Predicting Optimal Restaurant location in Chicago

Introduction

City of Chicago, is the most populous city in the U.S. state of Illinois and the third most populous city in the United States. Chicago is the principal city of the Chicago metropolitan area, often referred to as Chicagoland. At nearly 10 million people, the metropolitan area is the third most populous in the nation. Chicago is an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation. Chicago lays claim to many regional specialties that reflect the city's ethnic and working-class roots. Included among these are its nationally renowned deep-dish pizza.

As Chicago is known for its Pizza specialties, our customer is looking to open an upscale Pizza place in Chicago. The intention on this project is to collect and provide a data driven recommendation that can identify ideal location for opening an upscale Pizza place. Chicago Community areas are chosen for the ideal location. Core of Chicago is made of 77 communities, but we will focus on communities with larger population and higher per capita income as the restaurant is of a premium type.

Data Acquisition

This project will make use of the following data sources:

Chicago Community Area Census data

Data will be retrieved from Chicago open dataset
https://www.chicago.gov/city/en/depts/dcd/supp_info/community_area_2000and2010censuspopulationcomparisons.html

The original data source contains population data for each Chicago community areas for the years 2000 and 2010 and its percentage difference. For our project we will retrieve 2010 population data for each community areas.

Chicago Community Per Capita Income data

Data will be retrieved from Chicago Data Portal

https://data.cityofchicago.org/Health-Human-Services/Per-Capita-Income/r6ad-wvtk

This dataset contains a selection of six socioeconomic indicators of public health significance by Chicago community area, for the years 2007 – 2011. Here we use the per capita income data for each community.

Chicago Top Venue Recommendations from FourSquare API

We will be using the FourSquare API to explore Chicago community areas. The Foursquare explore function will be used to get the most common venue categories in each community, and then use this feature to group the communities into clusters. The following information are retrieved, Venue ID, Venue Name, Venue Coordinates and Venue Category.

Methodology

- Collect the Chicago Census data and Per capita income data.
- Wrangling the above data into single table and remove unnecessary features/attributes
- Identify the Lat/Long coordinates of Community areas
- Using FourSquare API find all venues for each Community areas.
- Cluster the community areas using K-means Clustering
- From the clustered community areas identify the best community to start the pizza place

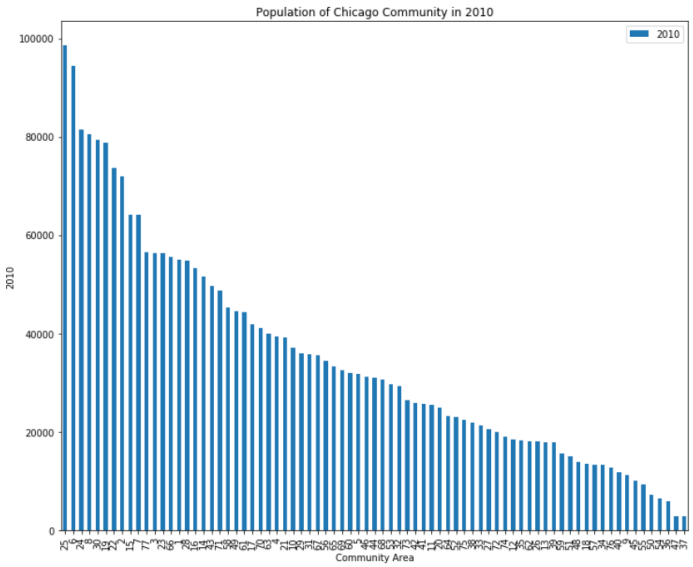
Data Wrangling

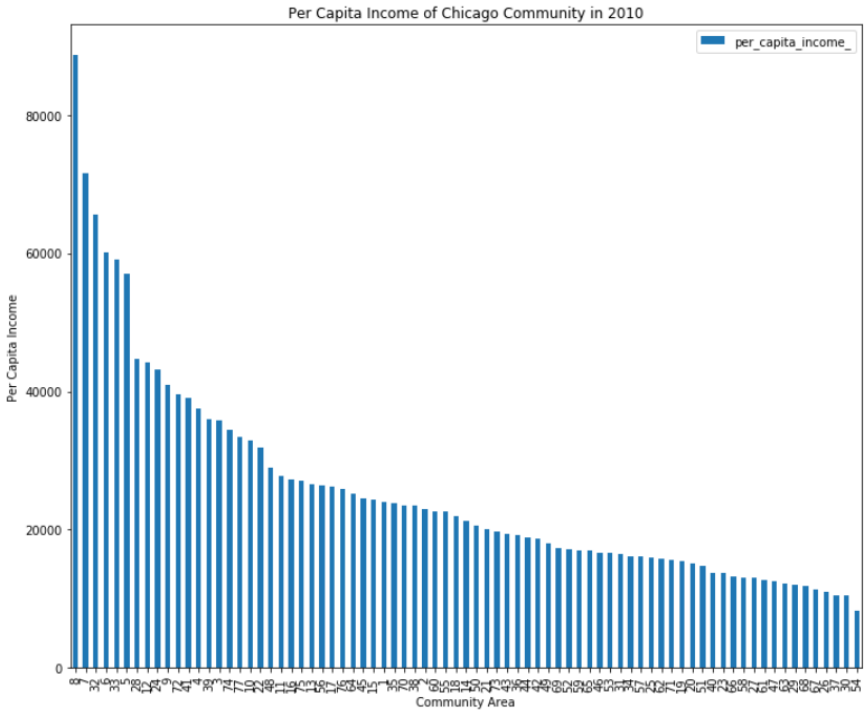
1.Data extraction: Data from multiple sources need to be combined in a single table. For Chicago census data, the data is in PDF format. To read PDF files we have installed 'Tabula' module in the notebook. This will help to extract the data from PDF. The data extracted from PDF needs to be further cleansed as there are spaces between the values. From the Census data we take the community and the 2010 population. For Per capita income data, we retrieve the community name and per capita features from the Chicago data portal. These two data sets are merged into a single data frame.

2.Retrieving community coordinates: 'Geocoder' module is used to retrieve the coordinates (latitude and longitude of each community areas.) After retrieving the Lat-Long co-ordinates we found that when the community name is not in standard form the data is not retrieved. We manually force fit the values into the data frame.

Analysis

Exploratory data analysis involved identifying the community with largest population and highest per capita income. Bar chart analysis of these two features are shown below.



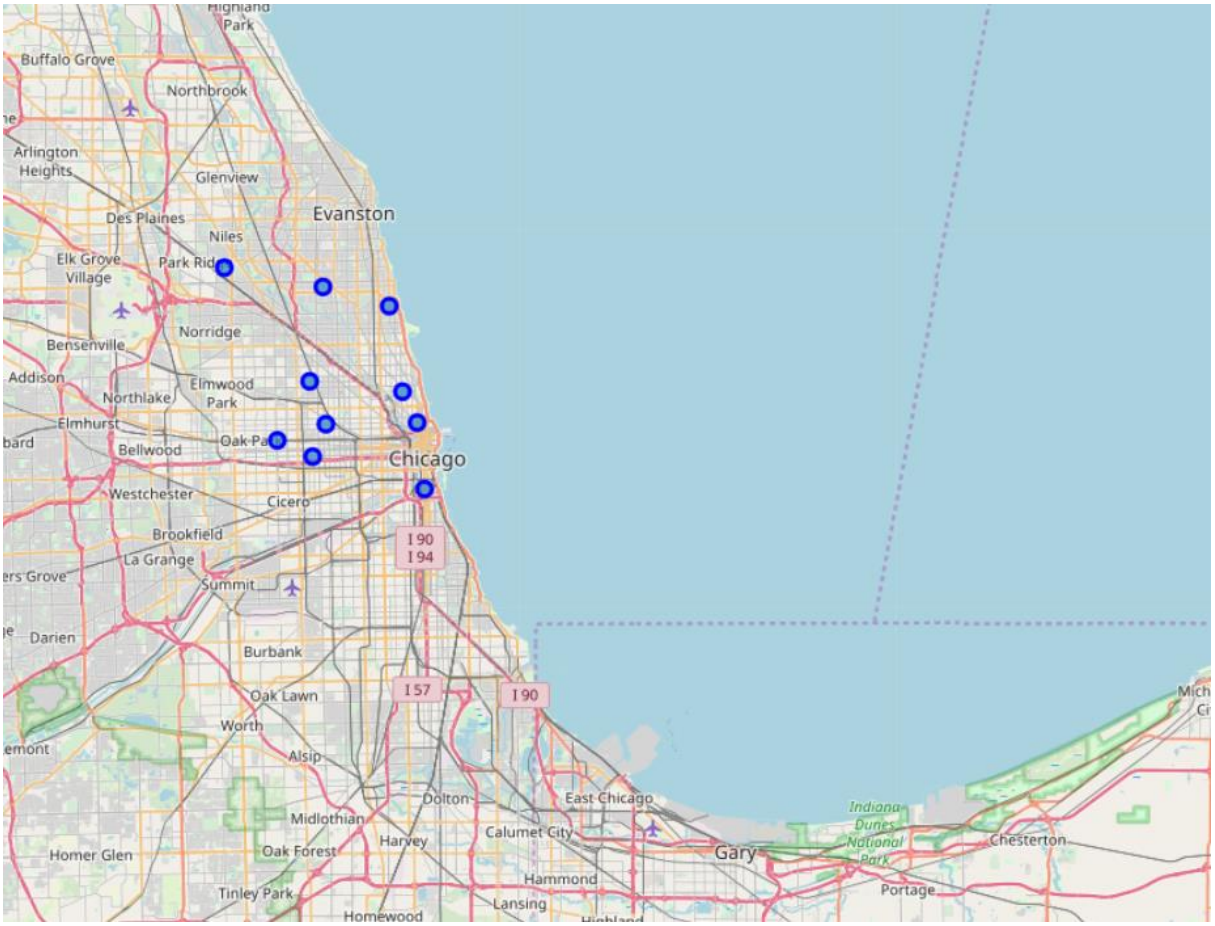


The charts indicate that the top population and per capita income are not same. Some Communities with higher per capita income may have low population and vice versa.

To identify the top Community areas, we standardize the features by dividing it wrt to the Median values. For the sake of analysis, we gave 60% preference to Population and 40% preference to Per capita income. From the scores of each feature, total score is computed by adding the Population score and Per capita income score. From the total score, top 10 Community area is identified as a potential location for opening the upscale Pizza place. The top 10 Community is shown below.

Num	Community Area	2010	per_capita_income_	Latitude	Longitude	pop_score	pci_score	tot_score
8	NEAR NORTH SIDE	80484	88669	42.007890	-87.813990	1.556349	1.663349	3.219699
6	LAKE VIEW	94368	60058	41.921840	-87.647440	1.824829	1.126633	2.951462
7	LINCOLN PARK	64116	71551	41.900150	-87.634330	1.239835	1.342231	2.582066
24	WEST TOWN	81432	43198	41.887740	-87.763920	1.574681	0.810355	2.385036
25	AUSTIN	98514	15957	41.877020	-87.730740	1.905002	0.299339	2.204341
22	LOGAN SQUARE	73595	31908	41.899070	-87.719470	1.423134	0.598565	2.021699
28	NEAR WEST SIDE	54881	44689	41.993736	-87.721421	1.061254	0.838325	1.899579
2	WEST RIDGE	71942	23040	41.981230	-87.660000	1.391169	0.432209	1.823379
19	BELMONT CRAGIN	78743	15461	41.928480	-87.734240	1.522683	0.290034	1.812717
32	LOOP	29283	65526	41.853880	-87.627110	0.566256	1.229208	1.795464

Map of Chicago with Community chosen for potential location



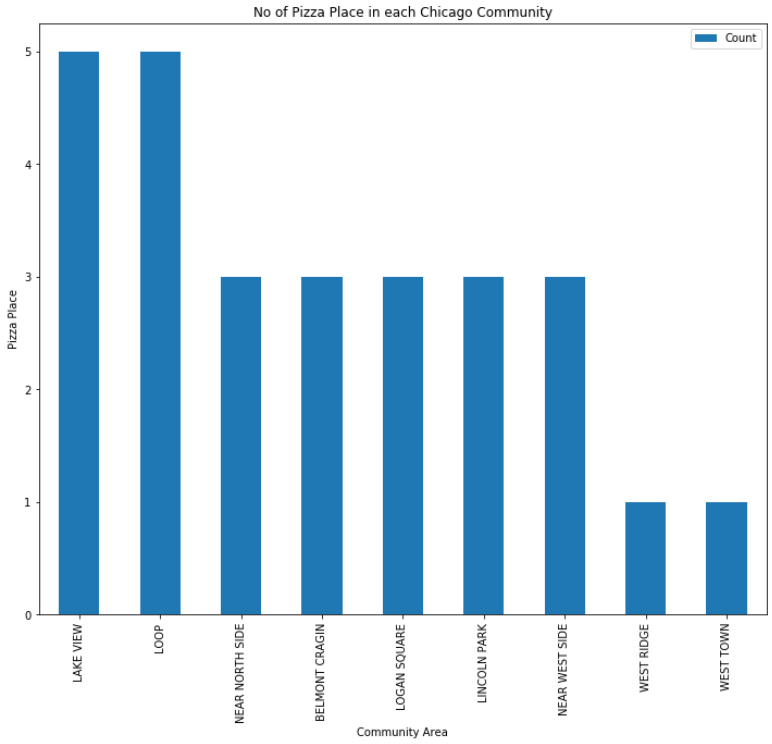
The map indicates that most of the community areas for identifying the potential location belong to north of Chicago. We will be using these 10 Community areas for further analysis and modelling

Extracting Venue data using FourSquare API:

Venue data for all these Community areas is provided by calling FourSquare API. For the sake of analysis, we have restricted 100 venues within the radius of 2000m. Once the venue details have been retrieved we filter the data to show all the restaurants.

	Count
Mexican Restaurant	27
Pizza Place	27
Coffee Shop	26
Sandwich Place	22
Italian Restaurant	22
Chinese Restaurant	21
Fast Food Restaurant	20
American Restaurant	17
Bar	16
Bakery	15
Café	13
Donut Shop	12

Restaurants count indicates that Pizza place have highest number compared to other kind of restaurants in these community areas. Further analysis wrt each Community shows that ‘Lake View’ and ‘Loop’ have 5 pizza place while ‘Austin’ community doesn’t have a pizza place.



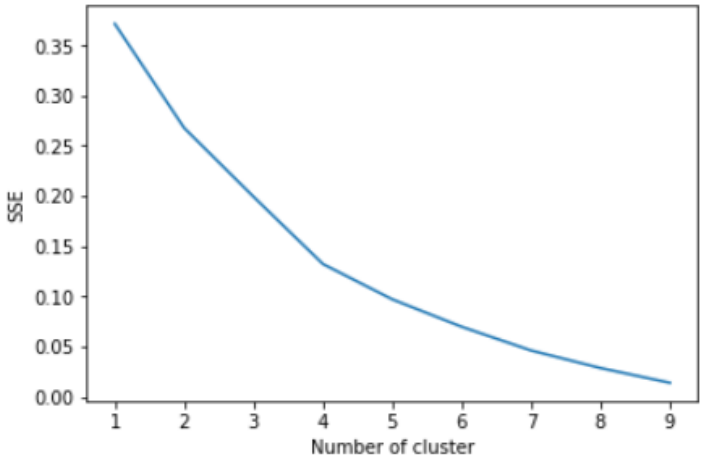
One hot encoding is done on the venues data. The Venues data is then grouped by the Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods.

	Community	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	AUSTIN	Fast Food Restaurant	Sandwich Place	Fried Chicken Joint	American Restaurant	Seafood Restaurant	Mexican Restaurant	Caribbean Restaurant	Food Truck	Burrito Place	Restaurant
1	BELMONT CRAGIN	Mexican Restaurant	Bar	Sandwich Place	Café	Brewery	Pizza Place	Coffee Shop	Latin American Restaurant	Taco Place	Diner
2	LAKE VIEW	Italian Restaurant	Pizza Place	Sushi Restaurant	Coffee Shop	Dessert Shop	Cupcake Shop	Bakery	Café	Mediterranean Restaurant	Indonesian Restaurant
3	LINCOLN PARK	Restaurant	Steakhouse	American Restaurant	Pizza Place	Coffee Shop	Italian Restaurant	New American Restaurant	Seafood Restaurant	Sushi Restaurant	Café
4	LOGAN SQUARE	Mexican Restaurant	Sandwich Place	Donut Shop	Fast Food Restaurant	Bar	Latin American Restaurant	Pizza Place	Dessert Shop	Hot Dog Joint	Ice Cream Shop
5	LOOP	Chinese Restaurant	Pizza Place	Mexican Restaurant	Asian Restaurant	Burger Joint	Cajun / Creole Restaurant	Italian Restaurant	Dessert Shop	Korean Restaurant	Dim Sum Restaurant
6	NEAR NORTH SIDE	Italian Restaurant	Sandwich Place	Coffee Shop	Fast Food Restaurant	Breakfast Spot	Pizza Place	Bakery	American Restaurant	Sushi Restaurant	Deli / Bodega
7	NEAR WEST SIDE	Coffee Shop	Pizza Place	Ice Cream Shop	Korean Restaurant	Afghan Restaurant	Burger Joint	Fast Food Restaurant	Fried Chicken Joint	Vietnamese Restaurant	Hot Dog Joint
8	WEST RIDGE	Vietnamese Restaurant	Coffee Shop	Breakfast Spot	Asian Restaurant	Thai Restaurant	Chinese Restaurant	Bakery	Mexican Restaurant	Bar	Italian Restaurant
9	WEST TOWN	Fast Food Restaurant	Donut Shop	American Restaurant	Seafood Restaurant	BBQ Joint	Sandwich Place	Coffee Shop	Fried Chicken Joint	Ice Cream Shop	African Restaurant

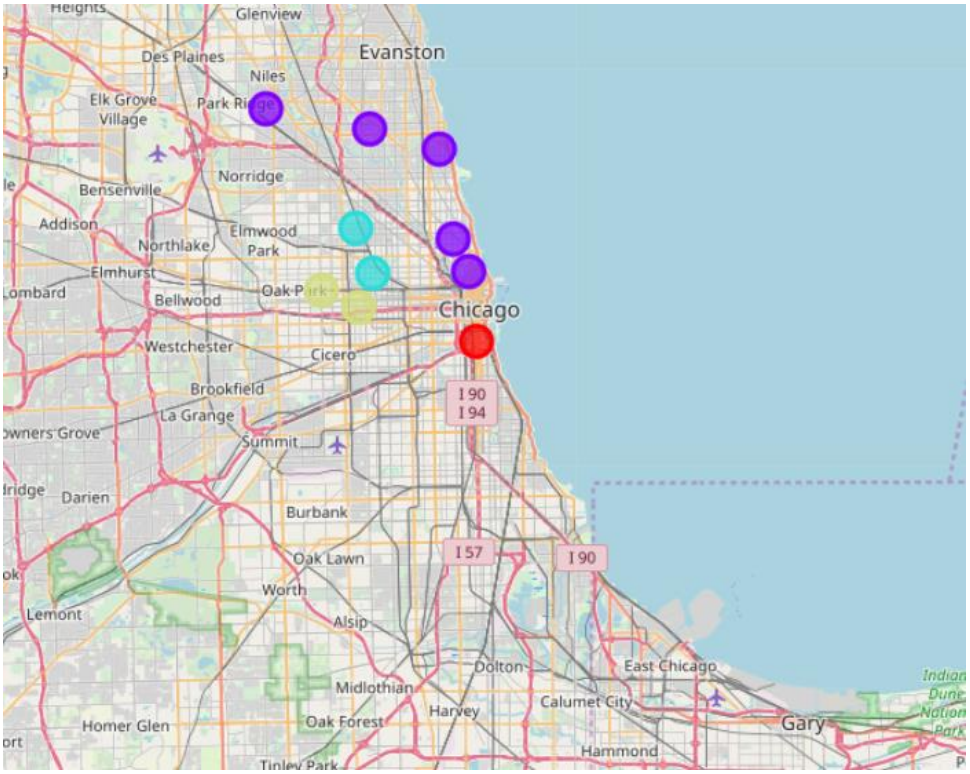
Modeling

To help people find similar community areas we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on similar venues.

To identify the optimal K- value we have run elbow method. The Elbow method is a heuristic method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset.



From the chart, it indicates that K = 4 is the optimal one. We will use a cluster size of 4 for this project that will cluster the 10 community areas into 4 clusters.



First Cluster:

Community Area	pop_score	pci_score	tot_score	Cluster Labels	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
LOOP	0.566256	1.229208	1.795464	0	Chinese Restaurant	Pizza Place	Mexican Restaurant	Asian Restaurant	Burger Joint	Cajun / Creole Restaurant	Italian Restaurant	Dessert Shop	Korean Restaurant	Dim Sum Restaurant

Second Cluster:

Community Area	pop_score	pci_score	tot_score	Cluster Labels	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
NEAR NORTH SIDE	1.556349	1.663349	3.219699	1	Italian Restaurant	Sandwich Place	Coffee Shop	Fast Food Restaurant	Breakfast Spot	Pizza Place	Bakery	American Restaurant	Sushi Restaurant	Deli / Bodega
LAKE VIEW	1.824829	1.126633	2.951462	1	Italian Restaurant	Pizza Place	Sushi Restaurant	Coffee Shop	Dessert Shop	Cupcake Shop	Bakery	Café	Mediterranean Restaurant	Indonesian Restaurant
LINCOLN PARK	1.239835	1.342231	2.582066	1	Restaurant	Steakhouse	American Restaurant	Pizza Place	Coffee Shop	Italian Restaurant	New American Restaurant	Seafood Restaurant	Sushi Restaurant	Café
NEAR WEST SIDE	1.061254	0.838325	1.899579	1	Coffee Shop	Pizza Place	Ice Cream Shop	Korean Restaurant	Afghan Restaurant	Burger Joint	Fast Food Restaurant	Fried Chicken Joint	Vietnamese Restaurant	Hot Dog Joint
WEST RIDGE	1.391169	0.432209	1.823379	1	Vietnamese Restaurant	Coffee Shop	Breakfast Spot	Asian Restaurant	Thai Restaurant	Chinese Restaurant	Bakery	Mexican Restaurant	Bar	Italian Restaurant

Third Cluster:

Community Area	pop_score	pci_score	tot_score	Cluster Labels	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
LOGAN SQUARE	1.423134	0.598565	2.021699	2	Mexican Restaurant	Sandwich Place	Donut Shop	Fast Food Restaurant	Bar	Latin American Restaurant	Pizza Place	Dessert Shop	Hot Dog Joint	Ice Cream Shop
BELMONT CRAGIN	1.522683	0.290034	1.812717	2	Mexican Restaurant	Bar	Sandwich Place	Café	Brewery	Pizza Place	Coffee Shop	Latin American Restaurant	Taco Place	Diner

Fourth Cluster:

Community Area	pop_score	pci_score	tot_score	Cluster Labels	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
WEST TOWN	1.574681	0.810355	2.385036	3	Fast Food Restaurant	Donut Shop	American Restaurant	Seafood Restaurant	BBQ Joint	Sandwich Place	Coffee Shop	Fried Chicken Joint	Ice Cream Shop	African Restaurant
AUSTIN	1.905002	0.299339	2.204341	3	Fast Food Restaurant	Sandwich Place	Fried Chicken Joint	American Restaurant	Seafood Restaurant	Mexican Restaurant	Caribbean Restaurant	Food Truck	Burrito Place	Restaurant

Results & Discussions

In the First cluster, we have 'Loop' which has 5 Pizza places in that Community. Also, it is one of the most common venue in that community. There are lot of competition in this cluster and it is not recommended to open the Pizza place here.

In the second cluster, we have 5 community places. These community places have higher Per capita income and larger population compared to other clusters. Their eating pattern and preference are similar. This cluster can be viewed as a potential cluster to open a Pizza place. Out of these 5 community areas 'West Ridge' doesn't have a Pizza place in the most common venue. Hence, we recommend 'West Ridge' as a potential candidate for opening the Pizza place.

In the third cluster, we have 'Logan Square' and 'Bellmont Cragin'. This cluster prefer Mexican restaurant over Pizza place as seen in the table. We can ignore this cluster

In the fourth cluster, we have 'West Town' and 'Austin'. In this cluster Pizza is not the preferred food. We don't have any Pizza place or Italian restaurant in the top 10 common venue. We can ignore this cluster too.

Conclusion

This capstone project provides us a small glimpse of how real-life data-science projects look like. We have used python libraries to extract the data, used Foursquare API to explore the common Venues and used K-Means clustering to cluster and identify the best community area as a potential candidate for Pizza place set up. For setup of other restaurants like Chinese restaurants, demographic data can be used to identify the community areas with respective demographic majority and then cluster the community.