IBM Data Science Capstone

Recommending Location for an Indian Restaurant in Chicago Using Data Science

Introduction:

Chicago is the third most populous city in the US. Located on the banks of Lake Michigan, Chicago is a major hub for various industries including finance, telecommunications, transportation, education etc. Chicago's major airport - O'Hare International airport is one of the busiest airports in the world. Chicago also has many national highways making traveling to this culture hub very easy for travelers. With Ivy league and reputed educational institutions set up in the Illinois and Chicago area, such as University of Illinois, University of Chicago etc., many international students, particularly from India are expected to be part of the student community. Being a major cultural hub, Chicago has a big restaurant industry with over 7,300 restaurants catering to a growing base of food lovers. Setting up a new restaurant in such a lucrative environment can be a very profitable venture. However, finding the best neighborhood based on cultural diversity, income level disparity and other demographics is a problem that needs a solution to set up a successful business. This project could find interest in individuals, small businesses and corporations who want to enter or expand in the restaurant/food services industry.

Business Problem:

The aim of this project is to find the best community area to open an Indian restaurant leveraging data analytics and machine learning techniques. Being a metropolitan city, Chicago is expected to have a diversified population with differences in income levels. In this project, we will try to gather data regarding income levels, crime data and existing restaurants in a neighborhood. With this data, we will use the Foursquare location data to find most common venues and identify the best neighborhood to set up a new Indian restaurant. The neighborhood needs to be safe and provide unique food choices to cater to the growing Indian community. This project could be helpful to clients who are looking to capitalize the ever-growing food & restaurant industry of the Chicago metro area by setting up an Indian restaurant.

Data:

1. Neighborhoods
2. Geocoding
3. Foursquare API - getting venues information for each neighborhood.

To solve the problem of identifying the best neighborhood to open a new restaurant, we will need to gather relevant data and analyze it.

The data collected for this project includes:

1. **Community area and per capita income (including hardship index)**

The first step is to understand the income level of each community along with the hardship index. This helps in assessing the scale of the restaurant.

```
df.head()
```

	COMMUNITY AREA NAME	PER CAPITA INCOME	HARDSHIP INDEX
0	Rogers Park	23939	39
1	West Ridge	23040	46
2	Uptown	35787	20
3	Lincoln Square	37524	17
4	North Center	57123	6

2. **Census data by community area to understand the demographics.**

Second step is to understand the population diversity of the communities. This helps in finding community with a higher Asian population to set up the Indian restaurant.

```
census.head()
```

	Community	Total Population	Percent Hispanic	Percent Non Hispanic Black	Percent Non_ Hispanic White	Percent Non_ Hispanic Asian	Percent Non_ Hispanic Other or Multiple Races
0	Rogers Park	53.470	24%	24%	45%	5%	3%
1	West Ridge	75.185	20%	13%	41%	21%	4%
2	Uptown	54.001	16%	19%	51%	11%	3%
3	Lincoln Square	46.881	18%	6%	62%	10%	4%
4	North Center	35.406	11%	9%	73%	4%	3%

3. Most recent crime data (prior twelve-month period) along with description and coordinate information

Third step is to understand the crime data of Chicago. Community level crime data was not readily available. Crime data with geographical coordinate information is used for the analysis. A geolocator program is used to find the address for the geo coordinates. From this address, the corresponding community name is captured.

```
geolocator = Nominatim(user_agent="locator")
count = 0

df_crime = pd.DataFrame(columns = ['Address', 'Latitude', 'Longitude'])

for i,j in zip(crime['LATITUDE'], crime['LONGITUDE']):

    x= str(i) + "," + str(j)
    count = count + 1
    location = geolocator.reverse(x)
    geo_string = location.address.replace(" ","").split(",")

    COMMUNITY = geo_string


    df_crime = df_crime.append({'Address': COMMUNITY, 'Latitude': i, 'Longitude':j}, ignore_index = True)
```

```
df_crime.head()
```

	Address	Latitude	Longitude
0	[Wendy's, 2215, NorthWashtenawAvenue, LoganSqu...	41.922170	-87.695539
1	[Popeyes, 7430, SouthStonyIslandAvenue, SouthS...	41.759448	-87.586156
2	[TCFBank, 1400-1408, WestFullertonAvenue, Linc...	41.925213	-87.663639
3	[McDonald's, 23, NorthWesternAvenue, NearWestS...	41.881855	-87.686448
4	[828-832, NorthStateStreet, NearNorthSide, Lin...	41.897674	-87.628228

```
crime.head()
```

	PRIMARY DESCRIPTION	SECONDARY DESCRIPTION	LOCATION DESCRIPTION	LATITUDE	LONGITUDE
0	BATTERY	SIMPLE	RESTAURANT	41.922170	-87.695539
1	CRIMINAL DAMAGE	TO PROPERTY	RESTAURANT	41.759448	-87.586156
2	ASSAULT	SIMPLE	RESTAURANT	41.925213	-87.663639
3	THEFT	$500 AND UNDER	RESTAURANT	41.881855	-87.686448
4	THEFT	OVER $500	RESTAURANT	41.897674	-87.628228

4. Active business licenses and location information

Final data point is finding active business licenses with location information. A geolocator program is used again to find the corresponding address for the geo coordinates and the community names are identified.

```
licences.head()
```

	ID	LICENSE ID	ACCOUNT NUMBER	SITE NUMBER	LEGAL NAME	DOING BUSINESS AS NAME	ADDRESS	CITY	STATE	ZIP CODE	LICENSE DESCRIPTION	BUSINESS ACTIVITY ID	BUSINESS ACTIVITY	LIC NUI
0	2483103-20190701	2664226	85443	277	ABM INDUSTRY GROUPS, LLC	ABM ONSITE SERVICES-MIDWEST, INC / PROFESSIONA...	1725 W HARRISON ST	CHICAGO	IL	60612	Valet Parking Operator	855	Valet Parking Operator	24
1	2583905-20190701	2664227	85443	284	ABM INDUSTRY GROUPS, LLC	ABM INDUSTRY GROUPS	1620 W HARRISON ST	CHICAGO	IL	60612	Valet Parking Operator	855	Valet Parking Operator	25
2	2583904-20190701	2664225	85443	215	ABM INDUSTRY GROUPS, LLC	ABM Parking Services	1611 W HARRISON ST	CHICAGO	IL	60612	Valet Parking Operator	855	Valet Parking Operator	25
3	2601954-20190716	2670145	428283	1	SMART VALET PARKING LLC	SMART VALET PARKING	940 W WEED ST	CHICAGO	IL	60642	Valet Parking Operator	855	Valet Parking Operator	26
4	2476627-20190701	2664242	216013	1	VERNON PARK TAP L.L.C.,	TUFANO'S/VERNON PARK TAP	1073 W VERNON PARK PL	CHICAGO	IL	60607	Valet Parking Operator	855	Valet Parking Operator	24

The final dataframe holds all the data points as stated above.

COMMUNITY AREA NAME	PER CAPITA INCOME	HARDSHIP INDEX	Latitude	Longitude	Total Population	Percent Hispanic	Percent Non Hispanic Black	Percent Non_ Hispanic White	Percent Non_ Hispanic Asian	Percent Non_ Hispanic Other or Multiple Races	License_Count	Crime_count
Rogers Park	23939	39	42.010531	-87.670748	53.470	24%	24%	45%	5%	3%	187.0	57.0
West Ridge	23040	46	42.003548	-87.696243	75.185	20%	13%	41%	21%	4%	253.0	58.0
Uptown	35787	20	41.966630	-87.655546	54.001	16%	19%	51%	11%	3%	212.0	59.0
Lincoln Square	37524	17	41.975990	-87.689616	46.881	18%	6%	62%	10%	4%	165.0	25.0
North Center	57123	6	41.956107	-87.679160	35.406	11%	9%	73%	4%	3%	223.0	41.0

Foursquare API calls and Venue Data:

Using the Foursquare API, corresponding venue names with category and location info is pulled and stored into a dataframe.

```
venues = getNearbyVenues(names = df['COMMUNITY NO SPACES'],
                                 latitudes= df['Latitude'],
                                 longitudes= df['Longitude']
                                )
```

```
len(venues['COMMUNITY NO SPACES'].unique())
```

```
77
```

```
venues.head()
```

	COMMUNITY NO SPACES	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	RogersPark	42.010531	-87.670748	El Famous Burrito	42.010421	-87.674204	Mexican Restaurant
1	RogersPark	42.010531	-87.670748	Morse Fresh Market	42.008087	-87.667041	Grocery Store
2	RogersPark	42.010531	-87.670748	Taqueria & Restaurant Cd. Hidalgo	42.011634	-87.674484	Mexican Restaurant
3	RogersPark	42.010531	-87.670748	Lifeline Theatre	42.007372	-87.666284	Theater
4	RogersPark	42.010531	-87.670748	Rogers Park Social	42.007360	-87.666265	Bar

Methodology:

One-hot method:

Once the nearby venue information is gathered, we then analyze communities by creating a dataframe with “zero” or “one” value for each venue category for a corresponding community.

```
# one hot encoding
onehot = pd.get_dummies(venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
onehot['COMMUNITY NO SPACES'] = venues['COMMUNITY NO SPACES']

# move neighborhood column to the first column
fixed_columns = [onehot.columns[-1]] + list(onehot.columns[:-1])
onehot = onehot[fixed_columns]

onehot.head()
```

	COMMUNITY NO SPACES	ATM	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argent Resta
0	RogersPark	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	RogersPark	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	RogersPark	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	RogersPark	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	RogersPark	0	0	0	0	0	0	0	0	0	0	0	0	0	

We then group the communities and find the mean score for each venue category.

```
grouped = onehot.groupby('COMMUNITY NO SPACES').mean().reset_index()
grouped.head()
```

	COMMUNITY NO SPACES	ATM	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Arge Rest
0	AlbanyPark	0.0	0.013514	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	
1	ArcherHeights	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	
2	ArmourSquare	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.035714	0.0	0.0	0.0	0.0	0.0	
3	Ashburn	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.090909	0.0	0.0	0.0	0.0	0.0	
4	AuburnGresham	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.032258	0.0	0.0	0.0	0.0	0.0	

```
grouped.shape
```

```
(77, 339)
```

K-means clustering (3 clusters)

Now, we train the data with the K-means clustering algorithm and find clustered communities. Number of clusters chosen were 3 for this analysis which yielded better classification.

```
df['Cluster Labels'].value_counts()
```

```
2     49
0     27
1      1
Name: Cluster Labels, dtype: int64
```

Top 10 most common venues

Now, we try to identify the top 10 most common venue categories for each community.

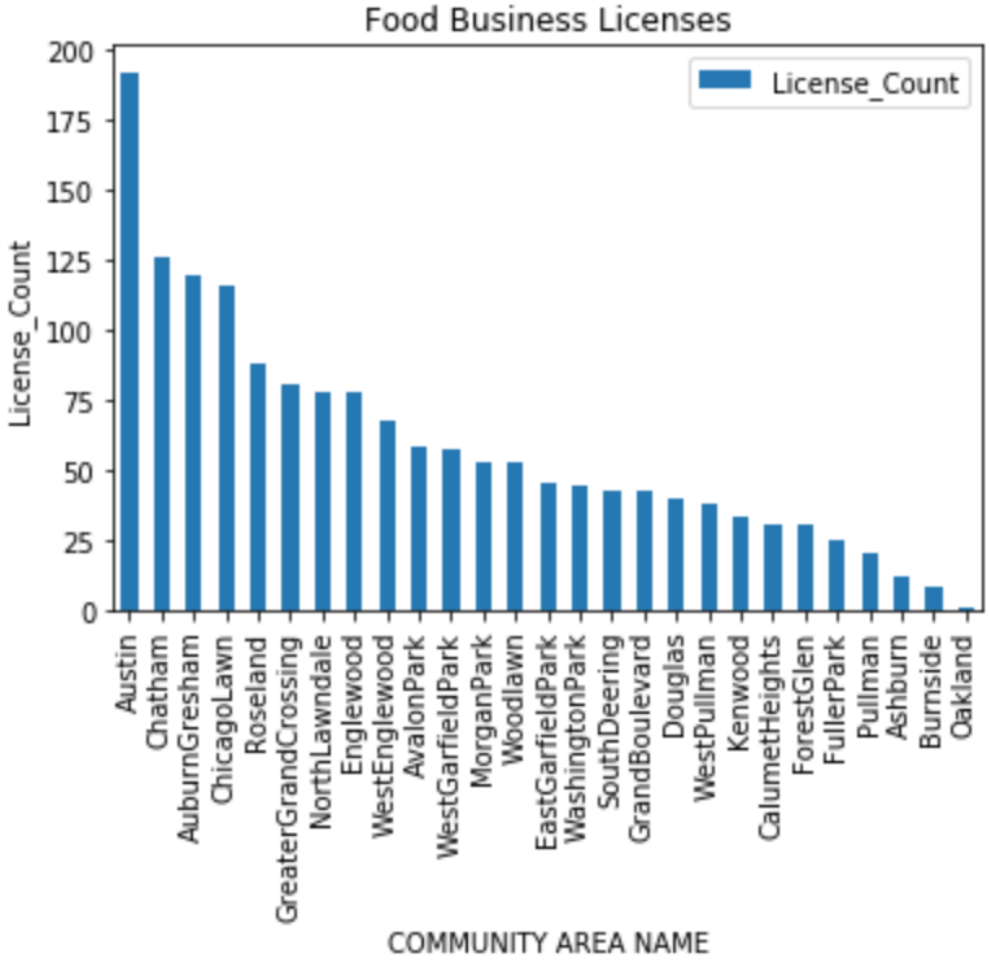
```
venues_sorted.head()
```

	COMMUNITY NO SPACES	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	AlbanyPark	ATM	Newsstand	New American Restaurant	Neighborhood	Nature Preserve	National Park	Nail Salon	Music Venue	Music Store	Museum
1	ArcherHeights	ATM	Noodle House	Non-Profit	Nightlife Spot	Newsstand	New American Restaurant	Neighborhood	Nature Preserve	Office	National Park
2	ArmourSquare	ATM	Organic Grocery	Optical Shop	Office	Noodle House	Non-Profit	Nightlife Spot	Nightclub	Other Great Outdoors	Newsstand
3	Ashburn	ATM	Optical Shop	Office	Noodle House	Non-Profit	Nightlife Spot	Nightclub	Newsstand	Organic Grocery	New American Restaurant
4	AuburnGresham	ATM	Optical Shop	Office	Noodle House	Non-Profit	Nightlife Spot	Newsstand	New American Restaurant	Organic Grocery	Neighborhood

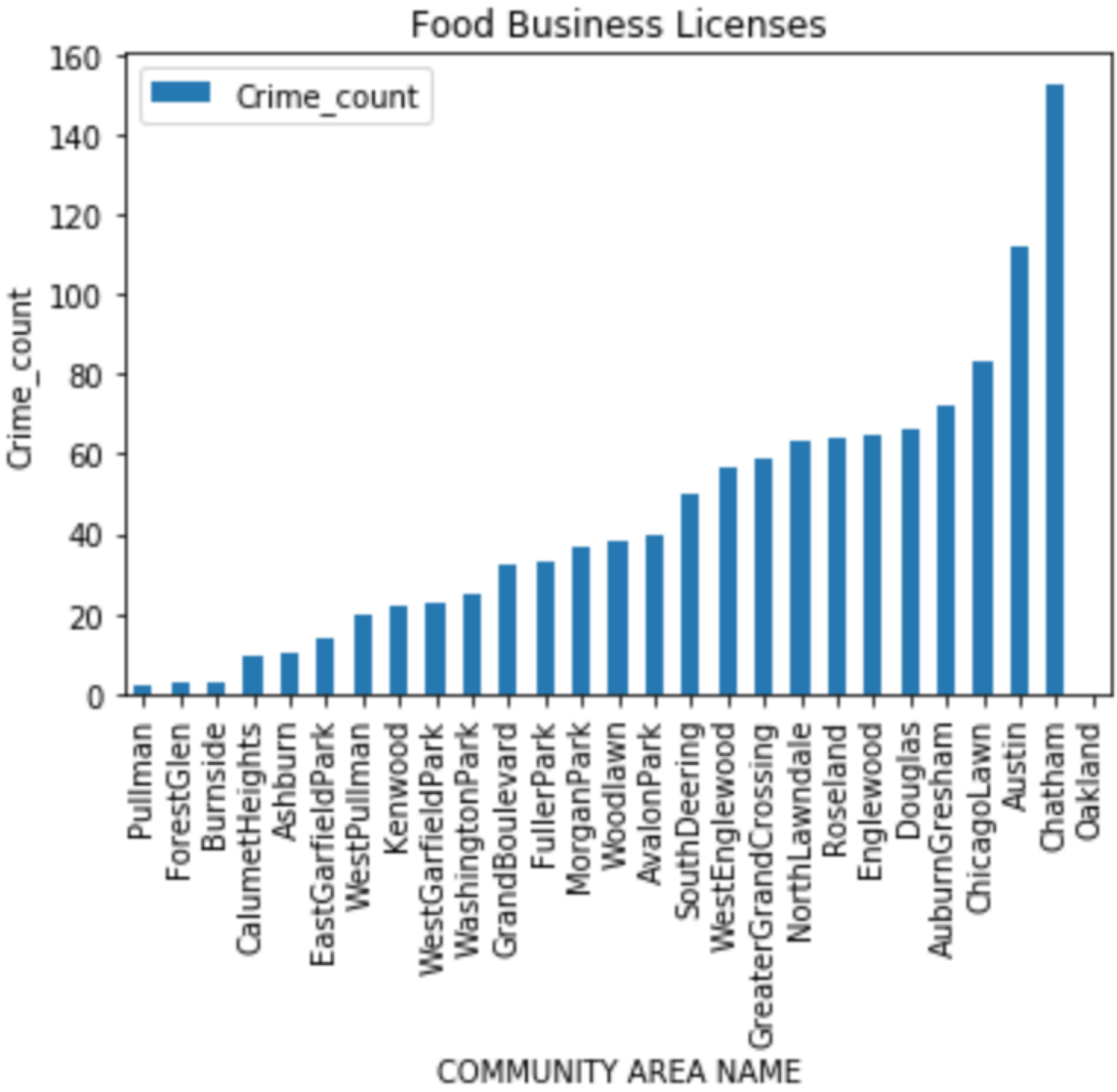
Visualization:

We try to visualize the community level information of business licenses, crimes, hardship index and non-Hispanic Asian population.

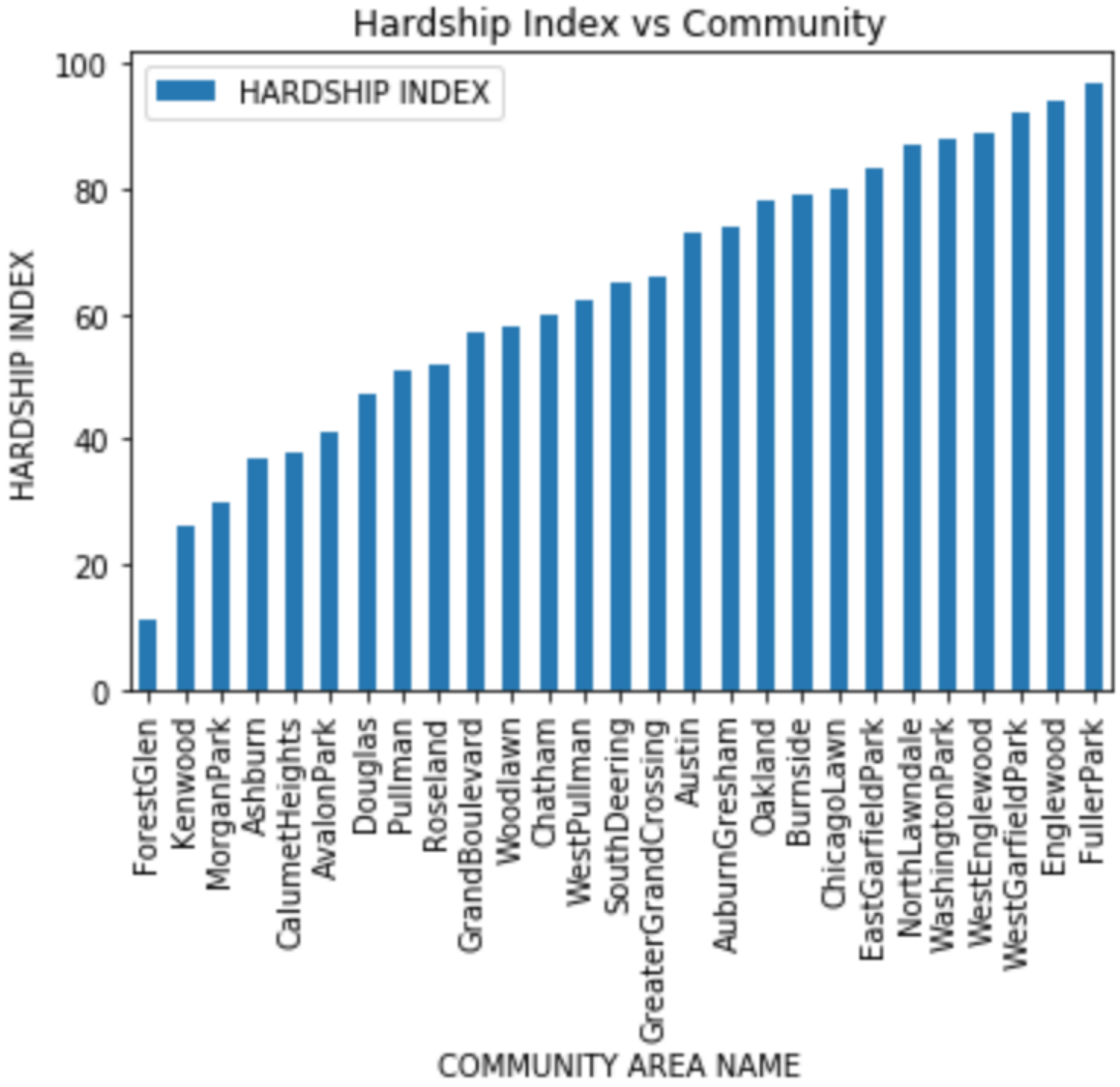
Community area vs license count



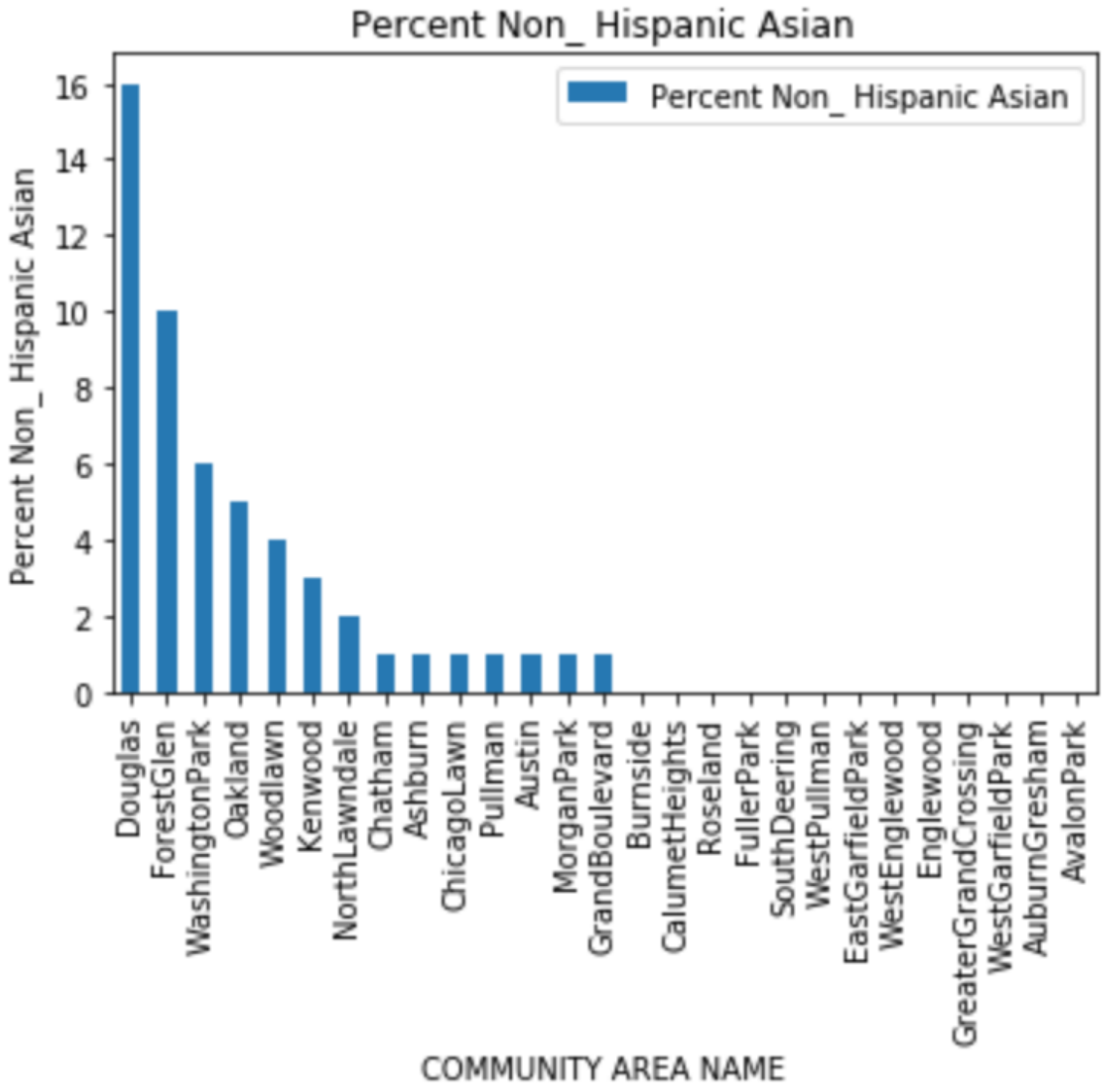
Community area vs crime count



Community area vs hardship index



Community area vs non-Hispanic Asian



Results:

Now, we identify the best communities with

- high business licenses (high business activity attracts higher workforce and customers for food business)
- Low crime count (safer community with lower insurance costs)
- Low hardship index (people have higher spending power)
- High Asian population (better chances of attracting people with similar food tastes and preferences)

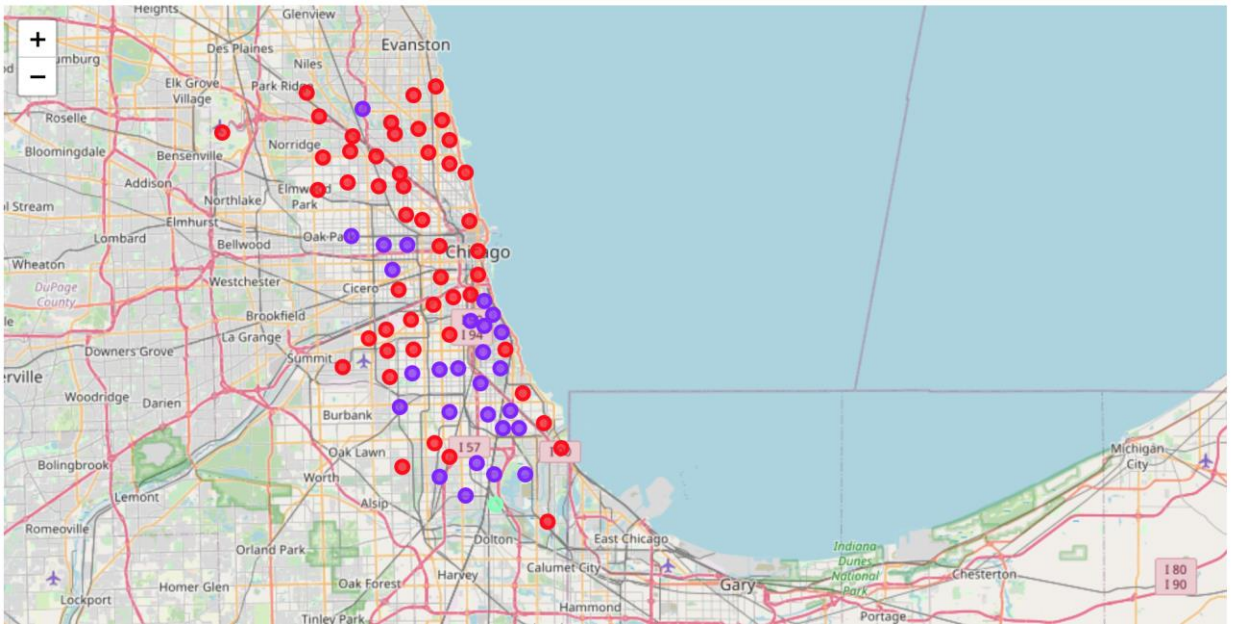
The resulting communities are:

	COMMUNITY NO SPACES	License_Count	Crime_count	HARDSHIP INDEX	Percent Non_ Hispanic Asian
0	MorganPark	53.0	37.0	30	1
1	Woodlawn	53.0	38.0	58	4

```
# coordinates of chicago

latitude = 41.881832

longitude = -87.623177
kclusters=3
```

```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=11)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]


colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(df['Latitude'], df['Longitude'], df['COMMUNITY NO SPACES'], df['Cluster Labels']):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster],
        fill=True,
        fill_color=rainbow[cluster],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

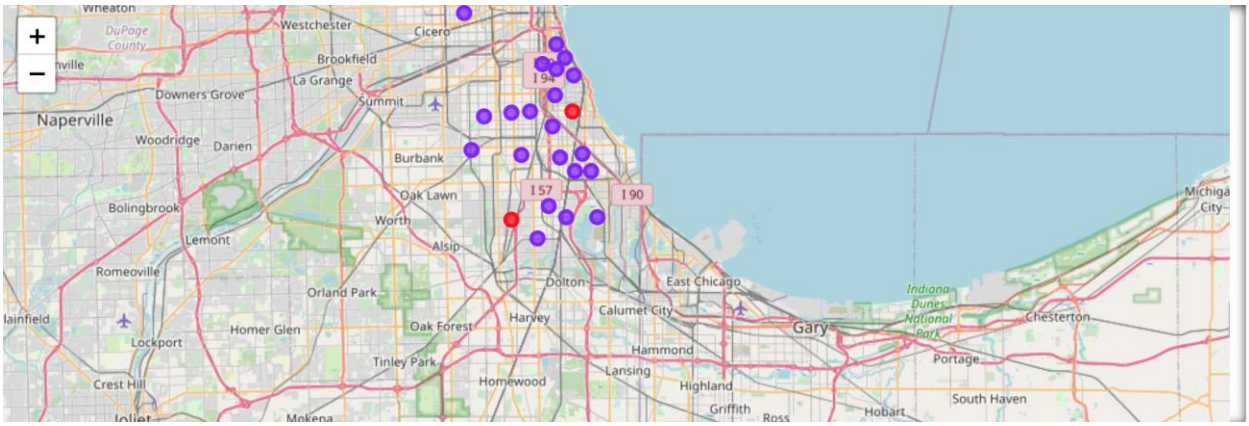
```
map_clusters
```



Recommendation:

Based on the analysis, it is beneficial to open a new Indian restaurant in either Morgan Park or Woodland communities. If we analyze these two communities, we note that these two communities are similarly positioned in the number of active businesses and crime counts in the areas. However, the hardship index is better in Morgan Park compared to the Woodlawn community and non-Asian population is higher in Woodlawn community. We can offer the client/ end customer with both options and decide a community based on the scale of the restaurant. If the client wants to set up a high scale restaurant, Morgan Park is the recommended community. If the plan is to set up a low/mid-scale restaurant, Woodlawn community is a better one.

Common communities are shown as pop-ups in the picture below.



Conclusion:

In this project, we have been able to understand the needs of the end customer and offer recommendations/ solutions using machine learning and proper data analysis techniques. This can be an iterative process with feedback and inputs for improvisation. There may arise a need to increase the cluster count or machine learning techniques depending on additional data.

References:

https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

http://www.actforchildren.org/wp-content/uploads/2018/01/Census-Data-by-Chicago-Community-Area-2017.pdf

https://data.cityofchicago.org/Public-Safety/Crimes-One-year-prior-to-present/x2n5-8w5q

https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses-Current-Active/uupf-x98q

https://www.chicago.gov/city/en/about/facts.html