IBM Applied Data Science Capstone

1. Introduction

This project is an implementation of the skills learned from this series of course. Though an entire data science pipeline including data collection, data cleaning, exploration data analysis, machine learning process, visualization and information delivery, we are able to solve the real business problems and present the meaningful insights derived from the data.

The topic of this project is to make a location recommendation for opening an Italian restaurant in Chicago central area. Chicago is the third-most-populous city in the United Status. As an international hub for finance, culture and technology, there are a lot of opportunities. However, starting a business is challenging especially at the beginning. The selection of location requires a comprehensive consideration of facts including population, transportation, security, rent and so on. An investigation of neighborhood environment is likely to provide meaningful insights for decision-making. Therefore, this project aiming to business structure in each neighborhood will be helpful to a business owner/ someone who is interested in opening a new business.

2. Data Preparation

1. The name of neighborhoods

Chicago is officially divided into 77 community areas. In this project, we will focus on three communities: Near North Side, Loop, and Near South side in the city center. This area containing 16 neighborhoods: https://en.wikipedia.org/wiki/Community_areas_in_Chicago. The name of these neighborhoods can be scarped from this website with the help of BeautifulSoup package.

2. The coordinates of neighborhoods

With the names of neighborhoods have been collected, the Geocoding can transfer the place names to coordinates. A data frame with the names and coordinates of the neighborhoods define the scope of the analysis.

3. The venues information for each neighborhood

In each neighborhood, there are venues in various categories. The information of these venues can be requested from the Foursquare API. In this project, we will utilize the venue name, location and categories information in the radius of 500 of each neighborhood.

After these processes, the data for analysis have been collected. It is noteworthy that the created data frame has only 14 neighborhoods. It is because a. The neighborhood “South Loop” occurs twice in the neighborhood name list; b. there is no exploration result in Foursquare for the neighborhood “Goose Island”.