Coursera Capstone

IBM Applied Data Science Capstone

Finding a Location for a New Italian Restaurant in Chicago, Illinois

1. Introduction

Chicago is the largest metropolitan area in the Midwest. With its large population, the city presents prime opportunities for young entrepreneurs to launch brand new businesses such as restaurants or cafes. Opening a new restaurant in Chicago can possibly prove to be a good decision depending on if the restaurant itself can attract customers. For this report, I'm exploring where the best locations would be to start a new Italian restaurant in a Chicago neighborhood. Properly choosing the correct location to start the restaurant is critical for attaining long-term success and keeping the restaurant open. When selecting neighborhoods to start the restaurant, there are a few key variables you have to take into consideration. These variables include which neighborhoods show a high demand of restaurants and what neighborhoods have surrounding venues that can attract customers to the area.

2. Business Problem

The primary goal of this project is to find the best possible neighborhoods to place a new Italian restaurant in Chicago. This will be done by using various methods previously learned in this course such as clustering and utilizing Foursquare location data. The main audience this project would apply to are entrepreneurs looking to start a new restaurant or current Italian restaurant owners who may be looking to relocate in the city of Chicago. This would apply to this audience

because it could give valuable insight on which neighborhoods have the most financial benefits.

3. Data Description

Multiple sources of data will be required to answer the business problem.

3.1 List of Neighborhoods

This project will require a list of all the neighborhoods in Chicago, Illinois for our dataset. The data will be obtained by scraping a list of all the Chicago neighborhoods from the Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago).

This page contains a list of all the neighborhoods in the city of Chicago.

3.2 Geocoder

Import the Python geocoder function to locate the geographical points of every neighborhood in the city of Chicago. This will provide every neighborhood's respective longitude and latitude values. After finding the longitude and latitude values for each neighborhood, they will be added to the dataframe.

3.3 Foursquare

Data about each venue in the city of Chicago will be needed for analysis. Foursquare API will be utilized to find the details about each venue such as their names and geographical points. The data collected from Foursquare will be used to find what neighborhoods have Italian restaurants.

4. Methodology

The required dataframe for this project will be acquired by scrapping the list of Chicago neighborhoods from the Wikipedia page(https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago). This will be done by using the Python requests to transform the data from the list into a dataframe. The longitude and latitude coordinates for each neighborhood will be discovered by using the geocoder function. These coordinates are required to utilize the Foursquare API later in the notebook. After the longitude and latitude values are found using the geocoder function, we will add the coordinates to our dataframe to match each respective neighborhood. A folium map will then be created using the dataframe. This folium map will help us visualize the city of Chicago and all of the neighborhoods in it. The map will also help by ensuring that the correct geographical coordinates were obtained for every neighborhood.

4.1 Venue Data

The Foursquare API location data will then be used to discover information about the venues of Chicago. Foursquare API can be accessed by creating a Foursquare account and generating a Foursquare ID and client key. Using both the ID and key, we'll have the ability to generate a list of the top 100 venues within a 500 meter radius of each neighborhood. Once we have this data, we'll convert it to a new dataframe containing the neighborhood names, neighborhood geographical coordinates, venue names, categories, and venue geographical coordinates. Using this data, we'll analyze our dataframe by grouping each neighborhood into rows and finding the mean of frequency of each venue category. The data will then be filtered by only including neighborhoods that have at least one Italian restaurant.

Having at least one Italian restaurant shows that there's a level of demand for Italian food in a neighborhood.

4.2 Clustering

K-means clustering will be used to cluster the neighborhoods. K-Means is a type of partitioning clustering which divides the data into non-overlapping subsets or clusters without any cluster internal structure or labels. K-means will be used on our dataset to divide each neighborhood into a cluster based on the number of Italian restaurants. We will divide the dataset into four clusters to evaluate which clusters have the highest frequencies of Italian restaurants.

4. Results

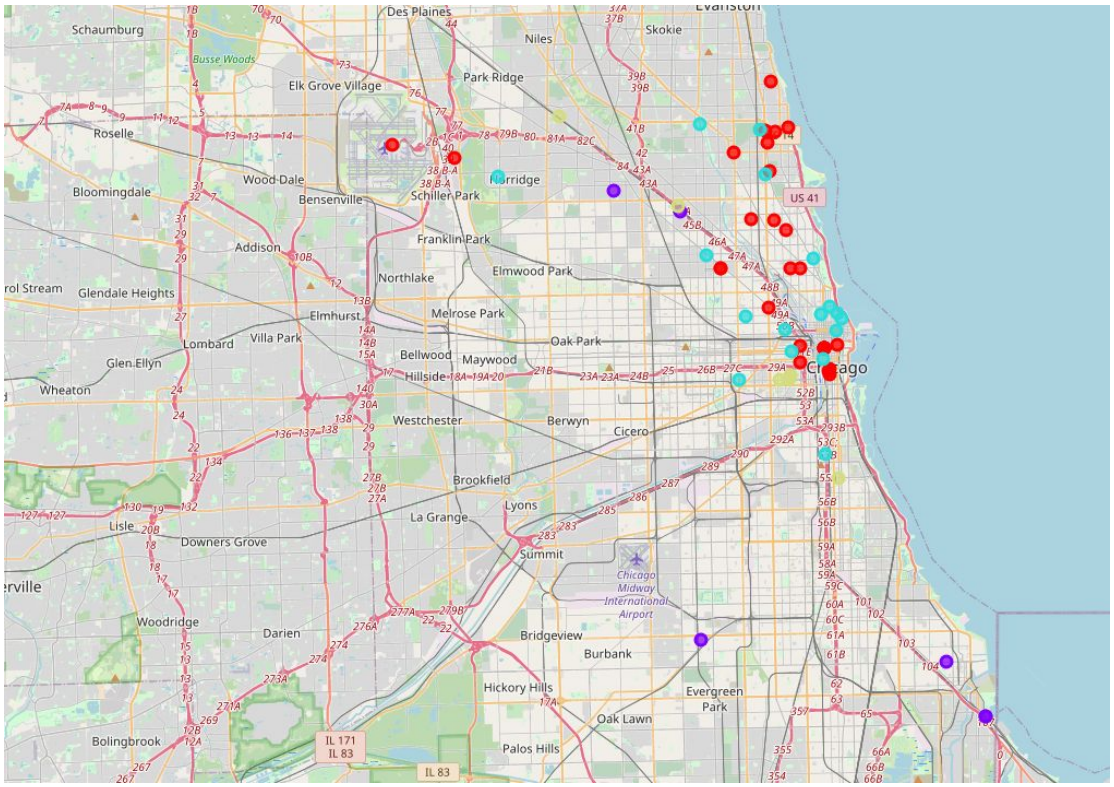
From dividing the neighborhoods into 4 clusters, the map below shows the clusters in different colors to distinguish the frequency of occurrence of Italian restaurants in each neighborhood.

Cluster 0 (Red): Neighborhoods with an extremely low number of Italian restaurants.

Cluster 1 (Purple): Neighborhoods with a high number of Italian restaurants.

Cluster 2 (Blue): Neighborhoods with a low frequency of Italian restaurants.

Cluster 3 (Light Green): Neighborhood with a moderate number of Italian restaurants.



5. Discussion

Analyzing the results, a lot of the Italian restaurants in Chicago are located in clusters 1 and 3. In contrast, clusters 0 and 2 have a very small number of Italian restaurants. Clusters 0 and 2 wouldn't be the most ideal neighborhoods to start an Italian restaurant due to the demand of Italian food likely being low in those neighborhoods. Clusters 1 and 3 would be the two primary options to start a new restaurant because they both have much higher chances of sustainability.

5.1 Recommendation

Cluster 3 would be the recommended cluster to start a new restaurant for two main reasons. First is the frequency of Italian restaurants in cluster 3. Observations from the notebook show that neighborhoods in cluster 1 have almost twice the amount of Italian restaurants in cluster 3. Although this shows that cluster 1 has a high demand for Italian food, it also means opening a new restaurant within a neighborhood in this cluster would prove to be a challenging task due to extreme competition. Cluster 3 still has a relatively high demand for Italian food while being in less competitive neighborhoods. The second reason cluster 3 would be the recommended cluster is due to the locations of the neighborhoods in the cluster. Half the neighborhoods in cluster 3 are located between the University of Illinois and the Illinois Medical District. Placing a Italian restaurant in a neighborhood between these two major institutions will give the restaurant immense potential to attract customers from multiple demographics.

6. Conclusion

In this project, I was able to analyze the results to answer the original business question. I was able to do this by gathering the necessary sources of data, scraping and cleaning the data, and using machine learning to cluster the data. Carefully

going through these steps allowed me to create observations from the results. Neighborhoods in cluster 3 would likely be the best destinations to build a new Italian restaurant in Chicago, Illinois. This is due to both the locations of its neighborhoods and the current demand for Italian restaurants in the area. The discoveries from this project should assist entrepreneurs in finding the most beneficial spot to place a restaurant in Chicago.