Coursera Capstone

IBM Applied Data Science Capstone

Opening a New BAKERY shop in Coimbatore, India

By: Anand Shanmugam

July 2020

Introduction

Coimbatore is the second largest city in the Indian state of Tamilnadu and is known for variety of bakeries and pastries. You will find bakeries in each and every corner of the city. I am passionate to open a bakery chain. The success of the chain depends on the location and competitors. If I choose a location that already has lot of bakeries, then my chain is likely to fail.

Business Problem

The objective of this capstone project is to analyse and select the best locations in Coimbatore to open a new bakery. 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 Coimbatore, India, if one has to open a bakery shop , where would you recommend to start?

Data description

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

- **List of neighbourhoods** in Coimbatore. This defines the scope of this project which is confined to the city of Coimbatore, second largest city of Tamilnadu, India.
- **Latitude and longitude coordinates** of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- **Venue data**, particularly data related to **Bakeries**. We will use this data to perform clustering on the neighbourhoods.

Methodology:

Data Source:

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Coimbatore) contains a list of neighbourhoods in Coimbatore, with a total of 36 neighbourhoods.

Data extraction:

We will use **web scraping** techniques to extract the data from the Wikipedia page, with the help of Python requests and **beautifulsoup** packages. We will wrangle the data, clean it, and then read it into a *pandas* dataframe so that it is in a structured format .

Latitude and Longitude:

We will get the geographical coordinates of the neighbourhoods using **Python Geocoder** package which will give us the latitude and longitude coordinates of the neighbourhoods and then visualize the neighborhoods in a map using **Folium** package

Venue Data:

After that, we will use **Foursquare API** to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data,.we are particularly interested in the **Bakery** category in order to help us to solve the business problem put forward. In this stage we prepare data for use in clustering. Since we are analysing the “Bakery” data, we will filter the “Bakery” as venue category for the neighbourhoods.

Clustering

Lastly, we will perform clustering on the data by using k-means clustering. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Bakery”. The results will allow us to identify which neighbourhoods have higher concentration of Bakery .Based on the occurrence of Bakeries in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new bakeries.