Clustering cities with more than 100.000 inhabitants based on their infrastructure (Foursquare venues)

Capstone Project - The Battle of Neighborhoods

Diego Carrasco Gubernatis

February 21, 2019

1. Introduction

1.1 Background

Families moving to a new country face many hardships, from understanding the culture, learning the language, finding a job and finding a place to live. To ease the migration they would like to find a city similar (or dissimilar) to the one they were living in their home country, with a similar or better infrastructure and quality of life. It would be helpful to have a comparison between the most important cities in their home country and they destiny, to find out which are similar.

1.2 Description of the problem

With so many alternatives and so many variables is easy to get lost and with so much information it's easy to misunderstand what the data is telling you. It would be great if there was a comparison of the different cities in a country of choice which gives you a general view of each place and it's pro's and contras.

Here the problem will be for a family to decide a city in Germany to move from a south-american country (Chile) to a European country (Germany), and the family would like to have a priority list of the places which better satisfy their requirements.

The requirements are:

- Schools
- Coffee

- Hospitals
- Nightlife
- Playground
- Shops
- Entertainment
- Restaurant

1.3 Interest

The findings of such an analysis would be of interest for families moving between Chile and Germany, and also for companies expanding their businesses to new cities.

2. Data Acquisition and cleaning

2.1 Data sources

I fount a list of the cities and they coordinates in Wikidata trough a query for data on cities with over 100.000 population, with labels and coordinates, which you can find here

I obtained the venues for each place using the Foursquare API for each city in both countries. I also saved this data as a csv file which can you can find here

To get the venues I created a new python package with a function which uses the *foursquare* package, then cleans the data and returns a pandas dataframe. Tha package is called foursquare*api*tools and can be found on its repository in GitHub here

2.2 Cleaning of the data

The data downloaded from Wikidata had to be cleaned by splitting the coordinates column into Longitude and Latitude, dropping *na* values and creating a new dataset with only the cities of Chile and Germay.

Then I got the venues for each city which was already cleaned by the function in the package I created, as it returns, in a dataframe, Name, City, Country, Latitude, Longitude, Category and Address from Foursquare. When getting the venues I added additional columns for country *(country2)*, city *(city2)* and search query *(query)* because I needed that information for the cluster analysis and foursquare returns empty data and differently written names (they are written by users). With the 2 additional columns I could get a proper dataset for the analysis.

After cleaning the data I had 2 Countries, 84 Cities and 31589 venues using Foursquare:
* 2072 venues in Chile
* 29517 venues in Germany
* 346 unique categories

3. Methodology

After creating the new dataset with cities form Chile and Germany I plotted the cities on a map using Folium to check if the coordinates were ok.

3.1 Venue categories for each city

After cleaning all the data and preparing the final datasets, I created a “One Hot encoding” for each category for each city base on the categories Foursquare returned and *city2* column I added. I added this column because, for example, for Antofagasta (a city in Chile) there were 4 different ways the users wrote its name, and it was the same for each of the 84 cities. Adding the *cities2* column allowed me to have the same city label for each venue.

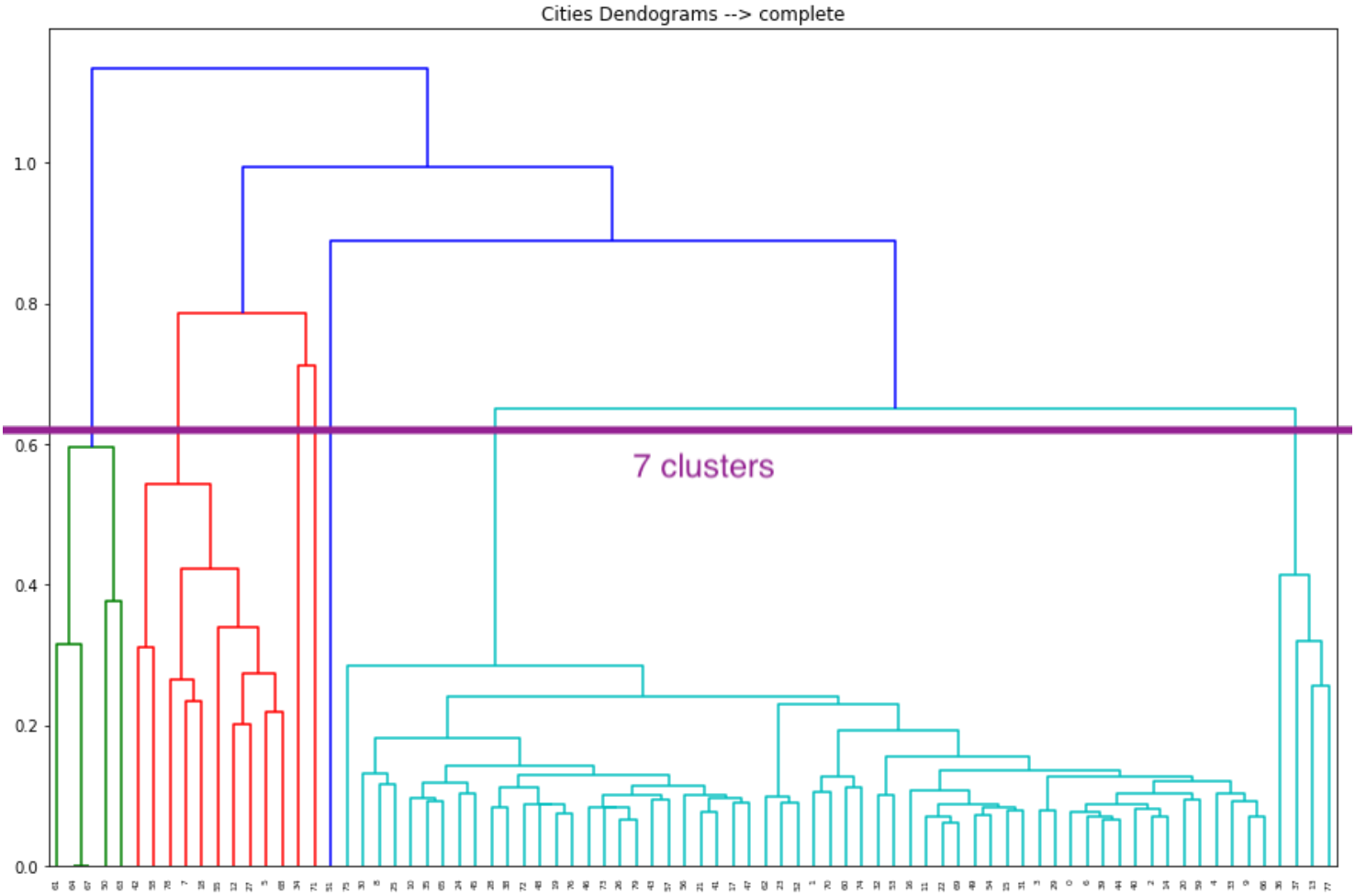
After doing the “One Hot encoding” I had a dataset *cities*onehot_ with shape (31589, 347), which I grouped by City with the category mean.

Then I created a new dataframe *city_venues_sorted* with the most common venues for each city.

3.2 Cluster Analysis

I used K-Means Cluster Analysis with K=5 and Hierarchical Cluster Analysis with 7 clusters and got similar results, I then plotted which on a Folium Map with Colors for each cluster.

I created a Dendogram with several linkage methods to see how the cities group. I used 7 clusters with the complete linkage method as per the following picture:



3.3 Comparing results

For K-Means Analysis I got the following results

- K-means for 0 has 58 cities
- K-means for 1 has 12 cities
- K-means for 2 has 3 cities
- K-means for 3 has 4 cities
- K-means for 4 has 3 cities

With Hierarchical Clustering with complete Linkage I got, for 7 Clusters:

- Hierarchical for 0 has 5 cities
- Hierarchical for 1 has 10 cities
- Hierarchical for 2 has 58 cities
- Hierarchical for 3 has 1 cities
- Hierarchical for 4 has 4 cities

3.4 Analyzing the results

I will analyze the most frequent venues for the 7 Hierarchical Clusters:

Cluster 0

venue	frequency
school	0.98
university	0.02

Cluster 1

venue	frequency
restaurant	0.31
shops	0.28
nightlife	0.15

Cluster 2

venue	frequency
restaurant	0.23
shops	0.23
nightlife	0.21

Cluster 3

venue	frequency
shops	0.40
restaurant	0.32
coffee	0.13

Cluster 4

venue	frequency
shops	0.65
restaurant	0.21
coffee	0.05

Cluster 5

venue	frequency
restaurant	0.30
university	0.20
shops	0.18

Cluster 6

venue	frequency
school	0.64
university	0.36

4. Conclusions

In this study I analysed the main cities in 2 countries based on infrastructure (venues) I got from Foursquare API.

K-Means and Hierarchical Clustering gave similar results, and because of the Dendogram for Hierarchical Cluster with *complete linkage* I decided for 7 clusters, even if cluster number 3 has only one city: “Mülheim an der Ruh” in Germany.

These city groups can be very helpful for families searching for a similar city in a new country and for companies looking for similar cities to expand.

5. Future directions

In this study, I compared the cities using foursquare data for the venues for only 2 countries. It would be very

interesting to compare all the biggest cities in the world also analysing variables such as rent cost, income and safety. It’s hard to get all that information for all the cities, as not all the countries have open data policies.

I also used 60 results in a 2000 meters radius for each city when looking for venues, and the results may change when using more data in a bigger radius.