Coursera Capstone Project : Applied Data Science

Sarthak Srivastava

Introduction

- Mumbai is India's commercial capital, its entertainment capital, and is famously known as the city that never sleeps. Mumbai is India's busiest and most-populous city, with the 2018 Census of India estimating that 12 million call it home.

A city that never sleeps, people are always working tirelessly for their work! And to minimize the time, many even try to grab their meals from restaurants near their work place to save time.

Business Problem

- Each person in the corporate sector is in a rush right from the morning. As a result, there are many restaurant outlets opening up in the proximity of such areas and locations.

- And this is exactly what my project is about! To open up a new Food Outlet in the best proximal region around a corporate sector. Let us call this restaurant “Mr. Brown”.

- Local train commuters rely heavily on their breakfast, sometimes even lunch, on light meals. These can be on the go sandwiches, fries, momos, pizza, burgers, some microwaveable or cold prepared meal along with beverages.

- Our goal is to find the optimal location where this restaurant can be set up and flourish with it’s light ready-to-go snacks.

- A location, in Mumbai City, where a food outlet can easily survive without much competition.

- A location, where food outlets are present in scarcity and are hugely needed.

- We can also try to find more than one location, and who knows we might be able to set up a ‘Mr. Brown’ chain of food outlets!

Data

Area, Location

The data of Area and Location in Mumbai can be extracted out by scraping the web using the BeautifulSoup library of python. The data used in the project is scraped from a Wikipedia page.

Geocoding

To get the latitude and longitude of the extracted area and location, I tried to use Foursquare API, but the results were mostly returned as ‘None’. I also tried to use the Google Maps Geocoding API, but was unsuccessful in the same way. Hence, the latitude and longitudes of the area are scraped from Wikipedia as well.

Venue Data

From the Area, Location and Region data, the venue data is found out by passing in the required parameters to the FourSquare API, and thereby creating a pandas DataFrame to contain all the venue details along with the location details.

How data will be used to solve the problem?

From the venues extracted using the Foursquare API, venues which are most demanded in the choice of people will be selected.

From the data analysis, we will group all these regions in different clusters according to category of venues.

The weakest cluster, i.e. the cluster which has met least demands of the people in the food category will be selected as our optimal region.