Capstone Project - The Battle of Neighborhoods

Location Analysis for Opening a New Restaurant in Cleveland

Chia-Hung Kuo 18 May 2020

1. Introduction

Business understanding

Introduction Greater Cleveland area is a fast-growing metropolitan area in northeast Ohio. Its 3.5 million residents make it one of the largest Ohio metro in giant Great Lakes Megalopolis. It is fast-growing fast because of the top-ranking universities, major healthcare systems and hospitals, and the balanced work-life lifestyle. The world-class museums and attractions draw approximately 2.5 million visitors annually. Therefore, it is a great place to start a new restaurant. Utilizing the analytic approach to conduct a market analysis is a great start.

Project objectives: Find a neighborhood with the population, household income, and lifestyle for starting a successful restaurant.

Targeted Audience: People who want to start a new restaurant in the greater Cleveland area.

Problem: Needs to be in a go-to neighborhood if people want to eat out and also have regular customers to support the business.

Clear Question: How to find the optimal location for the restaurant?
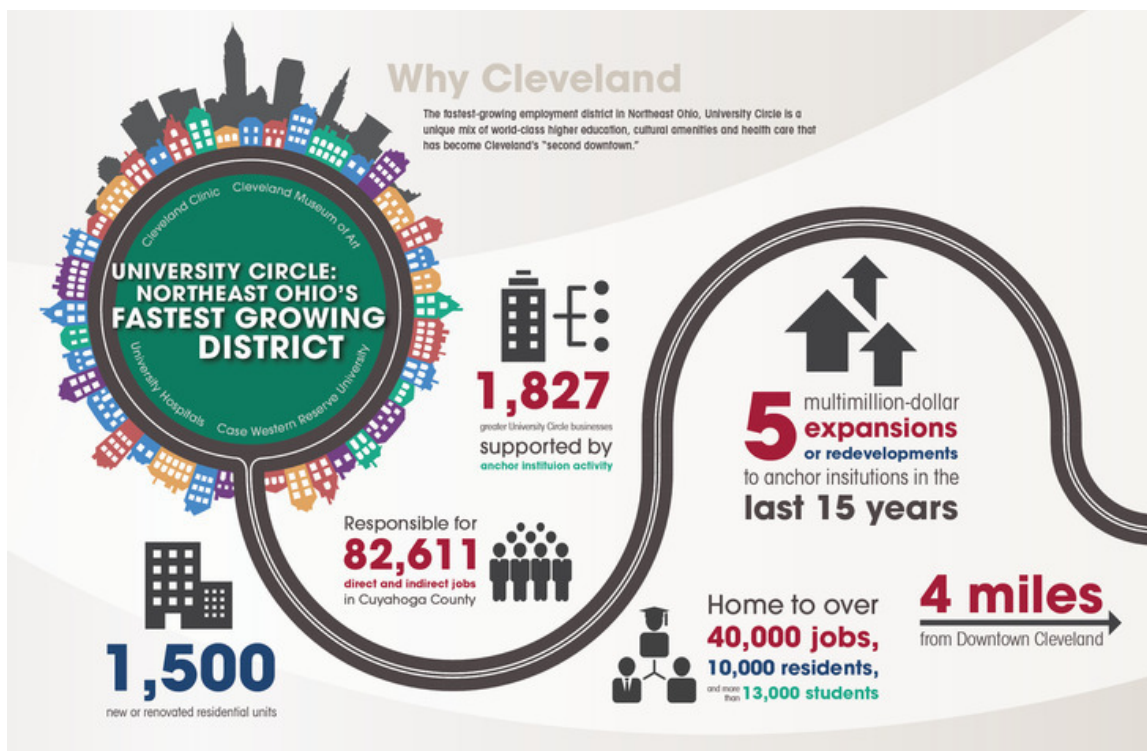
2. Analytic approach:

We first analyze the income, population, and similiar business in the borough to obtain the insight of the neighborhood. Since the borough contains many cities, we perform further detailed analyses on the cities from the selected borough.

Using K-Means clustering to find out the groups of neighborhoods that seems to behave similarly to each other in the scale of borough level, and then narrow down to postal code (ZIP Code) areas.

3. Data requirements

We need the county and city level of family income data and area population data with connected to county and city longitude and latitude information.

a. Data collections

The available data are obtained from Foursquare API, table from websites, and online databases:

i. Foursquare API for the listing of greater Cleveland neighborhoods information with clustered venues through.

ii. Ranking every Ohio city, county for median family income - Census Snapshot by Data Central
https://www.cleveland.com/datacentral/2017/12/ranking_every_ohio_city_county_2.html
Median family income in Ohio cities ranges from a high of $214,850 in the Columbus suburb of New Albany to below $40,000 in 14 separate cities, according to new estimates released this month by the Census Bureau.

iii. Great Cleveand Area Population Estimates by Cuyahoga County Planning Commission
https://www.countyplanning.us/resources/census-data/population-estimates/
The Census Bureau's Population Estimates Program produces population estimates on an annual basis for the nation, states, counties, and communities using components of demographic change.

iv. US Zip Code Latitude and Longitude by Open Data Soft
https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=OH
The data set provides the matching latitude and longitude for the cities and states.

v. ZIP Code Database by United States Zip Codes:
https://www.unitedstateszipcodes.org/zip-code-database/
This dataset provides a more detail insight into the demographics boundaries. It provides the boundary of Cuyahoga County with matched latitudes and longitudes, so that we could narrow down the searching area to only the northeast of Ohio state.

vi.   Libraries: Pandas, Urllib, Geopy, Requests, Scikit Learn, and Folium

b. **Data understanding/Visualization/Cleaning**
To have the data's initial insights, understand the descriptive statistics, and have the visualization, we performed three steps. First, we load the files to the format of datafames and visuall check the datasets' summary information. Second, we performed the data cleaning include removing the missing or invalid values and eliminating duplicate cells. Then we remove the columns that are not needed in this project and then removing the NaN values that might cause problems. Third, we did the visualize the datasets to further under the trend and to see if they are suitable for the project. After cleaning, we combined multiple data sources of the needed data with the matching reference to the city and zip code into a dataframe for use through the project.