Clustering capitals in Brazil

José Marcos

May 6, 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 acquisition and cleaning

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