IBM Data Science

Applied Data Science Capstone

Optimum relocation

Xavier Martínez

January 2020

TABLE OF CONTENT

1. INTRODUCTION

1.1 BACKGROUND

Several people nowadays relocate for different reasons. While undertaking a quick research to try to put a figure to the scale of this phenomenon, I found on a piece of information that, even it only relates to the USA, is quite indicative:

"Each year, roughly 40 million Americans, or about 14 percent of the U.S. population, move at least once". From the article *Population migration patterns: US cities we are flocking to* by Michael B. Sauter, published Oct 4, 2018.

The truth is that in many occasions the person relocating may not know the city in which he will be moving into; for example, because the relocation relates to the needs of the company where this person works for.

It could be the case that a person wants or needs to move, but that the neighbourhood or even the city, where this person would relocate is not determined.

The circumstances for relocating may be very diverse, but an element that probably is common in any case is that every individual relocating will have an idea of what kind of place it should be.

An easy way to describe the neighbourhood where one would like to relocate is by referring to a known one. This known neighbourhood could be the one where you live or the one where you like to live.

1.2 OBJECTIVE

The objective of this project is to develop a tool that would allow identifying the area or neighbourhood in a city or a group of cities that best matches the requirements of an individual that will be relocating.

The requirements would be expressed as an area or neighbourhood that the person relocating would indicate.

The criteria to be considered in comparing areas would include the presence of a wide range of venues such as social, cultural, entertainment, schools or natural environment; and other factors such as quality of life, cost of living, safety, health care, climate and pollution.

2. DATA

As indicated in the previous section, the criteria in comparing areas would include the presence of a wide range of venues as well as other factors such as quality of life, cost of living, safety, health care, climate and pollution. Therefore data related to all these factors will be required.

Also, along the process of finding venues, the conversion from an address to latitude and longitude will be needed.

And to present the locations considered, mapping them will be the best option. Therefore, map information will be required as well.

2.1 DATA SOURCES

2.1.1 Foursquare data

The data related to venues will be obtained from Foursquare (https://foursquare.com/), using their API. The data obtained will be for all, the neighbourhood that would represent the requirements of the person who relocates, and all the areas that will be compared to the former.

2.1.2 Numbeo data

The data related to the quality of life, cost of living, safety, health care, climate and pollution, will be obtained from Numbeo (https://www.numbeo.com/), using their API. Again, the data obtained will be for the neighbourhood that represents the requirements of the person who relocates and all the areas that will be compared to the former.

It needs to be noted here that the granularity of the information from Numbeo is different from the granularity provided by Foursquare. Numbeo provides information representative of the average of the city.

2.1.3 Google maps data

The data related to the transformation from address to geographical coordinates will be gathered from Google Maps API geocoding:

(https://developers.google.com/maps/documentation/geocoding/start?hl=en).

2.1.4 Folium library

Folium (https://python-visualization.github.io/folium/) will be the python library used in generating the maps that will show the areas considered.

2.2 DATA PREPARATION

Once the data is obtained from the different sources, it will be checked for adequacy and completeness.

There are some particular considerations for the Foursquare and Numbeo data exposed below.

2.2.1 Foursquare data

In some occasions, the data obtained from Foursquare may present a very large number of categories and subcategories. And given that different countries will be compared, it may happen that similar venue concepts are categorised differently (i.e. a Pub in the UK versus a Bar in Spain). To avoid this potential issue the information obtained for all the areas to be compared will be analysed and grouped in representative categories which will allow a more meaningful comparison.

In other occasions, the amount of information received may the scarce. That could lead to biased categorisation of an area, and a misleading comparison among areas. In those situations, those particular areas will be removed from the potential candidates.

2.2.2 Numbeo data

The data that can be obtained in Numbeo for some of the fields, such as the cost of living, can be either in absolute terms (i.e. cost of living expressed in monthly GPB, USD or EUR), or indexes. To facilitate the comparison between different cities, the use of indices is more convenient. The indices that will be considered are:

- Quality of life
- Purchase power including rent
- Safety
- Health care
- Climate
- Pollution

2.3 EXAMPLE IMPLEMENTED

The code developed will be presenting a particular case. And this case will be:

- Area to take as requirements: Richmond, Surrey, United Kingdom
- Cities where relocation will be considered:
    - Madrid, Spain
    - Frankfurt, Germany
    - Singapore, Singapore
    - New York, United States

3. METHODOLOGY

The methodology implemented in the notebook at a high level is gathering information from the different data sources, analyse and clean it, work with the data to extract the desired information and provide a recommendation.

The initial data provided by the user would be the city of reference and the list of cities that would like to be considered for relocation.

The steps in which the above methodology develops are:

1. Collect the names of the neighbourhood to be considered in each city
2. Collect the geographical coordinates of each neighbourhood
3. Collect the venues existing in each neighbourhood
4. Analise and redefine venues' categories
5. Cluster the neighbourhood based on venues' categories
6. Analise the clusters obtained (1st selection criteria)
7. Collect indices for each city
8. Define a distance function among cities and calculate
9. Analise distances obtained (2nd selection criteria)
10. Provide recommendation

The city of reference and all the neighbourhoods are integrated into one single dataframe to facilitate the later comparison among all the areas.

The following sections describe in more detail each of the methodology steps.

3.1 COLLECT NEIGHBOURHOODS NAMES

For this exercise, the names of the neighbourhoods for each city have been included in a list saved in a pickle file. The notebook loops over these files, opens and reads the lists, and appends the neighbourhoods to the dataframe.

3.2 COLLECT NEIGHBOURHOODS COORDINATES

In the same loop described above, the coordinates for each neighbourhood is gathered from Google Maps API geocoding and added to the dataframe.

At the end of this step, the dataframe contains the following data for each neighbourhood (including the city of reference): City, Neighbourhood, Latitude, Longitude.

A map of the any selected city with markings on the neighbourhoods is produced.

3.3 COLLECT VENUES

The next step is to collect all the venues in each of the neighbourhoods; this data is collected at the Foursquare API.

The limit of venues per neighbourhood is set to a value that should not limit the amount of information gathered. It has been set to 300 and later checked that no neighbourhood has reached this figure.

The radius considered around each neighbourhood is 1,400m. This radius is the one that will cover the whole Richmond area, that is the city of reference. Therefore, it seems logical to consider the same area in the other neighbourhoods.

The option of saving a file with all the data collected is provided at this moment. That would allow starting from this point in a later stage, without the need to re-collect all the data.

3.4 ANALISE AND REDEFINE VENUES' CATEGORIES

At this point, the venues information is analysed. Initially, the notebook checks the maximum and the minimum number of venues gathered in a neighbourhood per city, following a histogram of the number of venues is plotted. That informs whether the limit (300 in the example) has been reached (which is not the case) and whether there are neighbourhoods with not many data, in which case those should be removed from the list of candidates.

Based on the information above, a threshold is determined to decide what neighbourhoods will be removed because of no having enough data. The consideration here is to establish how many venues would we accept as enough in one area to accept that those are representative of the area.

After that, the focus is put on the categories. Having a large number of categories is not necessarily good for categorisation of the neighbourhoods. To better understand the character of the neighbourhood and make it comparable to others the categories are redefined, meaning that some subcategories (up to the 4^{th} level of subcategory, that is what the information gathered from Foursquare would provide), are grouped in their parent category.

For instance, all restaurants whatever the type of food served would have the category redefined to 'Restaurant'.

To tho that, a thorough analysis of all the near one thousand categories and subcategories that Foursquare considers have been undertaken, and various lists of recategorization have been prepared. Those lists are stored in a dictionary in a pickle file. Each dictionary key

corresponds to a parent category that will be used, and the list to which the key points, corresponds to the subcategories that will be changed that parent category.

3.5 CLUSTER THE NEIGHBOURHOODS

In this section, the clean data is analysed by calculating the frequency of occurrence of each category is calculated for each neighbourhood. It will be the key information to be used in the clustering process.

The ten most common venues are found for each neighbourhood, that provides the opportunity to both, check that the information collected for the city of reference corresponds to what was expected, as this is meant to be the known area; and see what all other neighbourhoods look like.

After that, the clustering process is undertaken by using the k-means method provided by sklearn. The number of clusters chosen is 7n.

3.6 ANALISE THE CLUSTERS OBTAINED

A histogram of the clusters obtained is plotted to analyse the results of the clustering process. The fact that most of the clusters have a good number of neighbourhoods would be indicative that the categories considered where able to identify the similarities and dissimilarities among the neighbourhoods.

It will constitute the first selection criteria. Hopefully, the cluster corresponding to the city of reference will have few other neighbourhoods. Those would be candidates for a relocation based on the existing venues criteria.

A world map is shown with all the clusters, zooming into each city the categorisation of the neighbourhood can be explored in detail. The notebook provides the alternative to jump straight to any of the cities, though.

3.7 COLLECT INDICES

The second selection criteria focused on the indices obtained in the Numbeo website. Those indices refer to the whole city rather than to neighbourhoods. Hence this criterion is analysed separately.

The indices considered for each city are:

- quality_of_life_index

- purchasing_power_incl_rent_index
- safety_index
- health_care_index
- climate_index
- pollution_index

The case that information for a city is available the user should consider whether indices for a nearby area would be representative or discarding the use of this second criteria.

In the particular example used in the notebook, it was found that Richmond (UK) has not information. However, as the area is integrated into the Greater London, the indices for London have been assumed as representative.

Similarly, as Queens is, in fact, part of New York, the indices for the later have been used in the example.

3.8 DEFINE A DISTANCE FUNCTION

To capture the differences between each potential relocation city and the city of reference a distance function has been defined. This distance is calculated as the weighted average of the difference of the indices.

It has to be noted that each of the components of this distance, each index, have a sign. Therefore, a higher index in a positive subject, such as quality of life, will be considered as a negative distance. Similarly, a lower index in a negative subject, such as pollution, will be considered as a negative distance. Positive distances would be found when positive subjects indices decrease or negative subjects indices increase.

The weighting of each os the components are subjective to the individual, therefore this information needs to be provided by the user.

3.9 ANALISE DISTANCES OBTAINED

As discussed above the distance function defined can provide both positive and negative values. Increasing positive values mean decreasing quality of the city-based the indices collected, therefore not desirable. On the other side, higher negative values show the increasing quality of the city, therefore desirable

In summary, the second selection criterion is the smallest distance (maximum negative value, or minimum positive value if no negative are found).

3.10 RECOMMENDATION

The final recommendation comes from the combination of the two selection criteria. The neighbourhoods recommended are those in the city which have the lowest distance, and that are in the same cluster than the city of reference.

4. RESULTS

The results obtained in the particular example are shown below, following the steps of the methodology.

4.1 COLLECT NEIGHBOURHOODS NAMES AND COORDINATES

The names are collected from the lists defined in the notebook, and coordinates collected from the Google Maps API geocoding. The information is stored in the dataframe shown below.

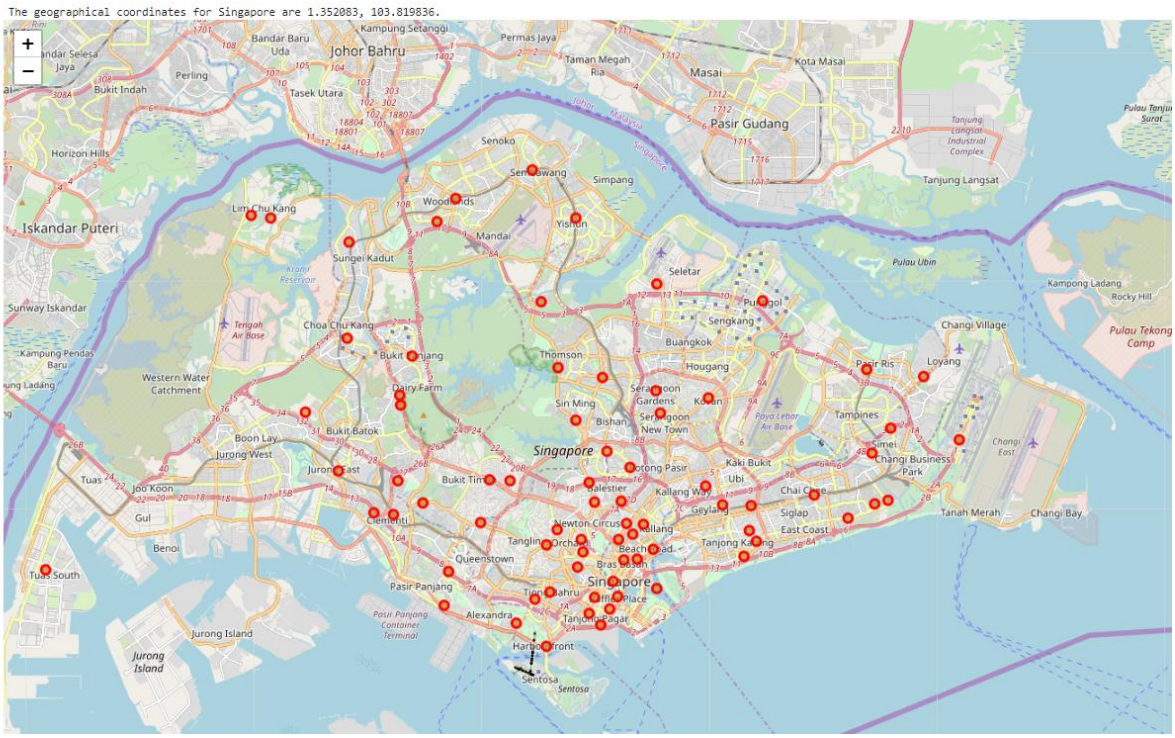
[4]:

	City	Country	Latitude	Longitude
0	Richmond	United Kingdom	51.461311	-0.303742
1	Singapore	Singapore	1.352083	103.819836
2	Madrid	Spain	40.416775	-3.703790
3	Frankfurt	Germany	50.110922	8.682127
4	Queens NY	United States	40.728224	-73.794852

The cities are shown on a world map for reference. The city of reference in red and the relocation candidates in blue.

[5]:

Then neighbourhoods coordinates are collected and stored in a dataframe. Maps of each of the cities generated. The one for Singapore is shown below.



4.2 COLLECT VENUES

The next step is to collect all the venues in each of the neighbourhoods, this data is collected at the Foursquare API and stored in a dataframe.

The image below shows the size of the dataframe generated and the initial registers on it, corresponding to the city of reference.

```
[20]: # size of the dataframe and a first glimpse:
print(city_venues.shape)
city_venues.head(10)
```

```
(5814, 8)
```

[20]:

	City	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Richmond	Richmond	51.461311	-0.303742	Richmond Green	51.461250	-0.305918	Park
1	Richmond	Richmond	51.461311	-0.303742	Al Boccon Di'vino	51.459607	-0.304772	Italian Restaurant
2	Richmond	Richmond	51.461311	-0.303742	Kiss the Hippo Coffee	51.460919	-0.304230	Coffee Shop
3	Richmond	Richmond	51.461311	-0.303742	Ole & Steen	51.460587	-0.304967	Coffee Shop
4	Richmond	Richmond	51.461311	-0.303742	Richmond Theatre	51.462130	-0.304009	Theater
5	Richmond	Richmond	51.461311	-0.303742	Gelateria Danieli	51.460967	-0.305259	Ice Cream Shop
6	Richmond	Richmond	51.461311	-0.303742	Franco Manca	51.459547	-0.305041	Pizza Place
7	Richmond	Richmond	51.461311	-0.303742	Waterstones	51.459532	-0.305984	Bookstore
8	Richmond	Richmond	51.461311	-0.303742	No. 1 Duke Street	51.461412	-0.303828	Restaurant
9	Richmond	Richmond	51.461311	-0.303742	Rustica	51.462546	-0.302594	Italian Restaurant

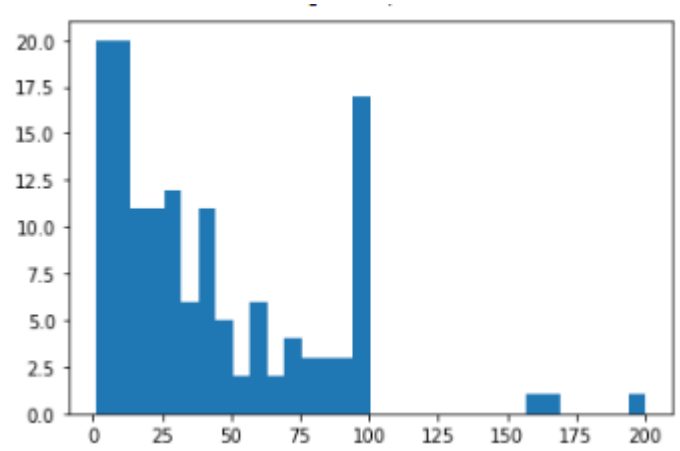
4.3 ANALISE AND REDEFINE VENUES' CATEGORIES

At this point, the venues information is analysed. The notebook checks the maximum and the minimum number of venues gathered in a neighbourhood per city, see image below.

```
[22]: # check by city what are the maximum and minimum number of venues in each neighbourhood
for city in city_list:
    count_max=city_venues[city_venues['City']==city].groupby('Neighbourhood')['Neighbourhood'].count().max()
    count_min=city_venues[city_venues['City']==city].groupby('Neighbourhood')['Neighbourhood'].count().min()
    print('The maximum and minimum counts in {} are: {},{}.'.format(city,count_max,count_min))
```

```
The maximum and minimum counts in Richmond are: 100,100.
The maximum and minimum counts in Singapore are: 100,1.
The maximum and minimum counts in Madrid are: 100,1.
The maximum and minimum counts in Frankfurt are: 100,2.
The maximum and minimum counts in Queens NY are: 200,4.
```

A histogram of the number of venues is ploted.



The plot shows a significant number of neighbourhoods with 50 or fewer venues; therefore the threshold would ideally be below that figure. I stablish is at 20. And remove registers corresponding to a neighbourhood with not enough data. The image below shows the number of registers removed and the list of neighbourhoods.

```
[27]: # This cell removes neighbourhoods where the number of venues is below the determined treshold
city_venues_clean = remove_registers(city_venues_clean,no_data_list,['Neighbourhood'])
```

```
Number of registers before removal:  5814
Searching for:  ['Telok Blangah', 'Hong Leong Garden', 'River Valley', 'Bukit Timah', 'Tanglin', 'Watten Estate', 'Thomson', 'Ser
angoon', 'Macpherson', 'Eunos', 'Upper East Coast', 'Eastwood', 'Kew Drive', 'Loyang', 'Changi', 'Simei', 'Tampines', 'Punggol',
'Upper Bukit Timah', 'Clementi Park', 'Ulu Pandan', 'Tuas', 'Dairy Farm', 'Bukit Panjang', 'Lim Chu Kang', 'Tengah', 'Kranji', 'W
oodgrove', 'Upper Thomson', 'Springleaf', 'Sembawang', 'Seletar', 'Fuencarral-El Pardo', 'Moncloa-Aravaca', 'Latina', 'Carabanche
l', 'Usera', 'Puente de Vallecas', 'Moratalaz', 'Hortaleza', 'Villaverde', 'Villa de Vallecas', 'Vicálvaro', 'Barajas', 'Flughafe
n', 'Niederrad', 'Oberrad', 'Queensbridge,', 'Blissville,', 'Point,', 'East']
Searcing in:  ['Neighbourhood']
Number of registers after removal:  5304
Number of registers removed:  510
```

After that, the focus is put on the categories. In the example, the number of potential categories will be reduced to 13. The image below presents the dictionary keys that will become the categories.

```
[3]: # Load the Category reduction criteria
     f = open('Cat_red_13.pckl', 'rb')
     cat_lists_dic = pickle.load(f)
     f.close()
```

```
[7]: cat_lists_dic.keys()
```

```
[7]: dict_keys(['Arts & Entertainment', 'College & University', 'Event', 'Food', 'Nightlife Spot', 'Outdoors & Recreation',
     'Professional & Other Places', 'Residence', 'Shop & Service', 'Travel & Transport', 'Restaurants', 'Cafe'])
```

Once applied the reduction, it appears that only 11 or the 13 are used, meaning that 2 of them have none venues assigned.

```
[31]: # Reduction of the categories
      for cat_name, cat_cont in zip(list(cat_lists_dic.keys()),list(cat_lists_dic.values())):
          for subcat_name in cat_cont:
              city_venues_clean.replace(to_replace=subcat_name,value=cat_name,inplace=True)

      print('There are {} unique categories.'.format(len(city_venues_clean['Venue Category'].unique())))
      city_venues_clean['Venue Category'].unique()
```

```
There are 11 unique categories.
```

```
[31]: array(['Outdoors & Recreation', 'Restaurants', 'Cafe',
             'Arts & Entertainment', 'Food', 'Shop & Service', 'Nightlife Spot',
             'Travel & Transport', 'Professional & Other Places', 'Residence',
             'College & University'], dtype=object)
```

4.4 CLUSTER THE NEIGHBOURHOODS

In this section, the clean data is analysed by calculating the frequency of occurrence of each category is calculated for each neighbourhood. The image below shows the head of the dataframe grouped by neighbourhood.

```
[33]: city_grouped = city_onehot.groupby('Neighbourhood').mean().reset_index()
      print('The grouped df shape is: ',city_grouped.shape)
      city_grouped.head()
```

```
The grouped df shape is:  (87, 12)
```

[33]:

	Neighbourhood	Arts & Entertainment	Cafe	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Restaurants	Shop & Service	Travel & Transport
0	Altstadt	0.070000	0.150000	0.0	0.080000	0.100000	0.120000	0.010000	0.0	0.280000	0.170000	0.020000
1	Amber Road	0.043478	0.043478	0.0	0.086957	0.086957	0.086957	0.043478	0.0	0.434783	0.130435	0.043478
2	Ang Mo Kio	0.033898	0.152542	0.0	0.305085	0.000000	0.050847	0.000000	0.0	0.288136	0.152542	0.016949
3	Anson	0.010000	0.180000	0.0	0.180000	0.050000	0.020000	0.010000	0.0	0.430000	0.060000	0.060000
4	Ardmore	0.054054	0.027027	0.0	0.054054	0.189189	0.027027	0.000000	0.0	0.324324	0.081081	0.243243

And the ten most common venues are found for each neighbourhood.

[36]:

	Neighbourhood	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	Altstadt	Restaurants	Shop & Service	Cafe	Outdoors & Recreation	Nightlife Spot	Food	Arts & Entertainment	Travel & Transport	Professional & Other Places	Residence
1	Amber Road	Restaurants	Shop & Service	Outdoors & Recreation	Nightlife Spot	Food	Travel & Transport	Professional & Other Places	Cafe	Arts & Entertainment	Residence
2	Ang Mo Kio	Food	Restaurants	Shop & Service	Cafe	Outdoors & Recreation	Arts & Entertainment	Travel & Transport	Residence	Professional & Other Places	Nightlife Spot
3	Anson	Restaurants	Food	Cafe	Travel & Transport	Shop & Service	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Arts & Entertainment	Residence
4	Ardmore	Restaurants	Travel & Transport	Nightlife Spot	Shop & Service	Food	Arts & Entertainment	Outdoors & Recreation	Cafe	Residence	Professional & Other Places
5	Arganzuela	Restaurants	Outdoors & Recreation	Nightlife Spot	Food	Shop & Service	Travel & Transport	Cafe	Residence	Professional & Other Places	College & University
6	Astoria	Restaurants	Food	Shop & Service	Nightlife Spot	Cafe	Outdoors & Recreation	Arts & Entertainment	Travel & Transport	Residence	Professional & Other Places
7	Astoria,	Restaurants	Food	Shop & Service	Nightlife Spot	Cafe	Outdoors & Recreation	Arts & Entertainment	Travel & Transport	Residence	Professional & Other Places

After that, the clustering process is undertaken by using the k-means method provided by sklearn, as shown below.

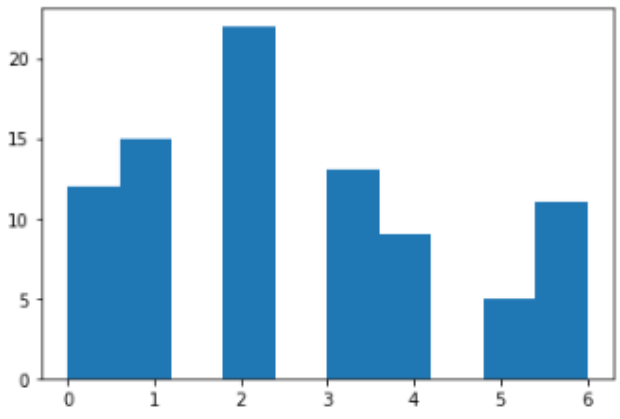
[37]:
```
# set number of clusters
kclusters = 7

city_grouped_clustering = city_grouped.drop('Neighbourhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(city_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

[37]:
```
array([0, 1, 6, 1, 5, 0, 2, 2, 5, 2, 1, 2, 2, 4, 3, 2, 2, 4, 1, 1, 3, 1,
       2, 0, 2, 2, 2, 5, 2, 3, 6, 1, 6, 2, 3, 3, 4, 3, 0, 6, 6, 2, 0, 0,
       0, 3, 3, 3, 4, 2, 5, 1, 3, 6, 0, 4, 1, 1, 6, 1, 4, 0, 0, 5, 6, 1,
       2, 0, 2, 0, 1, 2, 2, 4, 1, 6, 2, 2, 1, 4, 3, 2, 3, 3, 6, 6, 4])
```

4.5 ANALISE THE CLUSTERS OBTAINED

The histogram of the clusters obtained shows that the neighbourhoods are distributed across all the clusters; therefore the categorisation has worked satisfactorily.

```
plt.hist(kmeans.labels_)
plt.show()
```



Richmond has been allocated in the cluster number 1, along with other neighbourhoods from Singapore, Madrid and Frankfurt.

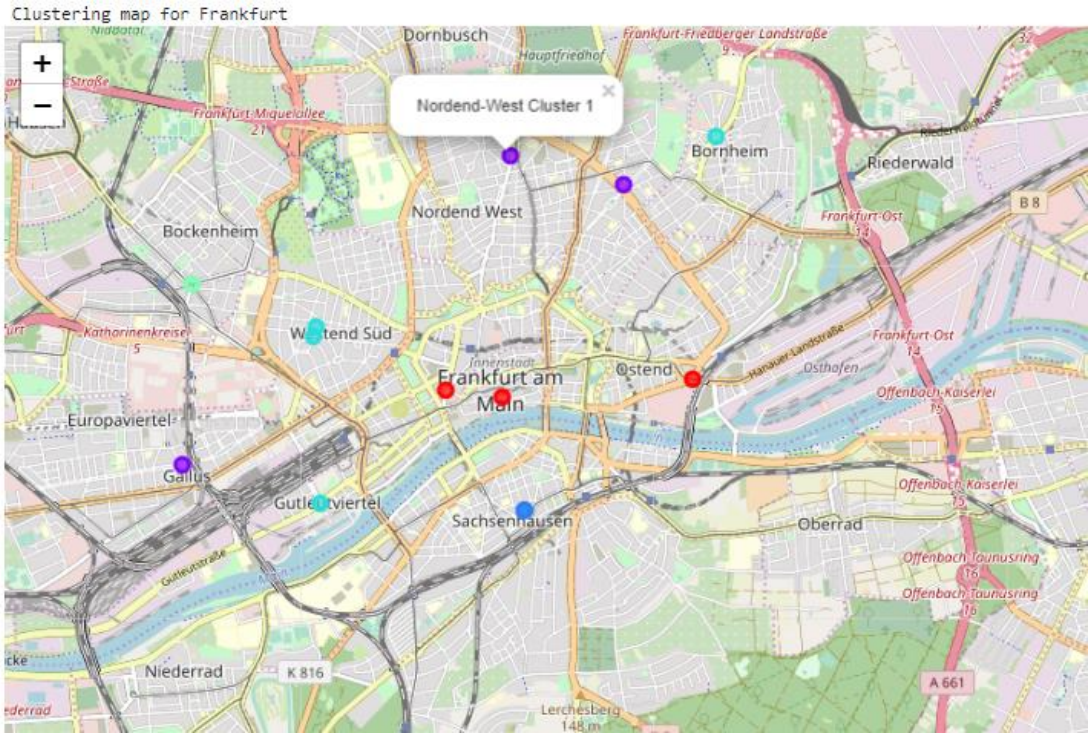
[51]:

	City	Neighbourhood	Cluster Labels	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	Richmond	Richmond	1	Restaurants	Nightlife Spot	Food	Shop & Service	Cafe	Outdoors & Recreation	Arts & Entertainment	Travel & Transport	Residence	Professional & Other Places
2	Singapore	Cecil	1	Restaurants	Cafe	Food	Nightlife Spot	Outdoors & Recreation	Travel & Transport	Shop & Service	Professional & Other Places	Arts & Entertainment	Residence
4	Singapore	People's Park	1	Restaurants	Food	Shop & Service	Nightlife Spot	Travel & Transport	Cafe	Professional & Other Places	Outdoors & Recreation	Arts & Entertainment	Residence
5	Singapore	Anson	1	Restaurants	Food	Cafe	Travel & Transport	Shop & Service	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Arts & Entertainment	Residence
6	Singapore	Tanjong Pagar	1	Restaurants	Nightlife Spot	Food	Cafe	Outdoors & Recreation	Travel & Transport	Shop & Service	Arts & Entertainment	Residence	Professional & Other Places
16	Singapore	Beach Road (part)	1	Restaurants	Food	Shop & Service	Cafe	Nightlife Spot	Travel & Transport	Arts & Entertainment	Outdoors & Recreation	Residence	Professional & Other Places
22	Singapore	Lavender	1	Restaurants	Food	Cafe	Travel & Transport	Shop & Service	Outdoors & Recreation	Nightlife Spot	Arts & Entertainment	Residence	Professional & Other Places
31	Singapore	Novena	1	Restaurants	Food	Cafe	Shop & Service	Travel & Transport	Outdoors & Recreation	Nightlife Spot	Residence	Professional & Other Places	College & University
42	Singapore	Amber Road	1	Restaurants	Shop & Service	Outdoors & Recreation	Nightlife Spot	Food	Travel & Transport	Professional & Other Places	Cafe	Arts & Entertainment	Residence
76	Madrid	Centro	1	Restaurants	Nightlife Spot	Food	Travel & Transport	Shop & Service	Outdoors & Recreation	Cafe	Arts & Entertainment	Professional & Other Places	Residence
82	Madrid	Chamberí	1	Restaurants	Nightlife Spot	Food	Shop & Service	Cafe	Arts & Entertainment	Outdoors & Recreation	Travel & Transport	Residence	Professional & Other Places
95	Madrid	San Blas-Canillejas	1	Restaurants	Travel & Transport	Outdoors & Recreation	Food	Cafe	Shop & Service	Nightlife Spot	Residence	Professional & Other Places	College & University
99	Frankfurt	Gallus	1	Restaurants	Food	Travel & Transport	Nightlife Spot	Arts & Entertainment	Shop & Service	Professional & Other Places	College & University	Cafe	Residence

It is remarkable that any of the neighbourhood from Queens NY has been allocated cluster number 1; therefore this city would be excluded from the potential cities to relocate.

The map below shows the clustering in Frankfurt; purple colour are cluster number 1.

[53]:



4.6 COLLECT INDICES, DEFINE A DISTANCE FUNCTION

The second selection criteria focused on the indices obtained in the Numbeo website. To capture the differences between each potential relocation city and the city of reference the distance ad defined previously is calculated. The table below shows the findings.

[45]:

	City	Quality of life	Purchasing power	Safety	Healthcare	Climate	Pollution	Distance
0	Richmond	119.244422	77.878674	47.712168	68.406884	88.254338	60.638066	0.000000
1	Singapore	144.358738	88.859402	69.464020	70.836230	57.453281	33.483621	-8.742190
2	Madrid	152.918820	74.211592	69.710169	79.313535	85.473048	53.293026	-13.245188
3	Frankfurt	174.257668	104.612589	59.782357	73.707203	84.716619	38.581419	-21.040778
4	Queens NY	141.932448	100.000000	55.561231	65.013747	79.661486	54.204299	-8.236424

4.7 ANALISE DISTANCES OBTAINED

As discussed above the distance function defined can provide both positive and negative values, being the most desirable the higher negative values, which show the increasing quality of the city.

Frankfurt is the city that presents a larger betterment based on the indices criteria.

4.8 RECOMMENTATION

The final recommendation comes from the combination of the two selection criteria. The neighbourhoods recommended are those in Frankfurt and in cluster number 1. The table below shows the recommended neighbourhoods for relocation.

[52]:

	City	Neighbourhood	Cluster Labels	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
99	Frankfurt	Gallus	1	Restaurants	Food	Travel & Transport	Nightlife Spot	Arts & Entertainment	Shop & Service	Professional & Other Places	College & University	Cafe	Residence
105	Frankfurt	Nordend-Ost	1	Restaurants	Nightlife Spot	Food	Cafe	Shop & Service	Travel & Transport	Outdoors & Recreation	Residence	Professional & Other Places	College & University
106	Frankfurt	Nordend-West	1	Restaurants	Cafe	Shop & Service	Food	Travel & Transport	Outdoors & Recreation	Nightlife Spot	Arts & Entertainment	Residence	Professional & Other Places

5. DISCUSSION

The criteria used in providing a recommendation and the recommendation itself has already been presented and exposed widely in the preceding sections. However, there are some aspects worth highlighting.

Firstly, the Category reduction aspect. It is a key aspect of obtaining a good and representative result in the clustering process. But at the same time has to be representative of the requirements. Following the example of the restaurants, in the example developed in the notebook, they have been all amalgamated. That has an underlying meaning concerning the individual preferences, that individual would like similarly any cuisine. The case that the individual had a strong opinion, positive or negative, of a particular cuisine it would be good that this is captured in a dedicated category.

Going a bit further in this subject, it could be analysed the possibility of altering the city of reference information to make it even more representative of the individual requirements. Giving more emphasis, therefore dedicated categories, to those aspects for which there is the strongest opinion; and amalgamating or removing aspects that should not guide a decision process.

Another option in defining requirements would be providing more than one city of reference and then calculating an average of them as a benchmark. This option could be combined with the previous one.

Secondly, the granularity of the indices. It has been used the information provided by Numbeo, with is city-based. But it is known, particularly in big cities, that they are very heterogeneous in their different neighbourhoods. The alternatives to overcome this limitation would be, the case that it exists, using a data source with more granularity.

Alternatively, it could be analysed what additional data could be used to modulate the indices across the cities. For instance, generally, the cost of housing would be higher in the city centre, or in particular areas. Using some local real estate information would be a way to capture this granularity and would allow modulating the cost of living, including rent across the city.

Thirdly, note that this exercise does not analyse in detail the number of clusters to be implemented. It has been considered that a small number of clusters, would provide a better understanding of the overall picture; rather than a large number scarcely populated. The

categories that will determine the clustering are at the same time quite broad; therefore, the approach seems coherent.

The case that more detailed requirements were available and that the amount of data was higher and more uniform, then a larger number of clusters would probably be required. I such instance, I would definitively recommend to analyse different scenarios and determine the optimum number of clusters.

Finally, the consideration of defined neighbourhoods for potential areas for relocations in some circumstances brings a non-uniform coverage of the cities. Stablishing a uniform grid of point to analyse would improve the area coverage.

6. CONCLUSIONS

Base on all the exposed above, I conclude that the tool developed is robust enough to undertake an initial analysis of the preferred relocation places. The methodology used can be applied to any other cities, while there is information available of them.

There are clear potential lines of further development of the tool, that would allow refining the selection.

Key ideas for this further development are:

- Capture individual requirements in more detail
    - Consider more than one city of reference
    - Alter the venues information gathered to highlight strong individual opinions
    - Customise to more detail the lists of categories
- Develop a more uniformly distributed grid or potential areas for relocating
- Considering not only the number of venues but also considering their scoring.
- Develop granularity of the indices across the cities.