Coursera Capstone

IBM Applied Data Science Capstone

Opening a Pizza Place in New York City, USA

By Arne Lachmann

Summer 2021

Introduction

Pizza is one of the favourite foods in the US and popular in many places all over the country, cp. https://www.thrillist.com/news/nation/most-popular-type-of-restaurant-in-every-state-map. It is available in many varieties, toppings ranging from classics such as Salami and Ham to sea food or even kebab variants and always a great deal regarding price and energy input. New York City, USA already offers many places to get some fresh, hot pizza, but this does not apply for the whole city. In this project I aim to find the best neighborhood to open a pizza place in NYC, USA.

Business Problem

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

Target audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new pizza places or Italian restaurants in New York City.

Data

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

- List of neighborhoods in New York City
- Latitude and longitude coordinates of those neighborhoods for plotting
- Venue data, particularly data related to pizza places and Italian restaurantsfrom foursquare.com

Data sources

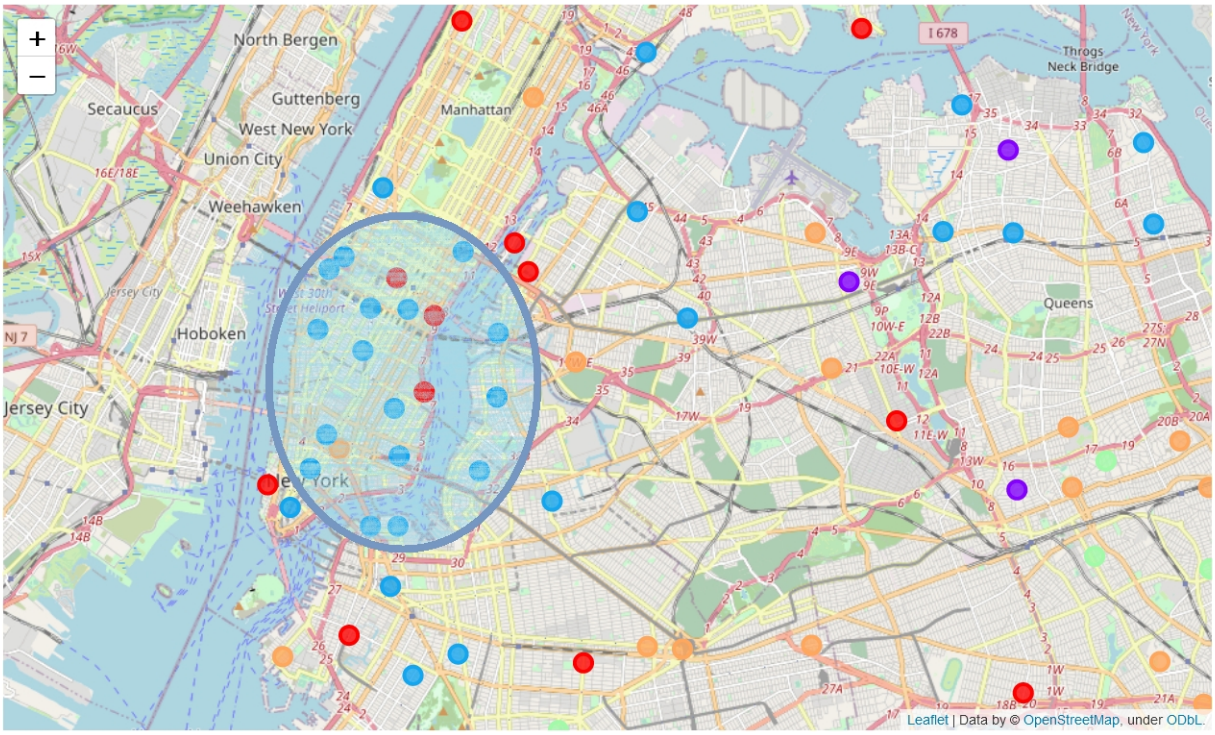
From (1) we get the list of all neighborhoods in New York City, counting 306. It also contains the information about latitude and longitude of these neighborhoods, stored in a JSON file. Afterwards, we will use the Foursquare API to get the venue data for those neighborhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the “pizza place” and “Italian restaurant” category in order to help us solving the business problem.

Methodology

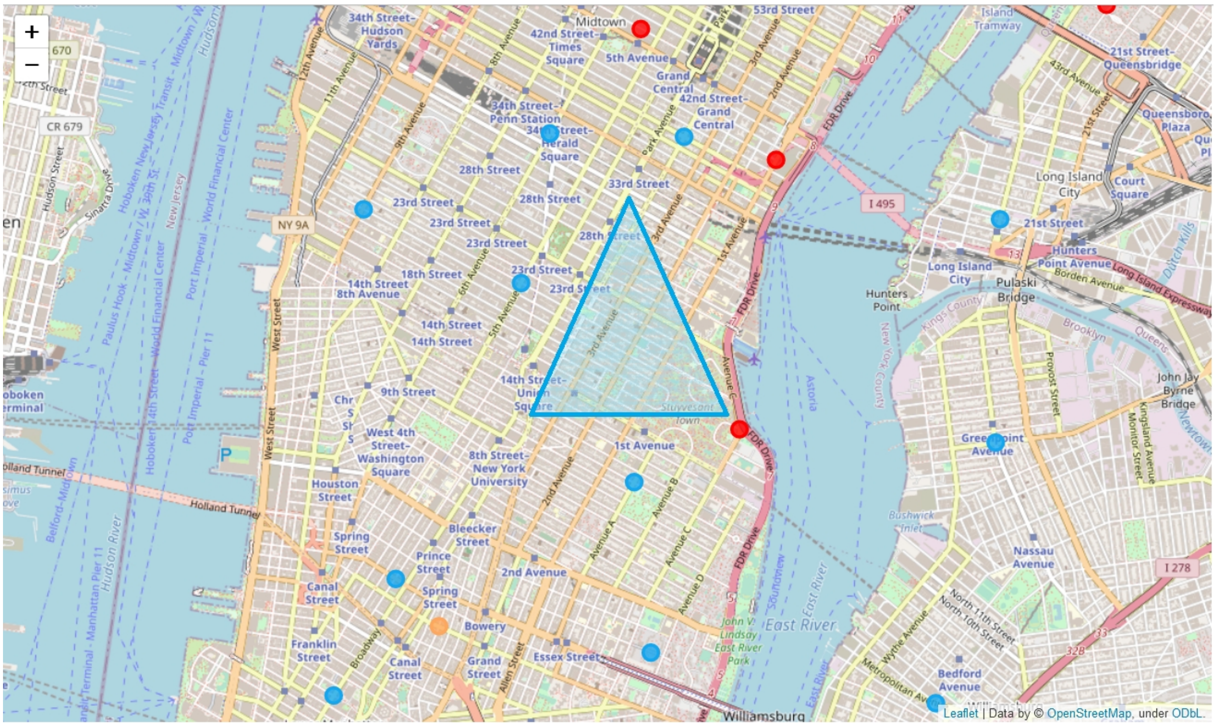
Firstly, we need to get the list of neighborhoods in New York City. Fortunately, the list is available in (1). This json file also contains the georaphical coordinates (longitude and latitude). Having the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using the Folium package. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and the Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods. Foursquare will return the venue data in JSON format and we will extract the venue name, the venue category, and venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. Lastly, we will perform 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. The clustering algorithm is performed on each of the five New York boroughs individually to find clusters of similar neighborhoods. It is also performed on New York overall. Then, of all neighborhoods in New York, the DataFrame is reduced to those neighborhoods that do not have a “pizza place” or “Italian restaurant” as one of their top 10 venues as both categories refer to “pizza”. This new DataFrame is clustered again to find the best spot to open a place to sell pizza in New York.

Results

The results from the k-means clustering show that Manhattan is the most homogeneous borough, with Brooklyn and the Bronx being a close second and the other two being relatively equivalently inhomogeneous. When ignoring those neighborhoods where the ten most common venues are neither a pizza place nor an Italian restaurant, the biggest and most dense cluster of similar neighborhoods happens to be in the south of Manhattan. The rest of New York is either not densely clustered enough to be recommended or the variety of clusters is too big.



Picture 1: Cluster of similar neighborhoods in the South of Manhattan



Picture 2: Zoomed in nearby 23rd Street and 2nd Avenue in Manhattan.

Discussion

The spot I will chose should be in a neighborhood where

- the surrounding neighborhoods are as homogeenous as possible,
- the surrounding neighborhoods have as little pizza places or Italian restaurants as possible,
- the neighborhood itself and the surrounding neighborhoods are as densely populated as possible,
- none of the top ten revenues is a pizza place and
- none of the top ten revenues is an Italian restaurant.

Nearby 23rd Street and 2nd Avenue in Manhattan is where I would recommend to open a spot to sell pizza since the results from the k-means clustering show that Manhattan is the most homogeneous borough, with Brooklyn and the Bronx being a close second and the other two being relatively equivalently inhomogeneous. When ignoring those neighborhoods where the ten most common venues are neither a pizza place nor an Italian restaurant, the biggest and most dense cluster of similar neighborhoods happens to be in the south of Manhattan. The rest of New York is either not densely clustered enough to be recommended or the variety of clusters is too big.

Limitations and suggestions

In this project, we only consider one factor, i.e. frequency of occurrence of pizza places and Italian restaurants, there are other factors such as population, income of residents and housing prices that could influence the location decision of a spot to sell pizza. 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 place to sell pizza.

Conclusion

In this project, I have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, clustering the data based on their similarities, and lastly providing recommendations to the relevant stakeholders regarding the best locations to open a new place to sell pizza. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in the south of Manhattan, nearby 23rd Street and 2nd Avenue are the most preferred locations to open a new place to sell pizza. 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 place to sell pizza.

References:

(1) https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json