Clustering sea cities of Spain

Alla Sidko

February 26, 2020

1. Introduction

1.1 Background

Every year or a few times a year almost all people have the same problem and the same dream. All the life people wait for vacation for rest and for relax. And the best way to relax is to go to the sea. All of us, at least once in a lifetime, dream about lying on the beautiful beach under the sun. Thus, every time we decide to go to sea, we have a problem with the choice of country and city.

As a rule, we want to go to a new place, to watch and try something new, but we don't know where our rest would be the best and would not cost so expensive. If you decide to travel to the Europe, one of the most popular sea country is Spain. Summer resorts and beaches were the first to be developed in Spain, and today, generate the most income for the Spanish economy. The mild climate during the whole year and the extensive sandy beaches of the Mediterranean and Atlantic Ocean as well as of its two archipelagoes have been attracting tourists from all of the world.

The most popular Spanish mainland coasts are on its Mediterranean side. There are two most visitable beach sides - Costa Brava and Costa Daurada, which are not so far from Barcelona. But for this work I want to explore Costa Brava. There are three main towns in this seaside : Lloret de Mar, Tossa de Mar, and Blanes. I would provide clustering of each of these towns to specify which town is the most suitable for different people and their needs.

1.2 Problem

There are a lot of websites and videos with information about Lloret de Mar, Tossa de mar, and Blanes, but often it is only general description without information about accommodation, stores, beaches and other facilities. And there is a problem to choose the most appropriate

town, that can satisfy your needs and possibilities. This project aims to describe three similar sea towns in Spain and which town would be suitable for different needs.

1.3 Interest

This information would be useful for people, who decide to go to the sea and can not choose a city. I've chosen such groups of people: family with children, couples, aged people, friends, people who want calm vacation or vice versa. I'm going to investigate which city is the most suitable for each group.

2. Data

2.1 Data sources

To solve this problem I will use the Foursquare location data. Foursquare is the most trusted, independent location data platform for understanding how people move through the real world. With Places Database, I can access precise, up-to-date community-sourced venue data. To gain my purpose I need data for Lloret de Mar, Tossa de mar, and Blanes. In every city I will explore their hotels, apartments, restaurants, cafes, groceries, markets, beaches, bus stations, train station, parks,, night life, playgrounds. I separately got all categories that I need for my exploration and joined into one table. So I choose such categories:

1. Hotel
2. Food
3. Food&Drink Shop
4. Bus Stop
5. Bus Station
6. Train Station
7. Spa
8. Nightlife
9. Outdoors&Recreation
10. Arts&Entertainment

2.2 Data Selection

I got data about venues for each of these cities in radius 1500 m. Then I stored it into dataframe and have such columns: *'id', 'name', 'categories', 'referralId', 'hasPerk', 'location.address', 'location.lat', 'location.lng', 'location.labeledLatLngs', 'location.distance', 'location.postalCode', 'location.cc, 'location.city', 'location.state', 'location.country', 'location.formattedAddress', 'location.crossStreet', 'venuePage.id'*. Column *categories* is in JSON format, so I select only name and filter the category for each row. After that I keep columns that include venue name, and anything that is associated with location and clean column names by keeping only last term. I suppose, that for my project aim I don't need columns as *postalCode, cc,* and *crossStreet,* so I decided to delete them. At the end I have this list of columns:

Column Name	**Type**
name	*object*
categories	*object*
lat	*float64*
lng	*float64*
distance	*int64*
city	*object*
state	*object*
country	*object*
formatted Address	*object*
address	*object*
Labeled LatLng	*object*
id	*object*

2.3 Data cleaning

Some rows can have *None* value in column “categories”, so I have to delete them first. Because for my exploration it is the main field and it can not be *None.* Also, some data have *NaN* value in another columns, but it don’t have influence on my research and I can leave such rows in tables.

As a result, I have three tables with venues of selected categories for each of towns. After data cleaning and data preparation I have such amount data:

- LLoret de Mar - 92 rows × 12 columns
- Tossa de Mar - 61 rows × 12 columns
- Blanes - 88 rows × 12 columns

And next I will analyze all of venues of these towns and compare them.