Capstone Project - The Battle of Neighborhoods

1.Introduction

Background

Chicago is an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation. The city has also been rated as having the most balanced economy in the United States, due to its high level of diversification. Chicago has been a hub of the retail sector since its early development. The city's overall crime rate, especially the violent crime rate, is higher than the US average. Chicago was responsible for nearly half of 2016's increase in homicides in the US, though the nation's crime rates remain near historic lows.

In order to open a retail store, decision on store-locations is one of the most important strategic decisions the retailer has to make for its long term success. Finding the right location for a new store is a process that takes careful consideration. Population, neighborhood demographics, income distribution, crime rate, local competition are all factors taken into consideration when grocery chains look for a new store location.

Problem

The aim of this project is to find location for opening of retail store, specifically grocery stores location across Chicago. This report will be targeted to stakeholder’s interest in opening grocery store in community area with low crime rate and higher per capita income.

2. Methodology

2.1 Data Acquisition

The data acquired for this project is a combination of data from three sources.
The first source of data is scraped from a Wikipedia page that contains the list of Chicago community area. This page contains additional information about the community area, the following are the columns:

Column Name	Description	Type
Serial Number		Number
Community Area Name		Plain Text
Neighborhood	Name of the neighborhood in the Community area	Plain Text

The second data source of the project uses a Chicago crime data that shows the crime per community area in Chicago. The dataset contains the following columns:

Column Name	Description	Type
ID	Unique identifier for the record.	Number
Case Number	The Chicago Police Department Record Number	Plain Text
Date	Date when the incident occurred.	Date & Time
Block	The partially redacted address	Plain Text
IUCR	The Illinois Unifrom Crime Reporting code.	Plain Text
Primary Type	The primary description of the IUCR code.	Plain Text
Description	The secondary description of the IUCR code.	Plain Text
Location	Description of the location	Plain Text
Arrest	Indicates whether an arrest was made.	Checkbox
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.	Checkbox
Beat	Indicates the beat where the incident occurred.	Plain Text
District	Indicates the police district where the incident occurred.	Plain Text
Ward	The ward (City Council district) where the incident occurred.	Number
Community Area	Indicates the community area where the incident occurred.	Plain Text
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).	Plain Text
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.	Number
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.	Number
Year	Year the incident occurred.	Number
Updated On	Date and time the record was last updated.	Date & Time
Latitude	The latitude of the location where the incident occurred.	Number
Longitude	The longitude of the location where the incident occurred.	Number
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal.	Location

Third data source is Chicago Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012. This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index," for each Chicago community area, for the years 2008 – 2012 . The dataset contains the following columns:

Column Name	Description	Type
Community Area Number		Number
COMMUNITY AREA NAME		Plain Text
PERCENT OF HOUSING CROWDED	Percent occupied housing units with more than one person per room	Number
PERCENT HOUSEHOLDS BELOW POVERTY	Percent of households living below the federal poverty level	Number
PERCENT AGED 16+ UNEMPLOYED	Percent of persons over the age of 16 years that are unemployed	Number
PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	Percent of persons over the age of 25 years without a high school education	Number
PERCENT AGED UNDER 18 OR OVER 64	Percent of the population under 18 or over 64 years of age (i.e., dependency)	Number
PER CAPITA INCOME	Community Area Per capita income is estimated as the sum of tract-level aggregate incomes divided by the total population	Number
HARDSHIP INDEX	Score that incorporates each of the six selected socioeconomic indicators (see dataset description)	Number

2.2 Data cleaning and processing

The data preparation for each of the three sources of data is done separately. The First data is scraped from a Wikipedia page using the Beautiful Soup library in python. Using this library we can extract the data in the tabular format as shown in the website. After the web scraping, string manipulation is required to get the names of the Community area in the correct form (see fig 2.1). This is important because we will be merging the two datasets together using the Community area.

	community area	Neighborhood
0	Albany Park	Albany Park
1	Riverdale	Altgeld Gardens
2	Edgewater	Andersonville
3	Archer Heights	Archer Heights
4	Armour Square	Armour Square

Figure 2.1

The second data from the Chicago crime data, the crimes during the most recent year (2020) are only selected. The major categories of crime are segregated group by community area code to get the total crimes per the community area (see fig 2.2).

```
chicago_crime.head()
```

	Primary Type	Location Description	District	Community Area
0	OTHER OFFENSE	STREET	15	25.0
1	OTHER OFFENSE	VEHICLE NON-COMMERCIAL	4	46.0
2	THEFT	DEPARTMENT STORE	8	57.0
3	MOTOR VEHICLE THEFT	STREET	14	23.0
4	CRIMINAL DAMAGE	RESIDENCE - PORCH / HALLWAY	12	24.0

Fig 2.2 Chicago crime data after preprocessing

From third dataset Chicago Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012 drop all null values and column consisting non desirable data, processed and a glimpse of processed data set has been shown in figure 2.3

	Community Area	community area	per_cap_income
1	1	Rogers Park	23939
2	2	West Ridge	23040
3	3	Uptown	35787
4	4	Lincoln Square	37524
5	5	North Center	57123

Fig 2.3 Chicago census data

The two datasets are merged on the Community area names to form a new dataset that combines the necessary information in one dataset (see fig 2.4). The purpose of this dataset is to visualize distribution of crime and per capita income across community areas and identify the Community area with the least crimes recorded and high per capita income during the year 2020

	community area	per_cap_income	Total_Case
0	Near South Side	59077	492
1	North Center	57123	379
2	Forest Glen	44164	153
3	Edison Park	40959	70
4	Beverly	39523	236

Fig 2.4 Chicago crime and per capita income

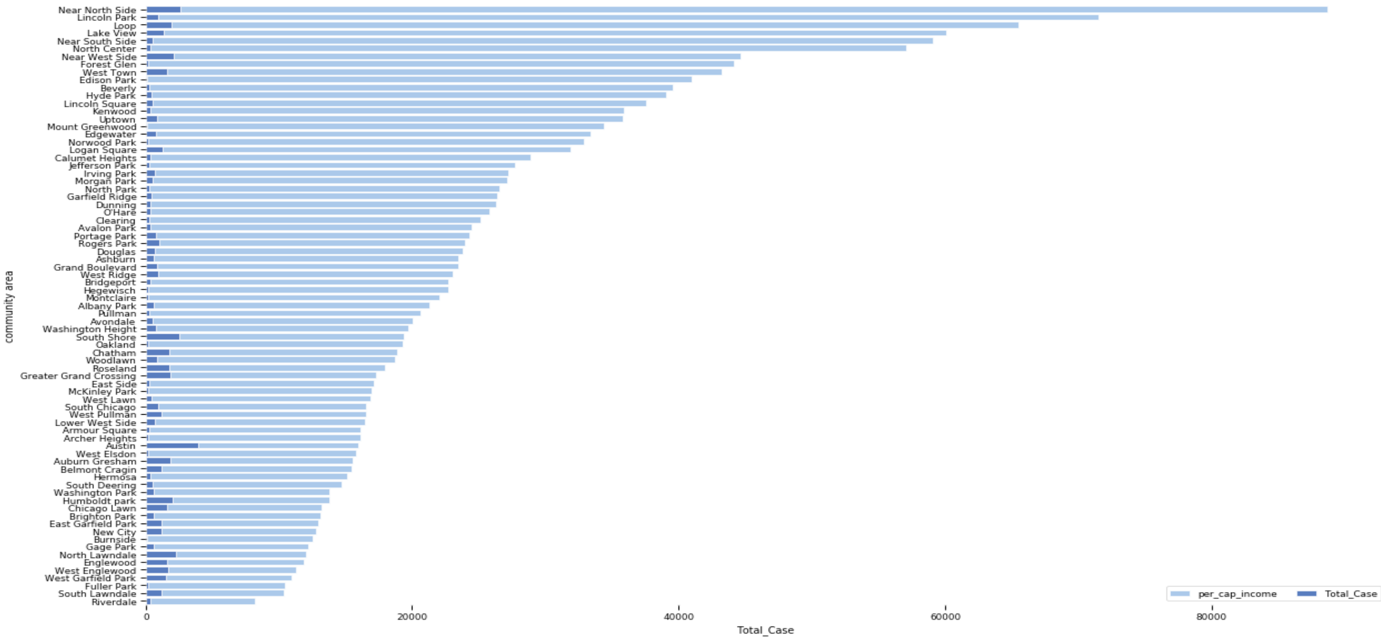


Figure 2.5 Distribution of per capita income community area wise

Range of Per capita income distributed randomly between $88,669 to $8201 with an average of $25,597 histogram and description of per capita income and number of crime recoded community area wise is given as

	Community Area	per_cap_income	Total_Case
count	77.000000	77.000000	77.000000
mean	39.000000	25563.168831	855.753247
std	22.371857	15293.098259	729.156599
min	1.000000	8201.000000	70.000000
25%	20.000000	15754.000000	314.000000
50%	39.000000	21323.000000	594.000000
75%	58.000000	28887.000000	1187.000000
max	77.000000	88669.000000	3923.000000

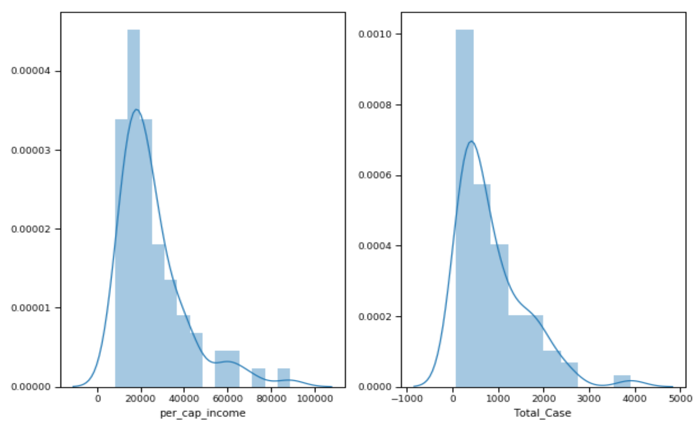


Figure 2.6 description of data frame **Figure 2.7 histogram of per capita income and total case recoded**

As per above plot distribution of total no of reported crime were distributed between 3932 to 70 with mean of 855 and median of 594 .For selection of community area with high per capita income and lowest crime reported, initial top 20 areas with largest per capita income were selected and then areas with less than median crime reported were selected among those community areas. After filtering data community area of data frame were merged with associate neighborhood. This dataset is created from scratch, the pandas data frame is created with the names of the neighborhoods and the name of the community area. The coordinates of the neighborhoods is be fetched using Open Cage Geocoder to create a final consolidated dataset of the Neighborhoods, along with their boroughs, crime data and the respective Neighborhood's co-ordinates.

	community area	Neighborhood	Latitude	Longitude	per_cap_income	Total_Case
0	Beverly	Beverly	41.718153	-87.671767	39523	236
1	Beverly	East Beverly	41.718153	-87.671767	39523	236
2	Beverly	West Beverly	41.718153	-87.671767	39523	236
3	Norwood Park	Big Oaks	41.885310	-87.622130	32875	221
4	Norwood Park	Norwood Park East	41.985590	-87.800582	32875	221

Figure 1.8 new consolidated dataset of the Neighborhoods, along with their community area, crime data and the respective Neighborhood's co-ordinates

Neighborhoods with high per capita income and lowest crime rate were selected. There are 35 neighborhoods which has been selected with above criteria, they are visualized on a map using folium on python (see fig 2.9)

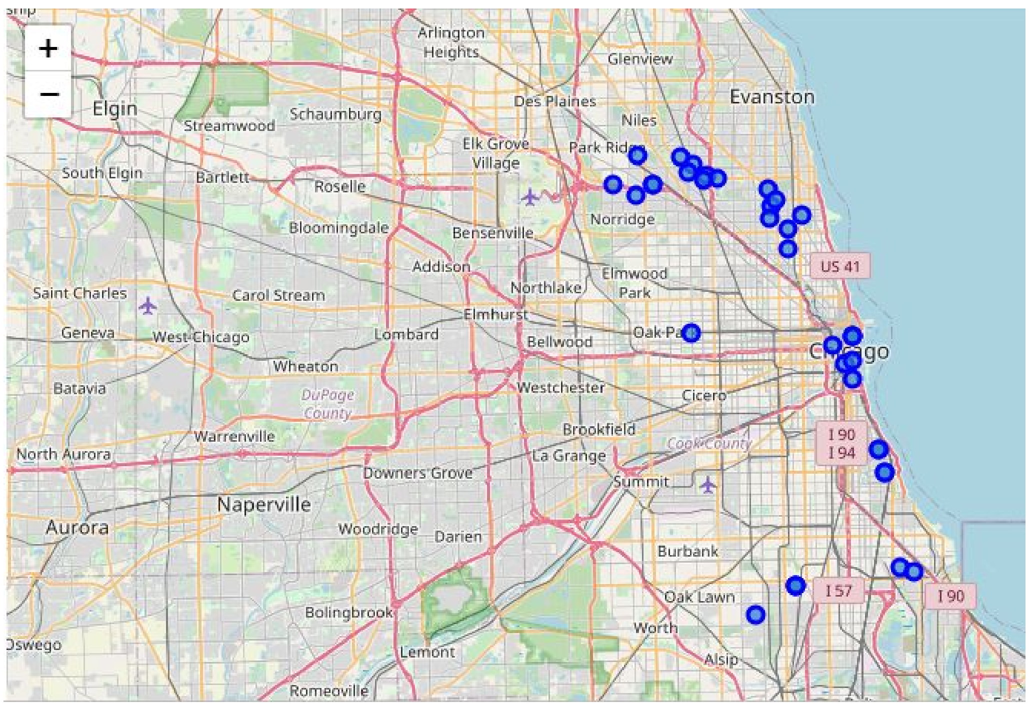


Figure 2.9 visualization of selected neighborhood

2.3 Modeling

Using the final dataset containing the selected neighborhoods along with the latitude and longitude, we can find all the venues within a 500 meter radius of each neighborhood by connecting to the Foursquare API. This returns a json file containing all the venues in each neighborhood which is converted to a pandas dataframe. This data frame contains all the venues along with their coordinates and category (see fig 2.10)

```
(1080, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Beverly	41.718153	-87.671767	Ridge Park	41.718378	-87.667921	Park
1	Beverly	41.718153	-87.671767	Jimmy Jamm's Sweet Potato Pies	41.721181	-87.669373	Bakery
2	Beverly	41.718153	-87.671767	Top-Notch Beefburgers	41.721281	-87.675382	Burger Joint
3	Beverly	41.718153	-87.671767	Southtown Health Foods	41.721257	-87.674822	Grocery Store
4	Beverly	41.718153	-87.671767	GBN Nail Salon	41.721371	-87.668500	Cosmetics Shop

Figure 2.10 Venue details of each Neighborhood

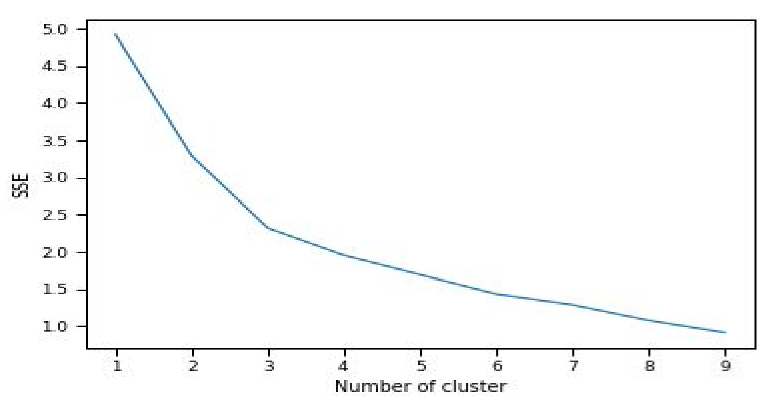
One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods.

```
neighborhoods_venues_sorted.head()
```

	Neighborhood	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	Beverly	Cosmetics Shop	Grocery Store	Park	Italian Restaurant	Farmers Market	Burger Joint	Boutique	Train Station	Bakery	Yoga Studio
1	Big Oaks	Hotel	Coffee Shop	Plaza	Steakhouse	Seafood Restaurant	Park	Sandwich Place	American Restaurant	Hotel Bar	Museum
2	Bowmanville	Sandwich Place	New American Restaurant	Ice Cream Shop	Bar	Filipino Restaurant	Garden	Dive Bar	Supermarket	Mobile Phone Shop	Coffee Shop
3	Budlong Woods	Bakery	Middle Eastern Restaurant	Karaoke Bar	Nightclub	Discount Store	Sushi Restaurant	Cajun / Creole Restaurant	Food & Drink Shop	Mexican Restaurant	Greek Restaurant
4	Calumet Heights	Bus Station	Gym / Fitness Center	Park	Yoga Studio	Garden	French Restaurant	Fountain	Football Stadium	Food Truck	Food Court

Figure 2.11: Ten most common venues in each neighborhood

To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use elbow criterion to find optimum number of cluster present in the dataset .



A cluster size of 5 for this project that will cluster the 35 neighborhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

3. Results

After running the K-means clustering we can access each cluster created to see which neighborhoods were assigned to each of the five clusters. Looking into the neighborhoods in the first cluster (see fig 4.1)

	Neighborhood	per_cap_income	Total_Case	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	Com V
0	Beverly	39523	236	0	Cosmetics Shop	Grocery Store	Park	Italian Restaurant	Farmers Market	Burger Joint	Boutique	Train Station	Bakery	S
1	East Beverly	39523	236	0	Cosmetics Shop	Grocery Store	Park	Italian Restaurant	Farmers Market	Burger Joint	Boutique	Train Station	Bakery	S
2	West Beverly	39523	236	0	Cosmetics Shop	Grocery Store	Park	Italian Restaurant	Farmers Market	Burger Joint	Boutique	Train Station	Bakery	S
30	Mount Greenwood	34381	130	0	Cosmetics Shop	Vineyard	Home Service	Park	Yoga Studio	Falafel Restaurant	French Restaurant	Fountain	Football Stadium	-

Figure 3.1: Cluster 1

The cluster one with 4 neighborhoods spared across mount Greenwood and Beverly community areas. Upon closely examining these neighborhoods we can see that the most common venues in these neighborhoods are Cosmetic Shop, Grocery Store, Parks, Restaurant, Bakery, and stadium. Looking into the neighborhoods in the second cluster clustered from, Forest glen, Calumet Heights, Near South Side which consist of venue such as Bus Station, Park, Nature Preserve

	Neighborhood	per_cap_income	Total_Case	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
14	Calumet Heights	28887	358	1	Bus Station	Gym / Fitness Center	Park	Yoga Studio	Garden	French Restaurant	Fountain	Football Stadium	Food Truck
15	Central Station	59077	492	1	Park	Bus Station	Intersection	Liquor Store	Gym	Train Station	Grocery Store	Shoe Repair	Donut Shop
24	Sauganash	44164	153	1	Park	Indian Restaurant	Asian Restaurant	Fast Food Restaurant	Pharmacy	Yoga Studio	Falafel Restaurant	French Restaurant	Fountain
26	Wildwood	44164	153	1	Nature Preserve	Trail	Theater	Park	French Restaurant	Fountain	Football Stadium	Food Truck	Food Court

Figure 3.2: Cluster 2

Third and fourth clusters, we can see these clusters have less than 3 neighborhood in each. This is because of the unique venues in each of the neighborhoods; hence they couldn't be clustered into similar neighborhoods (see figures 3.3 and 3.4). Neighborhood in cluster 3 is from same community area having similar common venues

	Neighborhood	per_cap_income	Total_Case	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
4	Norwood Park East	32875	221	2	Park	Yoga Studio	Farmers Market	Furniture / Home Store	French Restaurant	Fountain	Football Stadium	Food Truck	Food Court	Food & Drink Shop
5	Norwood Park West	32875	221	2	Park	Yoga Studio	Farmers Market	Furniture / Home Store	French Restaurant	Fountain	Football Stadium	Food Truck	Food Court	Food & Drink Shop

Figure 3.3: Cluster 3

The fourth cluster has one neighborhood which consists of Venues such as Golf Course, Yoga Studio and Gas Station

	Neighborhood	per_cap_income	Total_Case	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	10t Mos Commo Venu
25	South Edgebrook	44164	153	3	Golf Course	Yoga Studio	Gas Station	Furniture / Home Store	French Restaurant	Fountain	Football Stadium	Food Truck	Food Court	Food Drir Shc

Figure 3.4 Cluster 4

The cluster five is the biggest cluster with 24 of the 35 neighborhoods. Upon closely examining these neighborhoods we can see that the most common venues in these neighborhoods are Hotel Cafe, Restaurants, Football Stadium, Coffee Shop, and Yoga Studio

	Neighborhood	per_cap_income	Total_Case	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	
3	Big Oaks	32875	221	4	Hotel	Coffee Shop	Plaza	Steakhouse	Seafood Restaurant	Park	Sandwich Place	American Restaurant	
6	Old Norwood	32875	221	4	Hotel	Pizza Place	Café	Gym / Fitness Center	Parking	Metro Station	Grocery Store	Sandwich Place	Co
7	Oriole Park	32875	221	4	Football Stadium	Video Store	Gym	Park	Yoga Studio	Falafel Restaurant	French Restaurant	Fountain	F
8	Union Ridge	32875	221	4	Coffee Shop	Food Truck	Sandwich Place	Gym	Mediterranean Restaurant	Mexican Restaurant	American Restaurant	Salad Place	P
9	Bowmanville	37524	561	4	Sandwich Place	New American Restaurant	Ice Cream Shop	Bar	Filipino Restaurant	Garden	Dive Bar	Supermarket	Pr

Figure 4.5 Cluster 5

Visualizing the clustered neighborhoods on a map using the folium library (see fig 3.6)

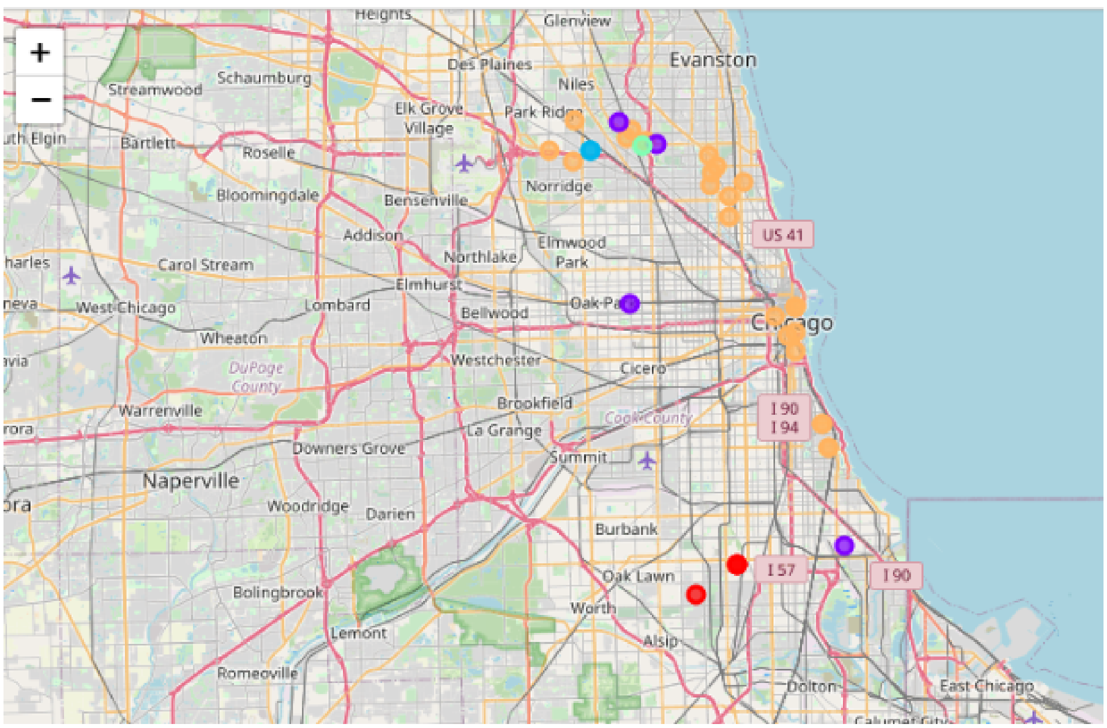


Figure 3.6 Clustered neighborhoods

Each cluster is color coded for the ease of presentation; we can see that majority of the neighborhood falls in the orange cluster which is the fifth cluster. The green cluster consists of 4 neighborhoods which is the 1st cluster. The purple cluster consists of 4 neighborhoods which is the 2nd cluster. Blue cluster consist of one neighborhood which is the 3rd cluster. The red cluster consists of two neighborhoods which is the 4th cluster.

4.Discussion

The objective of the business problem was to help stakeholders identify one of the safest Neighborhood with higher per capita income in Chicago Illinois, and an appropriate neighborhood within the community area to set up a commercial establishment especially a Grocery store. This has been achieved by first making use of Chicago crime data to identify a safe community area with considerably higher per capita income for any business to be viable. After selecting the community area it was imperative to choose the right neighborhood where grocery shops were not among top 10 most common venues in a close proximity to each other. We achieved this by grouping the neighborhoods into clusters to assist the stakeholders by providing them with relevant data about venues and safety and population with higher individual income of a given neighborhood.

5. Conclusion

We have explored the crime data to understand different types of crimes in all Community area of Chicago and later segregate them based on per capita income, this helped us for selecting area with high per capita income and lowest crime rate. Once we confirmed the community area the number of neighborhoods for consideration also comes down, we further shortlist the neighborhoods based on the common venues, to choose a neighborhood which best suits the business problem.