A battle of cities

Analyzing similarities between neighborhoods in two cities

MARCH 30

IBM Data Science Professional – Final Project
Authored by: Raúl Cano Argamasilla

Introduction

Business Problem

In the coming weeks, I am changing my job from Germany to Chile. Since I will need to relocate there, choosing a new apartment can be overwhelming given all the different districts in the new city, let alone the lack of knowledge of what it is like to live there.

Obviously, I would like to choose an area to live where I feel comfortable. An approximate solution is to find a neighborhood similar to the one I live in or any other I know in my current city, based on the services they offer.

By clustering neighborhoods in both cities, using their services as the features of the algorithm, we could find neighborhoods in the new city that are similar to the ones where I live currently. In this way, I will mitigate the uncertainty what I will find in my new location.

> ***"Finding similarities between neighborhoods in two cities can be a powerful tool for human resources departments when facing relocation of their employees worldwide"***

Audience

This situation is a very frequent one that employees all over the world have to face when they are relocated. If generalized to multiple cities, any human resources department could leverage the applicability of this algorithm to ease the on-boarding of employees moving to different cities.
In addition, I intend to use this project for my personal use to look for an apartment in Santiago.

Data

The needed data for this project is clear: postal codes of the cities to be analyzed.

Data processing

After an initial attempt to find location data on this two cities, there does not seem to be an official source with structured data. Therefore, the approach will be:

1. Extract the postal codes of each city.
2. Find the coordinates of each postal code using Nominatim or any other service offering such information.
3. Structure the extracted data coordinates properly

Data sources

The following sources of information for this phase have been identified:

Postal codes and city neighborhoods data

- https://www.geonames.org/postal-codes/
- For Santiago
  https://www.geonames.org/postalcode-search.html?q=santiago&country=CL
- For Darmstadt
  https://www.geonames.org/postalcode-search.html?q=Darmstadt&country=DE

Venues information

- https://developer.foursquare.com/

With this approach, I foresee that the solution is transformed in a general one with any tuple of cities. In order to execute successful query, we need:

- The name of the cities to search.
  For this data, one needs to type the names manually in the corresponding variable.
- The correct country codes that are used in the Geonames website, such as CL for Chile, or DE for Germany.
  As we see in the website, we can extract all available country codes by inspecting the dropdown menu.

```
<select name="country">
<option value="" selected=""> all countries</option>
<option value="DZ"> Algeria</option>
<option value="AD"> Andorra</option>
…
</select>
```

To obtain the postal codes, a query to Geonames including city and country, will output a HTML table with each postal code and coordinates in their rows, among other data. Our work will consist then in requesting such HTML and extract the relevant items from the table. For example, the URL https://www.geonames.org/postalcode-search.html?q=santiago&country=CL produces the following HTML table (I cut some rows to make visualization easier):

	PLACE	CODE	COUNTRY	ADMIN1	ADMIN2	ADMIN3
1	Santiago	8320000	Chile	Región Metropolitana	Provincia de Santiago	Santiago
	-33.454/-70.656					
2	Providencia	7500000	Chile	Región Metropolitana	Provincia de Santiago	Providencia
	-33.436/-70.609					
3	Las Condes	7550000	Chile	Región Metropolitana	Provincia de Santiago	Las Condes
	-33.421/-70.502					
4	Vitacura	7630000	Chile	Región Metropolitana	Provincia de Santiago	Vitacura

The data in that table has to be properly processed in order to obtain a similar table, where each line corresponds to one postal code and it lists its coordinates and other data in the same row. Here a glimpse of how the processed table should look like:

	POSTALCODE	BOROUGH	NEIGHBORHOOD	LATITUDE	LONGITUDE	CITY	COUNTRY
0	50823	Köln	Köln	50.951	6.926	Cologne	Germany
1	50667	Köln	Köln	50.939	6.955	Cologne	Germany
2	50668	Köln	Köln	50.950	6.963	Cologne	Germany
3	50670	Köln	Köln	50.950	6.950	Cologne	Germany
4	50672	Köln	Köln	50.944	6.936	Cologne	Germany

Methodology

In short, this project allows the user to specify any two cities all around the world, from which a visual grouping of its postal areas will be displayed according to the services on each one, such as shops, restaurants, etc.
That is, the services in the postal codes of each city will be analyzed and grouped in one of 5 clusters. Then, a map of each city will show the different clusters identified on each city, so the user can recognize quickly which areas are similar to each other.
This project offers a generic solution to any pair of cities as long as they are listed in the Geonames database:

Input

- *cities*: Dictionary with cities and country codes where the clustering will be done.
  For example:

```
cities = [
    {'city' : 'Cologne', 'country_code' : 'DE', 'country' : 'Germany',
'latitude' : '', 'longitude' : ''},
    {'city' : 'Santiago', 'country_code' : 'CL', 'country' : 'Chile',
'latitude' : '', 'longitude' : ''},

]
```

This algorithm will allow to enter as many cities as desired, but in this exercise we are focusing in only two: the city of origin and the city of destination.

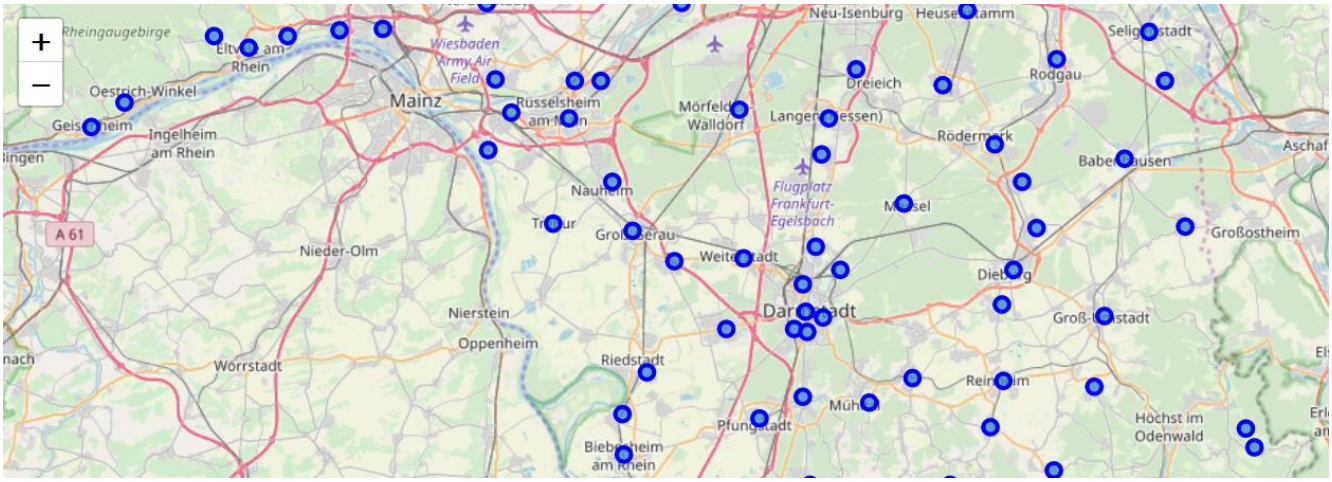
Output

- A list of clusters grouping areas in both cities, so one can inspect visually the similarity between them

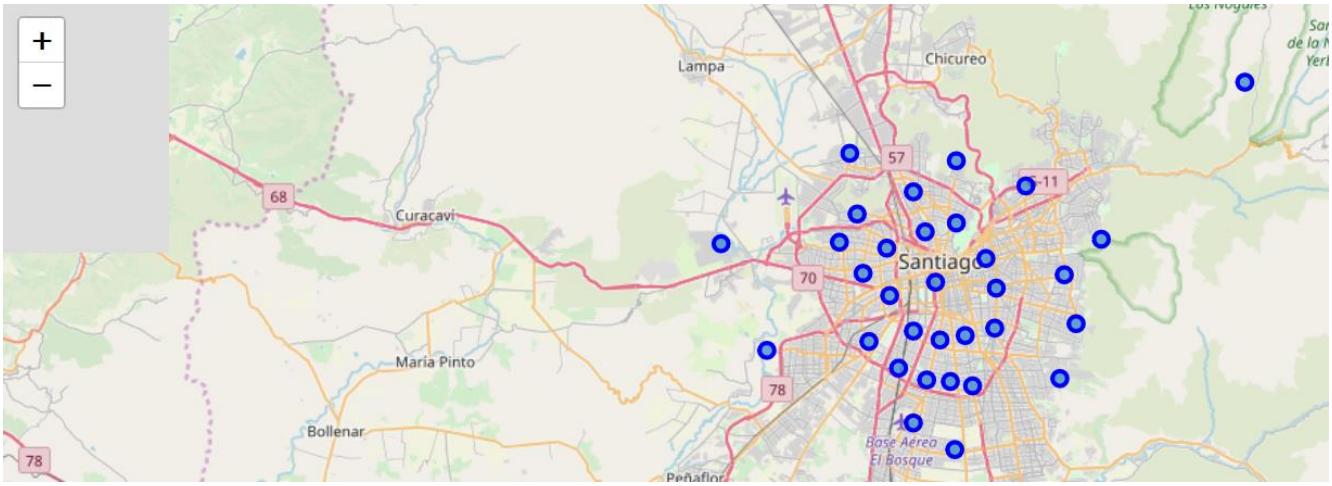
Clustering

The algorithm to cluster the different areas on each city is the K-means, with an initial selection of 5 clusters, though this can be changed to a different value within the code.

City of origin:



City of destination:



A first glimpse shows that the data we obtained includes the in the first city most of the surrounding areas of the city, not only the central neighborhoods. In principle, this should not affect the results, since it there is a big discrepancy between the populations of the two cities. That means that small populated areas outside the first city will be also be weighed in the algorithm.

After extracting the venues with the Foursquare API and grouping them to each neighborhood / postal code, we get a view of how they are distributed

	Neighborhood latitude	Neighborhood longitude	Venue	Venue latitude	Venue longitude	Venue category
ALSBACH-HÄHNLEIN	1	1	1	1	1	1
ALTENSTADT	4	4	4	4	4	4
BABENHAUSEN	5	5	5	5	5	5
BAD KÖNIG	3	3	3	3	3	3
BAD NAUHEIM	19	19	19	19	19	19
BAD ORB	4	4	4	4	4	4
BAD SCHWALBACH	6	6	6	6	6	6
BAD SODEN-SALMÜNSTER	4	4	4	4	4	4
BAD VILBEL	13	13	13	13	13	13

…

With further processing of the frequency of the venues, we build a dataframe including the most frequent categories per postal code.

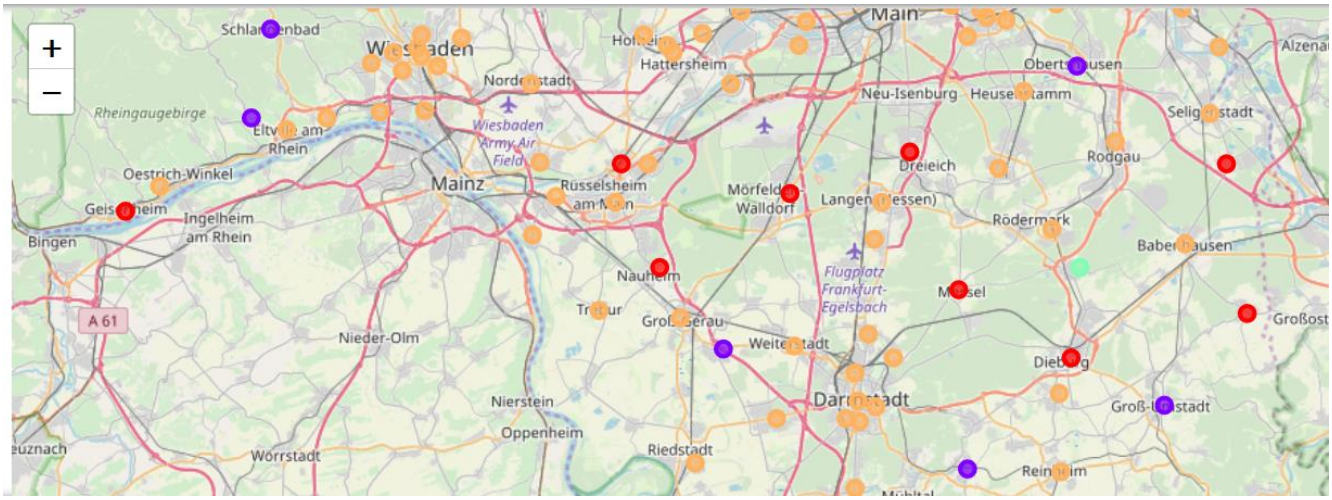
Postalcode	Borough	Neighborhood	Latitude	Longitude	City	Country	1st most common venue	2nd most common venue	3rd most common venue
60437	Frankfurt am Main	Frankfurt am Main	50.192	8.675	Darmstadt	Germany	Supermarket	Clothing Store	Ice Cream Shop
64291	Darmstadt	Darmstadt	49.911	8.657	Darmstadt	Germany	Supermarket	Café	Italian Restaurant
64297	Darmstadt	Darmstadt	49.819	8.645	Darmstadt	Germany	Supermarket	Café	Italian Restaurant
64283	Darmstadt	Darmstadt	49.872	8.648	Darmstadt	Germany	Supermarket	Café	Italian Restaurant

Finally, by running the clustering algorithm, we get a cluster label from 0 to 4, that will be the base for our resulting maps, allowing us to visualize the similarities among neighborhoods.

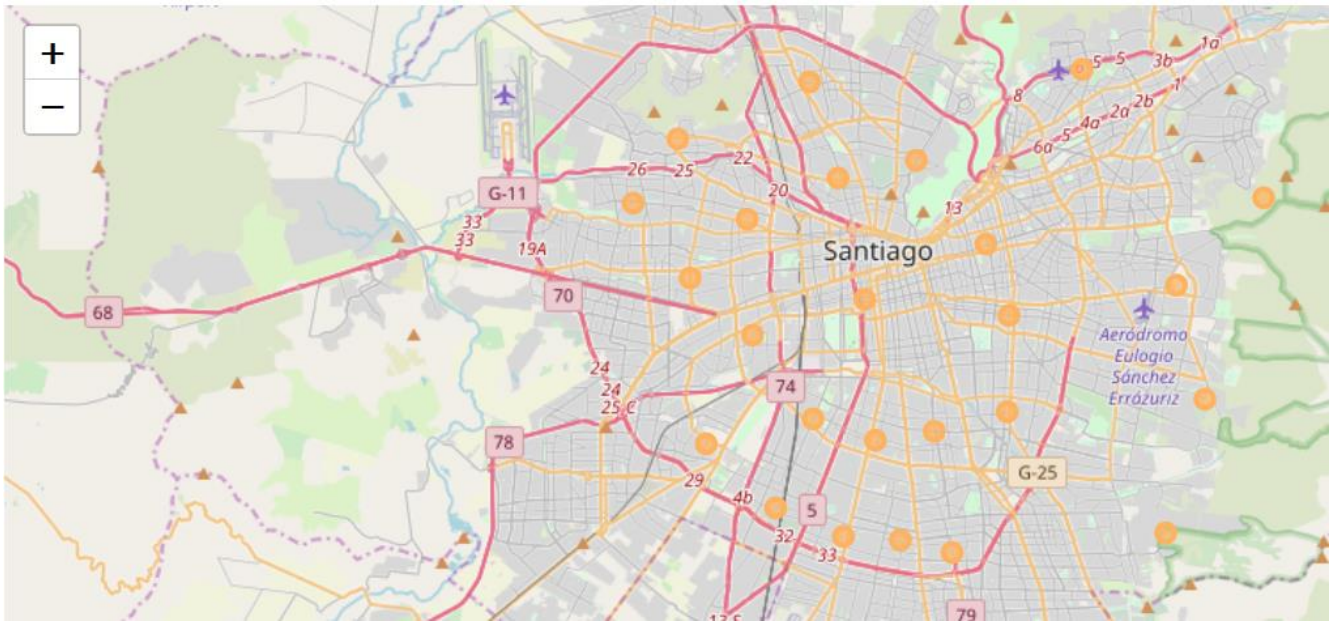
Results

After running the algorithm to plot each area with colors per cluster, we get the following maps:

City of origin:



City of destination:



Cluster 1 - Groceries

This first cluster has captured areas where basic services are more prominent, such as supermarkets.

	Borough	City	Country	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
16	Dreieich	Darmstadt	Germany	0.0	Supermarket	Hotel	German Restaurant	Italian Restaurant	Gastropub	Gas Station	Drugstore	Farmers Market	Farm	Yoga Studio
20	Dieburg	Darmstadt	Germany	0.0	Supermarket	Gas Station	Bank	Yoga Studio	Exhibit	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service
33	Pfungstadt	Darmstadt	Germany	0.0	Brewery	Supermarket	Fast Food Restaurant	Train Station	Event Space	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market
48	Altenstadt	Darmstadt	Germany	0.0	Train Station	Supermarket	Gas Station	Yoga Studio	Event Space	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service
57	Schöneck	Darmstadt	Germany	0.0	Supermarket	Yoga Studio	Football Stadium	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service	Fast Food Restaurant
59	Usingen	Darmstadt	Germany	0.0	Plaza	Supermarket	Gas Station	Café	Exhibit	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service
74	Nidda	Darmstadt	Germany	0.0	Supermarket	Drugstore	Plaza	Movie Theater	German Restaurant	Event Space	Food	Flower Shop	Flea Market	Financial or Legal Service
76	Fürth	Darmstadt	Germany	0.0	Chinese Restaurant	Supermarket	Train Station	Greek Restaurant	Yoga Studio	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service
78	Niedernhausen	Darmstadt	Germany	0.0	Supermarket	Bakery	Gas Station	Fast Food Restaurant	Train Station	Yoga Studio	Food & Drink Shop	Food	Flower Shop	Flea Market
82	Messel	Darmstadt	Germany	0.0	Supermarket	Soccer Field	Gastropub	Bank	Yoga Studio	Food Truck	Food & Drink Shop	Food	Flower Shop	Flea Market
83	Schaafheim	Darmstadt	Germany	0.0	Supermarket	Coffee Shop	Bar	Yoga Studio	Exhibit	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market

Cluster 2 - Hotels

This second grouping collects the areas mainly of hotels and other services related to dining or eating.

	Borough	City	Country	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
15	Friedrichsdorf	Darmstadt	Germany	1.0	Hotel	Supermarket	German Restaurant	Train Station	Drugstore	Bakery	Pet Store	Financial or Legal Service	Event Space	Farmers Market
18	Bensheim	Darmstadt	Germany	1.0	Hotel	Train Station	Drugstore	Café	Supermarket	Greek Restaurant	Plaza	Water Park	Hobby Shop	Electronics Store
36	Ober-Ramstadt	Darmstadt	Germany	1.0	Hotel	Gym / Fitness Center	German Restaurant	Train Station	Hobby Shop	Yoga Studio	Farm	Exhibit	Falafel Restaurant	Fast Food Restaurant
37	Groß-Umstadt	Darmstadt	Germany	1.0	Hotel	Vineyard	Café	Liquor Store	Drugstore	Plaza	Dive Bar	Falafel Restaurant	Food & Drink Shop	Food
56	Nidderau	Darmstadt	Germany	1.0	Hotel	Italian Restaurant	Asian Restaurant	Supermarket	Yoga Studio	Exhibit	Food Court	Food & Drink Shop	Food	Flower Shop
60	Kronberg im Taunus	Darmstadt	Germany	1.0	Hotel	Auto Dealership	Ice Cream Shop	German Restaurant	Lounge	Pizza Place	Restaurant	Castle	Café	Supermarket
66	Obertshausen	Darmstadt	Germany	1.0	Bakery	Hotel	Insurance Office	Supermarket	Fast Food Restaurant	Event Space	Food Court	Food & Drink Shop	Food	Flower Shop
99	Schmitten	Darmstadt	Germany	1.0	Hotel	Pharmacy	Cafeteria	Bar	Yoga Studio	Exhibit	Food & Drink Shop	Food	Flower Shop	Flea Market
103	Langenselbold	Darmstadt	Germany	1.0	Hotel	Bakery	Mexican Restaurant	Supermarket	Park	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service
112	Büttelborn	Darmstadt	Germany	1.0	Hotel	Sporting Goods Shop	Yoga Studio	Event Space	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service	Fast Food Restaurant
118	Schlangenbad	Darmstadt	Germany	1.0	Hotel	Pool	Park	Elementary School	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service	Fast Food Restaurant
124	Eppstein	Darmstadt	Germany	1.0	Light Rail Station	Hotel	Café	Castle	Trail	Bar	Exhibit	Food & Drink Shop	Food	Flower Shop

Cluster 3 – German dining

This cluster has captured the areas with a disproportionate presence of German restaurants, but also with other food related services, like “Food & Drink Shops”.

	Neighborhood	City	Country	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
38	Fischbachtal	Darmstadt	Germany	2.0	German Restaurant	Yoga Studio	Football Stadium	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service	Fast Food Restaurant
71	Maintal	Darmstadt	Germany	2.0	German Restaurant	Ice Cream Shop	Bakery	Football Stadium	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service
88	Hasselroth	Darmstadt	Germany	2.0	German Restaurant	Yoga Studio	Football Stadium	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service	Fast Food Restaurant
122	Idstein	Darmstadt	Germany	2.0	German Restaurant	Hotel	Restaurant	Scenic Lookout	Drugstore	Turkish Restaurant	Grocery Store	Dive Bar	Department Store	Flower Shop
131	Steinbach (Taunus)	Darmstadt	Germany	2.0	German Restaurant	Ice Cream Shop	Supermarket	Event Space	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service	Fast Food Restaurant
197	Fränkisch-Crumbach	Darmstadt	Germany	2.0	Construction & Landscaping	German Restaurant	Yoga Studio	Exhibit	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service

Cluster 4 – For kids

This cluster, capturing only 3 areas, seems to be good for children, since it has plenty of playgrounds and elementary schools, along with a fair share of food shops.

	Latitude	City	Country	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
58	50.230	Darmstadt	Germany	3.0	Playground	Arcade	Turkish Restaurant	Event Space	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service
84	49.951	Darmstadt	Germany	3.0	Playground	Yoga Studio	Elementary School	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service	Fast Food Restaurant	Farmers Market
181	49.739	Darmstadt	Germany	3.0	Playground	Yoga Studio	Elementary School	Food & Drink Shop	Food	Flower Shop	Flea Market	Financial or Legal Service	Fast Food Restaurant	Farmers Market

Cluster 5 – Mixed urban areas

Finally, the cluster capturing the biggest number of neighborhoods, seems to be the one related to city internal services, like supermarkets, cafes or bakeries.

	Borough	Longitude	City	Country	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	Frankfurt am Main	8.675	Darmstadt	Germany	4.0	Supermarket	Clothing Store	Ice Cream Shop	Gas Station	Italian Restaurant	Café	Bus Stop	Light Rail Station	Convenience Store	Sushi Restaurant
1	Darmstadt	8.657	Darmstadt	Germany	4.0	Supermarket	Café	Italian Restaurant	Bakery	Ice Cream Shop	Bus Stop	Tram Station	Gas Station	Sushi Restaurant	Coffee Shop
2	Darmstadt	8.645	Darmstadt	Germany	4.0	Supermarket	Café	Italian Restaurant	Bakery	Ice Cream Shop	Bus Stop	Tram Station	Gas Station	Sushi Restaurant	Coffee Shop
3	Darmstadt	8.648	Darmstadt	Germany	4.0	Supermarket	Café	Italian Restaurant	Bakery	Ice Cream Shop	Bus Stop	Tram Station	Gas Station	Sushi Restaurant	Coffee Shop
4	Darmstadt	8.681	Darmstadt	Germany	4.0	Supermarket	Café	Italian Restaurant	Bakery	Ice Cream Shop	Bus Stop	Tram Station	Gas Station	Sushi Restaurant	Coffee Shop
5	Darmstadt	8.645	Darmstadt	Germany	4.0	Supermarket	Café	Italian Restaurant	Bakery	Ice Cream Shop	Bus Stop	Tram Station	Gas Station	Sushi Restaurant	Coffee Shop
6	Darmstadt	8.637	Darmstadt	Germany	4.0	Supermarket	Café	Italian Restaurant	Bakery	Ice Cream Shop	Bus Stop	Tram Station	Gas Station	Sushi Restaurant	Coffee Shop
7	Frankfurt am Main	8.634	Darmstadt	Germany	4.0	Supermarket	Clothing Store	Ice Cream Shop	Gas Station	Italian Restaurant	Café	Bus Stop	Light Rail Station	Convenience Store	Sushi Restaurant
9	Griesheim	8.572	Darmstadt	Germany	4.0	Bakery	Light Rail Station	Café	Falafel Restaurant	Yoga Studio	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market

Discussion

After checking the results, we can say that the grouping has been disappointing, especially concerning city of destination. All the target postal areas have been grouped in the same cluster, something not specific enough to make a well informed decision to relocate somewhere else.
Further analysis shows that the reason for this is the lack of data on this city, that is, the ratio of available entries from the city of origin to the city of destination favors the former, which produces the following effect: neighborhoods in the first city are interestingly grouped, but in the second one, they are not.
It seems like this is a consequence an algorithm that is flexible enough to include almost any city (as long as it´s included in the Geonames database), at the cost of not being too specific if the data in that database does not go too deep in the areas of the city.
When running the algorithm using other pair of cities of which we find similar amount of data in the Geonames site, the areas of both cities seemed to be more relevantly grouped.

Conclusion

An algorithm for clustering neighborhoods in one city based on its venues, as we saw in a previous assignment, can be extended to more than one city, as to show the similarities between neighborhoods in any pair of cities.
This can be a powerful tool for human resources departments when facing relocation of their employees worldwide, since they could show visually to their employees how their favorite neighborhoods are similar to other areas in the city of destination. In this way, HR can mitigate the uncertainty of what employees will find in the new location.

We have created a flexible algorithm using the Geonames database for neighborhood coordinates data and put it to work with two cities: Darmstadt (Germany) and Santiago (Chile). The results in this instance have not been as useful as expected, since the data for the city of destination was not precise enough for
Even though the clustering seems like a good idea for this purpose, the data needed is still the most relevant piece on this puzzle and further research should be done to guarantee the best results.