Part I – Introduction

§1.1 - Description of a Business Problem and its Background

Since 1865 with the founding of the Union Stock Yards, Chicago has been at the heart of high-quality American meats. Steak aficionados have long acclaimed Chicago beefsteaks being among the top grade nationwide. Tourists are flocking in to landmark steakhouses in the Windy City. Although there are approximately 1,734 steakhouses in River North community alone, launching a competitive modern steakhouse in Chicago could still be lucrative and help enrich the city's culture and history.

While thinking about an idea like that can be exciting even when the COVID-19 Pandemic hasn't really become a thing of the past, implementing it would be a much different story. One particularly vexing problem is – where exactly to open it?

As you may have heard repeatedly, "The three most important words for real estate are - location, location, and location" – Yes, this is 100% true for a steakhouse as well. Location determines volume of customer, cost of space, safety, competition, and much more. Anyone who wants to open a steakhouse will have to take it very seriously.

Using unsupervised machine learning technologies recently acquired from the Applied Data Science Capstone Course, I conducted a thorough study in an effort to provide useful insights to this challenging problem.

§1.2 - Who will be Interested?

For someone who wants to open a steakhouse in Chicago but can't really figure out where to launch the business, this analysis can help. Systematic Machine Learning approaches are adopted throughout this project, one can be confident about the result and conclusion we are going to unveil.

Furthermore, if the business problem becomes, for example, where to open a hair salon or a fitness center (i.e., a different venue), the methodology and most of the Python codes in this project can be reused exactly in their entirety.

Part II – Data Description

Data used in this project are mainly derived from three major sources:

- Wikipedia Websites
- Python GeoPy Library
- Foursquare venue database

§2.1 - Wikipedia Websites.

Tons of Chicago related information can be found on the internet. After careful comparing and selecting, I finally choose to use the Wikipedia webpages for Chicago neighborhood and community information, such as the following:

- https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago
- https://en.wikipedia.org/wiki/Community_areas_in_Chicago

Being the nation's 3rd largest city, Chicago has 284 neighborhoods which combined to form its 77 official communities. While dealing with 284 neighborhoods isn't particular overwhelming for this project, most publicly available Chicago city data are on community level rather than on neighborhood level. Therefore, this analysis is based on community level data.

The 2nd link above presented a table that contains Chicago Community Name, Population, Area Size, Population Density, etc. It can readily be used. I attached an image of this table (Top 5 rows) after it has been read into Jupyter Notebook by Python, see Figure 1 below:

	Number[8]	Name[8]	2017 population[9]	Area (sq mi.)[10]	Area (km2)	2017 populationdensity (/sq mi.)	2017 populationdensity (/km2)
0	01	Rogers Park	55062	1.84	4.77	29925.00	11554.11
1	02	West Ridge	76215	3.53	9.14	21590.65	8336.20
2	03	Uptown	57973	2.32	6.01	24988.36	9648.06
3	04	Lincoln Square	41715	2.56	6.63	16294.92	6291.50
4	05	North Center	35789	2.05	5.31	17458.05	6740.59

Figure 1: Snapshot of Chicago Community Overview

§2.2 - Python GeoPy Library.

As one may have noticed, the community table in 2.1 doesn't include geographical coordinates for the communities. These coordinates are necessary when one needs to fetch venue information later on. Fortunately, the Python GeoPy library (in particular the geocode module) can be employed to obtain the latitude and longitude data for the city of Chicago as well as its 77 official communities. I will then create Python codes to concatenate these coordinates to the community table.

Figure 2 below shows what the community table looks like when concatenated with their corresponding geographical coordinates (Top 10 rows):

	Number[8]	Name[8]	2017 population[9]	Area (sq mi.) [10]	Area (km2)	2017 populationdensity (/sq mi.)	2017 populationdensity (/km2)	Latitude	Longitude
0	01	Rogers Park	55062	1.84	4.77	29925.00	11554.11	42.00897	-87.66619
1	02	West Ridge	76215	3.53	9.14	21590.65	8336.20	41.99948	-87.69266
2	03	Uptown	57973	2.32	6.01	24988.36	9648.06	41.96538	-87.66936
3	04	Lincoln Square	41715	2.56	6.63	16294.92	6291.50	41.97583	-87.68914
4	05	North Center	35789	2.05	5.31	17458.05	6740.59	41.95411	-87.68142
5	06	Lake View	100470	3.12	8.08	32201.92	12433.23	41.93982	-87.65682
6	07	Lincoln Park	67710	3.16	8.18	21427.22	8273.10	41.92184	-87.64744
7	08	Near North Side	88893	2.74	7.10	32442.70	12526.20	41.90034	-87.63433
8	09	Edison Park	11605	1.13	2.93	4235.40	1635.30	42.00789	-87.81399
9	10	Norwood Park	37089	4.37	11.32	8487.19	3276.92	41.98572	-87.80664

Figure 2: Snapshot of Chicago Communities with Geographical Coordinates

§2.3 - Foursquare API for venue information.

Venue information can be obtained via several different sources, and each has its own Pros and Cons. Foursquare is becoming increasingly popular and we just learned it in the Applied Data Science Capstone Course. So I decide to use Foursquare API to obtain venue information for each of the 77 Chicago communities. These venues will include basically everything – supermarkets, bars, restaurants, parks, gyms, libraries, etc.

Once obtained, the information is used to rank the popular venues for each community. Based on that, an unsupervised machine learning technology known as K-Means Clustering can be developed to group the communities into several clusters. The clusters are expected to contain indicative information as to whether or not a steakhouse is a likely fit here, and I will be able to eliminate all but one cluster.

Finally, I will use some other criteria (such as the ratio of population over # of steakhouse and published community reviews) to determine which community is the best candidate for opening a steakhouse.

Figure 3 below shows an example of some venues in one community (Top 10 rows):

	Community	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rogers Park	42.00897	-87.66619	Morse Fresh Market	42.008087	-87.667041	Grocery Store
1	Rogers Park	42.00897	-87.66619	Rogers Park Social	42.007360	-87.666265	Bar
2	Rogers Park	42.00897	-87.66619	Lifeline Theatre	42.007372	-87.666284	Theater
3	Rogers Park	42.00897	-87.66619	Rogers Park Provisions	42.007528	-87.666193	Gift Shop
4	Rogers Park	42.00897	-87.66619	Mayne Stage	42.007975	-87.665140	Concert Hall
5	Rogers Park	42.00897	-87.66619	J.B. Alberto's Pizza	42.007941	-87.665066	Pizza Place
6	Rogers Park	42.00897	-87.66619	The Common Cup	42.007797	-87.667901	Coffee Shop
7	Rogers Park	42.00897	-87.66619	Glenwood Sunday Market	42.008525	-87.666251	Farmers Market
8	Rogers Park	42.00897	-87.66619	The Glenwood	42.008502	-87.666273	Bar
9	Rogers Park	42.00897	-87.66619	Smack Dab	42.009291	-87.666201	Bakery

Figure 3: Snapshot of Chicago Community Venues Obtained via Foursquare API