**Analyzing The Community Areas of Chicago to start A New Restaurant**

Kerry Goode Jr

June 11, 2021

**Introduction**

The City ofChicago is the most populated city in the State of Illinois and the third most populated city in the United States. Chicago had an estimated population of 2.6 million people in 2019 which makes it the most populated city in the Midwestern United States and the fifth most populated city in North America.

The population of Chicago consist of people with various ethnicities from all over the world. The city has numerous restaurants which serve thousands of customers per day. Given the diversity in the population, there are a vast number of restaurants with different cuisines for everyone to enjoy. In this project we will examine the community areas in Chicago and make a recommendation on the location where a restaurant should be started.

**Business Problem**

A group of investors are looking to invest and start a franchise restaurant in Chicago. They have approached us to study the market and provide them with a recommendation on which community area in Chicago where they should invest and start their franchise. The main objective of this project is to obtain and analyze the right data about the various community areas in Chicago and recommend to our client a suitable location for their restaurant.

**Data**

The following data is needed to complete our analysis and to provide a recommendation:

1. Community Areas of Chicago
2. Geographical coordinates of the Community areas
3. Venue Data from FourSquare

**Community Area Data**

This data was extracted from the Community areas of Chicago page ([Category:Community areas of Chicago - Wikipedia](https://en.wikipedia.org/wiki/Category:Community_areas_of_Chicago)) using web scraping with BeautifulSoup library in Python. This will give us a specified list of community areas in Chicago.

**Geographical Coordinates of the Community Areas**

Geographical coordinates are needed for plotting maps to visualize the data. The geographical coordinates of the community areas were extracted using GeoPy library in python. Once we use GeoPy we will add two columns to our data frame with latitude and longitude information for each community area.

**Venue Data from FourSquare**

Using FourSquare data we extract Venue data. The venue data is used to study the venues in various community areas in Chicago. This data provides much needed details of the different restaurants in those areas to understand how competitive it is in each area. The addition of this data will help us make a more informed decision at the conclusion of the project.

**Methodology**

**Feature Extraction**

Feature extraction was completed through One Hot coding. During this method, each feature is a category that belongs to a venue which is then converted into binary. For example, 1 means this category is found in the venue and 0 means the opposite. All of the venues are grouped by the community areas and computed at the same time as the mean. The output will give us a venue for each row and each column will have the frequency of occurrence of that particular category.

Graphical user interface, text, application

Description automatically generated

**Unsupervised Learning**

Unsupervised learning was completed to find out the similarities between community areas. K-means (a clustering algorithm) was executed. In this scenario, K-means is used due to its simplistic nature and its similarity approach to find patterns.

Graphical user interface, text, application, email

Description automatically generated

Chart, line chart

Description automatically generated

**Plotting**

Various types of plotting techniques were used to visualize the data. By visualizing the data, it enables us to give a clear understanding of the data and it makes it easier to spot patterns. The Folium library was used to visualize the cluster date. The Folium library provided the capability to plot maps of the city of Chicago and the community areas within the city.

Map

Description automatically generated

**Results**

Based on the K-means cluster, the number of clusters that were derived from elbow method was 8. The code as well as plotting of the clusters can be seen below:

Graphical user interface, text, application, email

Description automatically generated

Map

Description automatically generated

After visualizing the clusters, each cluster was studied, and a conclusion was derived. The community area that had the greatest number of restaurants was cluster 4.

**Discussion**

It is recommended to the client that the community areas such as West Loop, River North, and Pilsen fall in cluster 4 and are good locations for starting their restaurant. The client can make a decision depending on other factors like the type of cuisine that will be served in their franchise and availability of of land plots which are outside of the scope of this project.

**Conclusion**

The data analysis and machine learning techniques used in this project are helpful determining solutions in certain business problems. Pythons libraries such as Geopy, Folium and BeautifulSoup make it easier for a data scientist to analyze a geographical location since those libraries allow for easy extraction of date that is available online. In this project, the community areas of Chicago were studied and a recommendation of the community areas where our client should start their franchise was provided.