Introduction

The owner of a successful French restaurant in New York is interested in opening a similar restaurant unit in Chicago, IL, USA. The first step he would like to take in this direction is to select a location in Chicago, but he does not know where to start. He is requesting some overall guidance on what areas to explore first.

His French restaurant is of medium-to-high price tier ($$$) and is considered more a formal dining restaurant than the average casual, fast food or bar experience. Consequently, this French restaurant is better designed to address fewer guests per day in longer dining experiences than high volumes of clients usually seen in very casual restaurants, fast food venues and bars. Hence, we should look for areas where there is some indication that similar sit-down dining experiences are of interest for local population.

Chicago is a large city with large diversity of people and interests. The city is divided into 77 different communities for statistical studies such as the Census. Some communities directly correspond to neighborhoods, others comprise of two or more neighborhoods.

The objective in this Data Science project is to identify which communities this French Restaurant owner should explore first when looking for a location to open his new restaurant in Chicago, IL, USA.

Data

Data requirements

For this analysis, I wanted to evaluate the differences between the various communities in Chicago by analyzing data related to population size and restaurants present in each of them. More specifically, the following are the variables (or features) I would like to leverage for this analysis:

	Numerical	**Categorical**
Population	Population size Population density	N/A
Ratio population/restaurants	People / Restaurant	N/A
Restaurants	Number of restaurants Price tier (on a 1$ to 4$ scale)	Restaurant categories

Data collection

After some online research, the below Wikipedia page proved to be a good source for three types of information:

Wikipedia page: https://en.wikipedia.org/wiki/Community_areas_in_Chicago

Relevant information:

- Names of Chicago's 77 communities
- Population in each community
- Population density per km2 in each community

The main table in this Wikipedia page, where the above information can be found, also provides duplicate or additional information that is not relevant for this analysis and will be removed after scraping the webpage, which I will explain in the Methodology section.

Continuing the data collection process, we need to get information on existing restaurants in each community and, for that, we were asked to leverage the Foursquare Places API. The following relevant variables are available through the Foursquare API

- Restaurant Names
- Restaurant Categories
- Restaurant Price Tier (on a 1 to 4 scale)

One limitation encountered in this data collection process is that Price Tier is considered a Premium endpoint in Foursquare which would require a Premium subscription, but our Budget for this project does not allow us to proceed with that subscription. So, we will drop this feature from the analysis and use only Restaurant Names and Category.

Finally, after collecting data from both Wikipedia and Foursquare, we can calculate two additional relevant variables:

- Count of restaurants per community
- Ratio average number of people per restaurant

With that, we conclude the process of collecting data and we can then proceed to cleaning, preparing and modelling the data, which will be discussed in the next section.