Segmenting and Clustering Community Areas in Chicago using Machine Learning for Opening a Restaurant

By: Asmat Ullah **June 05, 2020**

1. Introduction

1.1 Background

An entrepreneur is looking up for opening a restaurant in one of the seventy-seven community areas in the city of Chicago. The entrepreneur has not yet decided on the type of cuisine his restaurant will be serving but need analysis and recommendations based on overview of competition level, socio-economic conditions and suitability of community areas for opening a restaurant in the city of Chicago.

To do this analysis the following factors have been agreed to consider for each community area: population density, per capita income, count of crimes, restaurants and car parkings. We have summarized the importance of each factor to the analysis here. Areas with higher population generally leads to greater demand for consumption of goods and services. Another important factor that drives demand for goods and services is the purchasing power of the people. There will be lesser demand for goods and services if people do not have adequate disposable income to spend. So, we will prefer a community area in the city with higher population and wealth for opening a restaurant. Further, businesses thrive in areas with good law and order because a restaurant owner and its customers do not want to be robbed of their valuables and suffer physical and financial losses. Additionally, a business can generate higher profit if there is lesser competition. Higher competition narrows profit margin and a business may suffer losses and eventually go out of business. Last but not least, the availability of vehicle parking space is an important factor that customers take into account while visiting a restaurant. The availability of parking facility greatly increases the likelihood of customers visiting a restaurant.

1.2 Problem

To segment community areas for opening a restaurant in the city of Chicago based on population density, per capita income, count of restaurants, crime and car parkings.

2. Data

We obtained data about the number of communities, their population and population density per square mile from Wikipedia[1]. The population data is for the year 2017. We also obtained names of neighborhood in each community area from Wikipedia[2]. We used this data to determine population density per square and calculate the other ratios as defined below. We scrapped the data from the websites using beautiful soup libraries for Python.

We used Chicago government data portal to obtain data about crimes[3] and per capita income[4] of each community area. The crime data is for the year 2019 and per capita income is for the year 2012. We used

these data in conjunction with population data to compute Crimes per Hundred of Population and Income per Capita ratios as defined below.

To determine the number of restaurants and car parkings in each community, we used Foursquare's[5] API to explore each community area for restaurants and car parkings. The data from the API was used in conjunction with population data to calculate ratios of Restaurant per Thousand of Population and Car Parking per Thousand of Population as defined below.

It is tempting to use average rating of restaurants in each community area of the city that would indicate the quality of service offered by restaurants in the respective community and would be indicative of the level of competition an investor would face in the restaurant business in a community area. We tried to obtain rating of restaurants in the city of Chicago using Foursquare API but the API returned large number of restaurants without rating. Hence, we dropped the rating factor from our analysis. Had we included the rating result from Foursquare API, this would have caused noise in the analysis and clustering based on machine learning system that would not be useful for our purpose. Further, obtaining rating of venues from Foursquare is limited to five hundred venues per day only for standard account. There's cost associated with obtaining rating for more than five hundred venues.

3. Methodology

3.1 Data Refinement

We created a data table from the data consisting of the following twelve features of each community area:

i. Community Area Number
ii. Community Area Name
iii. Neighborhoods
iv. Latitude
v. Longitude
vi. Population
vii. Population Density (Sq./Mile)
viii. Per Capita Income
ix. Count of Crimes
x. Count of Restaurants
xi. Count of Parking

It is worth mentioning that community areas in Chicago are serially numbered ranging from 1 to 77 for identification purposes. We obtained the Latitudes and Longitudes of community areas using Geopy library. We passed each community area coordinates to Foursquare's API to explore them for restaurants and car parkings. The API returned maximum number of 50 venues for each community area for a given explore query. The returned data contained large number of duplicate results for venues as we had not passed on the radius parameter to the API within which to explore a given location. The purpose of this was to include as many restaurants as possible in our dataset. We addressed the issue by associating a venue to a community area based on nearest distance between a venue and a community area, and dropping rest of the duplicate venues data. We calculated the distance based on latitudes and longitudes of a venue and community areas. We used Haversine formula for calculating distance based on latitudes and longitudes. We obtained the code for the formula from Kite.com[6]. The returned data also contained venues that were not restaurants or car parkings. We removed these data from the dataset.

We then counted the number of restaurants and parkings in each community area and added it to our data table. We also counted the number of crimes in each community area using crime data from Chicago government data portal and this was then added to the data table.

3.2 Clustering

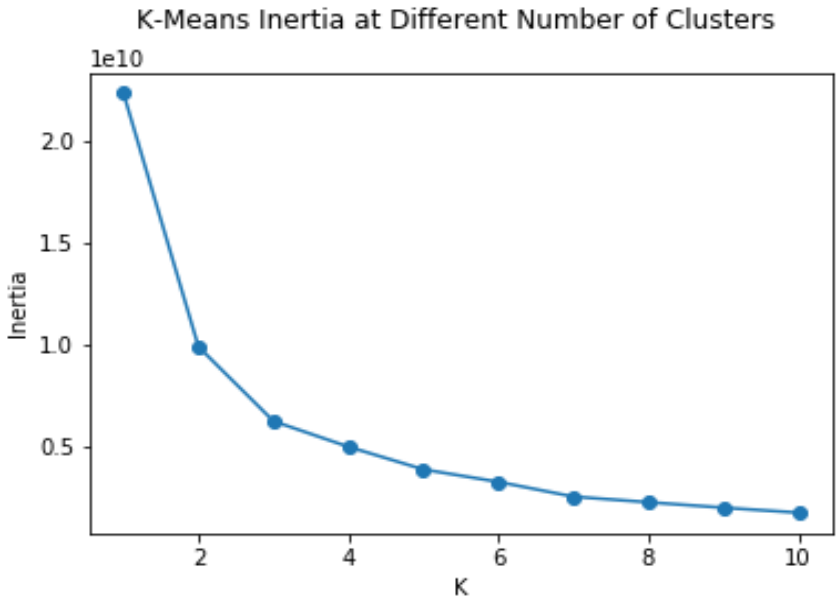
We need to use an unsupervised machine learning algorithm because the data is unlabeled and let the algorithm find clusters within the data. For this purpose, we selected K-Means algorithm to cluster communities in the city of Chicago based on Population Density per Square Mile, Income Per Capita, Count of Crimes, Count of Restaurants and Count of Car Parkings. K-Means algorithm groups data into specified numbers of clusters in which each data point belongs to a cluster. Data points are included in a cluster in such a way that the sum of the squared difference of data points and the cluster's mean is minimum.

Before performing K-Means algorithm, we standardize the data to remove the effects of size and difference in units of measurements of the features.

We used inertia method to find optimal number of clusters for K-Means algorithm. Inertia calculates square distance of data points within a cluster. Inertia decreases as number of clusters increases. The general rule of thumb is to select number of clusters at the elbow point from inertia graph because the additional decrease in inertia after this point is insignificant.

Figure 1 below shows graph of inertia level of K-Means algorithm for number of clusters ranging from 1 to 10:

Figure 1



Reading the above graph, we can see that decrease in inertia level is insignificant after K = 3. Based on this result, we selected number of clusters to be 3 for K-Means algorithm.

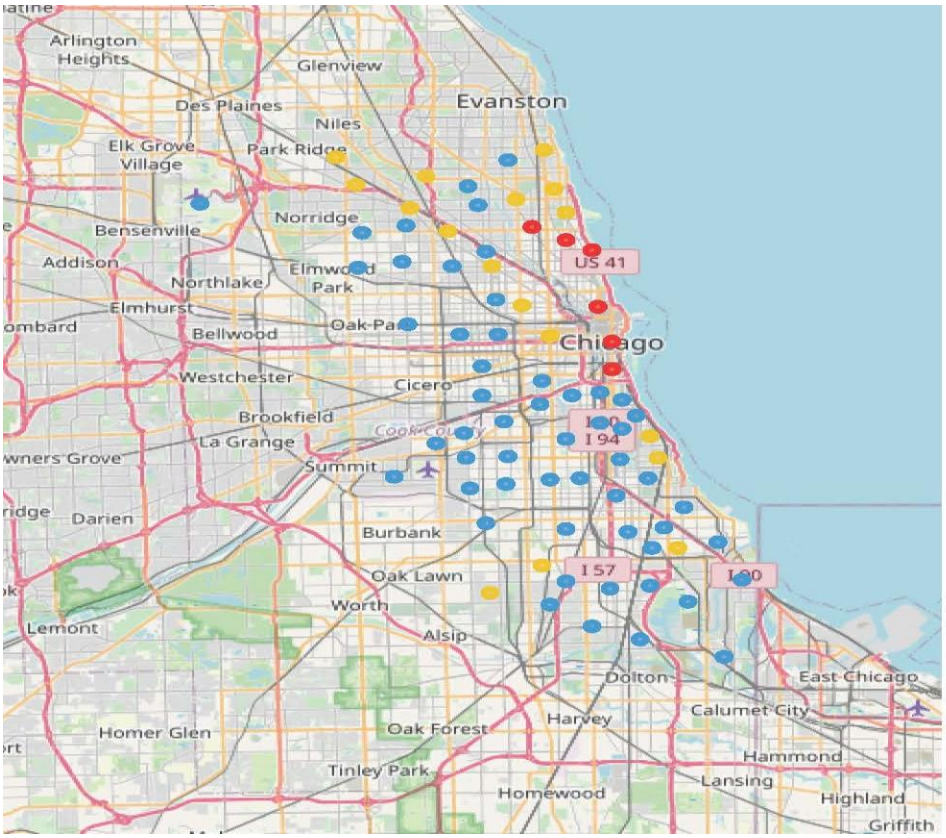
We then generated a map of Chicago city with markers displaying spread of the community areas of each cluster across the city of Chicago. To analyze the data, we generated side-by-side box plots of the clusters

for each feature of community areas and populated the box plots with scatter graphs to get a better understanding of the spread of community areas within each cluster and make comparison of the clusters based on the features. Box plots show five important summary statistics i.e. minimum value, first quartile, median, third quartile and maximum value. The box plots design was customized based on the ideas presented in the article by Ciarán Cooney[7].

4. Result

Figure 2 below shows the spread of community areas of each cluster across the city of Chicago based on the features we had used in K-Means clustering algorithm:

Figure 2

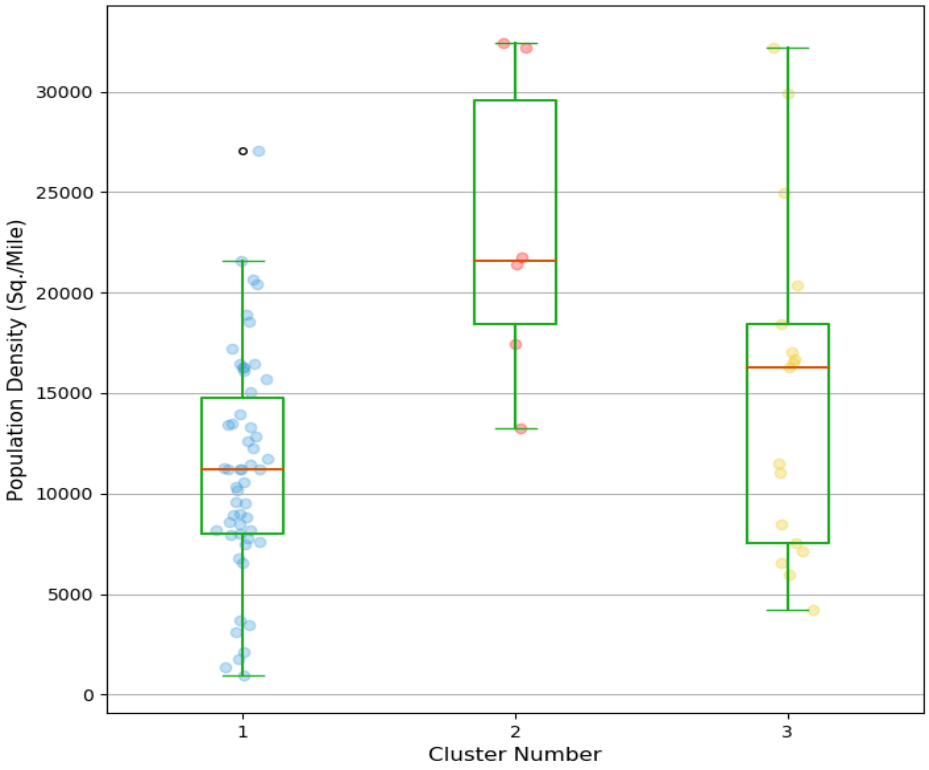


From the above map we can observe that most of the community area lies in the cluster with blue color followed by yellow and red clusters. To analyze the data further, we have to look at the box plots of each feature of the clusters. For comparison purposes we generated side-by-side box plots of each feature of the community areas clusters.

4.1 Population Density per Square Mile

Figure 3 below shows box plots that summarize the distribution of community areas within each cluster and across the clusters based on Population Density per Square Mile:

Figure 3

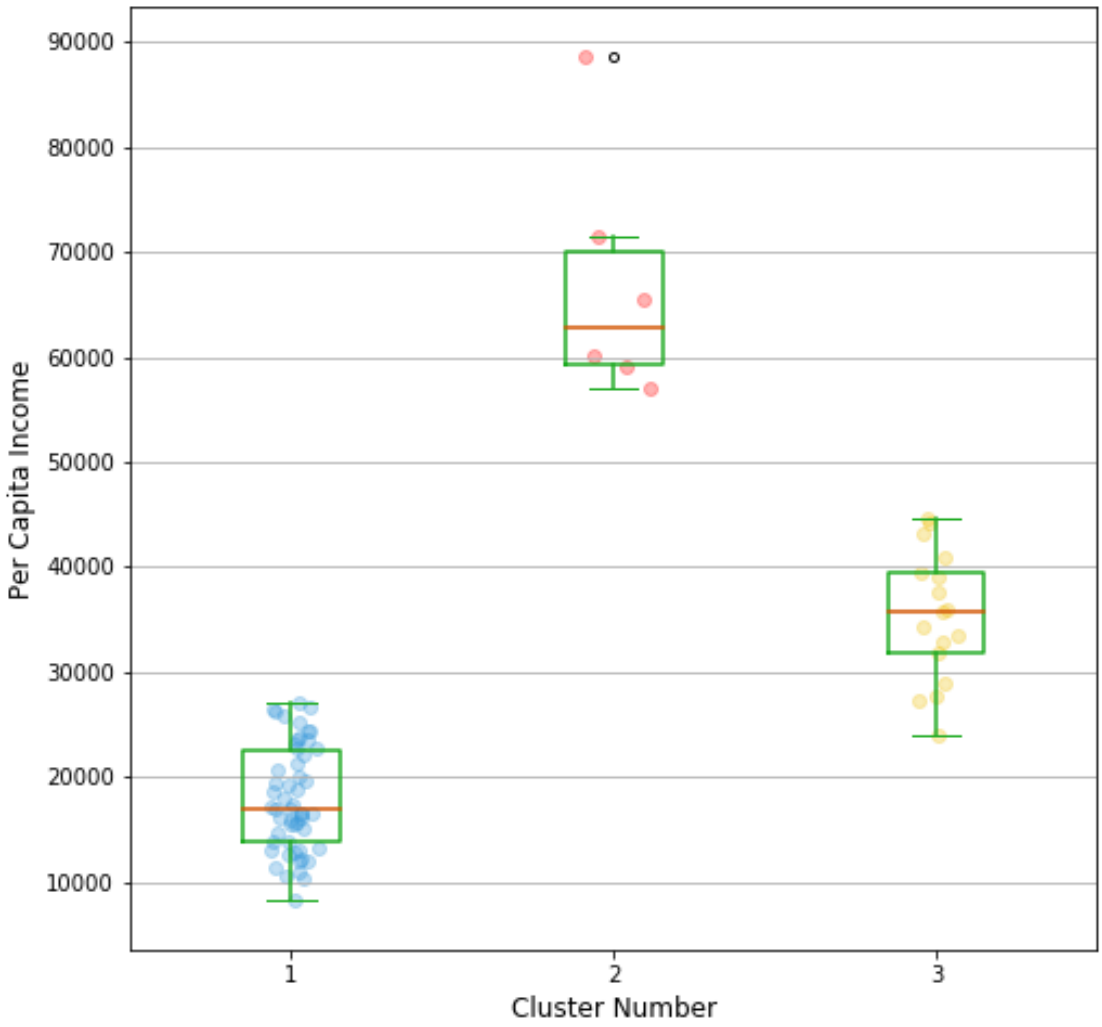


Based on the above box plots, we can generalize that cluster 1 consists of communities with lower Population Density per Square Miles with most of the community areas lying in the second and third quartile of the cluster. The second cluster consist of only six communities and has overall higher level of Population Density per Square Mile compared to cluster 1 and 3. The third cluster has mixed population density with one half similar to cluster 1 and the other half similar to cluster 2 in term of Population Density per Square Mile.

4.2 Per Capita Income

The following box plots in figure 4 summarize the distribution of Per Capita Income within each cluster and across the clusters:

Figure 4

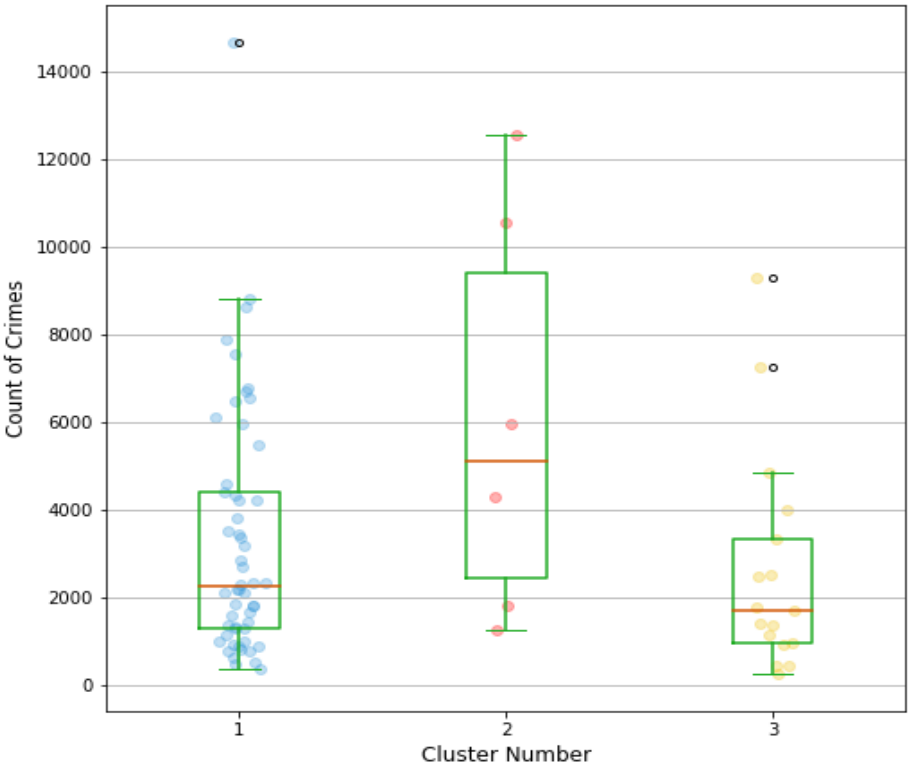


Based on the above box plots, we can say that community areas in cluster 1 has lowest Per Capita Income in the year 2012. Cluster 2 consists of community areas with highest Per Capita Income while cluster 3 lies in the middle of cluster 1 and cluster 2 in terms of Per Capita Income.

4.3 Count of Crimes

Next up is the distribution of number of crimes within each cluster and across the clusters. The box plots in figure 5 below summarize these distributions:

Figure 5

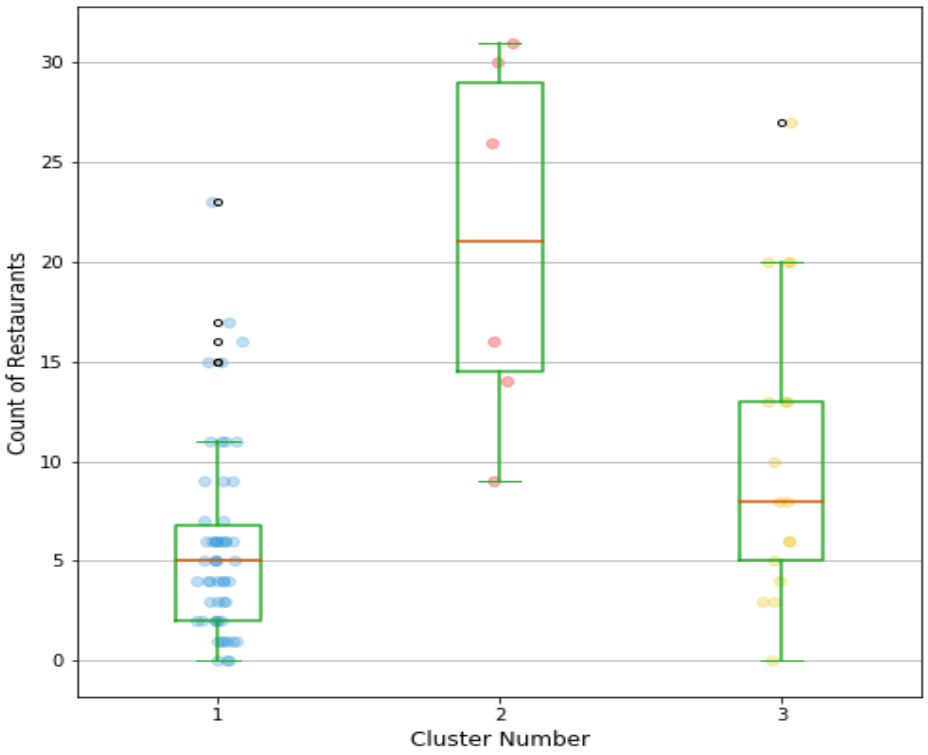


The above box plot of cluster 1 shows that half of its community areas lie in the lower crime count bracket i.e. community areas with crime count of approximately below 2200 for the year 2019 while the other half is widely dispersed between crime count of about 2200 and 9000. Crime in community areas in second cluster, on average, has higher count than community areas within the other two clusters. Community areas in cluster 3, on average, display lower crime rates than the other two clusters.

4.4 Count of Restaurants

Figure 6 below shows box plots of Count of Restaurants that summarize the distribution of community areas within clusters and across clusters based on number of restaurants in each community area:

Figure 6

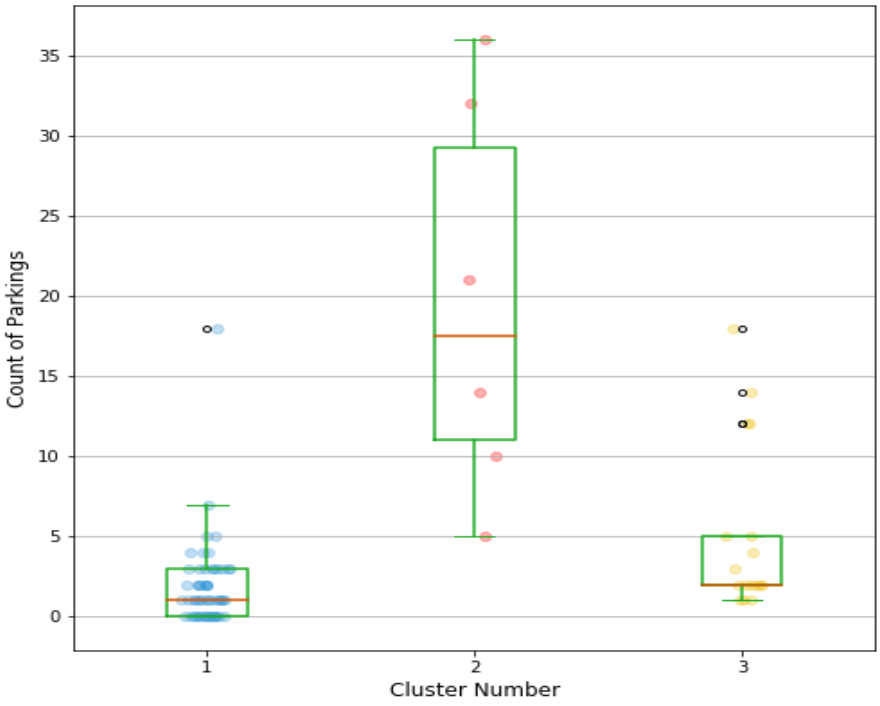


We can observe from the box plot of cluster 1 that community area in this cluster has generally lower number of restaurants than cluster 2 and 3 barred for some of the outliers in cluster 1. This indicates lesser competition of restaurants in community areas in cluster 1 compared to cluster 2 and 3. Cluster 2 consists of areas with highest number of restaurants, hence greater competition while cluster 3 falls below somewhere between cluster 1 and 3 in terms of competition between restaurants in the community areas.

4.5 Count of Parkings

Figure 7 shows box plots that summarize distribution of community areas based on number of car parkings within each cluster and across the three clusters:

Figure 8



From the above box plots, we can see that community areas in cluster 1 has, on average, lowest number of parkings available than community areas within the other two clusters. Communities in cluster 2 has generally higher number of parkings available while number of parkings in community areas in cluster 3 lies between cluster 1 and 2.

4.6 Summary of Result

We can summarize the above result in the following table:

Cluster Number	Population Density Per Square Mile	Per Capita Income	Count of Crimes	Count of Restaurants	Count of Parkings
1	Low	Low	Middle	Low	Low
2	High	High	High	High	High
3	Middle	Middle	Low	Middle	Middle

5. Recommendations

Based on the result above, community areas in cluster 1 are more suitable for restaurants with cheaper menu due to lower per capita income of the community areas in the cluster compared to cluster 2 and cluster 3. There also appears some leeway in charging higher price due to lower competition in the community areas in cluster 1 compared to the other clusters but significantly higher prices will deter customers due to low Per Capita Income of cluster 1. There are three community areas in this region with no restaurant. These are either industrial areas, areas with lower population or lower per capita income that are unlikely to be suitable for opening a restaurant. The lower Population Density per Square Mile explains the lower number of restaurants in the region. The lower Population Density per Square Mile reduces the chances for higher revenue and growth. Demand for restaurants in this region could also be impacted by the higher crime rate, despite having lowest Population Density per Square Mile, compared to cluster 2. Additionally, the region has also lowest number of car parkings compared to the other two region. Some community areas in this region have no parking at all. Lack of availability of car parking might deter customers from visiting a restaurant. The important thing to remember here is that if a nearby car parking is not available, a restaurant will have to bear additional cost to obtain area for car parking. The most appropriate strategy here will be to differentiate in terms of cost rather than quality.

Cluster 3 seems to be suitable for restaurants with moderate level of menu pricing justified by higher quality of service as the Per Capita Income in this region is higher than cluster 1. Competition level does not differ significantly from cluster 1 except for some community areas with higher level of competition. Population Density per Square Mile widely differs between the community areas within the cluster with some of them having Population Density per Square Mile similar to cluster 1 while others have Population Density per Square Mile similar to cluster 3. So, the level of demand and growth will depend on where a restaurant is located in cluster 3. Count of crime is lowest for this region making it attractive for customers visiting this area for restaurants. Car parking condition is mostly similar to cluster 1 except for three community areas where high number of car parkings are available.

Cluster 2 consists of lowest number of community areas i.e. only six but it is the cluster with scoring high on each feature. Highest Per Capita Income in this cluster of communities means they can comfortably afford higher priced menu. Highest number of restaurants in these areas mean highest competition compared to cluster 1 and 3, hence a restaurant will have to continuously improve its quality of service to stay ahead of its competitors. There are greater chances for higher level of demand and growth as the areas in this cluster have highest Population Density per Square Mile. Count of Crimes varies significantly among the community areas in this cluster with some of the community areas having highest number of crimes compared to cluster 1 and 3. The wealth of these areas might attract robbers and detract customers from visiting restaurants in the community areas with high crime rates. The higher crime rates in these areas may also, in part, be explained by higher population density of the areas. The cluster, on average, has higher number of car parkings compared to cluster 1 and 3 that can attract customers to this cluster of communities for eating out. In summary, the focus in this cluster will be mostly on quality of service rather than price. The appropriate strategy here will be how a restaurant differentiates itself from the rest of the restaurants in terms of quality. Higher prices will signify status quo that suits the needs of customers in these community areas.

6. Conclusion

The purpose of this report was to segment and cluster community areas in the city of Chicago based on population density, per capita income, count of crimes, restaurants and car parkings for opening a restaurant. We obtained and process the data to create a dataset of the relevant features of the community areas.

While exploring the community areas for restaurants and car parkings using Foursquare API, we encountered duplicate results of restaurants and car parkings with different community numbers because the radius, within which Foursquare explored a given location coordinates, overlapped between community areas causing duplicate venues. We removed the duplicate venues data by associating restaurants and car parkings to the nearest community area using Haversine formula. This may not be the accurate measure for computing distance between two locations, especially for populated areas, because it does not take into account structures, such as, buildings, houses, roads, streets, etc. More accurate measures may be to use map APIs, such as Google map, to find distance between two locations as these APIs take structures into account to find distance. However, there is cost associated with using these API. Hence, we relied on the Haversine formula for approximating distance between a venue and a community area.

As the dataset was unlabeled the appropriate technique to use was unsupervised machine learning algorithm. For this purpose, we used K-Means which is an unsupervised machine learning algorithm. We found out the appropriate number of clusters to use in K-Means is three. We did this by analyzing inertia of the algorithm at different number of clusters to decide on the appropriate number of clusters.

We used folium map to visualize the distribution of the community areas within the clusters on map. We then used box plots and scatter plots for each feature to analyze the distribution of community areas within each cluster and make comparison across clusters.

Cluster 1 consisted of community areas with lowest population density per square mile, per capita income, number of restaurants, number of car parkings and middle number of crimes. Based on these characteristics, the most appropriate strategy for a restaurant to use in these community areas is having a low-priced menu.

Features of community areas in cluster 3 ranked between cluster 1 and 2 except for count of crime that was the lowest among the three clusters. The appropriate strategy for a restaurant here will be more focus on quality with modestly priced menu.

Cluster 2, on average, ranked highest among the three clusters. The appropriate strategy to adopt here will be differentiation of service quality from the rest of competitors in the region. Price will not matter here as most of the customers will be high net-worth individuals for whom quality and status will matter more than price.

The above strategies are based on the broad characteristics of the cluster rather than an individual community area. A strategy might be refined after closely examining features of community areas within each cluster. This may be aided by performing further unsupervised machine learning on the three clusters separately to identify community areas with higher population density, per capita income, count of restaurants, count of parkings and lower count of crimes.

There are also other factors that may greatly affect the clustering of the community areas. One factor could be the rating of restaurants. Higher rating of restaurants in a community area will indicate greater level of competition. We did not include these data because of the cost associated with obtaining such data from third parties. Another factor could be the nature of crime committed in an area such as stalking and trespass

and where no arrest is made because of the minor nature of crime. Latest data about per capita income could also influence the segmentation and clustering of the communities.

The above analysis will help in narrowing down potential community areas for opening a restaurant. Additional on the ground research will be needed to find out taste of customers, cost and availability of resources such as rent, land, chefs, etc. and any other factor that may influence the final outcome.

7. Appendices

i. https://en.wikipedia.org/wiki/Community_areas_in_Chicago
ii. https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago
iii. https://data.cityofchicago.org/api/views/w98m-zvie/rows.csv?accessType=DOWNLOAD
iv. https://data.cityofchicago.org/api/views/kn9c-c2s2/rows.csv?accessType=DOWNLOAD
v. https://foursquare.com/
vi. https://kite.com/python/answers/how-to-find-the-distance-between-two-lat-long-coordinates-in-python
vii. https://towardsdatascience.com/scattered-boxplots-graphing-experimental-results-with-matplotlib-seaborn-and-pandas-81f9fa8a1801