**IBM Applied Data Science Capstone**

**Analyzing Neighborhoods in Chicago, IL**

**For a Retail Business**

Israel Tshikoji

1. Introduction
2. Background

Chicago is the most populous city in the US State of Illinois, and the third most populous in the United States. With an estimated population of about 3 million people, it is the most populous city in the US Midwest.

Chicago is an international hub for finance, commerce, industry, technology, etc.

Taking into consideration its big population size and its financial relevance, there is a definite opportunity to provide enough facilities (such as school, parks, restaurants, retail stores, etc.) to accommodate the needs of its residents and visitors.

Providing business amenities to a big population as the one under consideration is a business opportunity that any good investor would like to explore further. Although there are many types of businesses that can be studied and tested, this project will zoom in a retail business idea.

1. Business Problem

While it may be reasonably assumed that opening a retail store in the Chicago City area is a sound business idea, this project will try to locate which neighborhoods would offer a greater profitability for such an endeavor.

This project is carried out in order to advise a group of retail investors about the potential benefits of investing in the retail business in the Chicago city area.

1. Data

The list of Chicago neighborhoods will be scraped from the following website:

<https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago>

The raw data off the webpage will be transformed into a pandas dataframe.

**The Geopy library** will be used to get the latitude and longitude values of desired neighborhoods.

**The Foursquare API** will then be used to explore neighborhoods and segment them according to the most prominent business activity.

In analyzing the clustered neighborhoods, useful insights will be drawn to help recommend suitable locations for retail businesses.

1. Methodology

The neighborhoods of Chicago data were retrieved from the following url:

<https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago>

The data was scraped and converted into a dataframe. The **Geocoder library** was used to add the longitude and latitude of each neighborhood as shown below:

Table

Description automatically generated

The **Folium library**, along with longitude and latitude data, were used to visualize Chicago and its neighborhoods.

Map

Description automatically generated

The **Foursquare API** was then used to explore the neighborhoods and segment them into different venues. The limit value for the foursquare API was set to **100 venues** while the radius was set to **500 m**.

Graphical user interface

Description automatically generated with medium confidence

The Foursquare API returned a total of 347 unique venue categories. The table below shows 5 different neighborhoods with their respective 10 most common venues.

Graphical user interface

Description automatically generated with medium confidence

In order to better study and understand our dataset comprising different categories, the Kmeans machine learning algorithm was used to create clusters of different neighborhoods. The Kmeans is a simple, yet an effective machine learning algorithm for unsupervised data as in the project under study. The Kmeans divided our dataset into 5 different clusters as shown in the map below:

Map

Description automatically generated

1. Results

The project objective was to advise a group of stakeholders about starting a retail business in a Chicago neighborhood. While so many retail businesses can be conducted, a particular focus is given to starting a clothing store in a neighborhood where the market is favorable. The results below will show bar graphs of all the 5 clusters and the number of clothing stores (if any) in each respected cluster.

1. Cluster 1 (containing **4 clothing stores**)

Chart, bar chart, histogram

Description automatically generated

Figure Cluster 1 - Venue Category Distribution

1. Cluster 2 (containing **15 clothing stores)**

Chart, bar chart

Description automatically generated

Figure Cluster 2 - Venue Category Distribution

1. Cluster 3 (containing **0 clothing stores)**

Chart, bar chart

Description automatically generated

Figure Cluster 3 - Venue Category Distribution

1. Cluster 4 (containing **0 clothing stores)**

Chart, bar chart

Description automatically generated

Figure Cluster 4 - Venue Category Distribution

1. Cluster 5 (containing **0 clothing stores)**

Chart, bar chart, histogram

Description automatically generated

Figure Cluster 5 - Venue Category Distribution

To make our discussion complete, we included the table below to show the venue categories associated with cluster 5.

A picture containing table

Description automatically generated

1. Discussions

Looking at all the figures above, we noticed that cluster 1 and cluster 2 have 4 and 15 clothing stores, respectively. Cluster 3, 4, and 5 have no clothing stores in the listed venue categories.

From a business and investment prospective standpoint, we may assume that there might be a demand for such business activities. Providing such facilities to customers that need to drive across town to buy would be a good incentive for them. Keeping this in mind, we can shortlist the location of our business to include cluster 3, 4 and 5.

Having decided on locating our business on any of the forementioned clusters (namely 3, 4 or 5) is just the beginning of a thorough analysis. The next question that we will need to answer is which one of the three clusters makes a better location. A closer look to *figure 5* reveal 4 important features that may make cluster 5 more attractive for setting up our business. These are: intersection, bus station, train station and airport.

The train and bus stations are in the **Burnside neighborhood** (cluster 5), while the airport is located at the intersection of **Legends South, Sleepy Hollow and LeClaire Courts (**cluster 5**)**. The above depicted venues (airport, bus station, train station, intersection) are areas where traffic and thousands of people pass every day. This is a key element to having a big and constant flow of potential customers come in proximity with our business. Therefore, we would recommend considering cluster 5 as the ideal spot to implement our business.

1. Conclusion

This project objective was to identify different neighborhoods around the greater Chicago metropolitan area in order to locate a suitable location for a retail business.

Using Wikipedia, The Geopy and the Folium libraries, along with the Foursquare API, we were able to get the neighborhoods names, locate them on a map, and group them in 5 clusters comprising different venue categories.

After a thorough analysis of the segmented neighborhoods, we determined that cluster 1 and 2 were not considered to establish our retail business. Cluster 3, 4 & 5 were more prone to consideration since they each lacked retail business, in the line of clothing stores. A closer look into these three clusters (3, 4 & 5), revealed that cluster 5 was the better spot to implement our retail business. Cluster 5 was chosen due to its proximity with venues that gather large groups of people, such as airport, bus station and train station.

Future work may further investigate income levels in or around cluster 5, as well as crime rates, and other pertinent parameters necessary to the implementation of a retail business in an area that offers safety and prosperity for the merchants and the consumers.