# Bubble Tea Expansion

Shuaib Shameem

August 31, 2019

## 1. Introduction

### 1.1 Background

A bubble tea restaurant, based in Schaumburg, IL, is having great success since opening less than two years ago. With this success, they have managed to bring their accounts out of the red, and are already making a profit off the business. The owners of the business have decided to reinvest their profits into the business by expanding to a second location, this time in Chicago, IL. The owners believe that the business is running well due to the culture and socioeconomic status of the neighborhood around them.

### 1.2 Problem

Given the surroundings of the current business location, the owners would like to find a similar neighborhood in the Chicago area. This project aims to identify community areas in Chicago that feature similar businesses and socioeconomic status to the original location.

## 2. Data acquisition and cleaning

### 2.1 Data Sources

The established community areas of Chicago will form our list of candidate locations for the business expansion. The Chicago Metropolitan Agency for Planning (CMAP) provides these community areas, along with demographic information for each area, in yearly Community Data Snapshots. Data for Schaumburg is available in a separate Municipal Area Snapshot. CMAP provides a PDF breaking down the table labelling and data sources.

In order to provide map visualizations, we accessed the OpenStreetMap API via GeoPy by passing in the names of each location to get the coordinates.

Finally, we utilized the Foursquare API to retrieve local business information by passing in the coordinates provided by OpenStreetMap.

## 2.2 Data Cleaning

The CMAP tables provided a wide range of demographic information, most of which were not needed. From these tables, we scraped the names, total population, age cohorts, educational attainment, and household income. These data points, along with the coordinates from OpenStreetMap, were consolidated into one table. The data for age, education, and income were provided in terms of counts, which do not lend themselves to comparison. Therefore, these values were converted to percentages. When necessary for visualization, the median age and median income values were normalized.

| GEOG | TOT_POP | UND19 | A20_34 | A35_49 | A50_64 | A65_74 | A75_84 | OV85 |
|---|---|---|---|---|---|---|---|---|
| Edgewater | 55965 | 0.142142 | 0.290003 | 0.223265 | 0.198445 | 0.085142 | 0.039989 | 0.021013 |
| Lincoln Square | 41715 | 0.176339 | 0.321994 | 0.249982 | 0.159751 | 0.052235 | 0.023948 | 0.015750 |
| Logan Square | 73046 | 0.209293 | 0.377420 | 0.227186 | 0.118898 | 0.039537 | 0.020658 | 0.007009 |
| Near West Side | 62872 | 0.189226 | 0.426199 | 0.204527 | 0.107822 | 0.045744 | 0.021297 | 0.005185 |

An excerpt from the demographics table

We obtained Foursquare data by passing in coordinates for each location and a radius to search within, limited to 100 venues. A search radius of 1.5 km worked well for the Chicago community areas. However, due to its nature as a suburb, the search radius for Schaumburg was set to 3.3 km. The data from these requests was stripped down to the name, coordinates, and category for each business provided, and collected into one table. As we are only interested in the types of businesses in each community, this data was processed into a new table to provide percentages for the presence of each type of business in a given community. This allowed us to make a list of the top 10 types of businesses for each community.

| Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue |
|---|---|---|---|---|---|---|
| Albany Park | Pizza Place | Coffee Shop | Korean Restaurant | Chinese Restaurant | Mexican Restaurant | Sandwich Place |
| Archer Heights | Mexican Restaurant | Sandwich Place | Mobile Phone Shop | Bank | Discount Store | Bar |
| Armour Square | Chinese Restaurant | Bar | Pizza Place | Sandwich Place | Park | Bakery |

An excerpt from the business frequency table

# 3. Data analysis

## 3.1 Methodology

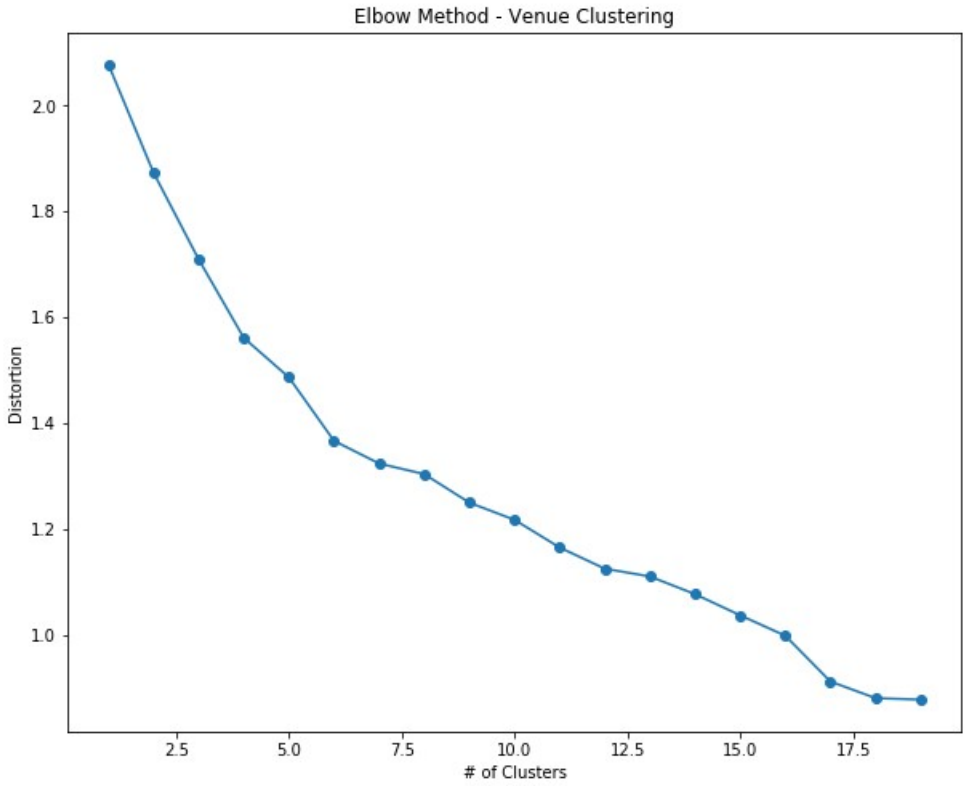
As previously stated, the aim of this project is to find communities that are similar to Schaumburg, the city where the original location is established. To this end, K-Means clustering was

deemed the best fit for processing the data. Clustering on all of the data at once is not viable due to the demographic data being numeric, while the business data is non-numeric.

Sequentially clustering on business types and then demographics (and vice versa) was considered, however, there would not be enough communities left after the first iteration to allow for a second. For this reason, the communities were clustered twice, once by business type and once by demographics. We then extracted communities that were present in the both outputs, and provided visualizations and tables for the remaining communities to build our conclusions on.

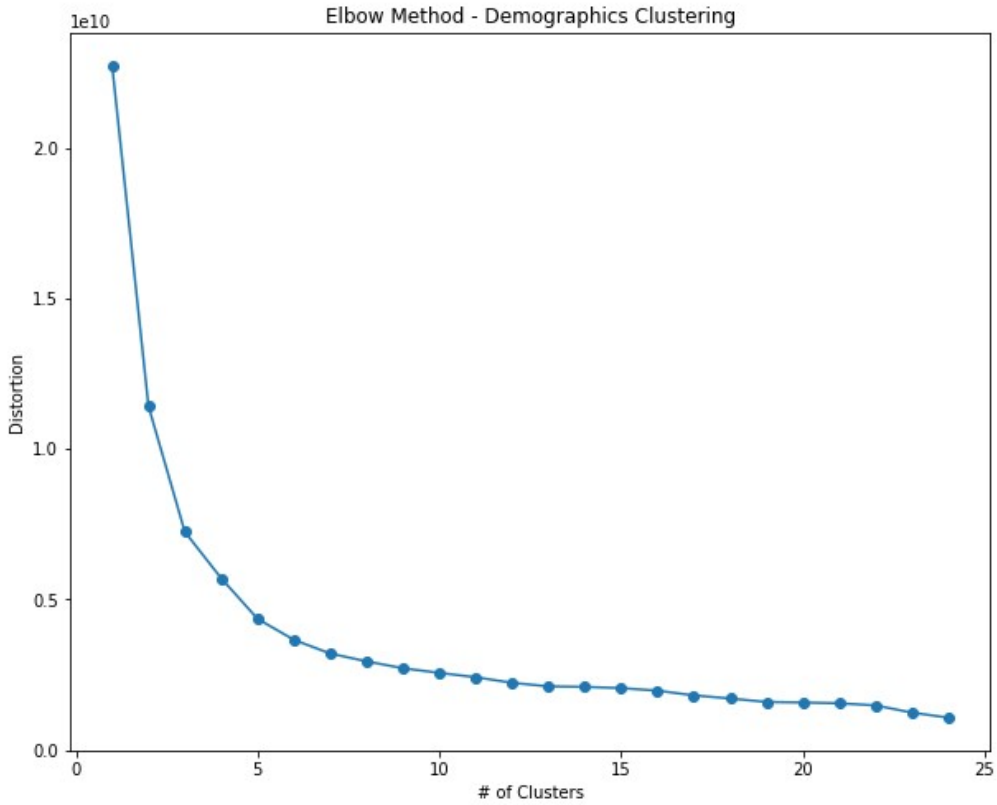
## 3.2 Clustering optimization

To determine the optimal number of clusters for each of the datasets, I utilized the elbow method. For the business data, this method did not yield a very useful plot:




There is no clear elbow with which to decide the number of clusters. I reasoned that there seems to be an elbow-like kink around 17 clusters, and distortion fell below 1.0 at 16 clusters. Therefore, I decided to use 17 clusters on the business data.

Applying the elbow method on the demographic data provided the following plot:

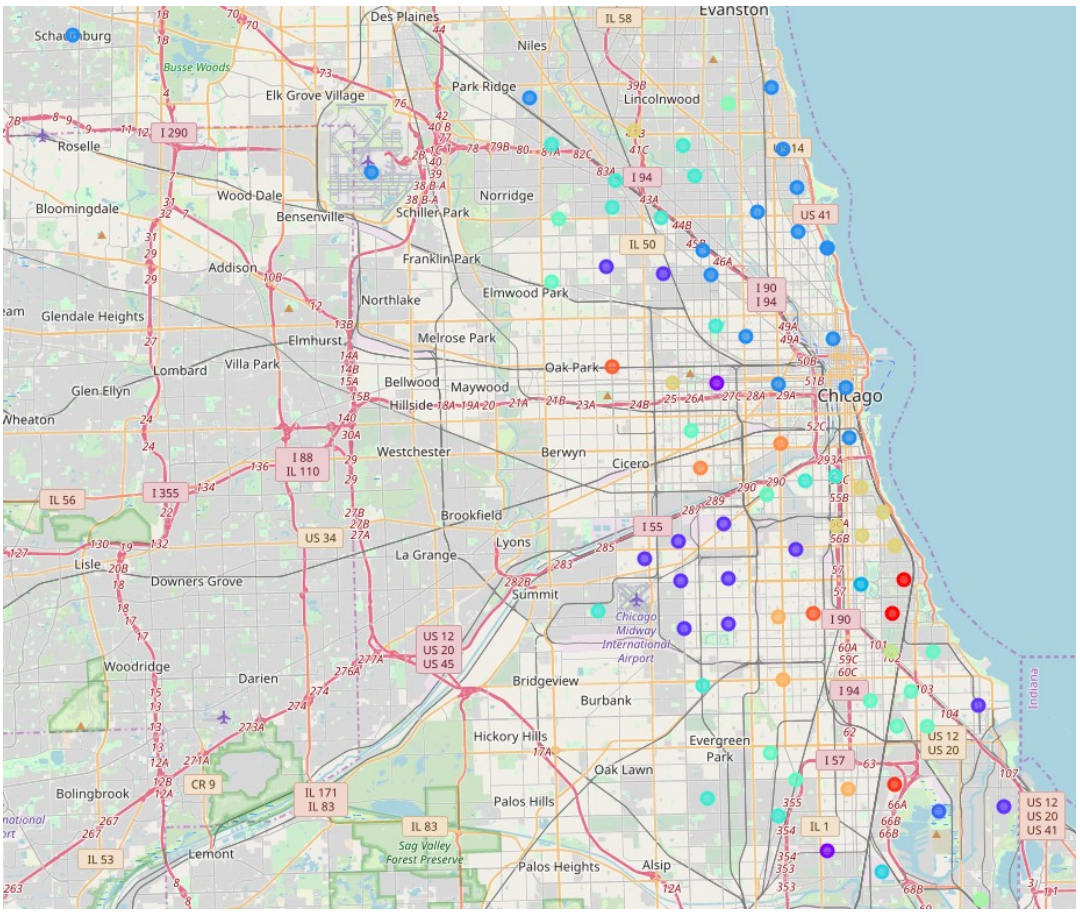



The elbow is much more defined, despite spanning several cluster options. Five clusters falls within that span, and provides an acceptably small distortion.

## 3.3 Clustering

### 3.3.1 Business type

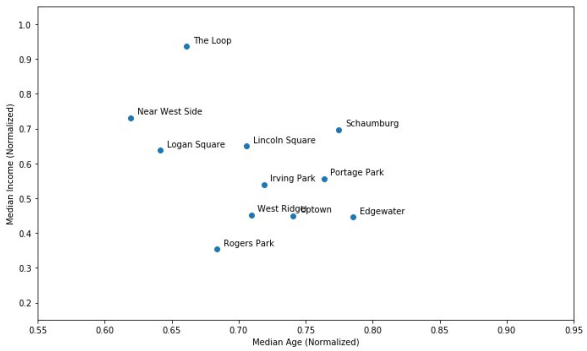
Utilizing 17 clusters, I ran K-Mean clustering on the business data. To visualize the result, each location was placed on a map with color-coded markers for each cluster:

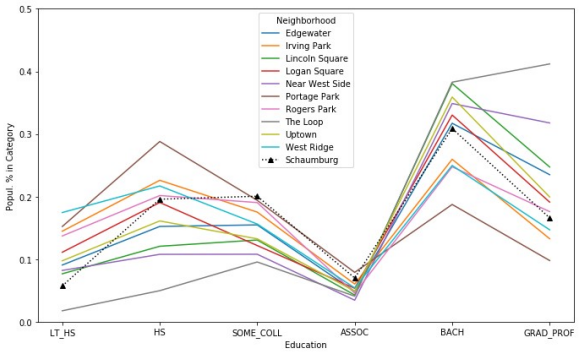


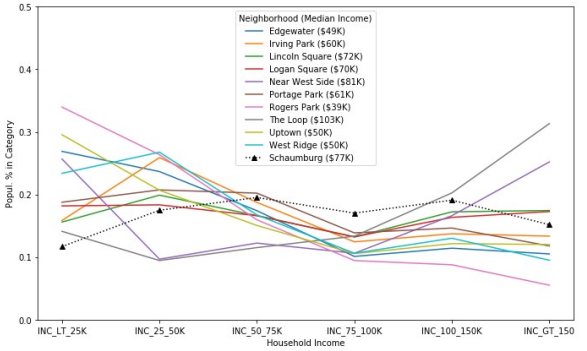

While the clustering of each community is not vital, we can see all the locations that are clustered with Schaumburg (Blue), such as O'Hare and other communities north of Chicago. In total, there were 16 community areas clustered with Schaumburg, out of the original list of 77.

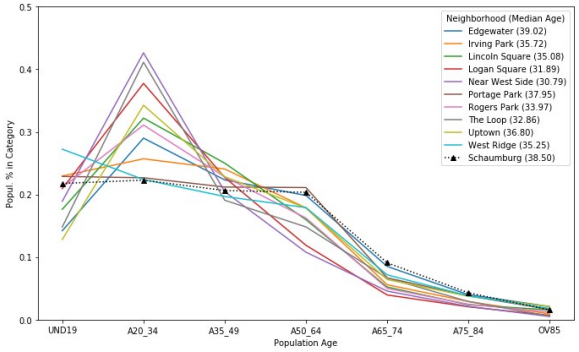
### 3.3.2 Demographics

Running K-Means on the demographic table with five clusters resulted in 10 community areas which were similar to Schaumburg. To visualize the similarities, I created a scatter plot of the Median Income vs Median Age, and three line plots for the education attainment, household income, and age for each of the communities.



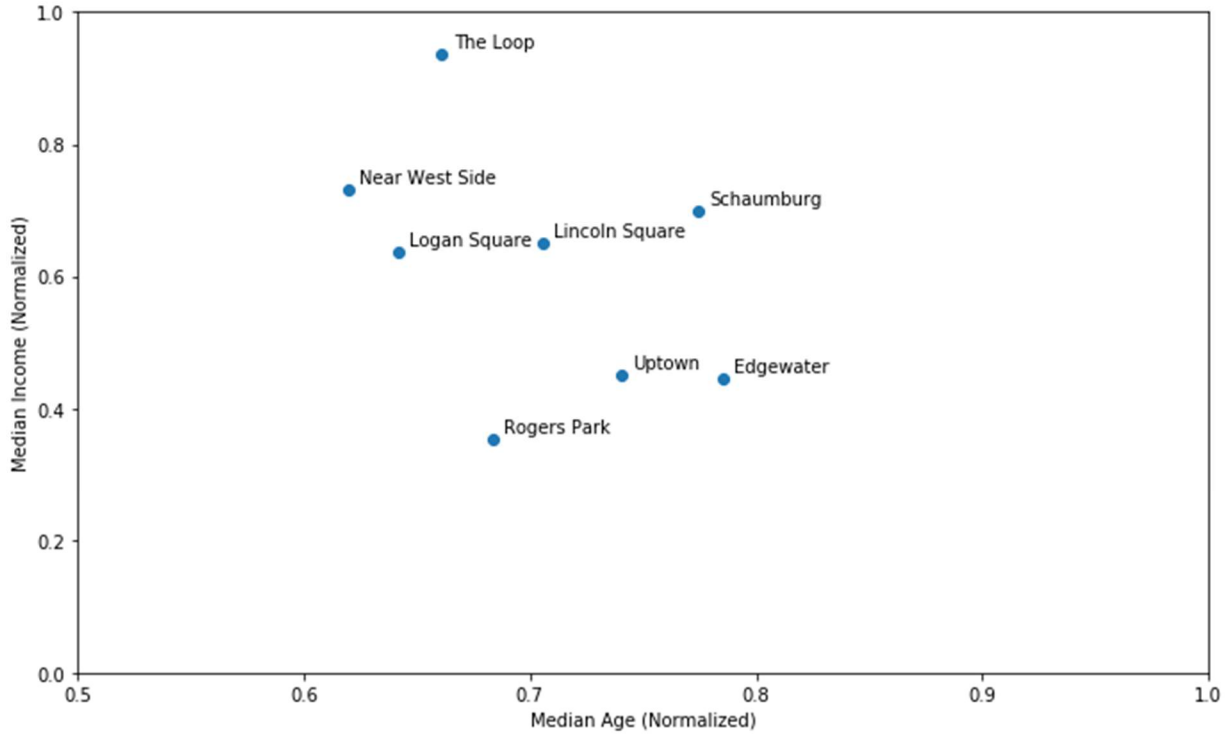



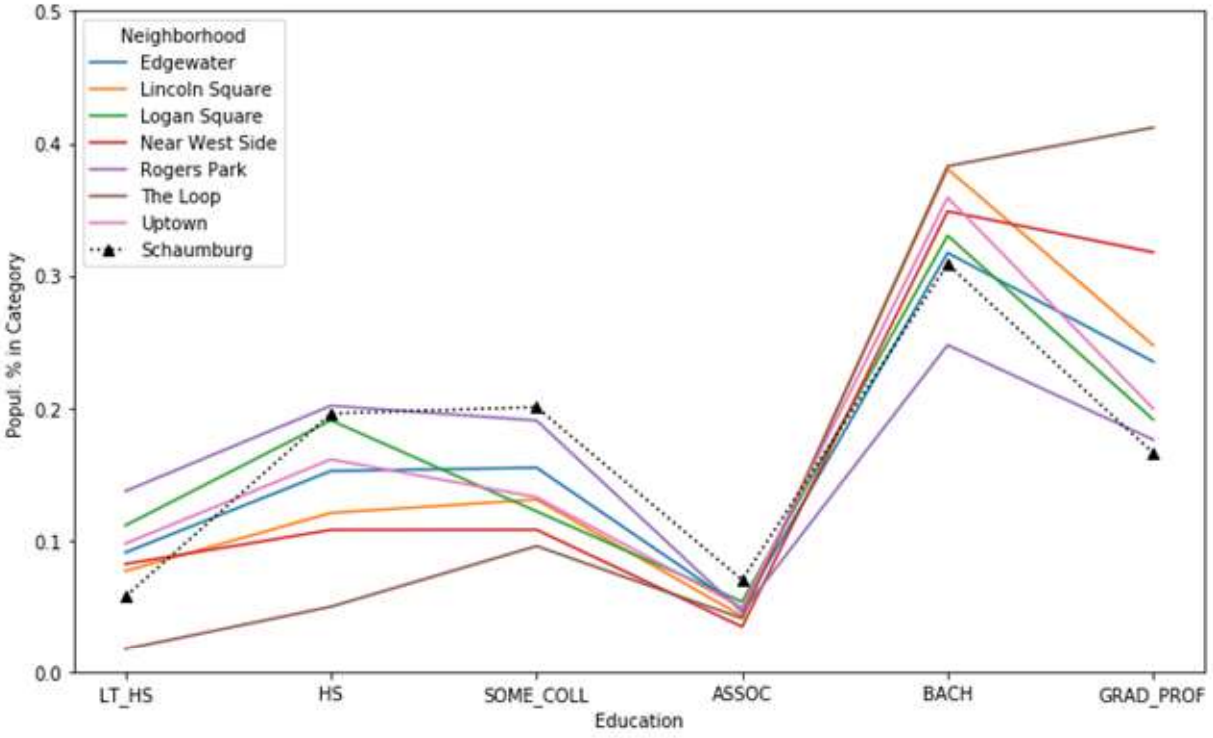

## 3.4 Results

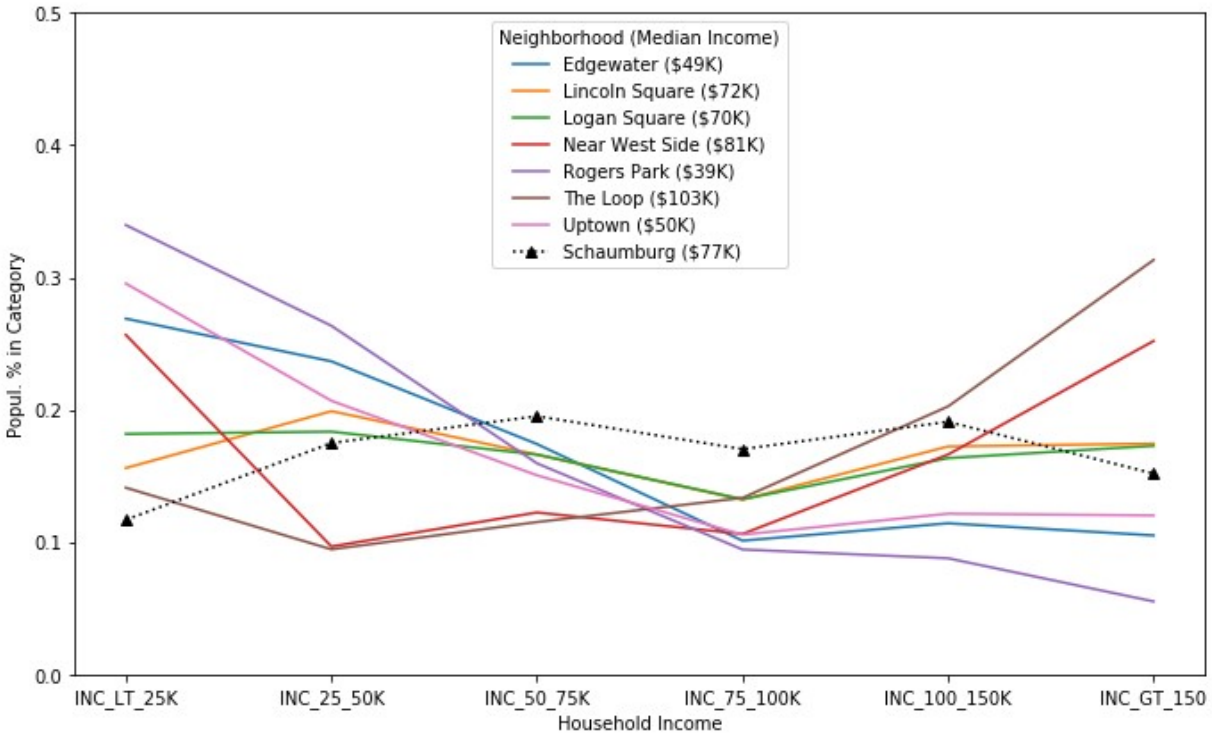
Given the output of the two clustering iterations described, we created a final list of candidate locations by extracting the data for only the community areas that are present in both outputs.

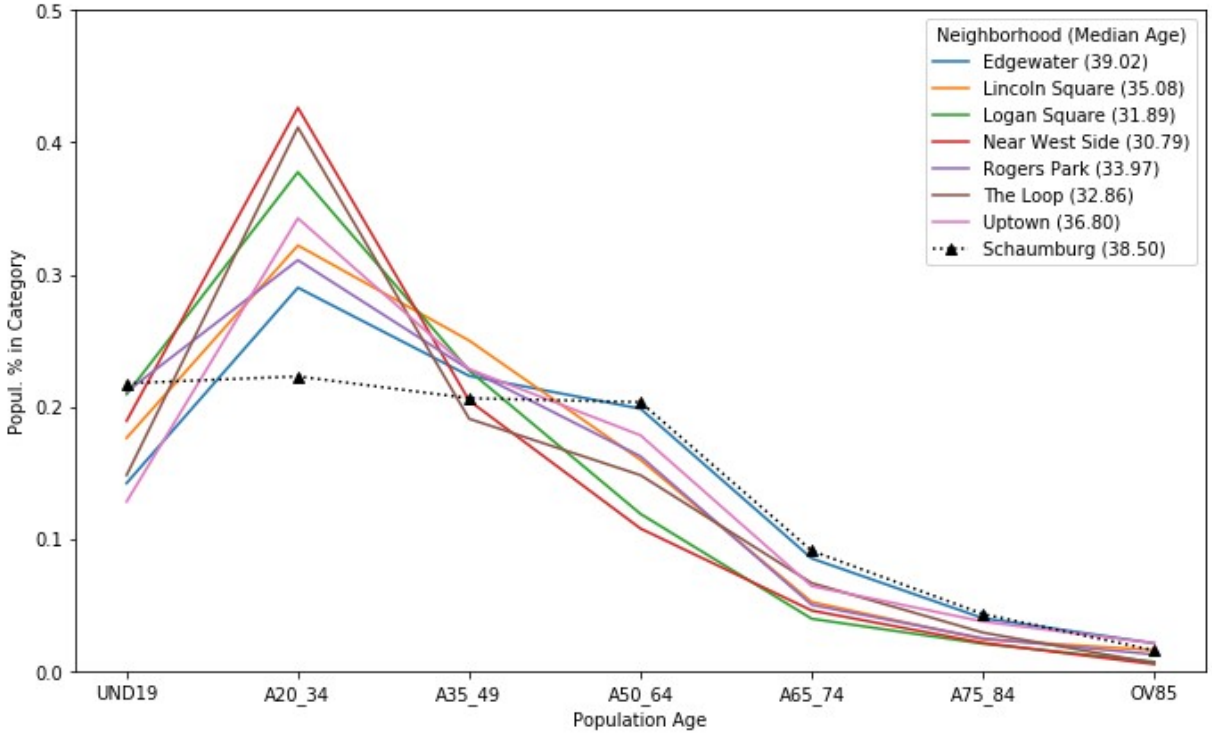
- Edgewater
- Lincoln Square
- Logan Square
- Near West Side
- Rogers Park
- The Loop
- Uptown

Then we put the final list to the same visualizations applied to the clustering outputs, where we can see how similar the community areas are to Schaumburg.


The Loop
Near West Side
Schaumburg
Logan Square
Lincoln Square
Uptown
Edgewater
Rogers Park
Median Income (Normalized)
Median Age (Normalized)
1.0
0.8
0.6
0.4
0.2
0.0
0.5
0.6
0.7
0.8
0.9
1.0


Neighborhood
Edgewater
Lincoln Square
Logan Square
Near West Side
Rogers Park
The Loop
Uptown
Schaumburg
Popul. % in Category
Education
LT_HS
HS
SOME_COLL
ASSOC
BACH
GRAD_PROF
0.5
0.4
0.3
0.2
0.1
0.0


Neighborhood (Median Income)
Edgewater ($49K)
Lincoln Square ($72K)
Logan Square ($70K)
Near West Side ($81K)
Rogers Park ($39K)
The Loop ($103K)
Uptown ($50K)
Schaumburg ($77K)
Popul. % in Category
0.0
0.1
0.2
0.3
0.4
0.5
INC_LT_25K
INC_25_50K
INC_50_75K
INC_75_100K
INC_100_150K
INC_GT_150
Household Income


Neighborhood (Median Age)
Edgewater (39.02)
Lincoln Square (35.08)
Logan Square (31.89)
Near West Side (30.79)
Rogers Park (33.97)
The Loop (32.86)
Uptown (36.80)
Schaumburg (38.50)
Popul. % in Category
0.0
0.1
0.2
0.3
0.4
0.5
UND19
A20_34
A35_49
A50_64
A65_74
A75_84
OV85
Population Age

| Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|
| Edgewater | Coffee Shop | Italian Restaurant | Mexican Restaurant | Asian Restaurant | Sandwich Place |
| Lincoln Square | Mexican Restaurant | Sushi Restaurant | Bakery | Coffee Shop | Vegetarian / Vegan Restaurant |
| Logan Square | Cocktail Bar | Coffee Shop | Café | Bar | Mexican Restaurant |
| Near West Side | Italian Restaurant | Pizza Place | Coffee Shop | Yoga Studio | Breakfast Spot |
| Rogers Park | Beach | Sandwich Place | Mexican Restaurant | Pizza Place | Café |
| Schaumburg | Fast Food Restaurant | Grocery Store | Hotel | Hookah Bar | Pizza Place |
| The Loop | Theater | Coffee Shop | Hotel | Park | Pizza Place |
| Uptown | Vietnamese Restaurant | Coffee Shop | Sushi Restaurant | Chinese Restaurant | Mexican Restaurant |

| 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|
| Breakfast Spot | Vietnamese Restaurant | Grocery Store | Sushi Restaurant | Burger Joint |
| Asian Restaurant | Grocery Store | Italian Restaurant | Gay Bar | Bar |
| Latin American Restaurant | Brewery | Pizza Place | Park | Ice Cream Shop |
| Brewery | Sandwich Place | Café | Park | Deli / Bodega |
| Fast Food Restaurant | Park | African Restaurant | Bakery | Bar |
| Gym | Coffee Shop | Bakery | Korean Restaurant | Bookstore |
| Museum | Snack Place | Bakery | Sandwich Place | Salad Place |
| Park | Grocery Store | Bar | Thai Restaurant | Breakfast Spot |

| GEOG | TOT_POP | MED_AGE | MEDINC |
|---|---|---|---|
| Edgewater | 55965 | 39.022775 | 49287.01181 |
| Lincoln Square | 41715 | 35.079103 | 71736.93086 |
| Logan Square | 73046 | 31.885020 | 70338.93986 |
| Near West Side | 62872 | 30.789626 | 80727.42475 |
| Rogers Park | 55062 | 33.972649 | 39106.18280 |
| The Loop | 35880 | 32.859723 | 103336.27930 |
| Uptown | 57973 | 36.801845 | 49680.90452 |
| Schaumburg | 74427 | 38.500000 | 77022.00000 |

From these visualizations, we can make a few generalizations. Firstly, educational attainment has a very similar spread amongst all of the community areas. Schaumburg has an even spread of income, whereas the Chicago communities tend to be more extreme on each end. All of the communities have a similar distribution of older people, while the Chicago communities have exceedingly more young adults and less children. Finally, we can see from the business types that these communities tend to have many restaurants, with a tendency towards Asian cuisine.

## 4. Conclusions

The visualizations do a good job of demonstrating the similarities of the communities we ended with compared to Schaumburg. We can use the points that differ to help us make a final judgement for where to build a new bubble tea restaurant.

On the topic of age demographics, the Chicago communities tend to have a higher percentage of young adults. This would work in favor of the business, as young adults are more likely to have disposable incomes, compared to children and adults who are more likely to have families (35-49 years). From this insight, we see that Near West Side, The Loop, and Logan Square have the highest percentage of young adults. Comparing the median income of these locations to rest, we can see that they are among the highest, with only Logan Square falling behind Lincoln Square and Schaumburg.

Removing Logan Square due to its substantially lower median income, we are left with Near West Side and The Loop as our top candidates.

## 5. Future directions

To continue where the conclusion left off, this business problem would require further data to choose the final candidate. These would include statistics such as average price per square foot of rental locations, average square footage of rental properties, and annual business turnover. Another item to consider is the type of foot traffic each location sees. For example, Near West Side is likely to see plenty of students from UIC during the week, but have less business on the weekend. In contrast, The Loop would see more business people during the week, who may be less inclined to order bubble tea, whereas the weekends would bring tourists and suburbanites.

For a better picture of demographics as they relate to the business, the business could implement a reward system. This reward system could gather age and gender demographics, while also allowing us to see which demographics are most likely to become return customers.

In reviewing the process, I noticed some decisions that could be improved upon. When choosing the number of clusters for using K-Means on the business data, six clusters could have sufficed given the change in the overall slope on either side. This would need to be tested to ensure it did not result in too many locations in the Schaumburg cluster. In addition, the business clustering may have improved if the dataset was trimmed down to the top five instead of top ten most common businesses. Finally, the education data did not seem to help with determining the best locations. Due to the use of K-Means clustering, we cannot be certain without testing.

Going forward, this methodology could be applied to a program to allow any business owner to determine where to expand. Doing so would require changing the demographic data source, as CMAP

only covers the Chicagoland area, and not the United States as a whole. The U.S. Census might be a viable option for such an application.