Capstone Week 4 Report

IBM Data Science Professional

Chicago Neighborhoods for Foodies

Author: Blake Rocheleau
Date: July 2020

INTRODUCTION

Moving to a new city can feel intimidating, given the thousands of housing options and venues that exist. Specifically, consider moving to Chicago, Illinois, the third most populated city in the United States. Chicago is home to many different parks, attractions, museums, gyms, theatres, schools, and even shopping malls. However, after considering all these venues, what if food was the most important factor in determining what part of Chicago you wanted to move to? There are over 200 different neighborhoods to choose from and more likely than not, some will have better or more restaurant options than others. This project will therefore utilize publicly available data to pinpoint the neighborhoods that someone who loves food above all else (also called a "foodie") should move to in the city of Chicago.

DATA

In order to determine what neighborhood would best suite food lovers, it is important to have knowledge of all the existing neighborhoods of Chicago. Therefore, a Wikipedia page containing all the neighborhoods of Chicago will be scraped (bulleted below) for this list. Additionally, the latitude and longitude coordinates of each neighborhood will be required to search for local venues. After coordinates are obtained, data on trending venues around each neighborhood will need to be extracted for each neighborhood in order to cluster and compare the best neighborhoods for foodies. Venue data will be obtained using the Foursquare API.

- List of Chicago neighborhoods

METHODOLOGY

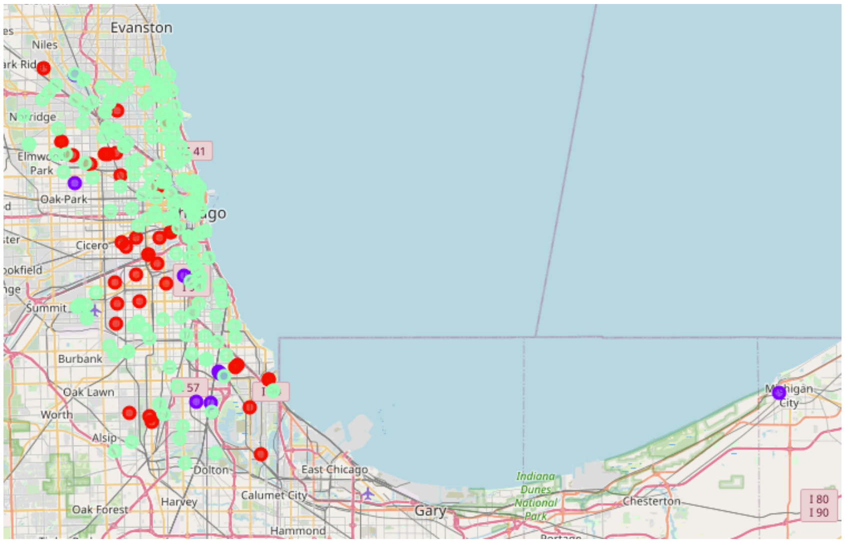
A list of neighborhoods in Chicago was scraped from a Wikipedia page (bulleted above) using the Beautiful Soup Python library. The data was converted from HTML data to a Pandas dataframe, which consisted of two columns for the neighborhood name and community area. Using the geocoder Python library, the latitude and longitude of each neighborhood was obtained and merged into the dataframe containing the list of neighborhoods. To confirm the successful capture of Chicago neighborhoods and coordinates, the neighborhoods were plotted on an interactive map of Chicago using the Folium Python library.

The next necessary step was to obtain the venues for each neighborhood so they could eventually be compared by means of food. The Foursquare API was used to obtained multiple venues for each neighborhood. Specifically, each API request contained a “Restaurant” query so that only food venues were supplied by Foursquare. The resulting list of venues was encoded using one-hot encoding so that each neighborhood could be assigned a 0/1 value for each unique venue category. This was necessary for clustering analysis.

In order to analyze the list of venues (and therefore the neighborhoods for food), k-means clustering was used. K-means clustering is a form of unsupervised machine learning that groups data into a predefined number of “clusters” based on their underlying similarities. Each cluster has no internal structure and is non-overlapping with other clusters. K-means is an appropriate method for solving this problem because it groups different neighborhoods into clusters that represent different densities of food options. For this problem, the number of clusters was arbitrarily chosen to be 3. We then were able to analyze the clusters by varying degree of food-option density.

RESULTS

The results from the k-means clustering algorithm are shown below in Figure (1). Cluster 0 can be seen in red, cluster 1 is in purple, and cluster 2 is in green.



Based on analysis from the cluster data, we define each cluster as follows:

- Cluster 0: Neighborhoods with a low number of food options
- Cluster 1: Neighborhoods with a moderate number of food options
- Cluster 2: Neighborhoods with a high number of food options

DISCUSSION

One observation that can be made from the data is neighborhoods closer to the water (Lake Michigan) tend to fall into cluster 2, which represents food-dense neighborhoods. This could explain a correlation with prime real-estate and food options. In that case, the availability of food options may be positively correlated with the amount of available housing (people tend to want to live closer to the water).

Another observation is cluster 2, which represents food-dense neighborhoods contains a large number of neighborhoods. This means for foodies looking to move to Chicago, there is a large selection of qualifying neighborhoods to choose from.

For further investigation, it would be useful to experiment with different radii when making the Foursquare API calls. The radius used in this experiment is 500 meters, which may not be enough to capture all the relevant food venues for each neighborhood. However, increasing the radius will possibly increase the number of overlaps in venues that get identified with multiple neighborhoods. Additionally, this report solely examines the frequency of available food venues. It would be interesting to factor in other variables, such as income and neighborhood population.

CONCLUSION

In this project, we have identified the problem of choosing an ideal neighborhood in Chicago to live in based on the availability of food venues. Neighborhood data was scraped from Wikipedia and captured for analysis in a Pandas dataframe. K-means clustering was applied to the dataset in order to group each neighborhood into unique clusters. The results were plotted on a map of Chicago using the Folium Python library. Three unique clusters of neighborhoods resulted, each with varying density of food venues. It was revealed that neighborhoods belonging to cluster 2 were optimal for people who prioritize food when moving to Chicago.