Coursera

IBM Applied Data Science Capstone

Opening a New Pizzeria in New York City

by: Huseyin Karakus

May 2020

A. Introduction

A.1. Background & Problem Description

New York City (NYC) is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles, New York is also the most densely populated major city in the United States. New York City's food culture includes an array of international cuisines influenced by the city's immigrant history. Italian immigrants brought New York-style pizza and Italian cuisine into the city. As of 2019, there are 27,043 restaurants in the city, up from 24,865 in 2017[1].

With 93% of Americans consuming pizza at least once a month, it is no surprise that pizza shops are one of the most popular types of restaurants. Opening a pizzeria can be a great investment, especially when you have prepared correctly. If you are wondering how to open a pizza shop, your first priority should be coming up with finding a great location for your pizzeria. What would you consider when selecting a location? By exploring the regional characteristics of pizza shops, I hope to figure out whether the neighborhood of pizza shops is an essential factor for the success of a pizza shop.

Business Problem

The objective of this capstone project is to analyze and select the best locations in New York City to open a new pizza shop. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In New York City, if an investor is looking to open a new pizza shop, where would you recommend that they open it?

A.2. Data Description & Preparation

Data Description

To solve the problem, we need the following data:

- List of neighborhoods in New York City. Wikipedia[2]

- Latitude and longitude coordinates of those neighborhoods. This is required to plot the map and to get the venue data.
- Venue data, particularly data related to pizza shops. We will use this data to perform clustering on the neighborhoods.

Data Preparation

We will use 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 restaurant 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).

I cleaned the data and reduced it to boroughs of NYC so that I can use it to find geological locations for further venue analysis.