Introduction

Boston is the capital and most populous city of the Commonwealth of Massachusetts in the United States, as well as the 21st most populous city in the United States. The city proper covers 48 square miles (124 km2) with an estimated population of 694,583 in 2018, making it also the most populous city in New England. The city is the economic and cultural anchor of a substantially larger metropolitan area known as Greater Boston. As a combined statistical area (CSA), this wider commuting region is home to some 8.2 million people, making it the sixth most populous in the United States.
Boston is the home of some of the top sporting teams like Redsox, Patriots, Celtics and Bruins. Also, Boston neighborhoods have some of the best colleges in the nation like Harvard, MIT etc. City is famous for some of the old food places and food varieties.

Business Problem

The objective of this capstone project is to analyze and recommend the best neighborhoods in the city of Boston, USA to open a new pizza place. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: Where would you recommend a new investor to open a new pizza place in the city of Boston?

Target Audience of this project

This project is useful for any investors who are willing to open a new pizza place in the city of Boston. Based on the rankings provided by TripAdvisor in 2018, Regina Pizzeria from Boston ranked as the #1 pizza place in USA. Boston neighborhoods already have a number of pizza chains, specialty pizza chains and local pizza places

Data

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

- List of neighborhoods in Boston. This defines the scope of this project
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to pizza places which is required to perform clustering on the neighborhoods.

Sources of data and methods to extract them

The Wikipedia page (https://en.wikipedia.org/wiki/Greater_Boston) contains a list of neighborhoods in and around Boston, with a total of 125 neighborhoods. We will use the web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and ***beautifulsoup*** packages. Then we will

get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases 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 Pizza places category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

References

- Top 5 pizza places in USA (pizzatoday.com/news/pizza-headlines/guide-to-the-2018-national-best-pizzas-lists/)
- Wikipedia