Clustering capital cities from Brazil

Author: Rodrigo Caetano Batista

Introduction/Business Problem

Brazil is a very large country and many enterprises have units in different cities, mostly capital cities, that have more money circulating. Almost all selective processes for new employees happens countrywide and the select must be allocated to one unit of the enterprise.

Moving from one city to the other can be difficult for a person, specially if accompanied by many other changes. In order to minimize the effects on the participant's behavior, enterprise tend to allocate employers in cities with some characteristics in common with the original city. One of this characteristic can be the venues on the city.

When the head of the select process allocate the new employees, a good way to do so would be to evaluate which city resembles the participant's original city the most. In order to help this allocation process, we will cluster the capital cities from Brazil. With the results, employers will be able to allocate the workers in a strategic way.

Data definition

To answer the business problem and deliver a practical guide to help employers to allocate new workers, we will make use of the Foursquare list of venues in each capital city from Brazil. It is supposed that cities with similar venues are similar in many other ways. Also for the employee that has to move to other city, he will be more comfortable if the new city have similar venues to his original city.

In addition, the latitudes and longitudes from each capital centrum was scrapped from a website used by GPS to improve map's precision.

Data acquisition and Analysis

Using the requests library, we searched in Foursquare for the 1000 venues closer to the city center from each city. The longitude and latitudes from each city was extracted from another website: "gps.pesquiza.com", this site is used by GPS devices to get those informations, so it is a reliable refference.

After that, we made a list of the ten most common venue categories in each city. This list was used to cluster the cities, those cluster will help in the decision making for the head of the selective process.

Methodology - Sample

At first, the latitudes and longitudes from every capital were scraped from the internet.

Our data comprehends the coordinates from 27 cities. Their coordinates can be accessed on “https://gps.pezquiza.com/apontamento-de-antena/latitude-e-longitude-das-capitais-brasileiras-para-usar-no-localizador-de-satelites/”.

Methodology - Sample

After that, we got the venues near those coordinates, from the Foursquare website. For that we used the request library. We also scrapped for the category of each venue. At the end we had a sample of approximately 500 venues from each city, as the top rows of the DataFrame below indicates:

:[257]:

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Aracaju – Sergipe	-10.5440	-37.0418	Museu da Gente Sergipana	-10.917090	-37.047771	History Museum
1	Aracaju – Sergipe	-10.5440	-37.0418	IL Sordo Gelateria	-10.930200	-37.050115	Ice Cream Shop
2	Aracaju – Sergipe	-10.5440	-37.0418	Forneria	-10.938256	-37.052529	Bakery
3	Aracaju – Sergipe	-10.5440	-37.0418	Sorveteria Castelo Branco	-10.932860	-37.079808	Ice Cream Shop
4	Aracaju – Sergipe	-10.5440	-37.0418	Parque da Sementeira	-10.943792	-37.053600	Park

Methodology - Cleaning the data and Analysis

In order to analyze the data, we got the number of each category that the city has and ordered them to get the 10 categories each city had the most. That will be the data used for making the clusters.

We clustered the data with the KMeans algorithm from the Scikit learn package, from Python. We defined the number of clusters as 5 because in Brazil there are 5 big regions and we hoped that by making those clusters we would achieve a good result.

Results

The results are shown in the map below. We can see that the clusters were spread on the territory which is good because big companies tend to have units in cities near each other (so this way we probably have a city in every cluster that has a unit). The only cluster that comprised only one region was the blue one.

Discussion

As highlighted in the results section, the clusters were well distributed on the entire map. As enterprises tend to have more units in the middle and southeast area, we hope that they can find a city in the same cluster as the city the employee is from. That is likely to happen for almost all cities, except those in blue, from the northeast.

Luckly, one city (Recife - Pernambuco) from the blue cluster has a huge industrial environment, so we hope that the enterprises have a unit there as well. If it doesn't happen, we can run the algorithm again and try to fit those cities in another cluster.

Conclusion

The main goal of this work was to cluster capital cities from Brazil to make employees that have to move when they are accepted in a job experience less changes. When the employee moves to a city very different from their own, it would take time before he gets used to the new environment, and that can cause a decrease in his performance on the job.

The clusters made here can be a good guide for allocating the new employees. We hope employers take in account the information here, and for more guidelines or services in this line, they can contact the author by his linkedin profile.

THANK YOU!