boroughs had to meet, i.e. specific desired venues. This filter set contained the following venues: *park, lake, gym, organic grocery, bakery, market, garden*. The number of occurrences of each keyword among each borough's venues was counted, generating a match score.

This approach provided a double selection method, whereby the first pass (clustering) provides a broader overview which helps to sub-select a group of boroughs based on their overall features, while the second pass uses a few specific criteria to further focus the search onto only the one or two highest matching boroughs.

Results

Crime rate varies across boroughs

In the first instance, we observed that the distribution of crime rates among boroughs is not uniform. This indicates that crime rate is a relevant parameter that could potentially add valuable information as a clustering feature and increase discriminability among clusters.

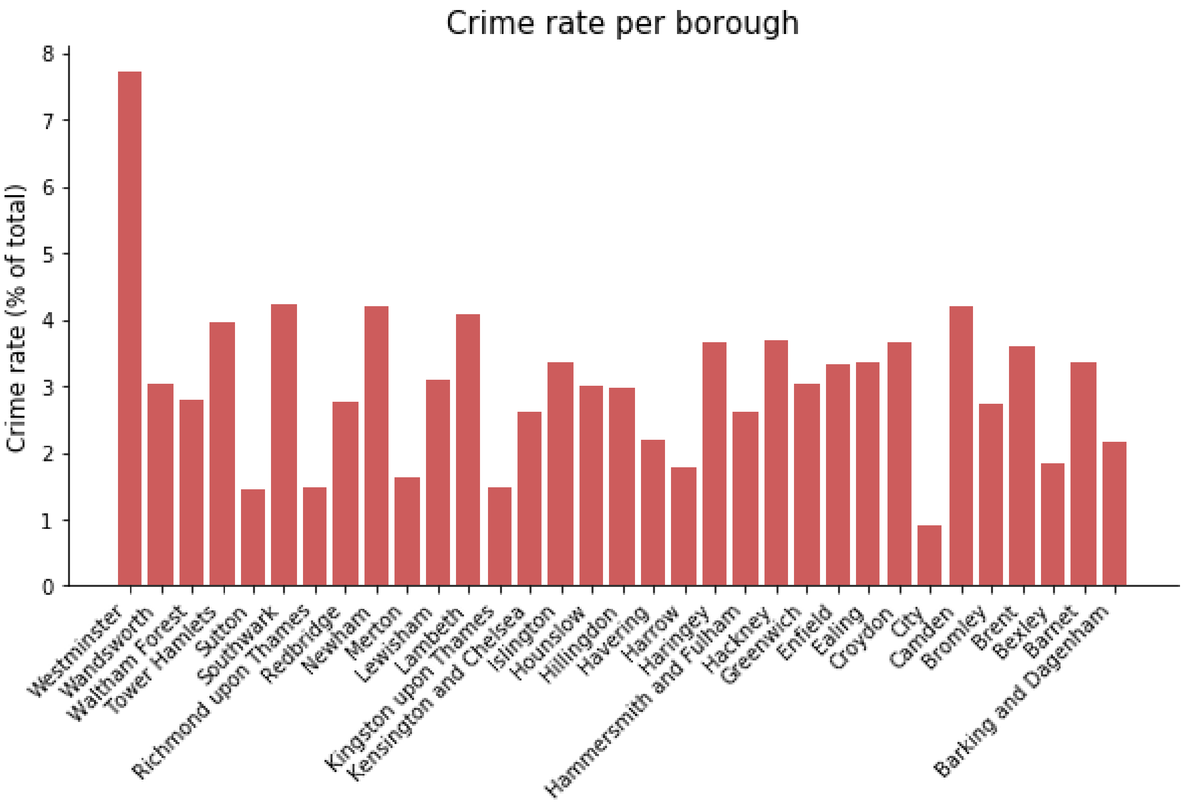

Figure 2. The distribution of crime rates across the 32 London boroughs.

Venue numbers and diversity vary across boroughs

Next, by looking more closely at the venue information retrieved for each borough, we find that both the numbers of venues and the number of distinct venue categories vary across boroughs. This indicates considerable differences among boroughs and supports testing the clustering approach.

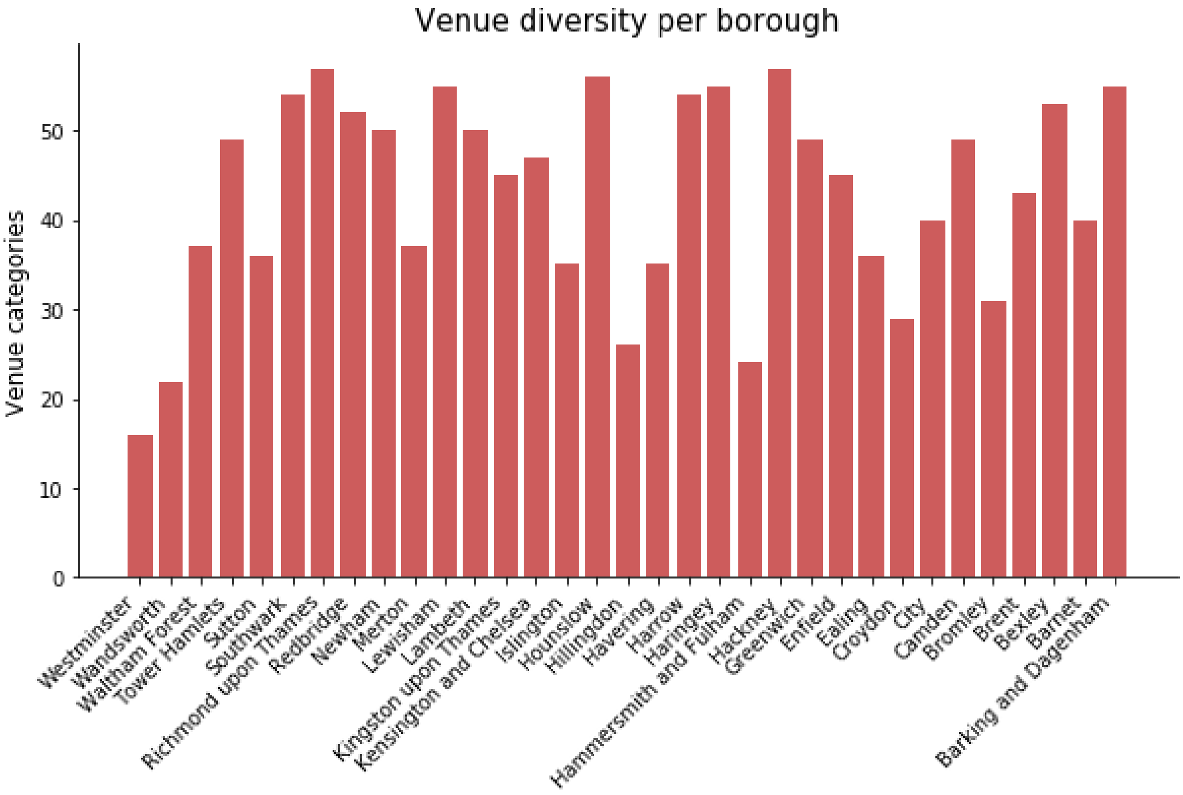

Figure 3. The number of unique venue categories retrieved by Foursquare for each of the 32 boroughs

Clustering generates distinct groups of boroughs with clear characteristic features

By visualising the most frequent features of each of the four obtained clusters, we can clearly observe their distinct properties and generate a description for each of them.

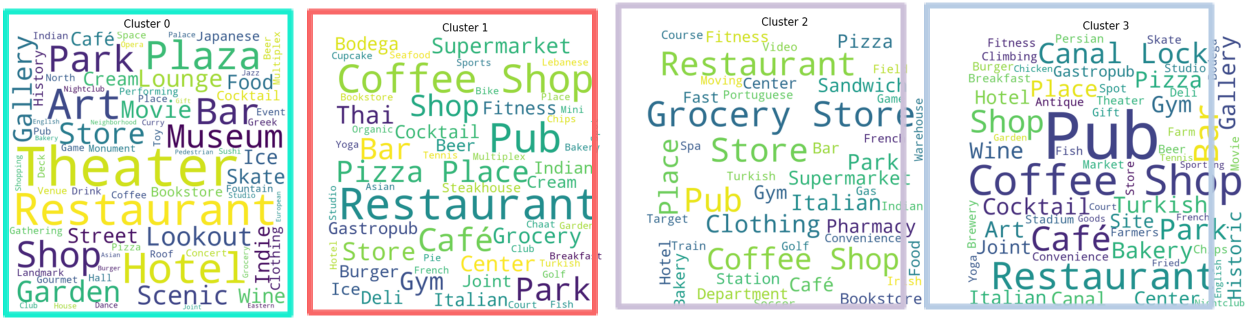

Figure 4. Word cloud representation of the characteristic venues of each of the four clusters generated by the k-means algorithm. Larger font sizes indicate a higher frequency of occurrence. The outline colours will be associated with the cluster number from here on.

Thus, based on its most prominent venues, cluster 0 indicates a lively area suitable for going out and with a high density of cultural venues and sightseeing attractions that would make it popular with tourists.
Cluster 1 stands out as being very food-centred, featuring a lot of eating out venues of many varieties, as well as pubs and bars. This suggests potentially popular areas for locals to meet and go out away from the tourist crowds.
Cluster 2 indicates slightly more residential areas, due to its seemingly lower diversity of venues, as well as its most common features being grocery shopping venues, cafes, supermarkets, gyms, convenience stores and some eating out venues.
Cluster 3 features a combination of recreational venues, culture, diverse eating out as well as other socialising locations such as cafes and pubs, suggesting perhaps up-and-coming, “hipster” areas.
The cluster information provides a clear separation of features, allowing us a first pass at narrowing down the search by identifying the more suitable areas according to our preferences. In this case, we selected cluster 3 as the best match for our requirements.

Additional keyword-based filtering provides further constraints
Once the cluster best reflecting our desired features has been identified, we can further narrow down the search. By employing an additional layer of specific keyword search among each borough’s venues we can restrict the search to as little as 1-2 most suitable areas with the highest match score (see Methodology).

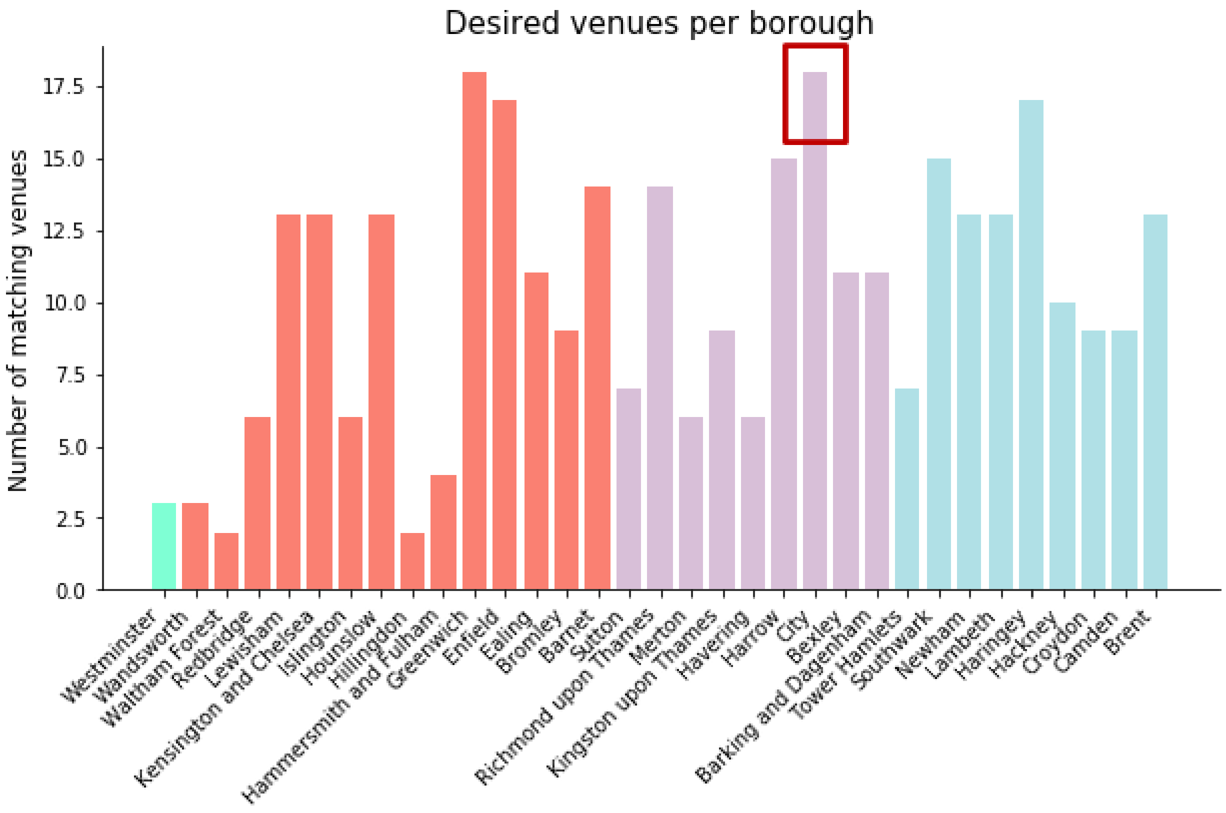

Figure 5. Distribution of match scores across the different boroughs, colour-coded according to the cluster they belong to as in the previous figure. The red rectangle highlights the best-matching borough.

Crime rate is an important factor and the relationship between it and the match score is of interest. Figure 6 presents the relationship between crime rate and match score for each borough. The boroughs are also color-coded according to the cluster they belong to, allowing the visualisation of this multi-feature interaction between crime rate, match score and cluster identity that would further inform the selection process. Interestingly, there is a strong and clear relationship between crime rate and cluster identity, less so than between cluster identity and match score.

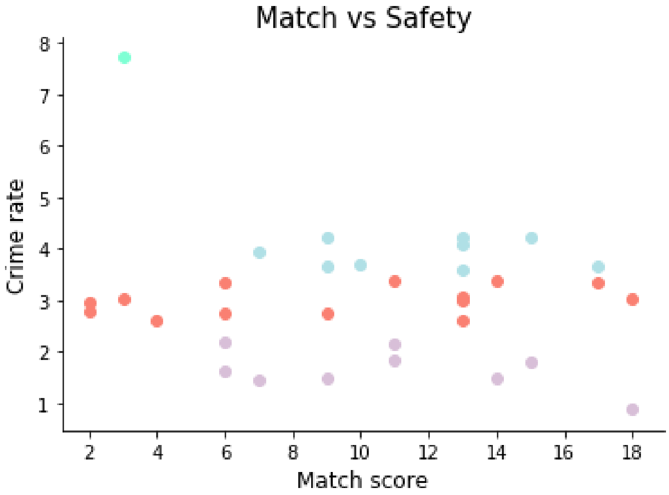

Figure 6. Relationship between clusters' (the colour groups) and individual boroughs' (points) match score (x axis) and crime rate (y axis)

The two-step process allows identifying the topmost suitable area

Figure 7 exemplifies the final output of this approach. Each borough is displayed on the London map with a marker size corresponding to the frequency of the desired keywords (i.e. its suitability based on our desired criteria), color-coded according to the cluster it belongs to. Additionally, a 10 km radius around an example target location is displayed on the map, to helps visualise which boroughs fall within a 40-45-minute commute distance, another potential criterion.

This results indicates that the current approach allows identifying the best suited area for home search according to our specific criteria, greatly speeding up and simplifying the search process.

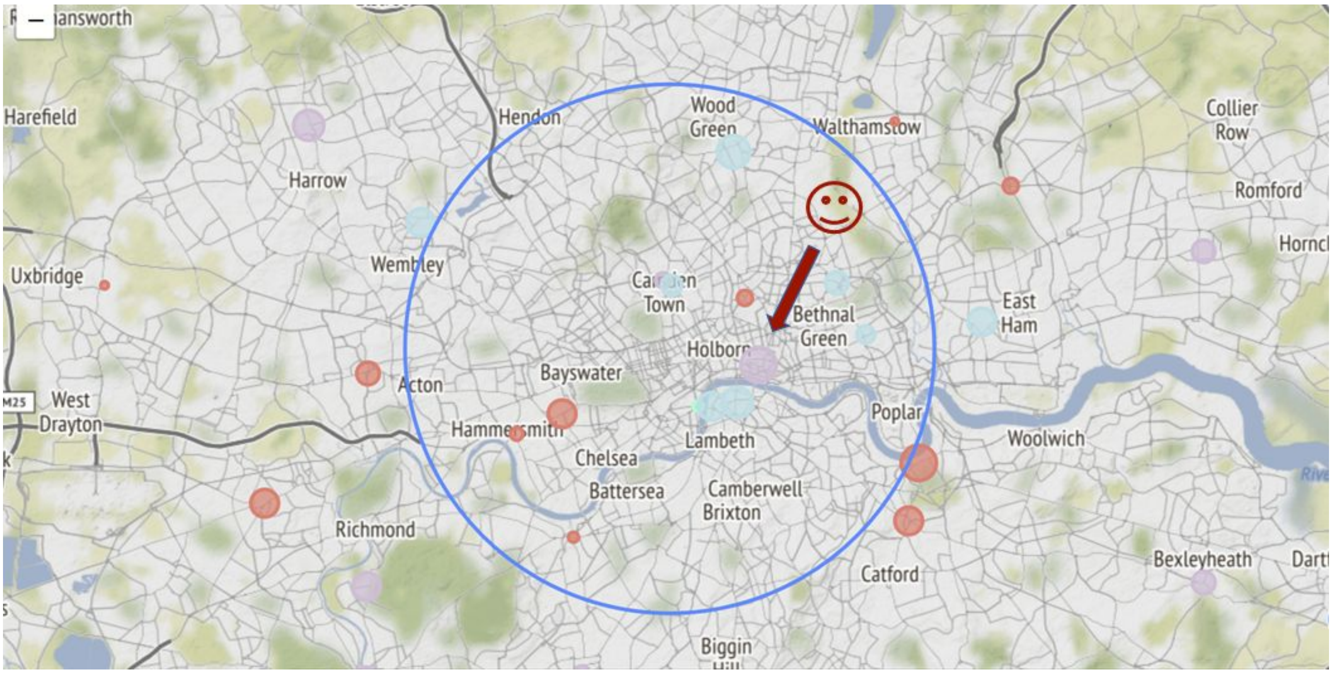

Figure 7. Map of London displaying part of the 32 boroughs. The borough markers are colour-coded by the cluster they've been assigned their diameter corresponds to the borough's match score. The large blue circle marks a 10 km radius around a central location chosen randomly. The red arrow indicates the topmost suitable borough that has been identified.

Discussion

We asked whether it was possible to make the home hunting process easier by narrowing down the search to only focus on a subset of neighbourhoods that fulfil criteria of our choice beyond location. The results outlined above provide convincing evidence supporting the successful implementation of this approach to address the problem outlined in the introduction.

Nevertheless, the current approach presents several limitations that should be addressed in order to improve the output of the algorithm. Subject to requirements, several parameters can be modified in order to achieve a finer and more accurate result.

A finer regional division should allow us to separate “hotspots” that very closely match our criteria from the larger boroughs.

The current borough coordinates retrieved using *Nominatim* do not necessarily match the geographical centre of the borough, which could significantly influence venue information. This also means that the search radius can partly overlap with that of other boroughs, leading to decreased discriminability/higher overlap, which could have an impact on the output of k-means clustering. In the present case, the venue search radius was selected based on the borough with the smallest surface. Adjusting the search radius for individual boroughs' surface could further improve accuracy and discriminability. The Foursquare limit of 100 venues per borough might further influence the results.

No other machine learning approaches have been tested against the current one, leaving open the very likely possibility that the algorithm could be improved or that more optimal results can be achieved differently.

Conclusion

While the method can unarguably be improved and refines, this project provides a proof of concept for a simple yet powerful approach to solving the common and far-reaching problem: the overwhelming number of choices getting in the way of finding our dream home.