# Capstone Project - The Battle of the Neighbourhoods

*The Applied Data Science Capstone is a part of **IBM Data Science Professional Certificate program**.*

# 1. Introduction

## 1.1 Background: City of Chicago [1]

Chicago, officially the City of Chicago) is the most populous city in Illinois, as well as the third most populous city in the United States. With an estimated population of 2,705,994 (2018), it is the most populous city in the Midwest.

Chicago is an international hub for finance, culture, commerce, industry, technology, telecommunications, and transportation. It is the site of the creation of the first standardized futures contracts at the Chicago Board of Trade, which today is the largest and most diverse derivatives market globally, generating 20% of all volume in commodities and financial futures. O'Hare International Airport is one of the busiest airports in the world, and the region also has the largest number of U.S. highways and greatest amount of railroad freight. In 2012, Chicago was listed as an alpha global city by the Globalization and World Cities Research Network, and it ranked seventh in the entire world in the 2017 Global Cities Index. The Chicago area has one of the highest gross domestic products (GDP) in the world, generating $680 billion in 2017. In addition, the city has one of the world's most diversified and balanced economies, not being dependent on any one industry, with no single industry employing more than 14% of the workforce.

[1] Wikipedia

## 1.2 Business Problem

Selection the best possible location for business applications is very important task for enterprises and business personnel. This is mainly because of high levels of competition.

This project is intended to recommend locations in the city of **Chicago** for various Business Solutions. The targeted audience for this project can be people interested in opening new businesses or in expanding pre-existing business setups, Government, Advertising Companies, Real estate, hospitality industries and many more.

# 2. Data

## 2.1 Requirements

1. List of neighbourhoods of Chicago.
2. Location coordinates of all the neighbourhoods.
3. Information of venues in each neighbourhood.

## 2.2 Sources

The list of neighbourhoods along with community area is available at this Wikipedia page: https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago.

Geopy libraries are used to acquire the location coordinates of each neighbourhood of Chicago.

Using FourSquare API, the venue information is obtained.

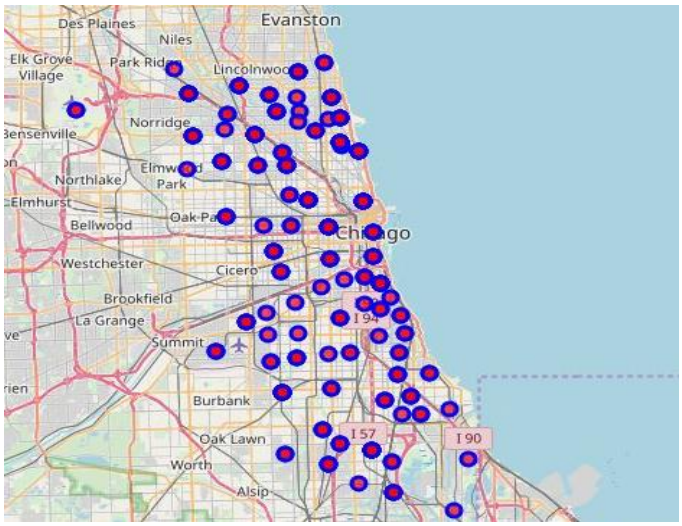
## 2.3 Data preparation and wrangling

With the use of Pandas library, the list of neighbourhoods and community areas is obtained from the wiki page and then this data is stored as a dataframe. Then using the community area names, the location coordinates are called using Geopy libraries creating two lists: “longitude” and “latitude”. It was found that the location coordinates of “Lincoln Square” which the Geopy library returned are not correct, therefore the rows containing “Lincoln Square” were dropped, thus cleaning the data. Then the two lists containing the location data were added to the dataframe as columns.

# 3. Methodology

## 3.1 Data analysis: Exploration

Using the Folium library, a map of Chicago is created using the location data provided by Geopy library. The by using the location data of each neighbourhood, Popup markers are plotted on the same map.



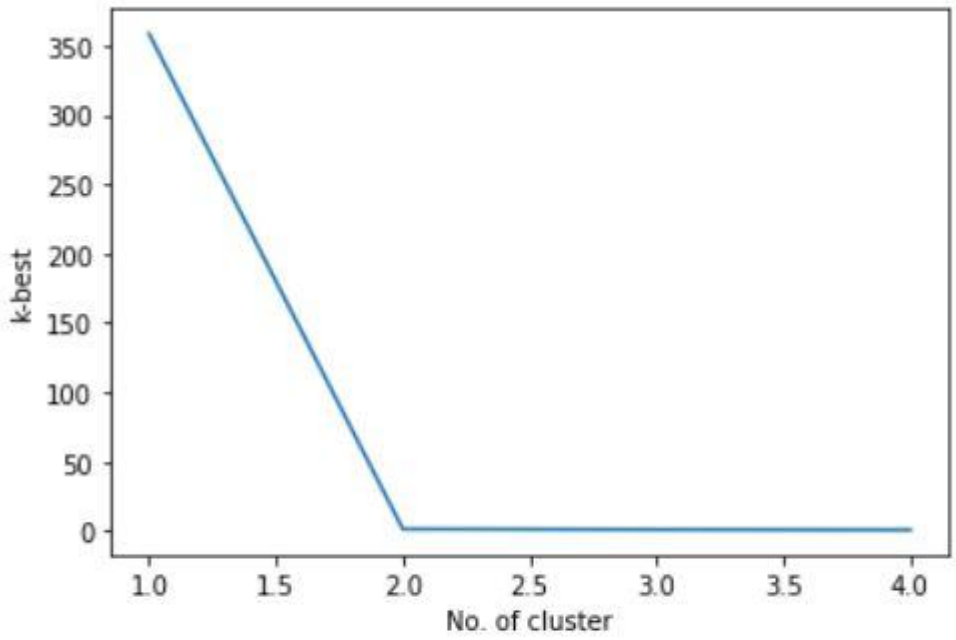
## 3.2 Data understanding

Once the dataframe contaning the list of Neibourhood  with their location data was ready, the venue data for all the neibouhood which is the target of this project.

For the venue data, Foursquare API was used. The API returned all the data regaurding the venues for each neighbourhood.

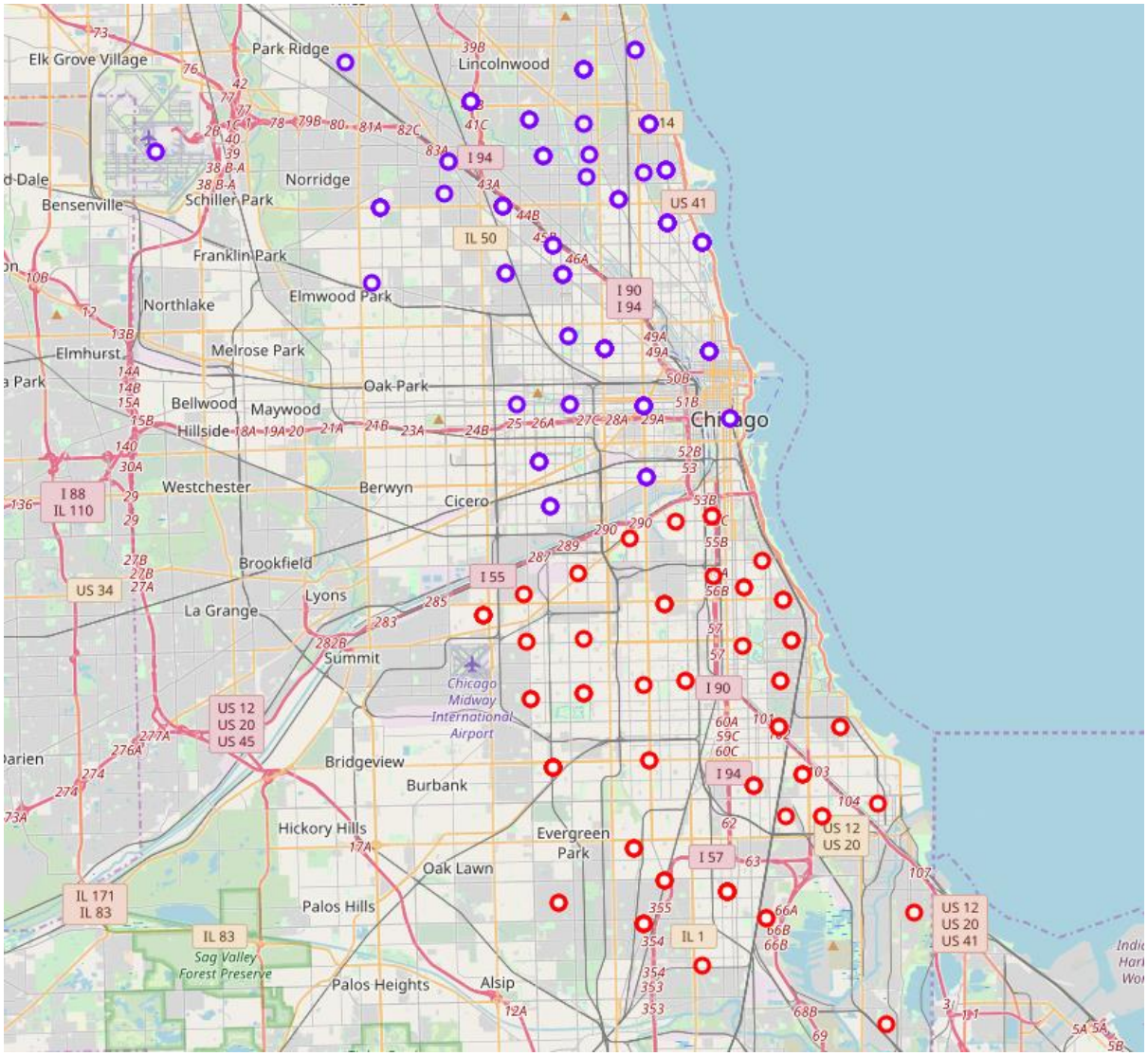
## 3.3 Clustering the Neighbouhoods

The data was divided into clusters on the basis of location coordinates. For clustering K-means clustering technique is used. Despite its simplicity, the K-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data. The best number of clusters was evaluated 2 as shown in the graph:




From the above graph, according to elbow method best value for K is 2

### 4.4 Visualizing clusters on map




# 4. Results

The final was dataset was divided into two clusters on the basis of location coordinates of the neighbourhoods. Both the clusters consist of equal number of neighbourhoods.

By examining the clusters, the interested personnel can comprehend the data of top 10 most common venues at each neighbourhood.

# 5. Discussion

The clustered dataset showed top ten most venues for each neighbourhood. The venues include

1. Food businesses like– various types of restaurants (Chinese, American, Italian, French, etc), food trucks, Bars, Desserts shops, food joints, Coffee shops etc.

2. Public places like– Parks, Bus stations, Banks, Gas stations etc.
3. Pharmacies and Doctors.
4. Shops and stores – like electronic stores, Supermarkets etc.
5. Various other businesses like– Construction, tailor shops, barber shops etc.
6. Gym/ Fitness Centres.

It was observed that in some neighbourhoods like Lincoln Park and North park there were no food businesses of any sort in the top ten categories.

In some neighbourhood top 3 most common venues were all related to fitness/health businesses. One such neighbourhood is Near North Side

Majority of the neighbourhoods have all the common venues related to food business.

# 6. Conclusion

This data can be used for various business application. For example, if a company wants to open a food joint which serves diet foods for customers who work-out and are diet conscious. This data model will help the company decide the location for the setup. Selecting a neighbourhood having most common venues as Gyms, fitness centres or yoga studios will be a good start.

Like wise this data model can be used by someone who is trying to relocate in the city. It will help to take decision giving the person an idea of the neighbourhood.

This model will be beneficial for all the location based business application in the City of Chicago.