Clustering capitals in Brazil

José Marcos

May 8, 2019

1. Introduction

1.1 Background

An ever-increasing number of destinations worldwide have opened up to, and invested in tourism, turning it into a key driver of socio-economic progress through the creation of jobs and enterprises, export revenues, and infrastructure development. Over the past six decades, tourism has experienced continued expansion and diversification to become one of the largest and fastest-growing economic sectors in the world. In 2017 a total of 1.326 billion international tourist arrivals were recorded in destinations around the world. International tourism receipts reached US$ 1,340 billion in that year. Considering this context, a market sector that can be very profitable is travel agencies.

1.2 Problem

Something that can be useful for those agencies is clustering the most important cities in a country, which are usually the destination of tourists, based on their predominant categories of venues so their clients could better decide which cities to visit. Considering it, this work aims to cluster the capitals of each state in Brazil, 27 in total, using data from the Foursquare API, but note that this could be generalized to any country.

1.3 Interest

Obviously, travel agencies would be very interested in presenting their clients a nice summary of the cities in a country, because it would improve their service and consequently increase their profits.

2. Data

2.1 Data sources

The only data we need besides those provided by Foursquare are the latitudes and longitudes of the capitals and it can be found here in json format, but containing all cities in Brazil.

2.2 Data cleaning

After downloading the data, it was converted into a Pandas DataFrame containing the following columns: latitude, longitude, name, capital (Boolean) and two useless ones. I’ve used the column capital to get only the cities that are capital of a state, which had True as values. After removing some columns and renaming others I’ve finally got the final dataset in the format shown below.

Table 1: Final dataset format of capitals' latitude and longitude

City	latitude	longitude
Aracaju	-10.90910	-37.0677
Belém	-1.45540	-48.4898
Belo Horizonte	-19.91020	-43.9266
Boa Vista	2.82384	-60.6753
Brasília	-15.77950	-47.9297

3. Methodology

3.1 Foursquare data

The data needed from Foursquare was collected by using the endpoint *explore*, which returns a list of recommended venues near a location. For each capital, were returned the top 100 recommended venues within a radius of 10 km, considering their latitude and longitude. After gathering the data, I decided to remove venues whose category was one of the following: Brazilian Restaurant, Restaurant, Bar, Café, Coffee Shop, Gym / Fitness Center, Pharmacy, Gym, Hotel, Ice Cream Shop. It was done because those categories are very common in any city, hence they wouldn’t be so useful for distinguishing the cities.

To use the data for training a cluster algorithm I had to convert the venue category column into a numerical feature, which was done by applying one hot encoding. After that, the rows were grouped by city and by taking the mean of the frequency of occurrence of each category, resulting in a dataset ready to be clustered as shown below.

Table 2: Dataset to be clustered

City	Acai House	Accessories Store	Adult Boutique	…
Aracaju	0.016667	0.000000	0.0	…
Belo Horizonte	0.000000	0.000000	0.0	…
Belém	0.029851	0.000000	0.0	…
⋮	⋮	⋮	⋮	⋮

3.2 K-means clustering

A number of clusters (k) has to be specified to train a k-means model, but the algorithm is somewhat naïve, it clusters the data into k clusters even if k is not the right number of clusters to use. Therefore, users need some way to determine whether they are using the right number of clusters when using k-means. It can be done by using the elbow method, whose idea is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 10), and for each k calculate the Sum of Squared Distances (SSD). Then, a line chart of the of the SSD for each value of k is plotted. If the chart looks like an arm, then the "elbow" on the arm is the value of k that is the best. The line chart obtained is shown below.

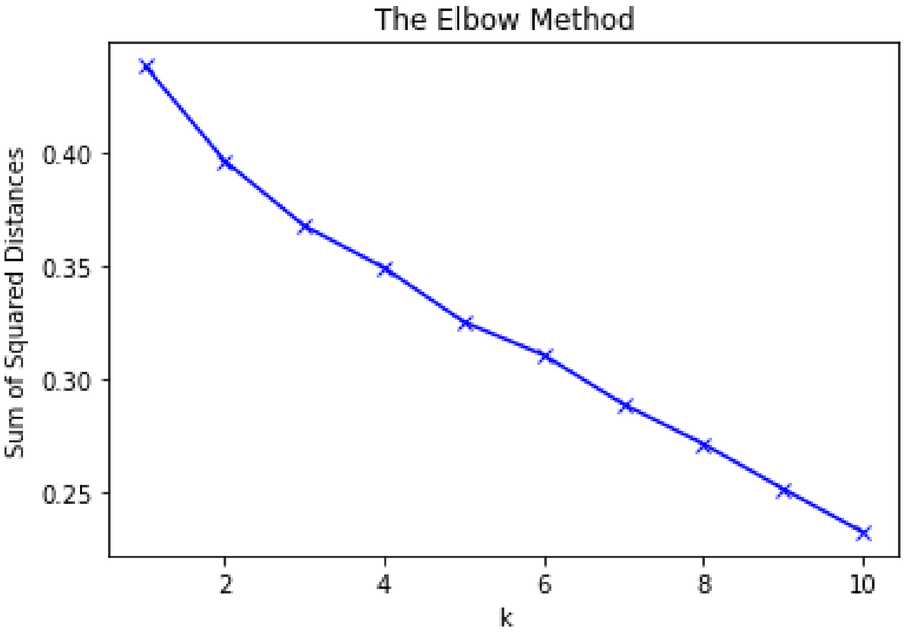

Figure 1: Line chart for the Elbow Method

Unfortunately, the result of this method was not conclusive, so I decided to assign the number of regions in Brazil, which is 5, to k.

4. Results

4.1 Visualizing the clusters

The map of Brazil was plotted, using the Folium library, with the capitals superimposed on top and their respective cluster identified by a color.

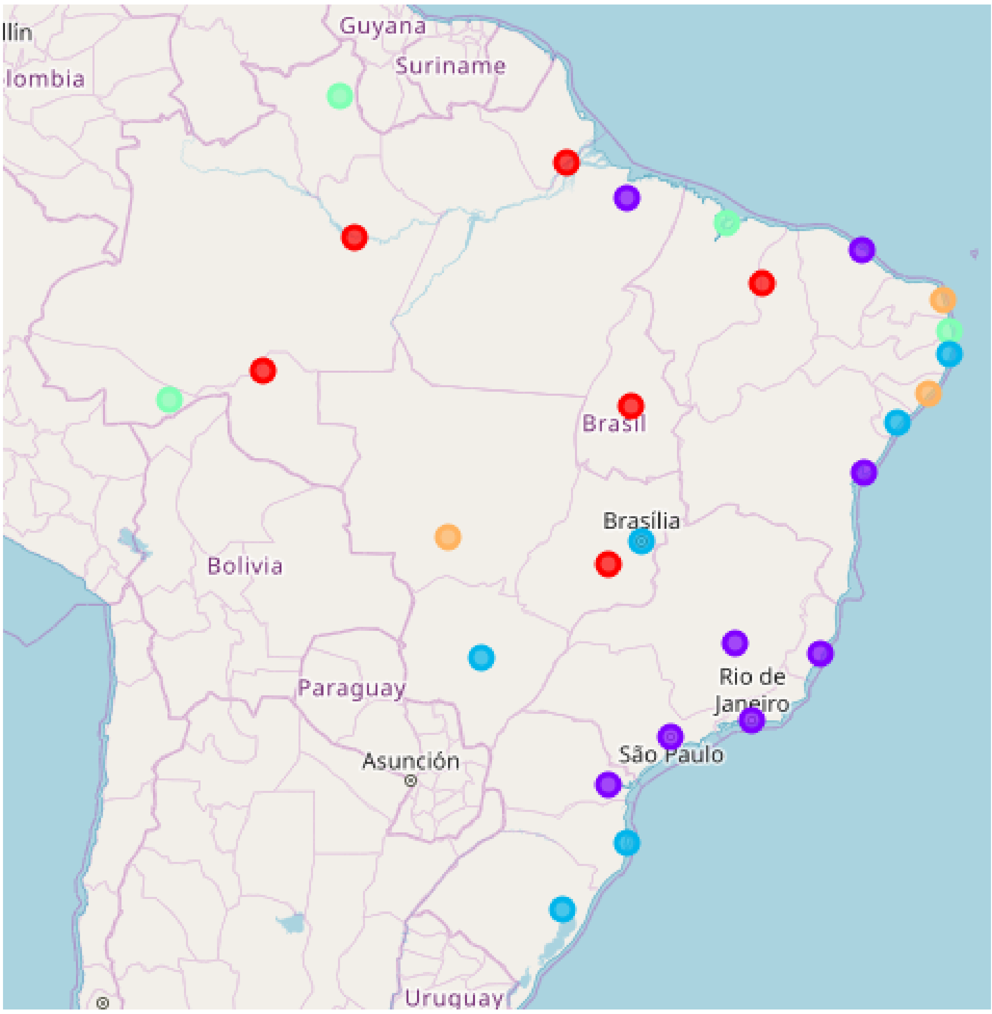

Figure 2: Map of Brazil showing the clusters obtained

4.2 Top 5 most common venues among the cluster

After the clustering process I was able to find the top 5 most common venues within each cluster, which can help summarizing the main characteristics of them.

Table 3: Top 5 most common venues of each cluster

<table>
<tr><th colspan="2">Cluster 1</th></tr>
<tr><td><b>Capitals</b></td><td>Goiânia, Macapá, Manaus, Palmas, Porto Velho, Teresina</td></tr>
<tr><td><b>Top 5 most common venues</b></td><td>Pizza Place, Snack Place, Dessert Shop, Burger Joint, BBQ Joint</td></tr>
<tr><th colspan="2">Cluster 2</th></tr>
<tr><td><b>Capitals</b></td><td>Belém, Belo Horizonte, Curitiba, Fortaleza, Rio de Janeiro, Salvador, São Paulo, Vitória</td></tr>
<tr><td><b>Top 5 most common venues</b></td><td>Italian Restaurant, Theater, Plaza, Pizza Place, Park</td></tr>
<tr><th colspan="2">Cluster 3</th></tr>
<tr><td><b>Capitals</b></td><td>Aracaju, Brasília, Campo Grande, Florianópolis, Porto Alegre, Recife</td></tr>
<tr><td><b>Top 5 most common venues</b></td><td>Burger Joint, Pizza Place, Bakery, Italian Restaurant, Northeastern Brazilian Restaurant</td></tr>
<tr><th colspan="2">Cluster 4</th></tr>
<tr><td><b>Capitals</b></td><td>Boa Vista, João Pessoa, Rio Branco, São Luís</td></tr>
<tr><td><b>Top 5 most common venues</b></td><td>Bakery, Plaza, Seafood Restaurant, Snack Place, Burger Joint</td></tr>
<tr><th colspan="2">Cluster 5</th></tr>
<tr><td><b>Capitals</b></td><td>Cuiabá, Maceió, Natal</td></tr>
<tr><td><b>Top 5 most common venues</b></td><td>Dessert Shop, Italian Restaurant, Bakery, Burger Joint, Japanese Restaurant</td></tr>
</table>

5. Discussion

The results obtained were very interesting, although there were some clusters that captured the similarities within an actual region in Brazil, there were others that were very heterogeneous regarding these regions. It suggests that cities located in the same region and that are thought to share the same culture can be, actually, very different.

Considering the table of most common venues within a cluster of the last section, we can point out the main aspects of each cluster and use it as a guide for clients deciding which cities to visit. That is shown in the table below.

Table 4: Main aspects of each cluster

	Recommendation
Cluster 1	Where you can try a really good meat
Cluster 2	Italian culture and theaters
Cluster 3	Diversified gastronomy
Cluster 4	It’s a must If you like seafood
Cluster 5	Very good dessert shops and Japanese restaurants

6. Conclusion

In this work, I analyzed the similarities of the capitals of each state in Brazil with the intention of clustering them to generate a summary that could help clients of travel agencies to decide which cities to visit. For this purpose I used the K-means algorithm with the number of clusters being the number of regions in Brazil. I found very useful insights showing that cities that belong to the same region are not necessarily similar to each other. As the final result, I could generate a recommendation guide that can be used by clients to choose the cities according to their preferences.

To improve this work data from other sources besides Foursquare could be used, because there are many other characteristics that a client would like to know about a place that he is intending to visit, such as attractions, prices, security and so on.