**IBM Applied Data Science Capstone – Coursera**



**Opening a new pizza place in Boston, USA**

**Submitted :** Anand Koottilingal

**Date** : 08/23/2019

**Table of Contents**

[Introduction 3](#__RefHeading___Toc174_2283293455)

[Business Problem 3](#__RefHeading___Toc176_2283293455)

[Target Audience of this project 3](#__RefHeading___Toc178_2283293455)

[Sources of data and methods to extract them 3](#__RefHeading___Toc180_2283293455)

[Results 5](#__RefHeading___Toc846_2283293455)

[References 6](#__RefHeading___Toc848_2283293455)

# 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 Red Sox, Patriots, Celtics and Bruins. Some of the best colleges Harvard, MIT etc. are located in Boston neighborhoods. City is famous for its food traditions and old food houses. All these factors are attracting new investors to Boston and understanding of the city neighborhood is a key before deciding the location for starting the business

# 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 Trip Advisor 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.

**Methodology**

First step on this analysis to get a list of neighbourhoods in the city of Boston. Fortunately, the list is readily available in the Wikipedia page (https://en.wikipedia.org/wiki/Greater\_Boston). Using the web scraping techniques supported by Python and beautifulsoup packages, all the neighborhood names are extracted. Before proceeding, the extracted list is compared to Wiki to confirm all neighborhoods are included in the result.

Next step is to find out the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, the Geocoder package is used to convert neighborhood address into geographical coordinates in the form of latitude and longitude. All gathered data is then populated to a pandas DataFrame and then visualized the neighborhoods through 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 Boston.

Using Foursquare API, captured a list of top 100 venues that are within a radius of 5000 meters. To do so, I first registered a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. Then added made API calls to Foursquare passing in the geographical coordinates of the neighbourhoods and the Foursquare credentials in a Python loop. Foursquare returned the venue data in JSON format and extracted the venue name, venue category, venue latitude and longitude. With the data, checked how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues.

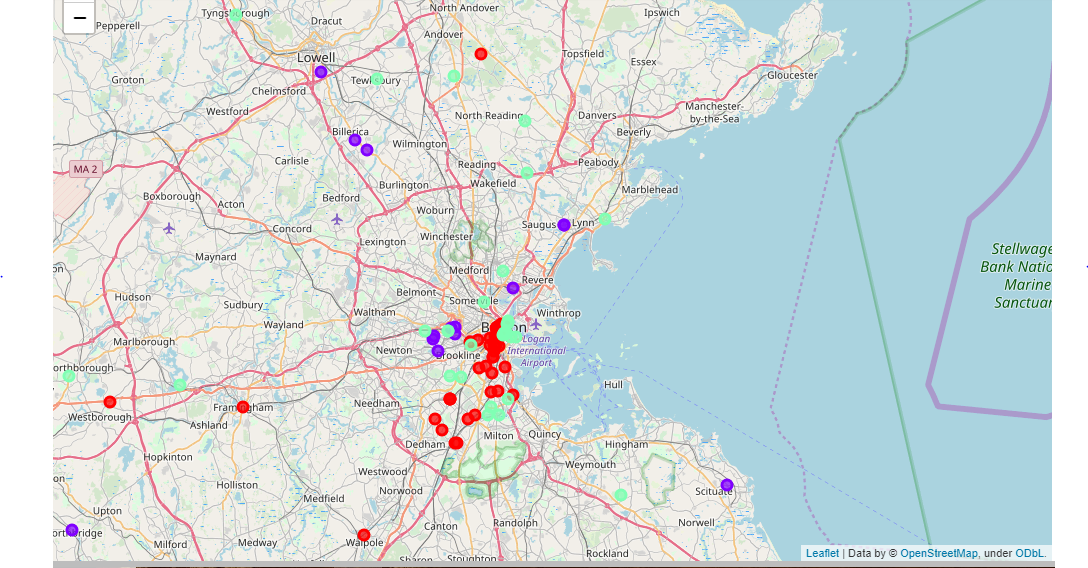
In the next step, analyzed each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Pizza Places” data, we will filter the “Pizza Places” as venue category for the neighbourhoods. Lastly, performed 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 “Pizza Places”. The results will allow us to identify which neighbourhoods have higher concentration of Pizza places while which neighbourhoods have fewer number of Pizza places. Based on the occurrence of Pizza places in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Pizza places

# Results

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

* Cluster 0: Neighbourhoods with high concentration of pizza places; around 51%
* Cluster 1: Neighbourhoods with low concentration of pizza places; around 18%
* Cluster 2: Neighbourhoods with medium concentration of pizza places; around 31%

The results of the clustering are visualized in the map below with cluster 0 in RED color, cluster 1 in PURPLE color, and cluster 2 in MINT GREEN color



**Discussion**

Analysis showing that most of the pizza places are concentrated in cluster 0 which include Boston downtown, Fenway, Backbay, South end and some of the financial district area (around 51%). Cluster 1 have around 18% concentration and cluster 2 have around 31%. This is showing that new pizza places can be opened in neighborhoods listed in cluster 1 as less competition compared to cluster 0 and 2. Meanwhile, pizza places in cluster 0 are likely suffering from intense competition due to oversupply and high concentration. Pizza places with unique menus and special varieties can survive from the competition and can open on these neighborhoods . Therefore, this project recommends that new investors looking to start pizza places in Boston neighborhoods to capitalize on these findings to open new pizza places in neighborhoods of cluster 1 unless there is a unique specialty on their menu

**Future Research**

In this project, only considered one factor i.e. frequency of occurrence of pizza places, there are other factors such as population, income of residents etc. can also influence the location decision of a new pizza places. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new pizza places. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

**Conclusion**

In this project, all steps in a data science analysis is applied like 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. investors regarding the best locations to open a new pizza places. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new pizza places. 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 pizza places.

# References

* Wikipedia - <https://en.wikipedia.org/wiki/Boston>
* Wikipedia - <https://en.wikipedia.org/wiki/Greater_Boston>
* Beautiful Soup documentation - <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
* Pandas documentation - <https://pandas.pydata.org/pandas-docs/stable/>
* Folium documentation - <https://python-visualization.github.io/folium/>
* Top 5 pizza places in USA (Article: pizzatoday.com/news/pizza-headlines/guide-to-the-2018-national-best-pizzas-lists/)