IBM Capstone Project

Travel Agency Tour Recommendation

Levan Gvalia

2020

Introduction

After enabling visa free entrance to EU from Georgia, in addition to introduction of cheap and popular airlines, tourism abroad has become much more available to masses than It ever was. As an analyst at Travel Agency in Georgia, I clearly see result of visa free travel and cheap airlines – more people tend to favor cheap and frequent travels.

The Travel Agency was focused on more expensive tours, with client tailored tour recommendations – the information of which was gathered manually by employees, through online searches and word of mouth. The problem is that, with recent changes, employees can't keep up with the requests of cheaper and more frequent travels, thus causing client churn rate to skyrocket. As company is not willing to give up on its main advantage over competition – client tailored tour recommendations – as well as miss an opportunity of cheap and frequent flights, some solution has to be offered.

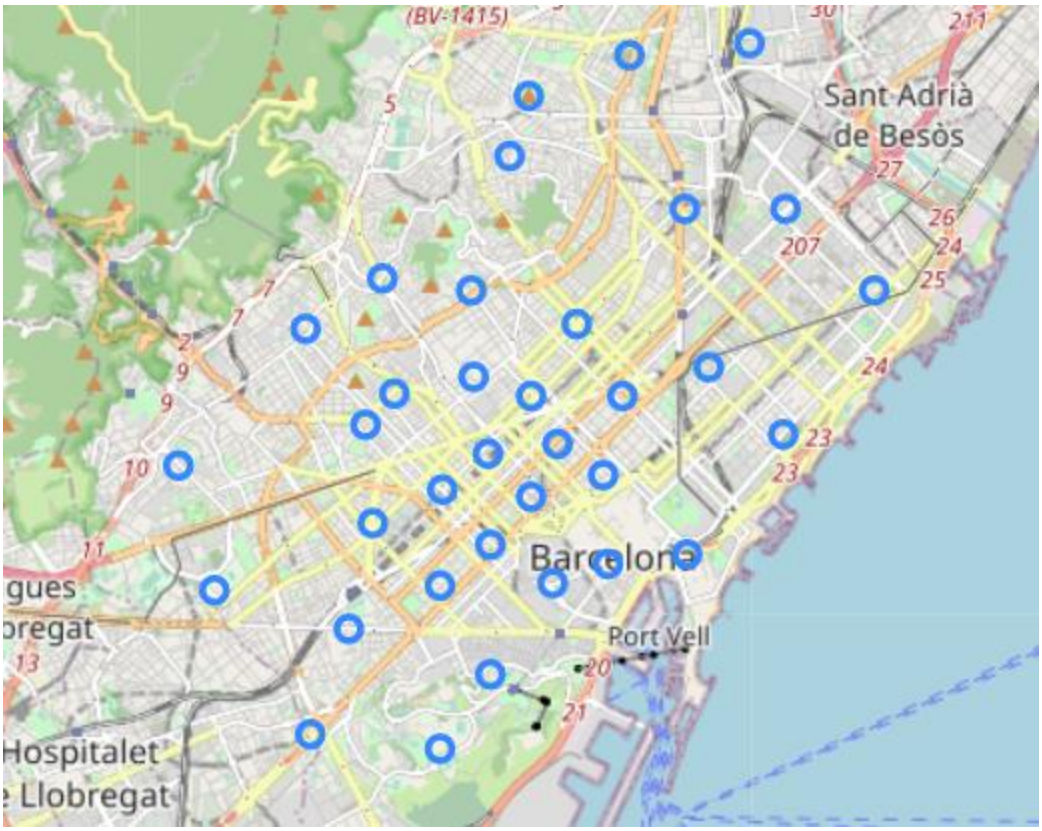
Business Problem

So, this is where I come in – I plan to use Machine Learning and Location Data to cluster neighborhoods depending on its venues on my own – the same process was previously done by several employees over several days. The scope of the project is, that I have to prove eligibility of my offered tool on one popular travel destination – Barcelona – if I am able to cluster neighborhoods appropriately, then management will approve the tool which then will be used on other travel destinations.

Data

Combination of several sources is the input data for the project:

1. Neighborhoods and Postal Codes of Barcelona
    - It is collected manually and imported as a data source into the project
    - Pandas library is used to import excel file
    - As more central parts of Barcelona are of most interest, some outskirts are removed.
2. Latitude and Longitude of Postal Codes
    - is collected through Arcgis of Geocoder package
    - map of Barcelona with neighborhoods superimposed



3. Venue Data of neighborhoods
    - Foursquare API is used to collect Points of Interest in proximity of Neighborhoods' location
    - Radius of 500 meters in addition to 100 venues limit per neighborhood is applied
    - Just head of the fetched data

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	el Raval	41.380145	2.168721	La Robadora	41.379500	2.170463	Gastropub
1	el Raval	41.380145	2.168721	Robadors 23	41.379581	2.170603	Jazz Club
2	el Raval	41.380145	2.168721	A Tu Bola	41.380096	2.169054	Tapas Restaurant
3	el Raval	41.380145	2.168721	33/45	41.381059	2.167399	Cocktail Bar
4	el Raval	41.380145	2.168721	Guixot	41.378509	2.167806	Spanish Restaurant

Methodology

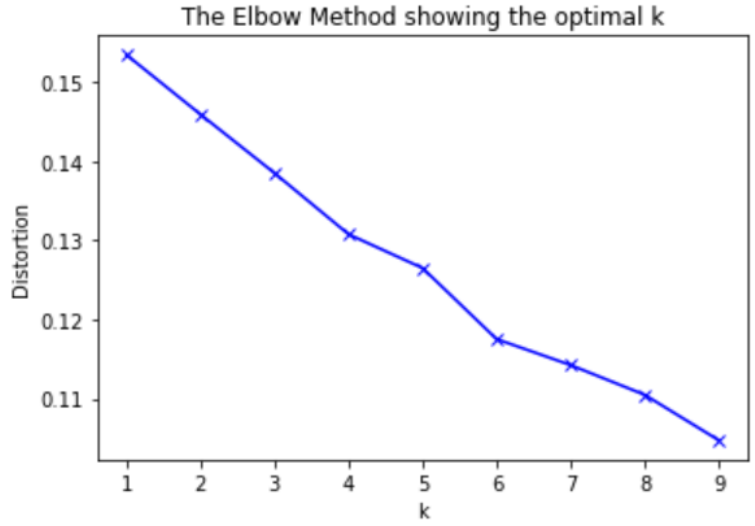
With simple exploratory analysis of venue counts, it seems that more than half of neighborhoods have more than 50 venues listed. Thus, sorted venues as most common venue categories in its respective neighborhoods and only considered top 10 venue categories per neighborhood, to make data more manageable.

This kind of data is perfect fit for k_means unsupervised clustering, in order to combine most related neighborhoods by top 10 venue categories.

In order to be able to cluster neighborhoods, I prepared data for analysis first. For that I used one hot encoding to get venue categories in columns and neighborhoods in rows. Than the date was grouped by mean aggregate function.

	Neighborhood	Zoo	Accessories Store	African Restaurant	American Restaurant	Aquarium	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...
0	Camp d'en Grassot i Gràcia Nova	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	...
1	Dreta de l'Eixample	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	...
2	El Besòs i el Maresme	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	...
3	El Putget i Farró	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.02	0.0	...
4	Fort Pienc	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.02	0.0	...

After sorting top 10 venue categories by neighborhoods, I used elbow method in order to get most appropriate count for clusters.



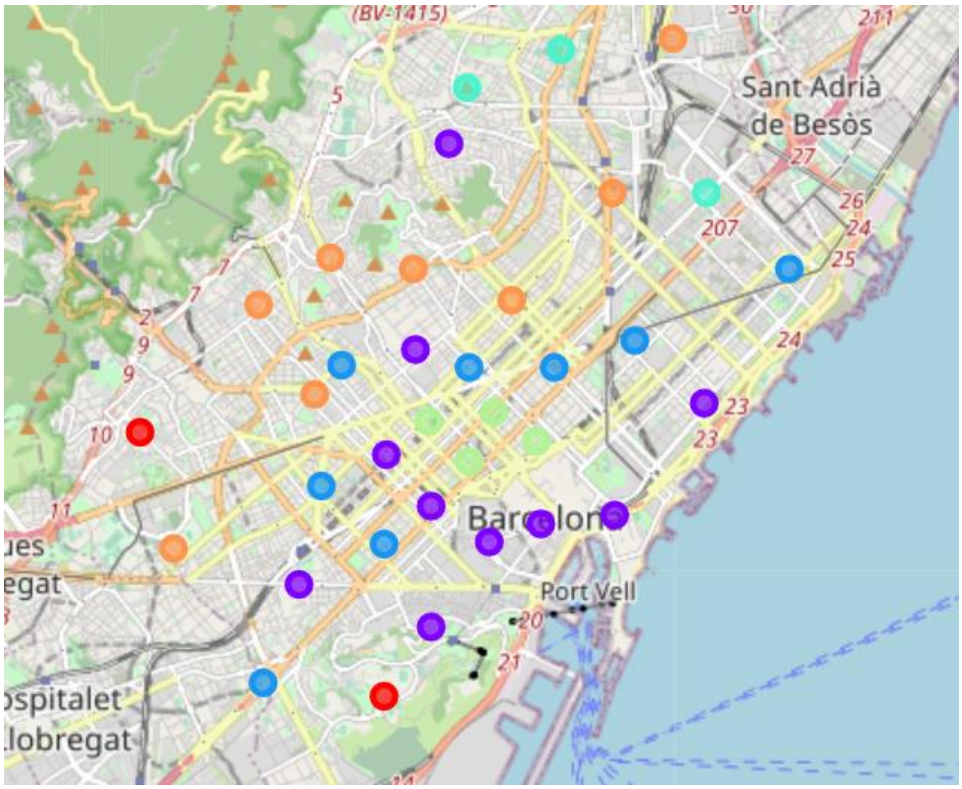
I assumed that 6 could be appropriate value for K-means clustering.

I applied K-means algorithm from scikit-learn, merged back cluster labels to main dataframe and checked results

Results & **Discussion**

Results were checked:

1. On the map to identify proximity to central regions



2. Checked individually by neighborhoods

	PostCode	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	8001	1	Spanish Restaurant	Mediterranean Restaurant	Tapas Restaurant	Cocktail Bar	Pizza Place	Plaza	Bar	Hotel	Bookstore	Donut Shop
1	8002	1	Tapas Restaurant	Spanish Restaurant	Plaza	Hotel	Ice Cream Shop	Italian Restaurant	Mediterranean Restaurant	Wine Bar	Bar	Coffee Shop
2	8003	1	Tapas Restaurant	Bar	Seafood Restaurant	Mediterranean Restaurant	Spanish Restaurant	Ice Cream Shop	Burger Joint	Restaurant	Hotel	Food & Drink Shop
3	8004	1	Mediterranean Restaurant	Pizza Place	Park	Theater	Restaurant	Plaza	Spanish Restaurant	Brewery	Tapas Restaurant	Palace
4	8005	1	Beach Bar	Spanish Restaurant	Tapas Restaurant	Mediterranean Restaurant	Gastropub	Bakery	Vegetarian / Vegan Restaurant	Restaurant	Pizza Place	Ice Cream Shop
10	8011	1	Café	Mediterranean Restaurant	Tapas Restaurant	Dessert Shop	Japanese Restaurant	Bakery	Coffee Shop	Spanish Restaurant	Cocktail Bar	Argentinian Restaurant
11	8012	1	Plaza	Mediterranean Restaurant	Bar	Tapas Restaurant	Pizza Place	Arts & Crafts Store	Restaurant	Ice Cream Shop	Café	Cocktail Bar
13	8014	1	Tapas Restaurant	Bar	Wine Bar	Mediterranean Restaurant	Italian Restaurant	Middle Eastern Restaurant	Hotel	Pie Shop	Pizza Place	Plaza

3. And grouped cluster labels and sorted by venue category counts

	Cluster Labels	Venue Category	Venue
503	5	Bakery	23
603	5	Spanish Restaurant	22
606	5	Supermarket	18
550	5	Grocery Store	17
558	5	Hotel	17

Combination of all of these three was used to make assumptions about cluster types

According to map:

1. as it seems clusters 1 & 2 occupy most of the city map, more concentrated in city center and scattered around the city too
2. Cluster 5 is third most common of the clusters and is mostly located around the uptown of the city
3. Cluster 4 is located across the city center, but takes less impressive count of locations spots.
4. Cluster 3 seems to be concentrated in more outskirts of the city
5. Cluster 0 seems to be outlier in terms of map occupation, but it might be interesting to check this in more details

According to individual neighborhoods and cluster aggregated counts:

1. As it seems Cluster 1 is more concentrated around Tapas restaurant, which is specialty of Barcelona and Spain itself. It should be interesting location for tourist to check in, while traveling to Barcelona, to taste local cuisine
2. Cluster 2 seems to be focused more on international food, then local cuisine. This might seem interesting for tourists, as they might be interested in tasting other foods too, after trying out local cuisine. Mediterranean food seems more frequent in addition to Italian food and some Japanese restaurants.
3. Cluster 4 has higher concentration of hotels/hostels, which should be interesting info for tourist who is in search of hotels, or trying to avoid places with high concentration of hotels
4. Cluster 5 seems to be more residential cluster, as grocery shops and supermarkets have higher count. This seems to be less interesting locations, but some specific tourists want check residential areas of city and not only touristic places, to feel real city vibe.
5. Cluster 0 seems to be concentrated on sport activities, for those who want to take a break from local or international cuisine tasting and lose several calories on the way, or just relax.
6. Cluster 3 does not seem to be of particular interest as it does now show any trend of particular places. As seen on the map, this cluster is located in less central areas of the city.

Conclusion

As I mentioned in the introduction, whole point of this project is to prove machine learning capabilities for location/neighborhood auto recommendation to customers. All the steps executed required little human intervention and this method can be used for other locations too.

There is space for improvement, of course. Other machine learning algorithms can be added for getting better results. Or algorithms can be tailored for individual or specific groups of customers, who are more interested in some particular areas of tourism.

Also some human interactions can be reduced too. As I used local file for Barcelona districts, web scrapping can be incorporated in this part. Also paid features of Foursquare API can be used to enhance venue data to better fit needs.

Decision should be made bt stakeholders, but to me this method is out of competition. Still it won't be easy to prove to stakeholders, but with some other locations and improvements mentioned above success should be easily achieved