IBM Data Science Professional Certificate
Applied Data Science Capstone

Rafael Henrique da Silva Santos
rafacto@gmail.com

June 2020

Moving Abroad - Choosing a new city to live

1. Introduction

Moving abroad and choosing a new city to live is not easy. To reduce the number of possibilities, one can look at the sister cities of his/her current city. Sister cities are cities that establish a bond of cooperation on many factors like culture, health, education, transport, and economic development. Often the cities are located in different countries, developing a paradiplomacy relation, a relationship that does not depend on federal governments (which is what designates diplomacy). Typically, to become sisters two cities need to have similar features like the number of residents, historical facts, or economic sector. Thus, moving to a sister city can be easier, since they share some features and have political facilitators.

Different factors like cultural life, attractions, language, climate, jog market, and life cost can impact the life experience in the new city. This study aims to be a tool for the process of finding a new place (city and neighborhood) to live, taking into consideration all those factors. The idea is to analyze, compare, and segment the sister cities given some living parameters (Cost of living, Rent, Property Price, Crime, Health Care, Pollution, Traffic, and Climate). With the cities separated in groups defined according to relevant factors that influence the life experience, one can easily decide which one to move in.

After the most proper city being chosen, it is time to select its more suitable neighborhood. In this process, all the neighborhoods of the city are compared and segmented according to their most popular venues (restaurants, bars, museums, parks, beach, cinema, etc), allowing the characterization of each group. Finally, information about safety per neighborhood is used to generate the final decision of which place to live in. To illustrate this whole process, the Brazilian city of Recife is chosen as the hometown and start point of the analysis described.

2. Data

To obtain the Recife sister cities, Wikipedia is consulted. Any city page in Wikipedia presents a section called "sister cities", however, it can vary depending on the language chosen. Naturally, there is more information in the language of the country the city belongs to. In the case of Recife, Wikipedia in English lists only three sister cities while the Portuguese version shows eight. The latter was chosen.

The living parameters are obtained from Numbeo, the world's largest database of user-contributed data about cities and countries worldwide and provides current and timely information on world living conditions including cost of living, housing indicators, health care, traffic, etc. They formulated an index for each of those parameters: Cost of living, Rent, Property Price, Crime, Health Care, Pollution, Traffic, and Climate. The access to Numbeo's API is not free, but it has a page with all the information about those living indexes. As the amount of data is little and the API access is paid, the desired information was obtained accessing manually the Numbeo page and building an excel file with it.

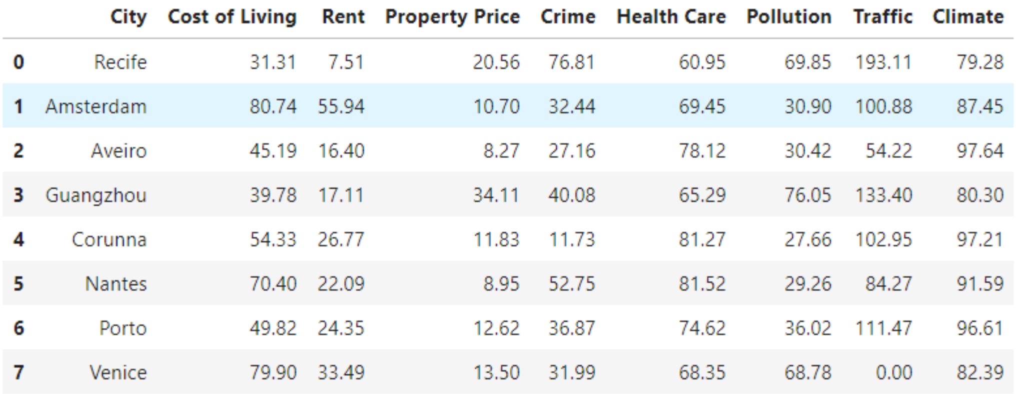
To obtain the top venues of the neighborhoods, Foursquare is consulted. Foursquare is a local search-and-discovery mobile app that provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history. The app collects information about all sorts of venues: restaurants, bars, cafes, museums, art galleries, parks, clubs, universities, schools, markets, services like laundry, etc. Each venue has a page with, among other information, a rate (from 0 to 10), description, photos, and user tips. Foursquare provides an API that allows application developers to interact with the Foursquare platform and to retrieve, among others contents, all sorts of information about the venues near a location. The result of the requests to Foursquare API is a json file.

To obtain the list of the chosen city's neighborhoods, their coordinates, boundaries, and safety index, the municipal page was consulted. All that information can be easily found in tables and json format.

3. Methodology

The first step is to analyze, compare, and segment, through exploratory analysis and clustering, the hometown's sister cities according to living parameters to decide which city to move in. The second step is to segment the neighborhoods of the chosen city according to their top venues, aiming the grouping and further characterization of them. Finally, the safety information of each neighborhood is joined to help in the final decision of each neighborhood to move in.

The list of Recife sister cities was obtained from Wikipedia Recife page: Amsterdam, Aveiro, Guangzhou, Corunna, Nantes, Porto, Venice, and Vitória. This last city was removed for being a Brazilian city and the study being about moving abroad. After consulting the living indexes at Numbeo page, we obtained the following table.



	City	Cost of Living	Rent	Property Price	Crime	Health Care	Pollution	Traffic	Climate
0	Recife	31.31	7.51	20.56	76.81	60.95	69.85	193.11	79.28
1	Amsterdam	80.74	55.94	10.70	32.44	69.45	30.90	100.88	87.45
2	Aveiro	45.19	16.40	8.27	27.16	78.12	30.42	54.22	97.64
3	Guangzhou	39.78	17.11	34.11	40.08	65.29	76.05	133.40	80.30
4	Corunna	54.33	26.77	11.83	11.73	81.27	27.66	102.95	97.21
5	Nantes	70.40	22.09	8.95	52.75	81.52	29.26	84.27	91.59
6	Porto	49.82	24.35	12.62	36.87	74.62	36.02	111.47	96.61
7	Venice	79.90	33.49	13.50	31.99	68.35	68.78	0.00	82.39

Table 1. Sister cities and their living indexes

Cost of Living and Rent indexes are calculated in relation to New York City, which means that for New York City, each index should be 100(%). If another city has, for example, a rent index of 120, it means that on average in that city rents are 20% more expensive than in New York City. If a city has a rent index of 70, that means on average rent in that city is 30% less expensive than in New York City. Each of the other indexes is calculated in a very specific way that can be verified at the Numbeo webpage.

3.1 When cheap becomes expensive

Cost of living and rent are the main factors that affect one expanse (here they are called **cost indexes**) and, thus, are essential in the decision of choosing a new city to move in. However, problems like crime, pollution and traffic can negatively influence the life experience in the new place and may be relevant to care about, thus, here they are called **negative indexes**. At the same time, the health care system and the climate are positive features that can also help in the decision (**positive indexes**). The following graphs compare those aspects among the cities.

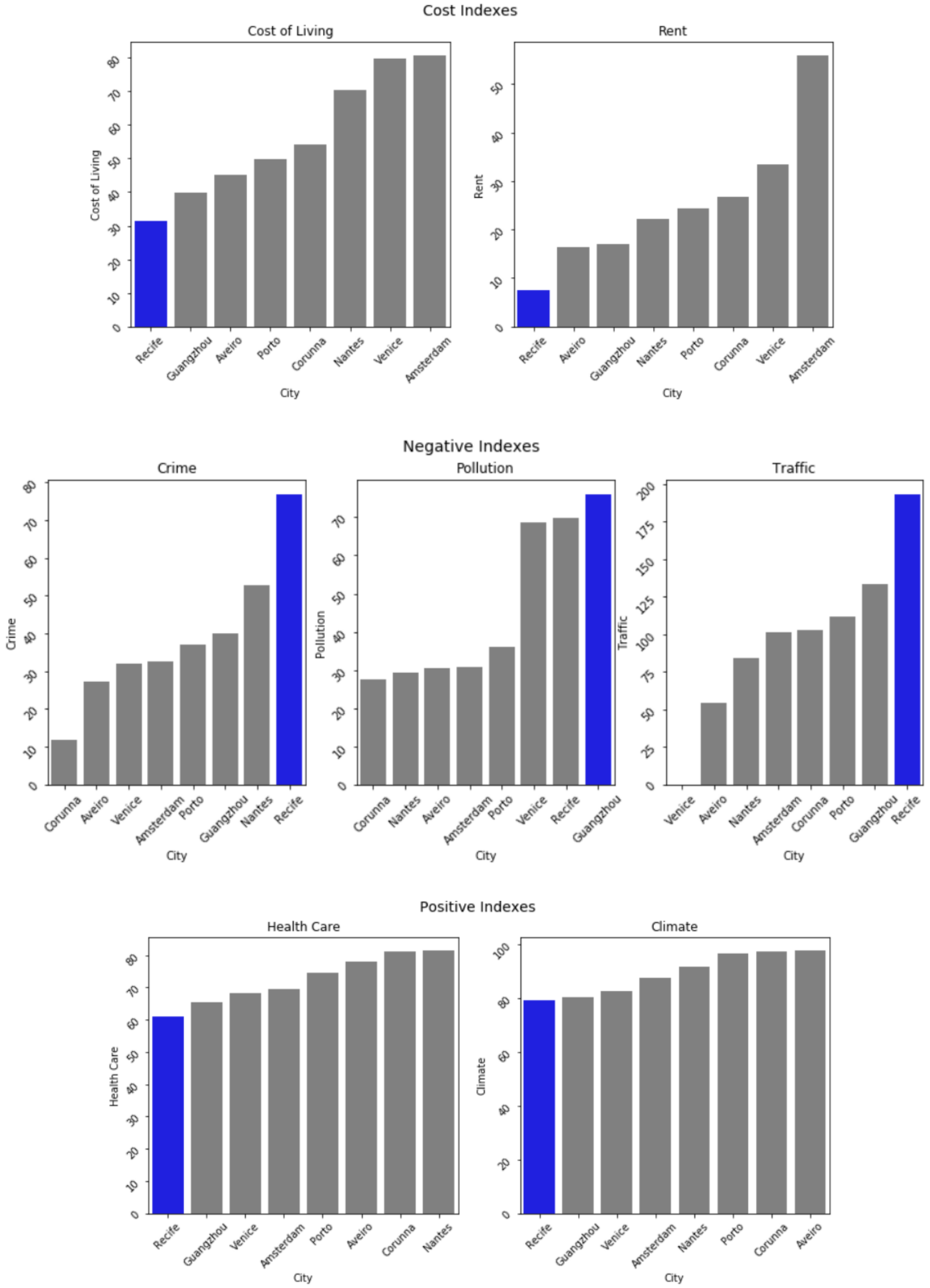


Figure 1. Living indexes by city

As can be seen, Recife is the cheapest city to live in. However, this comes with a price. It presents the highest crime and pollution indexes, the second worst pollution index, and the lowest health care and climate indexes. At first look, one can imagine that the more expensive the city the better the quality of life indicators, but this relation is not that direct and simple.

Venice and Amsterdam, for example, are those with the highest cost of living and rent, however, they do not present the lowest crime, traffic (in the case of Amsterdam), and pollution nor the best health care system and climate. Some graphs can attest to those weak correlations among the features.

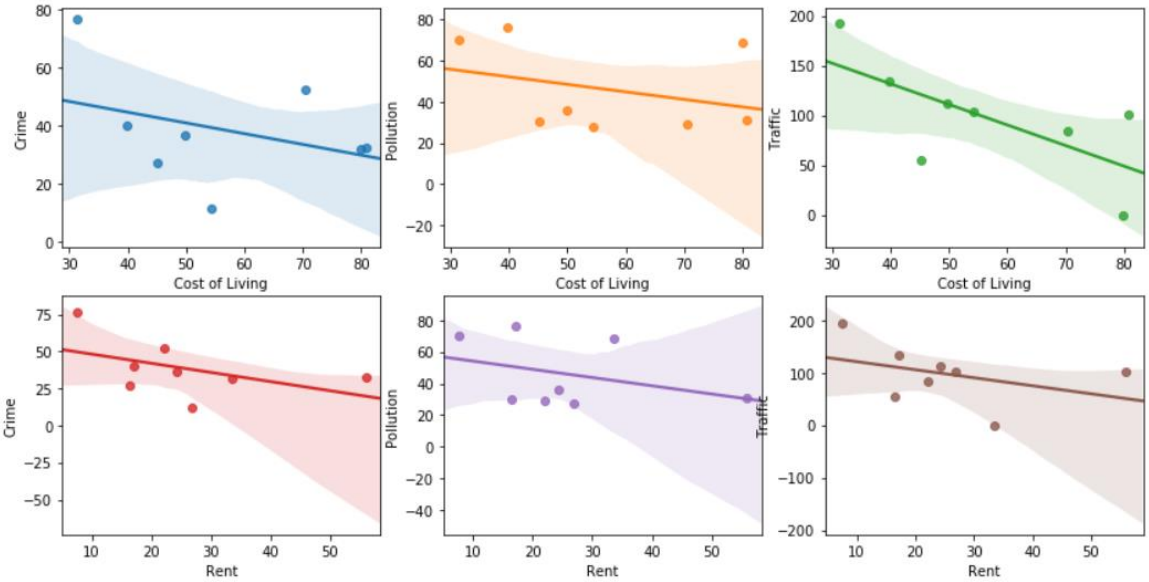


Figure 2. Correlation among Cost indexes and Negative indexes

Although the relationships between the cost indexes and the negative indexes are clearly negative, the slope of the line is not steep, which means weak correlations. The same (weakness) happens with the relationships between cost indexes and positive indexes., with correlations even weaker.

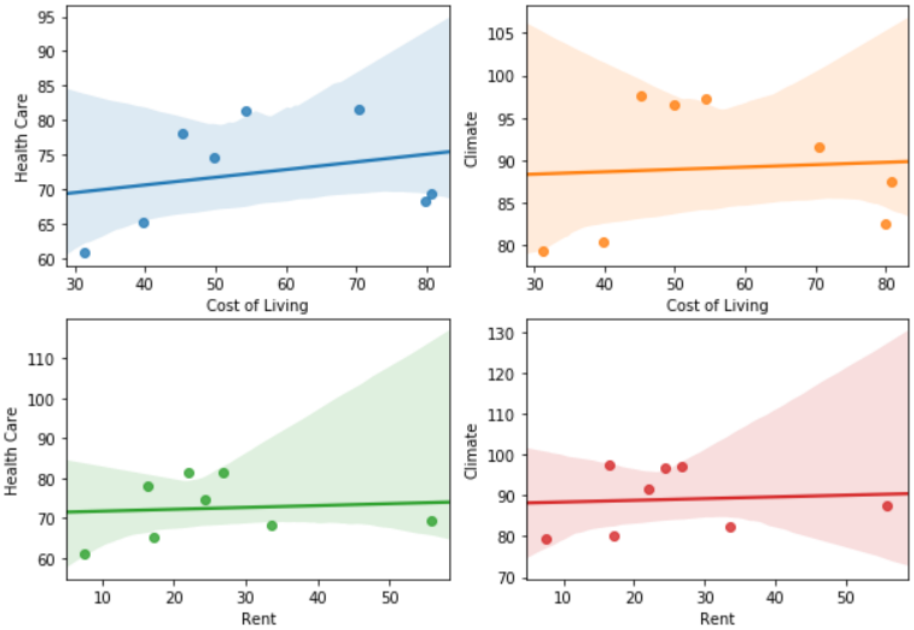
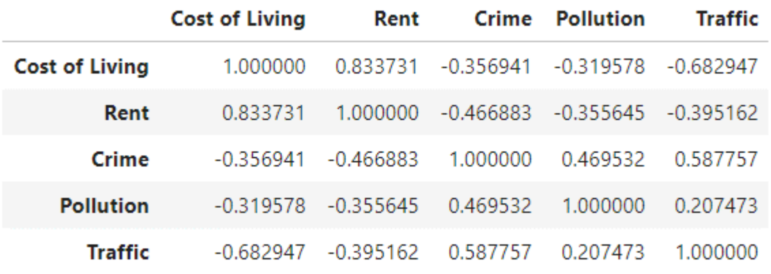


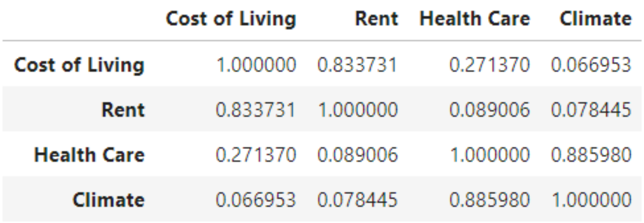
Figure 3. Correlation between Cost indexes and Positive indexes

The weaknesses of those relationships (Cost Indexes x Negative Indexes and Cost Indexes x Positive Indexes) can be confirmed in the correlation matrices bellow.



	Cost of Living	Rent	Crime	Pollution	Traffic
Cost of Living	1.000000	0.833731	-0.356941	-0.319578	-0.682947
Rent	0.833731	1.000000	-0.466883	-0.355645	-0.395162
Crime	-0.356941	-0.466883	1.000000	0.469532	0.587757
Pollution	-0.319578	-0.355645	0.469532	1.000000	0.207473
Traffic	-0.682947	-0.395162	0.587757	0.207473	1.000000

Table 2. Correlation matrix Cost indexes X Negative indexes



	Cost of Living	Rent	Health Care	Climate
Cost of Living	1.000000	0.833731	0.271370	0.066953
Rent	0.833731	1.000000	0.089006	0.078445
Health Care	0.271370	0.089006	1.000000	0.885980
Climate	0.066953	0.078445	0.885980	1.000000

Table 3. Correlation matrix Cost indexes X Positive indexes

The correlation varies from -1 to 1 with -1 being a strong negative correlation and 1 being a strong positive correlation. Cost of Living is weakly correlated to two of the negative indexes (Crime: -0.36 and Pollution: -0.32) and Rent is weakly correlated to each of the negative indexes (Crime: - 0.47, Pollution: -0.37, Traffic: -0.40). For the positive indexes the correlation is even weaker: Cost of Living (Health Care: 0.27, Climate: 0.07), Rent (Health Care: 0.09, Climate: 0.08).

Since the relationships among the variables are not trivial, to help in the decision of choosing a new city to move in, a clustering algorithm is applied to segment the cities according to the similarity of those indexes, allowing the grouping/characterization of the neighborhoods and a deeper comparison among them. K-Means, an unsupervised technique, is going to be used to this task since there are no labels to a previous training step.

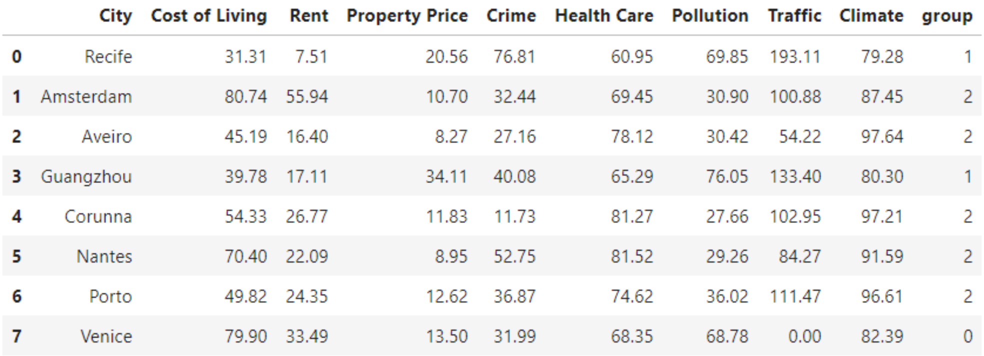
To allow K-Means to interpret equally features with different magnitudes and distributions, first, they are normalized. Then, an initial k of 3 is chosen. The idea is to identify one group of cities similar to the current city (Recife) and at least 2 other groups to have different types of options to choose between.

Once the city is chosen, its neighborhoods are clustered taking into consideration their venues. The idea is to group the neighborhoods according to the type of venues aiming the categorization of each group and the identification of similar neighborhoods, which, disposed in a map, will give to the seeker a simple tool for supporting his/her decision. The top venues within a radius of 1km in each neighborhood are obtained consulting the Foursquare API. The result is a json file that can be easily converted into a pandas DataFrame to be explored. For requesting the top venues, it is necessary the coordinates (latitude and longitude) of each neighborhood. This information was obtained from the city municipal webpage.

Finally, a choropleth map of Amsterdam neighborhoods is generated taking into consideration the safety index, which measures how safety each neighborhood is. This map is then merged with the map obtained in the previous step resulting in a final map that shows the neighborhoods segmented by venues similarity, their top ten venues, and their safety index. Thus, visualizing this map, one can decide which neighborhood to live.

4. Results

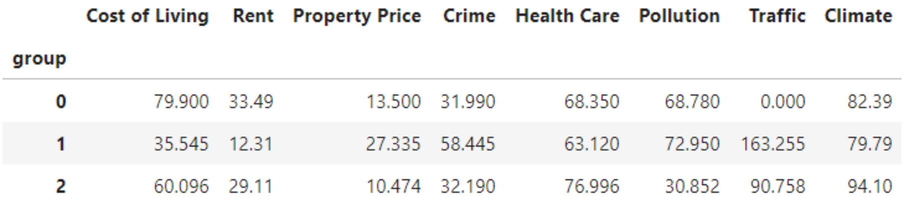
The K-Means with k=3 was applied to the data of Table 1 and the resulting groups can be seen in the following table.



	City	Cost of Living	Rent	Property Price	Crime	Health Care	Pollution	Traffic	Climate	group
0	Recife	31.31	7.51	20.56	76.81	60.95	69.85	193.11	79.28	1
1	Amsterdam	80.74	55.94	10.70	32.44	69.45	30.90	100.88	87.45	2
2	Aveiro	45.19	16.40	8.27	27.16	78.12	30.42	54.22	97.64	2
3	Guangzhou	39.78	17.11	34.11	40.08	65.29	76.05	133.40	80.30	1
4	Corunna	54.33	26.77	11.83	11.73	81.27	27.66	102.95	97.21	2
5	Nantes	70.40	22.09	8.95	52.75	81.52	29.26	84.27	91.59	2
6	Porto	49.82	24.35	12.62	36.87	74.62	36.02	111.47	96.61	2
7	Venice	79.90	33.49	13.50	31.99	68.35	68.78	0.00	82.39	0

Table 4. Sister cities clustered according to living indexes

Now it is possible to create a profile for each group considering the common characteristics of each cluster. Grouping Table 4 by the column group and taking the mean, we have the following:



	Cost of Living	Rent	Property Price	Crime	Health Care	Pollution	Traffic	Climate
group								
0	79.900	33.49	13.500	31.990	68.350	68.780	0.000	82.39
1	35.545	12.31	27.335	58.445	63.120	72.950	163.255	79.79
2	60.096	29.11	10.474	32.190	76.996	30.852	90.758	94.10

Table 5. The mean of each living indexes for each Sister cities' clusters

Venice is a peculiar city and it was already expected to be alone in a group. It is highly touristic, predominantly historic, which makes the rent high, and presents 0 traffic due to the canals and the blocking of vehicles. Recife and Guangzhou are cheap cities to live, but present high values of negative indexes and low values of positive indexes. The others are European cities with moderate to high cost indexes but that provide a notable life quality, presenting high values of health care and climate and low to moderate values of crime, pollution, and traffic.

- Group 0 (Venice): peculiar city with particular features;
- Group 1 (Recife and Guangzhou): cheap but mediocre quality of life;
- Group 2 (Amsterdam, Aveiro, Corunna, Nates, and Porto): moderate to high life cost but good life quality.

If you are not an artist or don't wish to work with tourism, Venice is not the best city to move in. On the other hand, Guangzhou presents similar problems to Recife. Thus, the chosen city should be from Group 2. Among them, Aveiro presents the lowest Cost Indexes, high values of positive indexes, and low values of negative indexes. However, other factors can influence the decision, like the job market and the size of the city. Thus, instead of choosing Aveiro, Amsterdam was the choice.

Now that the city is chosen, it is time to decide which neighborhood to live in. The 99 Amsterdam's neighborhoods data (name and coordinates) was obtained from the Amsterdam municipal webpage and disposed in the following map.

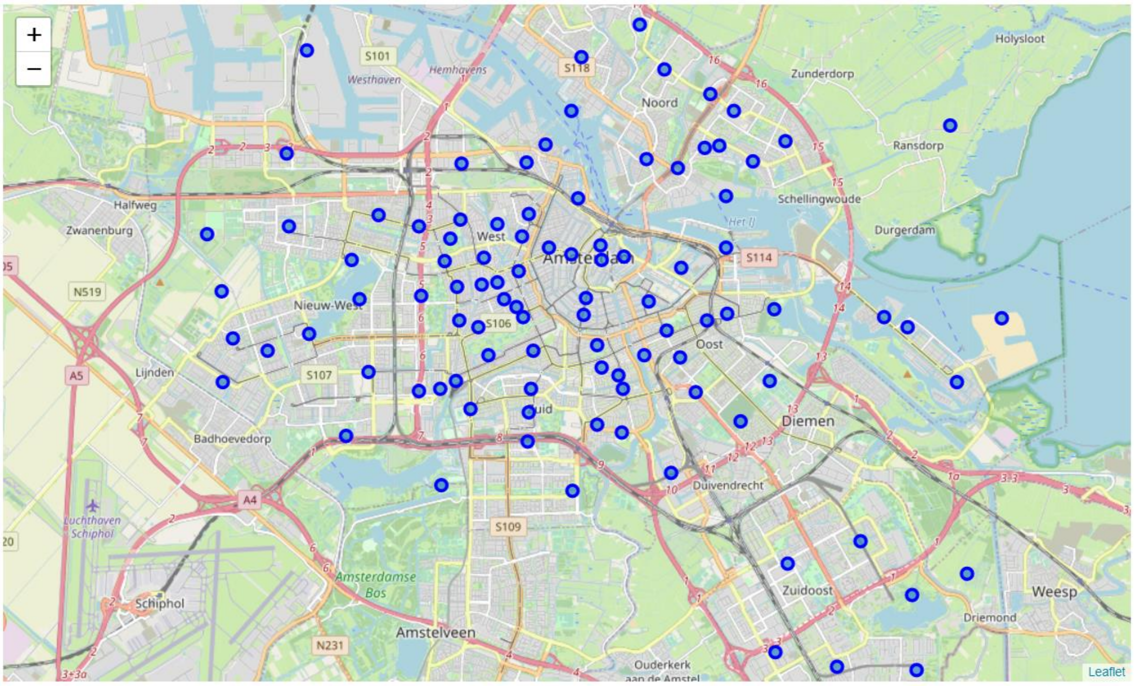
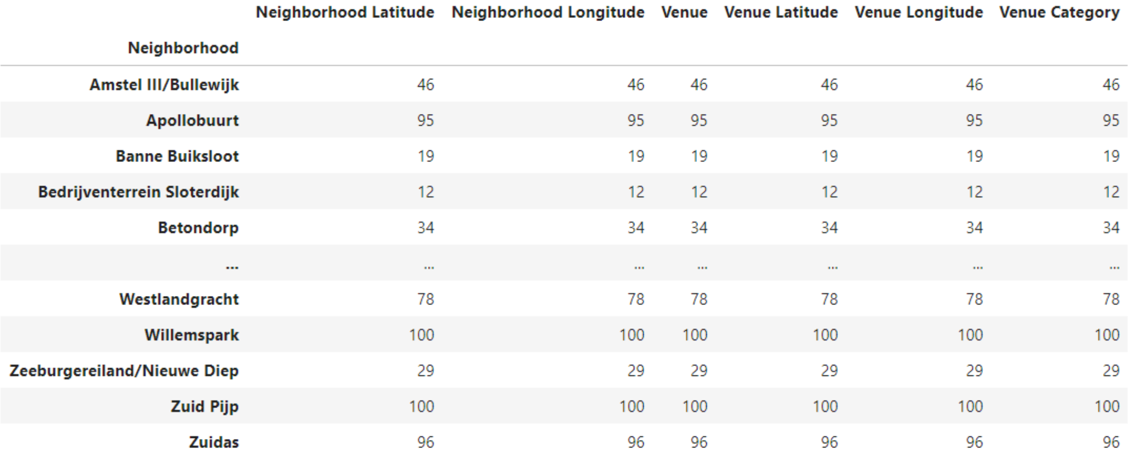


Figure 4. Map of Amsterdam and its neighborhoods

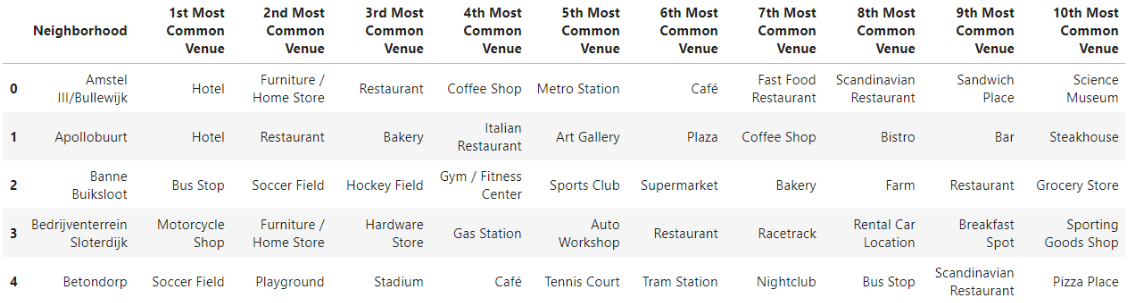
The top 100 venues of each neighborhood within a radius of 1 km are then obtained. The result does not present 100 venues for each neighborhood since some neighborhoods don't have 100 venues registered in Foursquare. The following table shows the number of venues returned for each neighborhood. A total of **319** unique venue categories were obtained.



	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Amstel III/Bullewijk	46	46	46	46	46	46
Apollobuurt	95	95	95	95	95	95
Banne Buiksloot	19	19	19	19	19	19
Bedrijventerrein Sloterdijk	12	12	12	12	12	12
Betondorp	34	34	34	34	34	34
...	...	...	...	...	...	...
Westlandgracht	78	78	78	78	78	78
Willemspark	100	100	100	100	100	100
Zeeburgereiland/Nieuwe Diep	29	29	29	29	29	29
Zuid Pijp	100	100	100	100	100	100
Zuidas	96	96	96	96	96	96

Table 6. The total number of venues per neighborhood

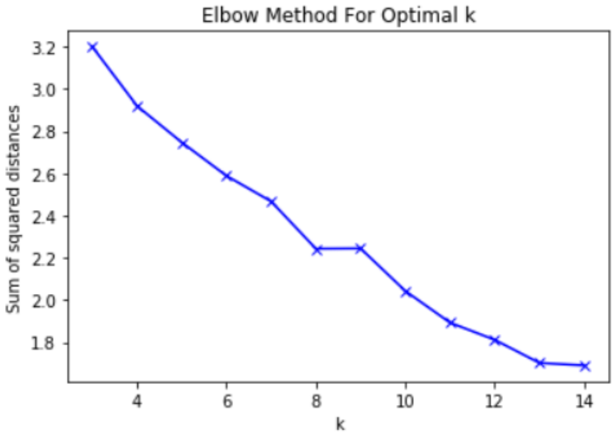
The frequency of those 319 venue categories for each neighborhood was calculated and a final table with the top 10 most common venue categories per neighborhood was generated to be used in the clustering of the neighborhoods given the top venue categories. The first five rows of that table can be seen below.



	Neighborhood	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	Amstel III/Bullewijk	Hotel	Furniture / Home Store	Restaurant	Coffee Shop	Metro Station	Café	Fast Food Restaurant	Scandinavian Restaurant	Sandwich Place	Science Museum
1	Apollobuurt	Hotel	Restaurant	Bakery	Italian Restaurant	Art Gallery	Plaza	Coffee Shop	Bistro	Bar	Steakhouse
2	Banne Buiksloot	Bus Stop	Soccer Field	Hockey Field	Gym / Fitness Center	Sports Club	Supermarket	Bakery	Farm	Restaurant	Grocery Store
3	Bedrijventerrein Sloterdijk	Motorcycle Shop	Furniture / Home Store	Hardware Store	Gas Station	Auto Workshop	Restaurant	Racetrack	Rental Car Location	Breakfast Spot	Sporting Goods Shop
4	Betondorp	Soccer Field	Playground	Stadium	Café	Tennis Court	Tram Station	Nightclub	Bus Stop	Scandinavian Restaurant	Pizza Place

Table 7. Top 10 most common venue categories per neighborhood

To segment the neighborhoods according to the similarity of venue categories KMeans was leveraged. To find the best K, the elbow method was applied. This method runs the KMeans for each K value in a specified range (in this case, from 3 to 15) and calculates a metric of accuracy for clustering. This metric can be the distance between data points and their cluster's centroid, which indicates how dense our clusters are. Then, those values were plotted and the elbow point (where the rate of decrease sharply shifts) was determined: k=8.



K-Means was then executed with k=8 and the neighborhoods were plotted (Figure 5) along with their groups.

To complement the previous map, information about the safety of each neighborhood was obtained to create a choropleth map. Amsterdam municipal page presents information about the Safety Index, which measures how safe a neighborhood is. The lower the index value the safer the neighborhood. The GeoJSON file that defines the areas/boundaries of Amsterdam neighborhoods and that is necessary for the creation of a choropleth map was also obtained from the Amsterdam municipal webpage.

The choropleth map was then merged with the map of the clustered Amsterdam neighborhoods. The top 10 venues were added to the label of each point to concise all the needed information in one map (Figure 6), facilitating the decision about which neighborhood to live.

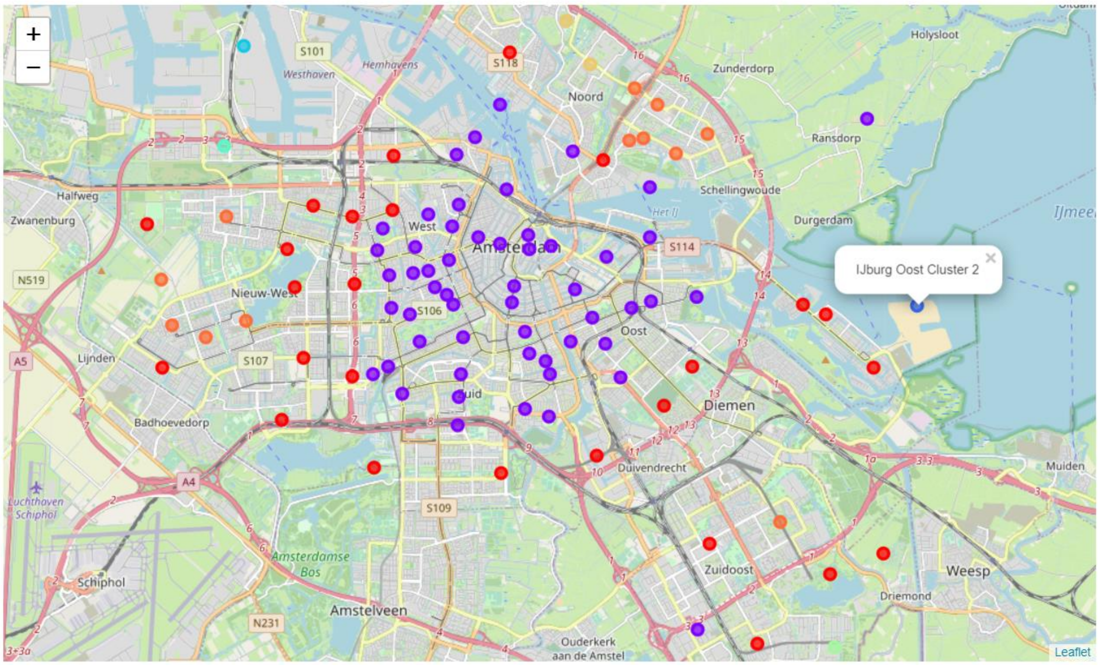


Figure 5. Amsterdam neighborhoods clustered by their top 10 venue categories

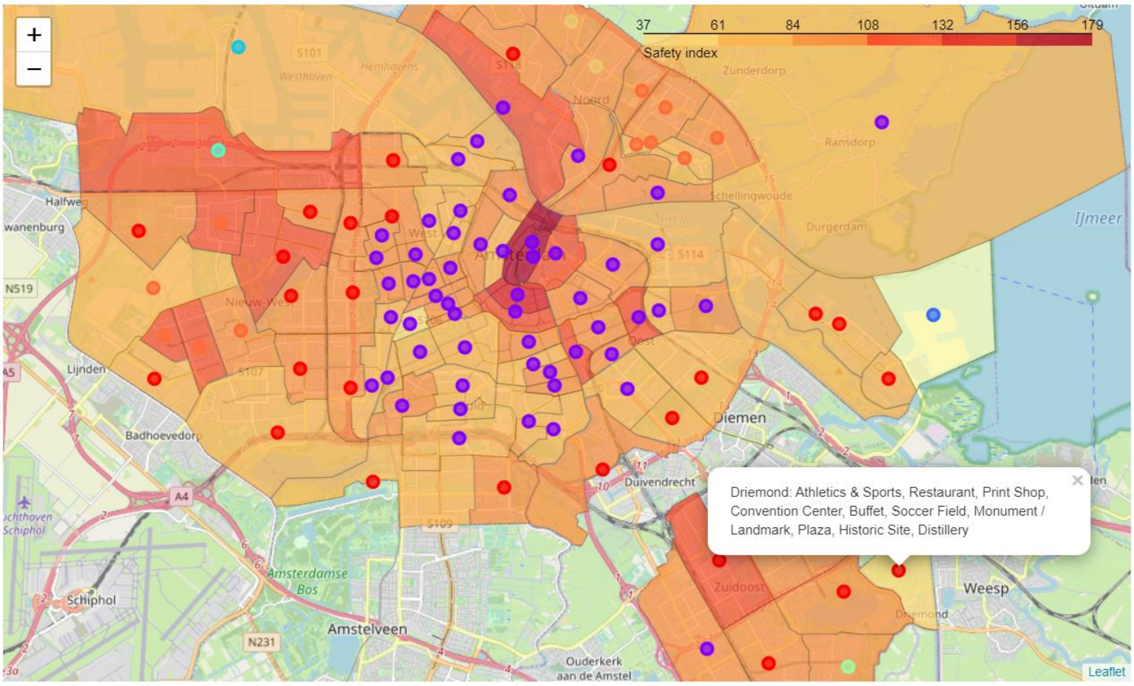

37
61
84
108
132
156
179
Safety index
Driemond: Athletics & Sports, Restaurant, Print Shop, Convention Center, Buffet, Soccer Field, Monument / Landmark, Plaza, Historic Site, Distillery
Leaflet

Figure 6. Amsterdam neighborhoods clustered but their top 10 most common venues "choroplethed" by Safety Index

5. Discussion

The clustering of the 99 neighborhoods considering the top 10 most common venue categories resulted in 8 groups that were categorized to facilitate the processes of choosing the most suitable neighborhood.

- Cluster 1 (26) - Area with many restaurants, cafes, and supermarkets. Some hotels but not many, indicating relative proximity of touristic areas. Some bars and clubs, indicating certain nightlife;
- Cluster 2 (54) - Many restaurants, cafes, and supermarkets. Many hotels and museums, indicating an area close to touristic attractions. Many bars and clubs suggesting a vivid nightlife. It seems to be the cultural center of the city;
- Cluster 3 (1) - Beach area;
- Cluster 4 (1) - Isolated neighborhood probably rich/expensive since it presents a heliport and a harbor;
- Cluster 5 (1) - Isolated regular area with venues specialized in vehicles;
- Cluster 6 (1) - Isolated area with nature-related attraction;
- Cluster 7 (2) - Area with venues for the practicing of different sports;
- Cluster 8 (13) - A regular residential area far from the center (the main spots are bus stops and tram stations), but with all basic stores (supermarket, bakery, restaurants, etc.). It also presents options for recreation and sports.

Given the clusters' classification, the choice depends on each person and his/her tastes and lifestyle. Here one of them is chosen as an example: the neighborhoods of **Cluster 1.** They present all sorts of attractions and utilities and are not in the center, although close to it, so it looks not so quiet nor not so noisy.

But the cluster 1 is composed of 26 neighborhoods. Which one to choose? The map of Figure 6 can be used to make this final decision since it shows the safety index and the top 10 most common venue categories for each neighborhood.

6. Conclusion

Moving abroad and choosing a new city to live is not easy. The present study applied data analysis and clustering to help in the decision processes of finding a new city and neighborhood to move in. First, the target city was chosen among the sister cities (cities that establish a bond of cooperation on many factors) of the current city (Recife, Brazil). They were compared and segmented taking into consideration indexes like the cost of living, rent, pollution, crime, health care, etc.

Second, after the most proper city being chosen (Amsterdam), it was time to choose the neighborhood. In this process, all the neighborhoods of Amsterdam were compared and segmented according to the most popular venues (restaurants, bars, museums, parks, beach, cinema, etc.) existent in each one. Each neighborhood group was labeled according to the type of venues, allowing a characterization of each group, and then displayed in a map to facilitate the decision.

Finally, a choropleth map of Amsterdam neighborhoods was generated taking into consideration the safety index, which measures how safety each neighborhood is. This map was then merged with the map obtained in the previous step resulting in a final map that shows the neighborhoods segmented by venues similarity, their top ten venue categories, and their safety index. Thus, visualizing this map, one can decide which neighborhood to live.