COURSERA CAPSTONE

IBM Applied Data Science Capstone

Opening a New Restaurant Chain in Ahmedabad, India

By: **SUPAN SHAH**

February, 2020

INTRODUCTION

Ahmedabad is the largest city in the state of Gujarat in India. It is also known as the “Manchester of India”. Ahmedabad has emerged as an important economic and industrial hub in India.

Gujarati people are known for their love of food (after the fact that they also love travelling, of course!). The people here are extremely fond of food, so much that on average there are more number of foodies in a household than there are households - well, obvious, isn’t it?

As a result, the food industry in the pertaining area has never been out of business. Opening a restaurant is one of the most profitable businesses in this part of the world. You can find restaurants that have been managed by single families, since ages! Hence, opening restaurants allows firms and entrepreneurs to earn consistent income. Of course, with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

BUSINESS PROBLEM

The objective of this capstone project is to analyse and select the best locations in the city of Ahmedabad, Gujarat, India to open a new restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: ***In the city of Ahmedabad, if a company / an entrepreneur is looking to open a new restaurant (chain), where would you recommend that they open it?***

TARGET AUDIENCE

This project will be particularly useful to companies as well as entrepreneurs and investors looking to open or invest in new restaurants in the city of Ahmedabad.

DATA

To solve the problem, we will need the following data:

- List of neighbourhoods in Ahmedabad. This defines the scope of the project which is confined to the city of Ahmedabad, Gujarat in India.
- Latitude and Longitude coordinates of those neighbourhoods. This is required in order to plot the map and also get the venue data.
- Venue data, particularly data relating to restaurants. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods employed:

This Wikipedia page - (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Ahmedabad) contains a list of neighbourhoods in the city of Ahmedabad, with a total of 81 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and the beautifulsoup package. Then we will get the geographical coordinates of the neighbourhoods using the Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use the Foursquare API to get the venue data for those neighbourhoods. Foursquare API will provide many categories of the venue data; we are particularly interested in the Restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling to machine learning (K-means clustering) as well as map visualization (Folium).

METHODOLOGY

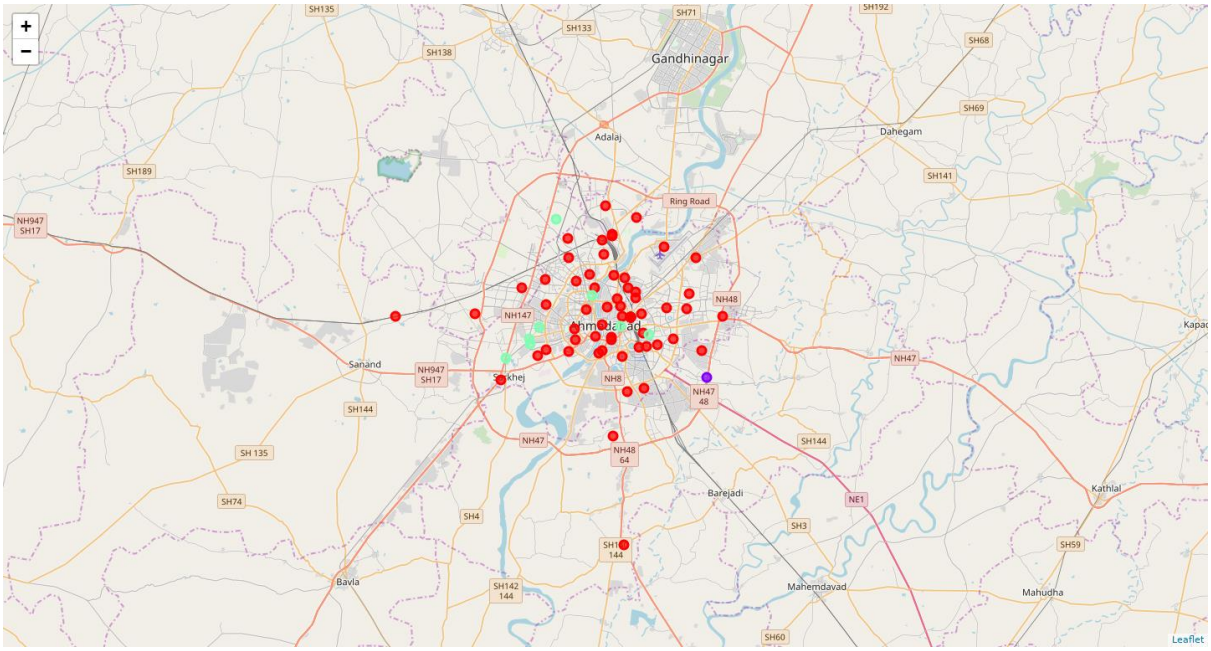
Firstly, we need to get the list of neighbourhoods in the city of Ahmedabad. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Ahmedabad). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Ahmedabad. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Restaurant” data, we will filter the “Restaurant” as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Restaurant”. The results will allow us to identify which neighbourhoods have higher concentration of restaurants while which neighbourhoods have fewer number of restaurants. Based on the occurrence of restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new restaurants.

RESULTS

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Restaurant”:

- Cluster 0: Neighbourhoods with low number to no existence of restaurants
- Cluster 1: Neighbourhoods with moderate number of restaurants
- Cluster 2: Neighbourhoods with high concentration of restaurants

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour



CONCLUSION

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. companies, entrepreneurs and investors regarding the best locations to open a new restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant.

REFERENCES

Category: Suburbs in Ahmedabad - Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Ahmedabad

Foursquare Developers Documentation - Foursquare. Retrieved from https://developer.foursquare.com/docs

APPENDIX

Cluster 0 (Low Density of Restaurants):

- Agol
- Vastrapur
- Kalyanpura
- Kharna
- Khodiyarnagar
- Khokhra
- Lambha
- Maninagar
- Memnagar
- Mithakali
- Sardarnagar
- Saraspur
- Sabarmati
- Shardanagar
- Ghodasar
- Girdharnagar
- Gita Mandir
- Godhavi
- Gomtipur
- Jamalpur
- Isanpur
- Cantonment
- Motera
- Naranpura
- Naroda
- Nava Vadaj
- Odhav
- Paldi
- Polarpur
- Ranip
- Vastral
- Usmanpura
- Juhapura
- Kalupur
- Jivraj Park
- Calico Mills
- Bahiyal
- Dabhoda
- Chandkheda
- Dariapur
- Jholapur
- Jawahar Chowk
- Ellis Bridge
- Ghatlodiya
- Ujedia
- Thakkar Bapanagar
- Thaltej
- Shubhash Bridge
- Kabirchawk
- Shastrinagar
- Shahpur
- Shahibaug
- Sarkhej
- Bopal
- Bhojva
- Bhairavnath Road
- Behrampura
- Bareja
- Chandlodiya
- Bapunagar
- Asarwa
- Asarwa Chakla
- Amraiwadi
- Ambawadi
- Alam Roza

Cluster 1 (High Density of Restaurants):

- Ramol

Cluster 2 (Moderate Density of Restaurants):

- Khadia
- Makarba
- Gota
- Navjivan
- Rajpur
- Jodhpur
- Anand Nagar
- Vejalpur