Mumbai Snack Place Analysis

Rootvij Bage | IBM Data Science Certificate (Coursera) | 7/29/19

Introduction

I am looking to find which area of Mumbai would be the best for me to open a snack place. In India, many people are below the poverty line and opening small businesses is the most popular way for citizens to provide for themselves and their families. Knowing where to open their store can help them run a better business as they are setting up shop in an area where they are more likely to succeed

Data

I will be using wikipedia to webscrape the list of suburbs in Mumbai and merge it with location data of their longitude and latitude using geocoder. I will then use the Foursquare API to see what kind of venues are in each area to see where most of the snack shop density is.

Methodology

To begin, we gather the list of suburbs in Mumbai from Wikipedia and webscrape the site using the beautifulsoup package in Python to get a list of the names. We then use the geocoder package to obtain the latitude and longitude of the data so we can use the Foursquare API. After we obtain this data, we create a dataframe and visualize the points on a map using the Folium package.

We then use the Foursquare API to see the top 100 venues within a 2000 meter radius. The API returns to us a json of the name, category, latitude, and longitude. We now check how many unique venues we have and then group the neighborhoods, taking the mean of the frequency of each venue category. We do this to prepare the data for the k-means clustering. Since we are trying to find where to open a snack shop we filter the data to "Snack Place" only.

Now, we perform k-means clustering on the data and choose to begin with 3 clusters, based on the frequency of occurrence. The distribution on the map should help us analyze and make a decision on where to open a snack shop.

Results

From the k-means clustering we see that we can categorize the city in to 3 areas based on the frequency of occurrence for "Snack Place". Cluster 0 has a moderate amount of snack places, Cluster 1 has little to no snack places, and Cluster 2 has a high amount of snack places. You can see a visualization of the clustering in the map and the different colors.

Results

From the k-means clustering we see that we can categorize the city in to 3 areas based on the frequency of occurrence for “Snack Place”. Cluster 0 has a moderate amount of snack places, Cluster 1 has little to no snack places, and Cluster 2 has a high amount of snack places. You can see a visualization of the clustering in the map and the different colors.

Discussion

In the clustering we saw 3 different clusters: no snack shops, moderate amount of snack shops, and high amount of snack shops. In the area where there are no snack shops, this might be because there is no demand so nobody wants to open a shop in this area. The high amount areas may be too crowded to open a shop as maybe there is a high amount of lower middle class families and a lot of small businesses being opened. There seem to be a decent amount of moderate amount areas.

Conclusion

After performing the clustering we have decisions to make on where we want to open a new snack place. We can open in Cluster 0 but we run the risk of there being little to no demand for a snack place. However, since there are not many in the area, we could capture the entire market. Opening in cluster 2 is out of the question because there is way too much competition in that area. Therefore, we will decide to open our shop in cluster 1 because we see that there is enough interest there while not being over throttled by competition.