Report on

Migrating to Spain – Exploring Similar Neighborhood

Table of Contents

1. Introduction

Migrating from one city to another is many a times a hectic process. New place, new people, new culture, and most importantly, new neighborhood. So exploring the new place is, thus, a new beginning from square one. It would really help one if he/she could find the amenities or restaurants or the venues just like the ones in their current location, in the city where they are migrating.

Here, I am assuming that I am migrating from my current city, Pune, India to city of Madrid, Spain. In this capstone, I will attempt to apply the techniques learned throughout the Data Science courses to explore the neighborhoods in the capital of Spain that is city of Madrid.

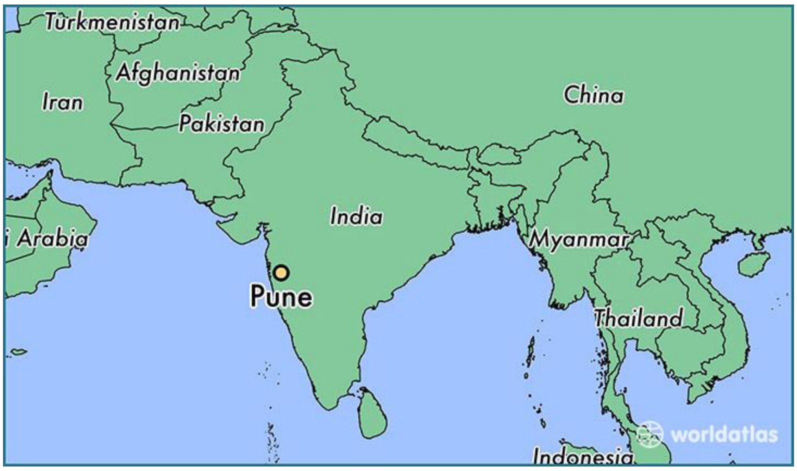
I will acquire my places of interest in my current location using the Foursquare API. I will then use the same API and explore the similar kind of venues in the city of Madrid.

2. Data

Let us discuss the data that I will be using for this project.

2.1 Data for current location

As discussed in the introduction, my current location is the city called 'Pune' in India.



Coordinates for Pune: 18.5203062 73.8543185

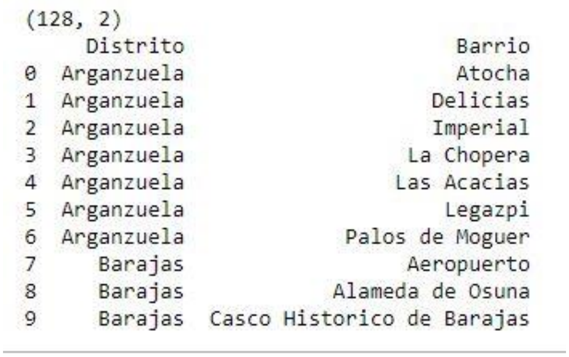
By using the Fousquare API with its explore endpoint and limiting the result to 80 venues and radius as 1000, I was returned with the following result:

```
47 venues were returned by Foursquare.
There are 27 unique categories.
```

2.2 Data for city of Madrid

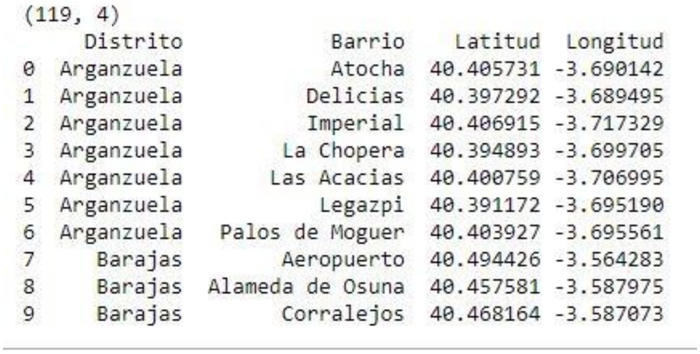
Now let us get the data for neighborhoods in Madrid. For that, I am using the data from Portal de datos abiertos del Ayun- tamiento de Madrid. Download the csv file titled Relación de barrios (superficie y perímetro).

This file is a list of 128 districts and neighborhoods called as ‘Distrito and Barrio’ in Madrid. Following are the first ten records from the file:



```
(128, 2)
     Distrito                     Barrio
0  Arganzuela                     Atocha
1  Arganzuela                   Delicias
2  Arganzuela                   Imperial
3  Arganzuela                 La Chopera
4  Arganzuela                Las Acacias
5  Arganzuela                    Legazpi
6  Arganzuela            Palos de Moguer
7     Barajas                 Aeropuerto
8     Barajas            Alameda de Osuna
9     Barajas  Casco Historico de Barajas
```

Let’s get their geo coordinates. For that I am using **Nominatim** from **geopy.geocoders**. The list returned 119 records. Here are the coordinates for the first ten records:



```
(119, 4)
     Distrito           Barrio    Latitud  Longitud
0  Arganzuela           Atocha  40.405731 -3.690142
1  Arganzuela         Delicias  40.397292 -3.689495
2  Arganzuela         Imperial  40.406915 -3.717329
3  Arganzuela       La Chopera  40.394893 -3.699705
4  Arganzuela      Las Acacias  40.400759 -3.706995
5  Arganzuela          Legazpi  40.391172 -3.695190
6  Arganzuela  Palos de Moguer  40.403927 -3.695561
7     Barajas       Aeropuerto  40.494426 -3.564283
8     Barajas Alameda de Osuna  40.457581 -3.587975
9     Barajas       Corralejos  40.468164 -3.587073
```

By using the Fousquare API with its explore endpoint and limiting the result to 80 venues and radius as 1000, I was returned with the following result:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category

Atocha	40.4054769	-3.68979999	Bodegas Rosell	40.403802520	-3.6906202941	Spanish Restaurant
Atocha	40.4054769	-3.68979999	Only You Hotel Atocha	40.407160659	-3.6884378646	Hotel
Atocha	40.4054769	-3.68979999	Running	40.40671358	-3.686904474	Sporting Goods
…	…	…	…	…	…	…

The dataframe is 3446 rows and 254 unique categories.

3. Methodology

To find similar neighborhoods in Madrid to my neighborhood in Pune city, we use kmeans to group similar neighborhoods in cluster. One of these cluster includes neighborhood in Pune city, so other neighborhoods in the same cluster will be similar to mine. First I transformed the data so all attributes are numeric. For this purpose I followed the following steps:

Let us transform data into numeric form so that we can apply Kmeans. For this purpose, follow the following steps:

- Put together the location data of State with those of the neighborhoods of Madrid (variable geo_barrios)
- Collect the data of places of interest of State with those of Madrid (variable madrid_venues)
- Use "onehot encoding" to transpose the categories of the places of interest and convert them to numerical values
- Group the resulting matrix by neighborhood, using the average value of each category
- Applying kmeans using clusters (10)

It turned out that best value for K was 4.

4. Analysis

For each neighbourhood, look for what are the 12 most frequent venue categories.
For limiting the candidate neighborhoods further, I used the distance for all points in cluster 2, which is the one that neighborhood in Pune city, to its centroid. Then I picked the closest neighborhoods in Madrid to mine.

- Distances of all the points to the cluster where the Pune neighborhood is located
- Adding cluster label to each point
- Sorting rows by cluster and distance
- Keeping only row for the cluster (where the Pune neighborhood is located
- Get the index of Pune in this last dataframe
- Select the neighborhoods closest to the one in Pune, according to the distances to the centroid

5. Result and Conclusion

The result of this entire project shows that by using this model, it is possible to help people who have to move to another city and want to find conditions similar to those in their current residential location using public data available through the Foursquare API.

One of the difficulties in kmeans algorithms is the choice of the value for K. To decide what value to use, we executed the algorithm with different K values and, for each case, and calculated the Silhouette Coefficient and the Calinski-Harabaz Index.
These are 2 metric that allow us to decide if we obtain dense and well separated clusters. With both indicators the best value for K was 4.

The characteristics that distinguish my neighborhood, according to the results of Foursquare, is the diversity of places to eat, shops and places to exercise. These same characteristics are present in almost all the selected neighborhoods.