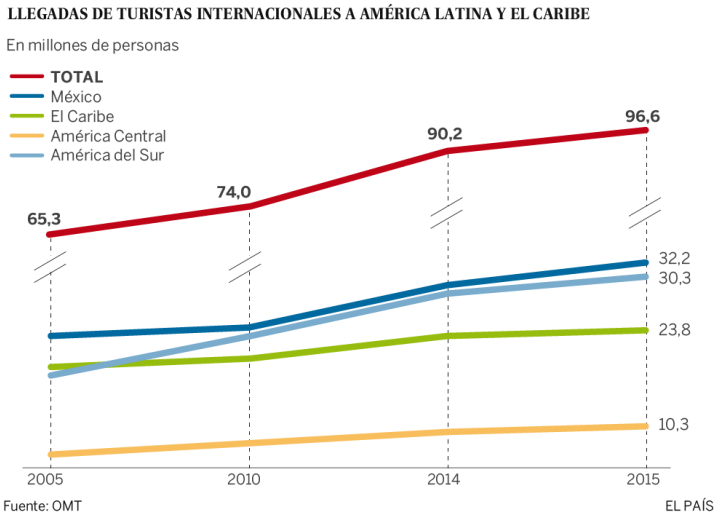
Similar cities in South America

1. Introduction

1.1 Description of the problem

The problem would be to determine which cities in South America are similar, so that it is an additional information for tourists or agencies, when making a trip.



According to the graph in recent years the trend in the number of tourists visiting South America is increasing, which is why it is necessary to present additional information that may be of interest when visiting a city.

1.2 Interested in the project.

One of the interested parties would be travel agencies, since they could recommend South American cities to tourists that are similar to what they are looking for at a lower price.

The tourists themselves, as they could build an application with the data from Foursquare, and make comparisons of cities, for example, according to restaurants, hotels, museums, etc

2. Data

The two main sources are:

2.1 Foursquare:

Offers the Place Database, you can access accurate and up-to-date data on places of community origin. The large selection of rich location data and releases the potential to enhance your application or website with the ability to describe locations, analyze trends, and improve the user experience. With the API that Foursquare offers we'll get the data from restaurants, plazas, museums, etc., in the cities we're interested in analyzing.

2.2 GeoData:

This is a website (https://www.geodatos.net/poblacion), where it shows the ten most populated cities for each country, as well as the longitude and latitude, for example, for Peru, my country.

País	Ciudad	Población	Coordenadas
Perú	Lima	7,737,002 habitantes	-12.043, -77.028
Perú	Arequipa	841,130 habitantes	-16.399, -71.535
Perú	El Callao	813,264 habitantes	-12.057, -77.118
Perú	Trujillo	747,450 habitantes	-8.116, -79.03
Perú	Chiclayo	577,375 habitantes	-6.771, -79.841
Perú	Iquitos	437,620 habitantes	-3.749, -73.254
Perú	Huancayo	376,657 habitantes	-12.065, -75.205
Perú	Piura	325,466 habitantes	-5.194, -80.633
Perú	Chimbote	316,966 habitantes	-9.085, -78.578
Perú	Cusco	312,140 habitantes	-13.523, -71.967

Figure 1: 10 most populated cities in Peru
Source: GeoData

To solve the problem, we are only interested in the countries of South America:

- Peru
- Colombia
- Ecuador
- Argentina
- Chile
- Uruguay
- Paraguay
- Brazil
- Bolivia
- Venezuela

3. Methodology:

3.1 Web scraping:

Web scraping is a technique used by software programs to extract information from websites. Usually, these programs simulate human navigation on the World Wide Web either by using the HTTP protocol manually, or by embedding a browser in an application. This technique will be used on the website https://www.geodatos.net/poblacion, where the data mentioned above will be obtained.

3.2 K-means

K-means is an unsupervised classification (clustering) algorithm that groups objects into k groups based on their characteristics. The grouping is done by minimizing the sum of distances between each object and the centroid of its group or cluster. The quadratic distance is usually used.

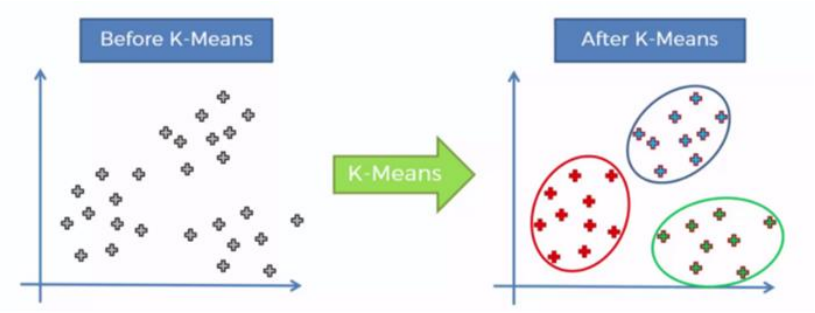


Figure 2: K-means

The cluster number can be determined by inertia

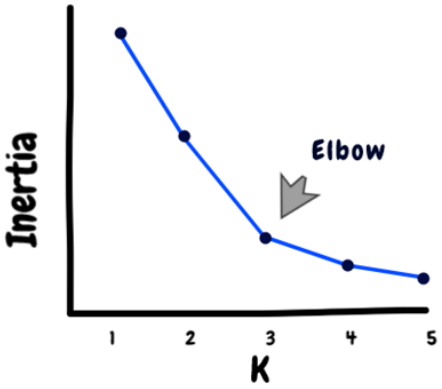


Figure 3: choice of the number of groups through inertia

4.1 Exploratory análisis

In the following chart you can show all the cities you are going to buy, to determine which cities are similar to each other.

South American cities

Figure 4: analyzed cities in South America

In addition, the following image shows the number of populations by each city, in which it is appreciated that countries like Peru the most populated cities are concentrated in the north, on the contrary, in Brazil the most populated cities are dispersed

Population of the cities

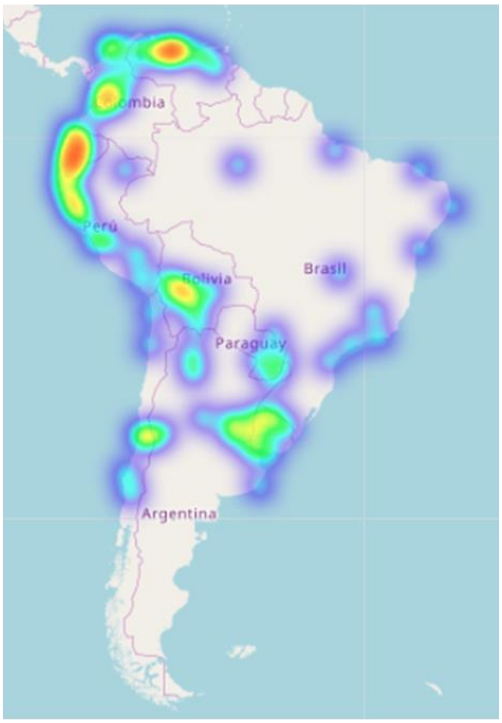


Figure 5: Population of the analysed cities

The most populated city is in Argentina followed by Brazil and Peru.

	pais	ciudad	poblacion	coordenadas	num_poblacion	latitude	longitude
80	Argentina	Buenos Aires	13,076,300 habitantes	-34.613, -58.377	13076300	-34.613	-58.377
40	Brasil	São Paulo	10,021,295 habitantes	-23.548, -46.636	10021295	-23.548	-46.636
0	Perú	Lima	7,737,002 habitantes	-12.043, -77.028	7737002	-12.043	-77.028
20	Colombia	Bogotá	7,674,366 habitantes	4.61, -74.082	7674366	4.61	-74.082
41	Brasil	Río de Janeiro	6,023,699 habitantes	-22.906, -43.182	6023699	-22.906	-43.182

4. Results

Data from Foursquare, all places such as restaurants, squares, etc., from all cities obtained through web scraping. The places that stand out are the restaurants, hotels and squares for, public places.

	Venue Category
Restaurant	329
Hotel	309
Pizza Place	257
Café	222
Ice Cream Shop	198
Plaza	189
Bar	186
Coffee Shop	178
Bakery	165
Burger Joint	133

You have selected 6 clusters according to the inertite graph.

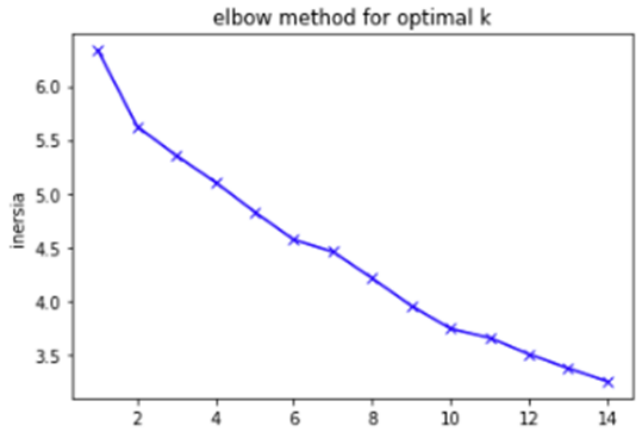


Figure 6:number of cluster whit inertia

In which it was obtained that the cluster 4 has more cities followed by the cluster 0, in addition it is appreciated that 3 cities exist that are totally different to the others, they could be considered atypical cities.

	Cluster Labels
4	69
0	24
3	4
5	1
2	1
1	1

The following groupings were obtained for the most important cities in South America, where it can be seen that the cities with the characteristics of cluster 4 are present in all the countries of the continent. It is highlighted that the two atypical cities are located in Bolivia, this means that tourists looking for new things could be recommended Bolivia, and another of the atypical cities is located in Uruguay.

Figure 7: Cluster of the cities

5. Discussions and recommendations

You could reduce the number of clusters and see the results, as we got three cities in one group. In addition, it would be grouped by means of oreo algorithm like DBscan, Hierarchical Cluster.

It would also be possible to compare these cities with the rest of the world and observe similarities between different continents, since the cultural differences are well marked.

According to the results it is recommended, for example, to the tourists, to visit Bolivia since they present cities with different characteristics, for the tourists who are already used to certain places and do not want to change they are recommended Brazil, Chile or Colombia.

It is recommended to generate an application that can compare cities and obtain those that are similar by grouping kmeans clusters and Foursquare data.

6. Conclusion

In conclusion, there are two well-defined groups, and the rest would indicate cities that are interesting to analyze.

Generating this type of alternatives in more detail would be interesting for tourists who know nothing about South America.