Data Science Capstone Project

Finding suitable locations for opening a cafe

By Aayush Kamath

Introduction

- The number of cafes opening in Mumbai are on the rise.
- People flock to cafes because it provides a certain level of comfort through its cozy ambience.
- Cafes are also a hotspot for a lot of working professionals who find the environment comfortable enough to carry out their work while maybe enjoying a cup of tea or coffee.
- Running a café is a business and for any business belonging to the restaurant/café industry location plays an important factor in whether or not the business succeeds in the long run.

Business Problem

- Having identified location as a key component in running a café it is necessary that through careful analysis we arrive at the optimum location.

- Objective: To analyze suburbs/neighborhoods in Mumbai, and come up with location of suburbs/neighborhoods which might be ideal for opening a café.

- The stakeholders for this business problem could be the people who are looking to open a café or property brokers who would help their clients find the most suitable location for opening a cafe.

Required Data

1. The list of suburbs/neighborhoods in Mumbai, India.

2. The geographical coordinates, that is, the latitude and longitude of the neighborhoods in Mumbai.

3. Data of venues near all neighborhoods, especially data of cafes.

Data Sources

- The list of suburbs/neighborhoods in Mumbai is scraped from the wikipedia page of suburbs in Mumbai. The scraping of data is done using the Beautiful Soup library in Python.
- Once we have the list of all neighborhoods, we find the latitude and longitude for each of them using the Geocoder package in Python.
- The venue data for all neighborhoods are obtained by using the Foursquare API.

Methodology

- Data is scraped from the wikipedia page of Suburbs of Mumbai to gather a list of all neighborhoods.
- The latitude and longitude of every neighborhood is found using the Geocoder package
- An API call for data of the top venues (limit 100) in a 1km radius of every neighborhood is made.
- A json file is returned by foursquare from which we extract the venue name, its longitude and latitude and the venue category. Then, we analyze the frequency of every type of venue category for a neighborhood.

Methodology (Continued)

- Now, that we have all frequencies, we extract the frequency of cafes for each neighborhood as our business problem involves finding suitable locations for a café.

- Based on these frequencies, we divide all the neighborhoods into 3 different clusters using k-means clustering method.

- The result obtained helps us analyze about the suitability of each neighborhood as a location for a new café.

Results

We obtained 3 clusters:

- Cluster 0 had neighborhoods with high frequency of cafes.

- Cluster 1 had neighborhoods with moderate frequency of cafes.

- Cluster 2 had neighborhoods with low frequency of cafes

Discussion

- Neighborhoods in Cluster 0 show the highest frequency of cafes. Neighborhoods belonging to cluster 0 might not be the best choice for business as there are already a lot of cafes in those neighborhoods.
- Cluster 1 shows neighborhoods with moderate frequency of cafes. Opening a cafe can be considered as the market isn't saturated and at the same time there seems to be a decent amount of demand for cafes.
- Cluster 2 has neighborhoods with either low frequency of cafes or zero cafes. Opening a cafe in any one of these neighborhood presents a high risk high reward scenario. It is too unpredictable.

Limitations

- The factor taken into account here is frequency of cafes in a neighborhood. But various factors such as population, availability of real estate for opening a café, and average income of the neighborhood would also play a key role in deciding about the suitable locations for a café.

Conclusion

- In this Project we have identified a business problem, provided a brief introduction, identified the data requirements and the data sources, extracted data from those sources and prepared it and clustered it in order to gain the necessary insights and accordingly, provide recommendations.
- Finally, the inference made from this project could be that cluster 2 is saturated due to high frequency of cafes, cluster 0 is unpredictable due to low frequency of cafes and cluster 1 is probably the most suitable group of neighborhoods where a café could be opened as the frequency of cafes in those neighborhoods are moderate.

Thank You!