IBM DATA SCIENCE CAPSTONE

WHERE TO LIVE FOR A YOUNG PROFESSIONAL IN CHICAGO, IL

BY: SHELLEY LEUNG

JUNE 2020

TABLE OF CONTENTS

Introduction

Chicago is one of the most populous and growing US cities in the country. It is the home to plenty of arts, entertainment, sports, education, research and much more. No wonder Bean-town is attracting plenty of young professionals moving into the busy city each year. Chicago is a relatively young city with the person's average age being 34.9 years old and 63% of the population being single out of 2.7 million residents and growing.

With a growing city full of young people moving to Chicago looking for entertainment and adjusting to life in a new city, there has been demand on creating programs to research and provide information to those new to the city, create guides for tourists and keep the city's economy booming.

In this project, we will use location data and clustering techniques to determine which areas of Chicago has the most activity and human traffic. This can help a new transplant select which is an accessible and quiet neighborhood to potentially select their new home.

Business Research

In this research project, we will explore the different neighborhoods in the Greater Chicago area. However, due to the vast amount of different community areas in Chicago, we will focus solely on the inner city areas. The Windy City has attracted plenty of tourists all year round and because of job opportunities, it has also attracted many young professionals to the city. We will explore the different neighborhoods and describe which is best for young professionals to settle into their new big city life.

Target Audience

As mentioned earlier, because of the vast amount of career opportunity Chicago has proven to provide to young professionals, it is one of the most desired cities for career growth. With this attractive offer, many young professionals have flocked here. Every new person in the city would not know the neighborhoods well. Therefore, this research is to provide an overview of what each neighborhood in Chicago can provide for their new resident.

Data

To solve the problem, the following data sources were used:

- List of neighborhoods in Chicago: This helps define all the different neighborhoods within the great area.
- Latitude and Longitude coordinates of the neighborhoods: This is required to create a visual of the city and determine which neighborhoods sit in the urban parts of Chicago.
- Venue data: Populating the different types of venues within each urban neighborhood will help determine the convenience of the area, the entertainment, restaurants and more.

Data Source:

- *Wikipedia:* This wiki page contains a list of every neighborhood in the Greater Chicago area. There are about 246 neighborhoods in the Greater Chicago area. We will use web scraping techniques with the *BeautifulSoup* library in python and *Geocoder* to obtain the geographical coordinates of the neighborhoods.
  https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago

- *Foursquare API:* Foursquare API location data is a large database with over 105+ million places worldwide. Foursquare will be able to provide the different venues around the inner city neighborhoods in Chicago and the category types of each.

Methodology

1. Web scrape data from Wikipedia to create a dataframe for the list of neighborhoods (sample view):

	Neighborhood
0	Albany Park
1	Altgeld Gardens
2	Andersonville
3	Archer Heights
4	Armour Square

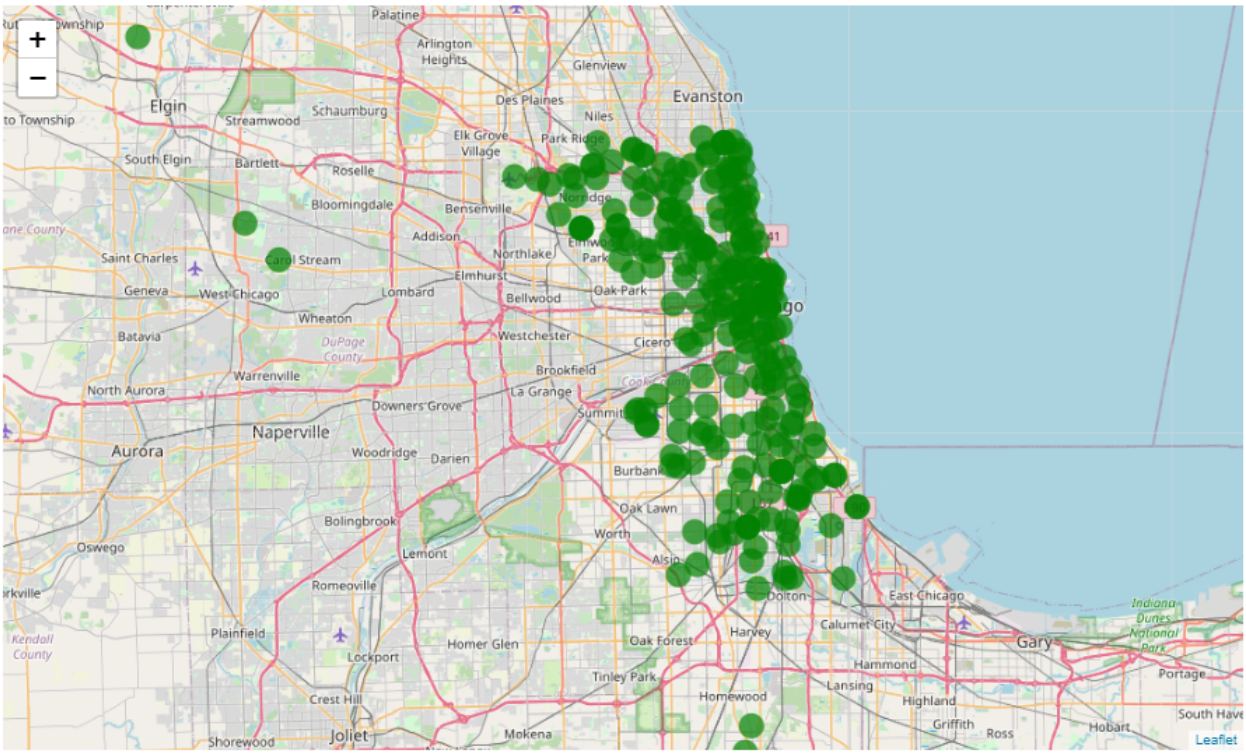
2. Get geographical coordinates of the neighborhoods to complete the set of data required to map the neighborhoods in Chicago (sample view):

	Neighborhood	Latitude	Longitude
0	Albany Park	41.96829	-87.72338
1	Altgeld Gardens	41.65441	-87.60225
2	Andersonville	41.98046	-87.66834
3	Archer Heights	41.81154	-87.72556
4	Armour Square	41.83458	-87.63189

3. Obtain the coordinates of Chicago and plot all 246 neighborhoods using folium maps:

```
41.8755616 -87.6244212
```

(map_chi)

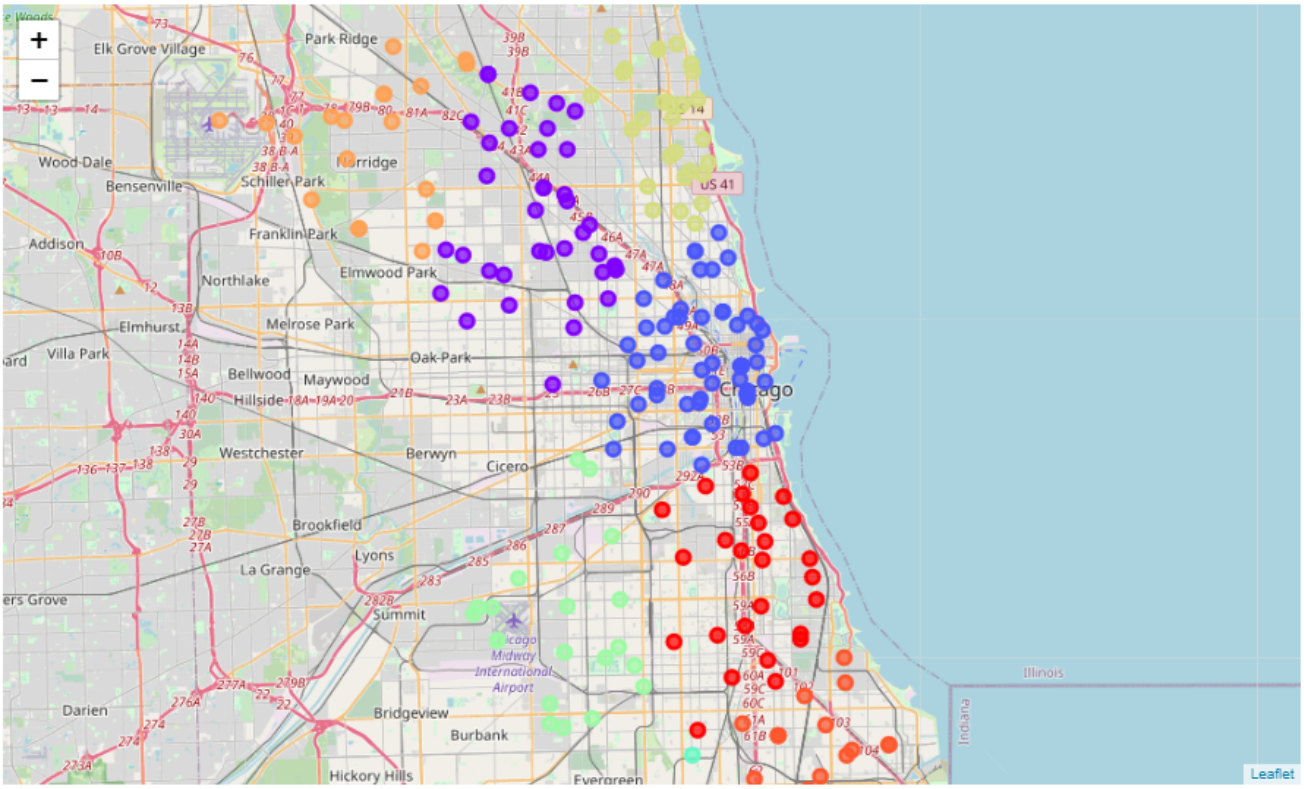


Based on the view of the map and the abundant amount of neighborhoods, it is difficult to determine which neighborhoods belong in an urban area. However, we do know that in the center of the city, that is an approximate location of the urban areas.

4. We will use a machine learning method, k-Means clustering, to group all 246 neighborhoods to their 10 different clusters.

	Neighborhood	Latitude	Longitude	Cluster Labels
0	Albany Park	41.96829	-87.72338	1
1	Altgeld Gardens	41.65441	-87.60225	5
2	Andersonville	41.98046	-87.66834	7
3	Archer Heights	41.81154	-87.72556	6
4	Armour Square	41.83458	-87.63189	0

K-Means clustering is a form of unsupervised machine learning that identifies the number of centroids and allocates each data point within close proximity to the centroid. This is a great method to identify all of the neighborhoods and the distance to the center of urban Chicago.
(map_clusters)



5. We noticed that a blue cluster (Cluster 2) contains a group of neighborhoods centered closest or in the city center of Chicago. We will filter out for only Cluster 2 using boolean methods and focus on the 60 neighborhoods in the cluster by creating a new dataframe for Cluster 2.

```
(60, 4)
There are 60 neighborhoods in the city area of Chicago based on the total in Cluster 2.
```

	Neighborhood	Latitude	Longitude	Cluster Labels
64	Fifth City	41.884250	-87.632450	2
122	Lower West Side	41.852240	-87.671150	2
219	Wacławowo	41.884250	-87.632450	2
128	Marynook	41.846188	-87.653358	2
233	West Loop	41.883040	-87.653350	2

6. Now, we will use Foursquare API data to obtain venue data and merge into a single dataframe with Cluster 2 neighborhoods. Then group each venue with their respective neighborhood. Due to the limitations with Foursquare Developer, the maximum retrieved were 100 venues per neighborhood.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
5946	East Garfield Park	41.87863	-87.70514	Tony's Italian Beef 4100 W Madison	41.881121	-87.728168	Fast Food Restaurant
5947	East Garfield Park	41.87863	-87.70514	Baba Pita	41.868758	-87.685328	Middle Eastern Restaurant
5948	East Garfield Park	41.87863	-87.70514	Iron Mountain	41.864173	-87.690909	Business Service
5949	East Garfield Park	41.87863	-87.70514	Mason (Elizabeth) Park	41.883345	-87.728445	Park
5950	East Garfield Park	41.87863	-87.70514	McDonald's	41.881077	-87.727853	Fast Food Restaurant

7. After obtaining the venues data and selecting the proper neighborhoods, we will analyze each neighborhood by grouping each by venue categories and the top 10 most frequent categories per neighborhood.
(Top 10 most popular venues in Cluster 2)

	Count
Italian Restaurant	246
Mexican Restaurant	237
Coffee Shop	232
Hotel	213
Park	187
Pizza Place	175
New American Restaurant	146
Bar	144
Grocery Store	138
Theater	133

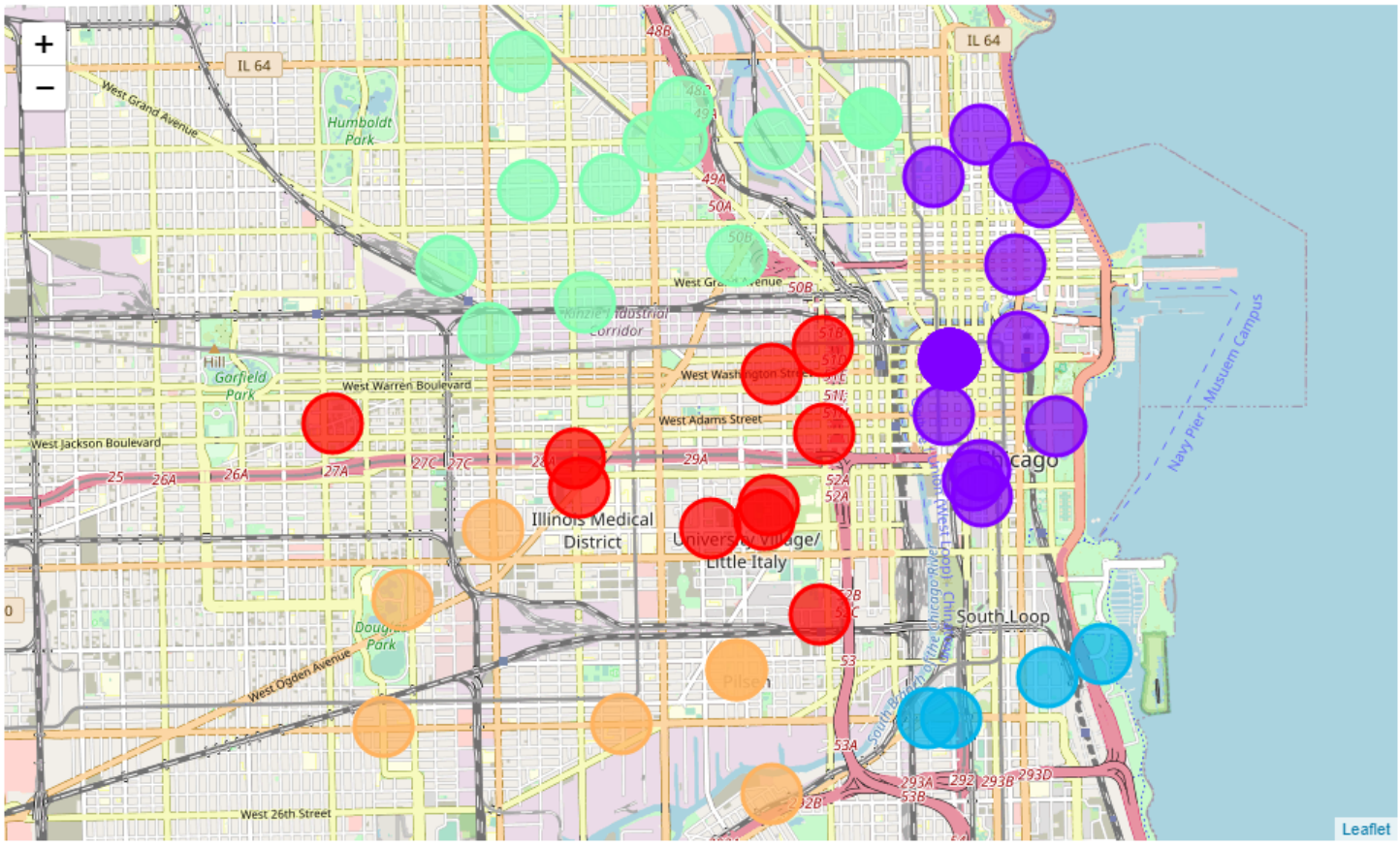
8. After performing this analysis, we revisit the same machine learning method, k-Means clustering, to combine the venue data and the neighborhood data in order to examine each neighborhood closely. For this time, the clustering methods will focus on the venue data.

Results: Analyzing Each Neighborhood

(Sample view of the final results. Total number of clusters: 5)

	Neighborhood	Latitude	Longitude	Cluster Labels	Cluster_Label 2	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
109	Lake View East	41.885757	-87.624328	2	1	Hotel	Theater	New American Restaurant	Park	Italian Restaurant	Coffee Shop	Steakhouse
195	Sheridan Park	41.870150	-87.654250	2	0	Italian Restaurant	Coffee Shop	Grocery Store	Hot Dog Joint	Sandwich Place	Park	Mexican Restaurant
198	Smith Park	41.892410	-87.691670	2	3	New American Restaurant	Coffee Shop	Restaurant	Breakfast Spot	Brewery	Deli / Bodega	Flower Shop
127	Marshall Square	41.851993	-87.699064	2	4	Mexican Restaurant	Sandwich Place	Italian Restaurant	Taco Place	Bank	Gas Station	Fast Food Restaurant
124	Magnificent Mile	41.900580	-87.624210	2	1	Hotel	American Restaurant	Italian Restaurant	Coffee Shop	Café	Grocery Store	Pizza Place

- ***Cluster 0 (Red):*** Contains the most Italian, Mexican, fast food, and New American restaurants. There is also an abundant number of coffee shops located in this cluster.
- ***Cluster 1 (Purple):*** Contains the most hotels and theaters.
- ***Cluster 2 (Aqua Blue):*** Contains the most Chinese restaurants and parks.
- ***Cluster 3 (Light Green):*** Contains the most diverse number of venues that range from bars, restaurants, gyms, coffee shops, grocery stores and more.
- ***Cluster 4 (Orange):*** Contains the most Mexican restaurants, Sandwich places and bars.



Observations and Recommendations

After analyzing each neighborhood, here are some additional observations and recommendations:

- If an abundant choice of restaurants such as Italian, Mexican, American and coffee shops is a priority, Cluster 0 contains one of the most in those areas.
- Cluster 1 contains the most hotels and theaters. Therefore is most likely a tourist or business destination for visitors.
- Cluster 2 contains the most parks and also the home to Chicago's two Chinatowns. Therefore this area contains many chinese restaurants, dim sum places and asian restaurants.
- Cluster 3 is the most diverse among the group with a variety of top venues in each neighborhood within the cluster. This is the best recommendation because not only does it contain entertainment, it also has plenty of grocery stores, gyms, barber shops and other necessities to choose from.
- If Mexican restaurants, bars and sandwich shops are a priority, neighborhoods in Cluster 4 would be an excellent choice.

Limitations & Suggestions for further Research

Through this research project, there were some limitations in order to keep the focus of the research on venue data in the Windy City. Despite this, there are also opportunities to further expand the research in these areas to create a better experience for the target audience.

- Foursquare was only able to retrieve 100 venues per neighborhood for my data unless I pay for an upgrade. Therefore the venue data was limited.
- Crime rates in each neighborhood were not calculated and built into this research project. Main reason is to focus on the activities and venues in each neighborhood.
- Affordability was not factored in due to the large range of income each individual may have and the difference in lifestyles for the target audience.
- Census statistics for age groups residing in these neighborhoods were not obtained due to the focus on venues and not census information.

Conclusion

Excluding any limitations to the research, neighborhoods in cluster 3 seems to be the most diverse and ideal area to settle in for a new transplant in the Windy City. There are 14 neighborhoods in this cluster:

1. Smith Park	2. Pulaski Park	3. The Villa
4. Noble Square	5. Old Town	6. East Village
7. Old Town Triangle	8. Wicker Park	9. Fulton River District
10. Bucktown	11. Ukranian Village	12. Goose Island
13. West Town	14. Heart of Chicago	

References

1. Census data reference: https://censusreporter.org/profiles/16000US1714000-chicago-il/
2. Wikipedia: https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago
3. Foursquare Developer API