SEGMENTING AND CLUSTERING COMMUNITY AREAS IN CHICAGO FOR OPENING A RESTAURANT USING MACHINE LEARNING

By: Asmat Ullah

INTRODUCTION

- The purpose of this study is to segment and cluster community areas in the city of Chicago for opening a restaurant using machine learning algorithms.

- The factors used to segment and cluster the community areas include Population Density Per Square Mile, Per Capita Income, Count of Crimes, Count of Restaurants and Count of Car Parkings.

INTRODUCTION

- **Population Density Per Square Mile:** The level and growth of demand is affected by the population of an area, the larger the population the greater will be the demand and vice versa.
- **Per Capita Income:** The higher the disposable income of people in a region the higher will be the spending on goods and services.
- **Count of Crimes:** People tend to avoid visiting areas with higher crime rates causing decrease in demand for goods and services in that area. Investors don’t want to invest in areas with higher crime rates and lesser demand.
- **Count of Restaurants:** The greater the number of restaurants in an area the higher will be the competition in that region and lower profit margins.
- **Count of Car Parkings:** Customers tend to prefer restaurants with car parking available.

DATA SOURCES

- Census and community areas data was obtained from Wikipedia. Census data pertains to year 2017. These data was used to obtain Population Density Per Square Mile for each community area.

- Per Capita Income data for the year 2012 and crime data for the year 2019 was obtained from Government of Chicago data portal

- Restaurant and parking data for each community area was obtained using Foursquare API.

DATA SOURCES

- Census and community areas data was obtained from Wikipedia. Census data pertains to year 2017. These data was used to obtain Population Density Per Square Mile for each community area.

- Per Capita Income data for the year 2012 and crime data for the year 2019 was obtained from Government of Chicago data portal

- Restaurant and parking data for each community area was obtained using Foursquare API.

Methodology

- Created a data table from the data consisting of the following twelve features of each community area:
  - Community Area Number
  - Community Area Name
  - Neighborhoods
  - Latitude
  - Longitude
  - Population
  - Population Density (Sq./Mile)
  - Per Capita Income
  - Count of Crimes
  - Count of Restaurants
  - Count of Parkings

Methodology

- Obtained the Latitudes and Longitudes of community areas using Geopy library.
- Passed each community areas coordinates to Foursquare's API to explore them for restaurants and car parkings.
- Removed 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.

Methodology

- Counted the number of restaurants and parkings in each community area and added it to the data table.

- 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.

Methodology - Clustering

- Used unsupervised machine learning algorithm, as the data was unlabeled, to find clusters within the data

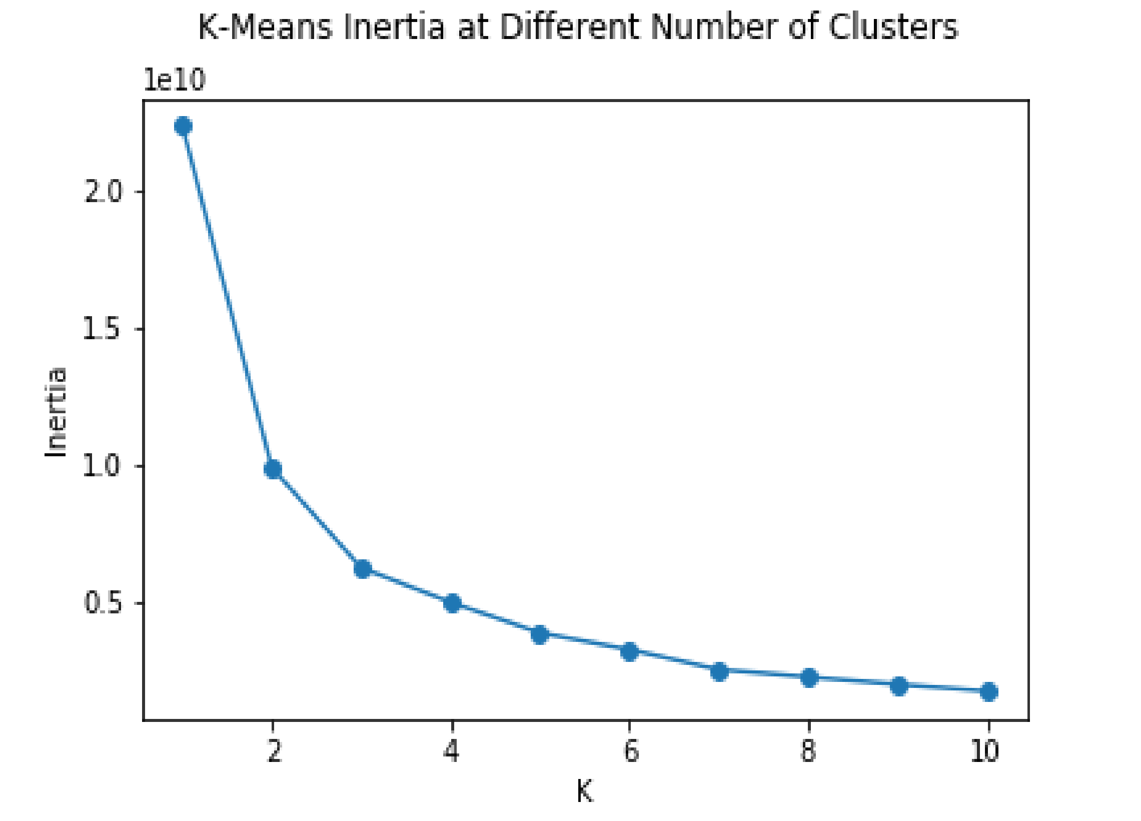
- K-Means, an unsupervised machine learning method, was used to group data into specified numbers of clusters in which each data point belongs to a cluster.

- 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.

- Standardized the data to remove effects of size and difference in units of measurements of the features before running the algorithm.

Methodology - Clustering

- Used inertia method to find optimal number of clusters for K-Means algorithm.
- Inertia calculates square distance of data points within a cluster.
- 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.
- Looking at the graph, we can observe decrease in inertia level is insignificant after K = 3.



Methodology - Clustering

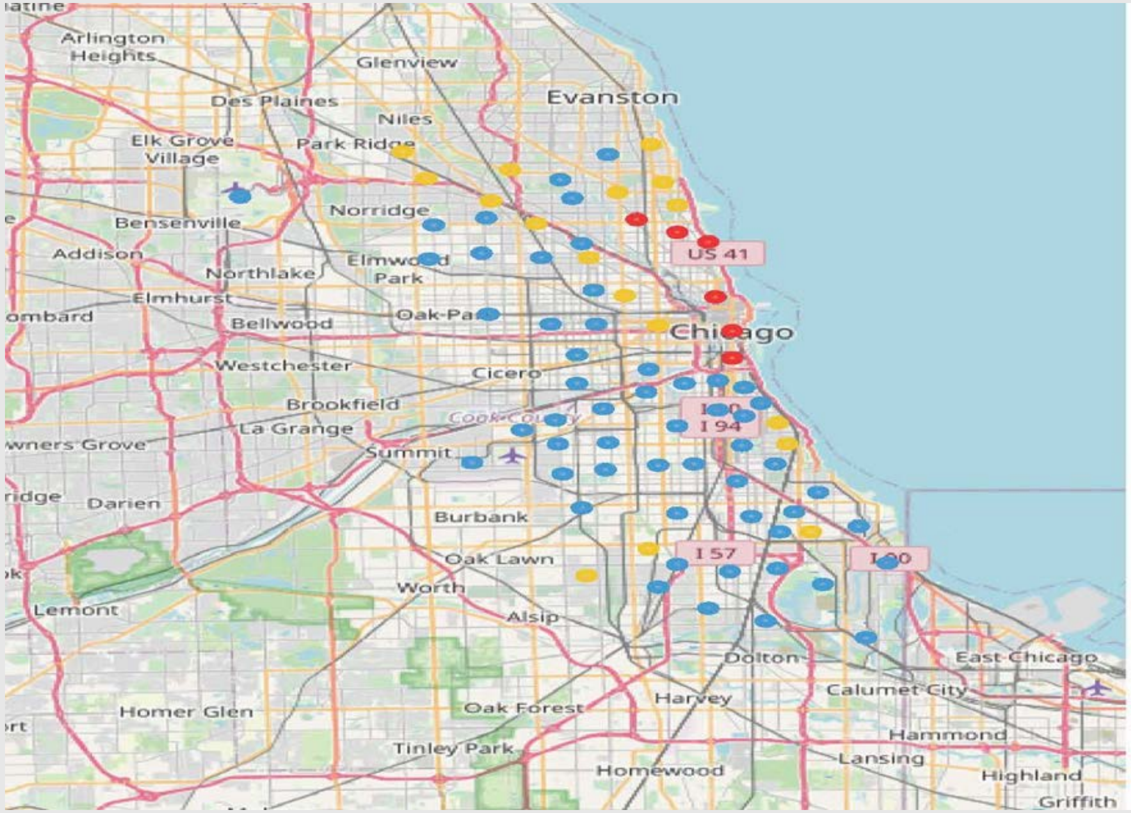
- Generated a map of Chicago city with markers displaying spread of the community areas of each cluster across the city of Chicago

- 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 clusters based on the features.
- Box plots summarizes data distribution by graphically showing the min, first quartile, median, third quartile and max values of data.

Result – Map

- The map 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.

- From the map we can observe that most of the community area lies in the cluster with blue color followed by yellow and red clusters.

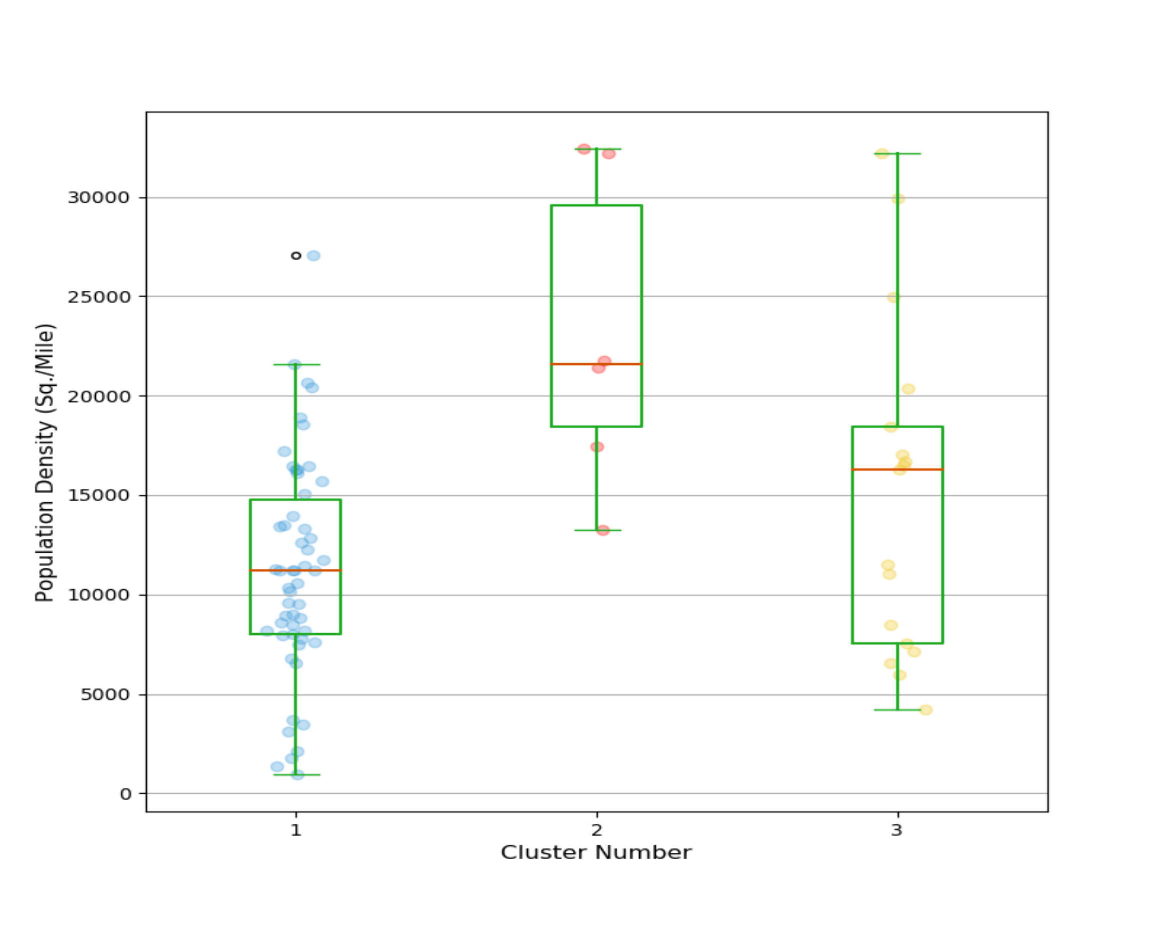


Result – Box Plots

Population Density per Square Mile

- The box plots summarize the distribution of community areas within the clusters and across the clusters based on Population Density per Square Mile.

- Cluster 1 overall 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.

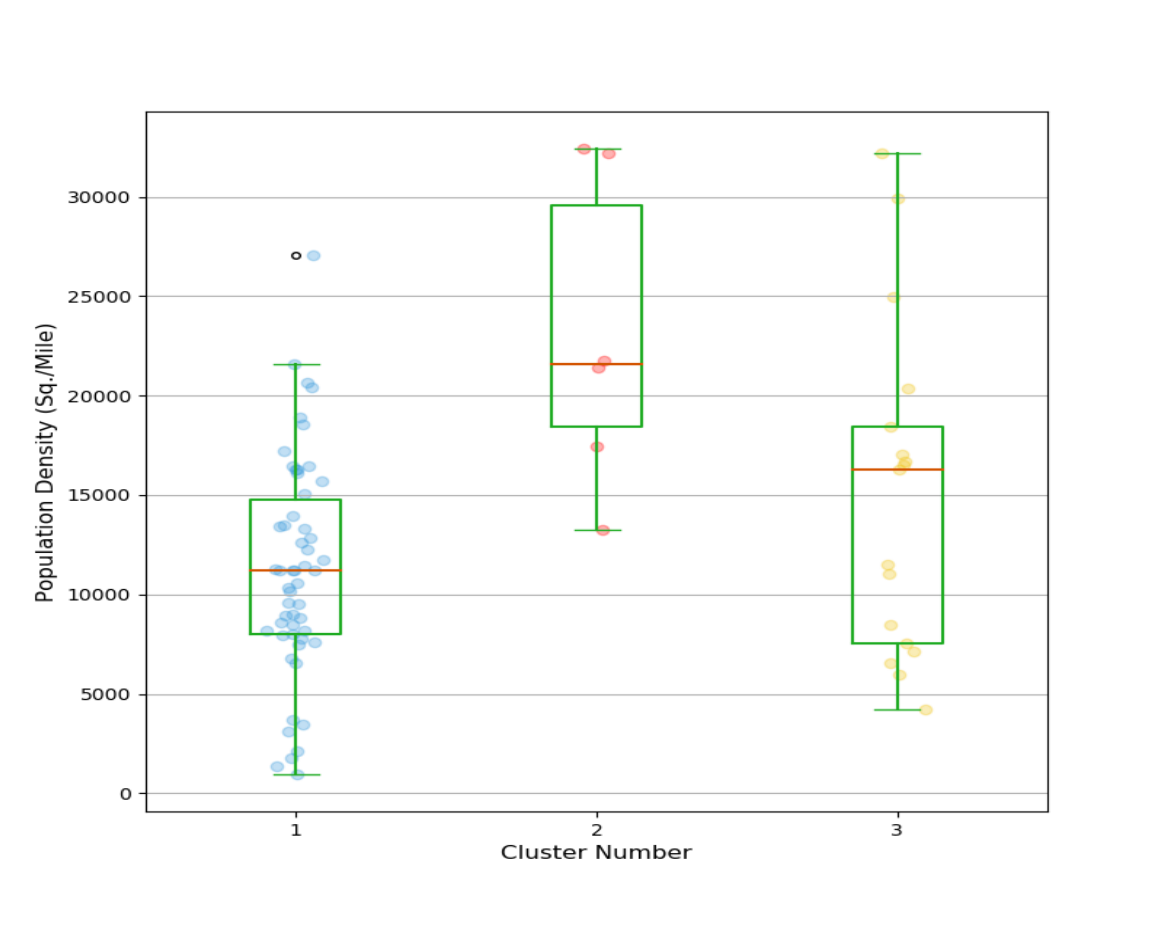


Result – Box Plots

Population Density per Square Mile

- 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.

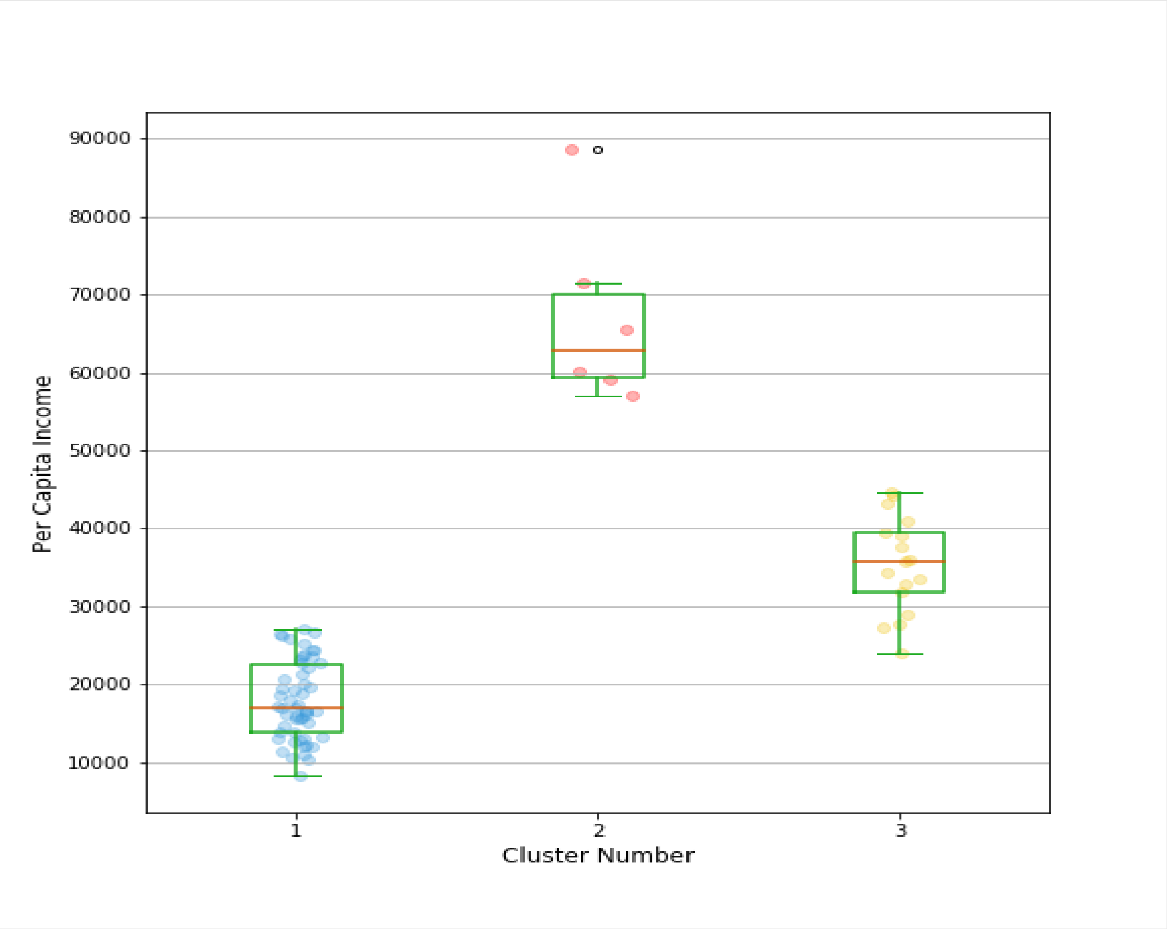
- The third cluster has mixed population density with one half similar to cluster 1 and the other half similar to cluster 2 in term of Population Density per Square Mile.



Result – Box Plots

Per Capita Income

- The box plots in figure 5 below summarize these distributions
- 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.

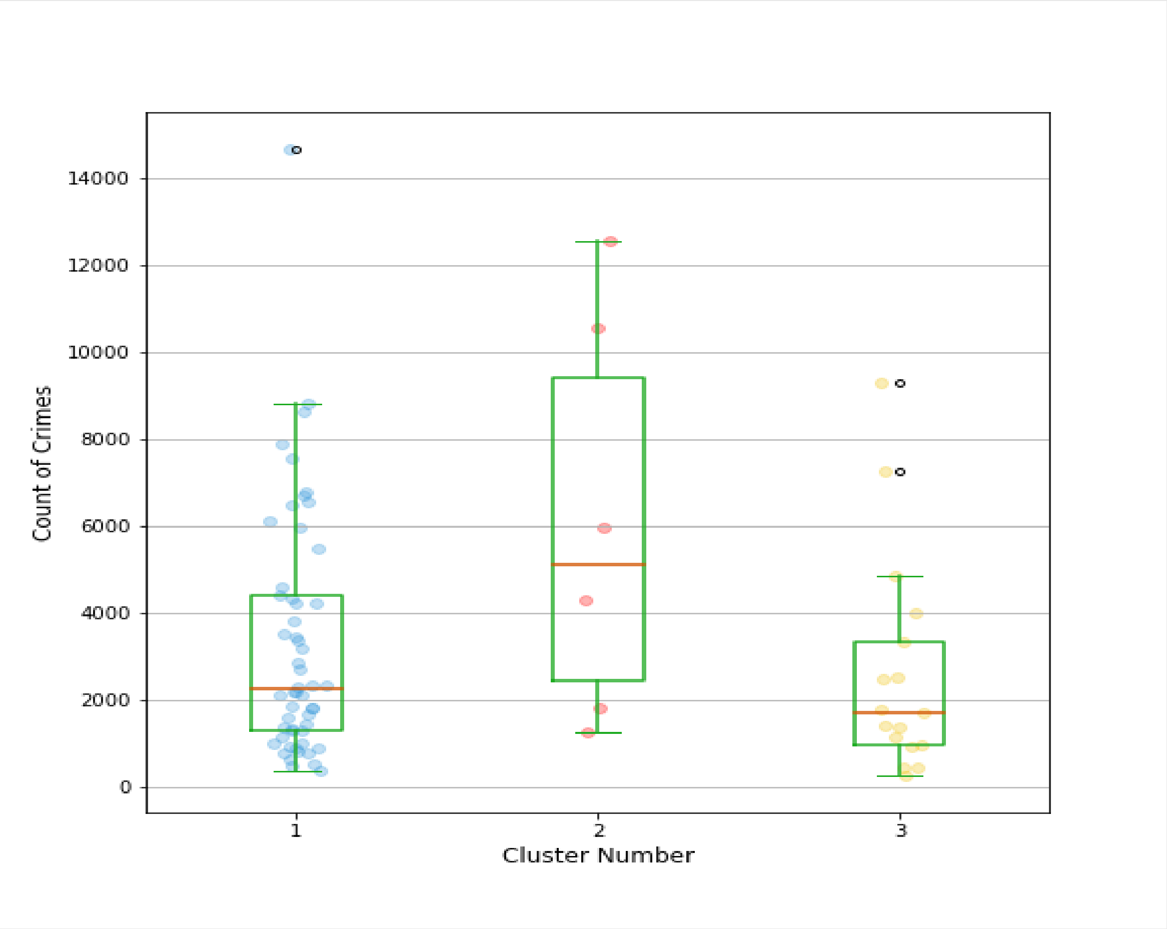


Result – Box Plots

Count of Crimes

- The box plots summarize the distribution of Count of Crimes within each cluster and across the clusters.

- 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.

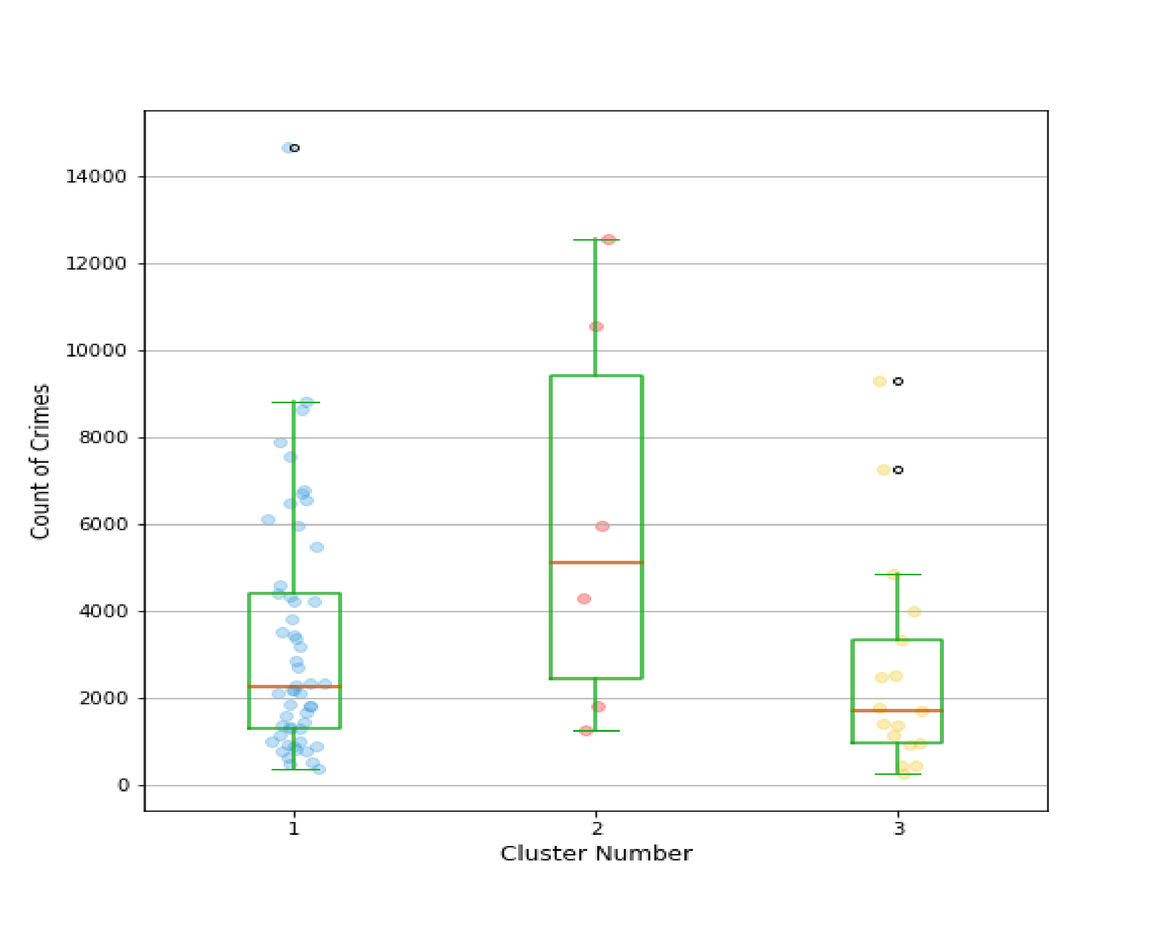


Result – Box Plots

Count of Crimes

- Crime count 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.

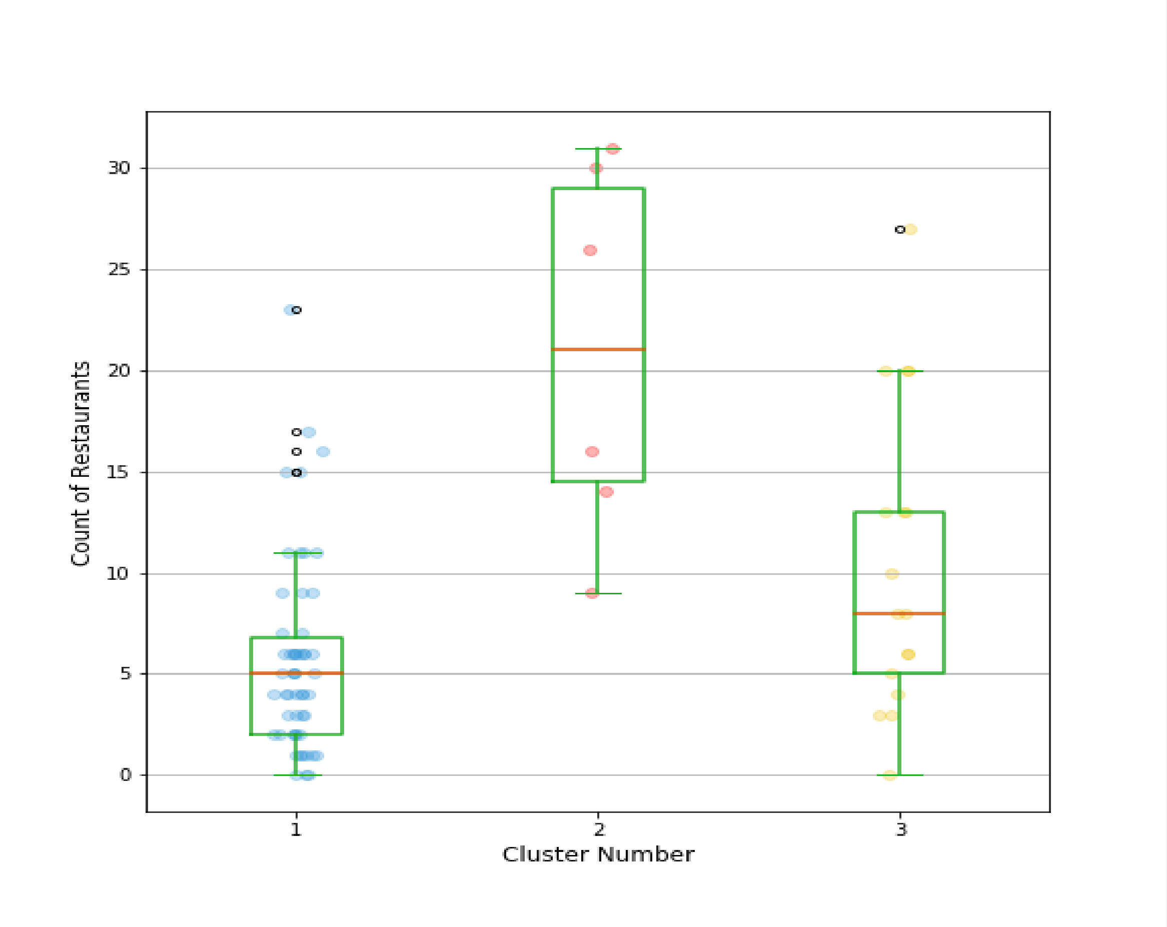


Result – Box Plots

Count of Restaurants

- The box plots summarize the distribution of community areas within clusters and across clusters based on number of restaurants in each community area.

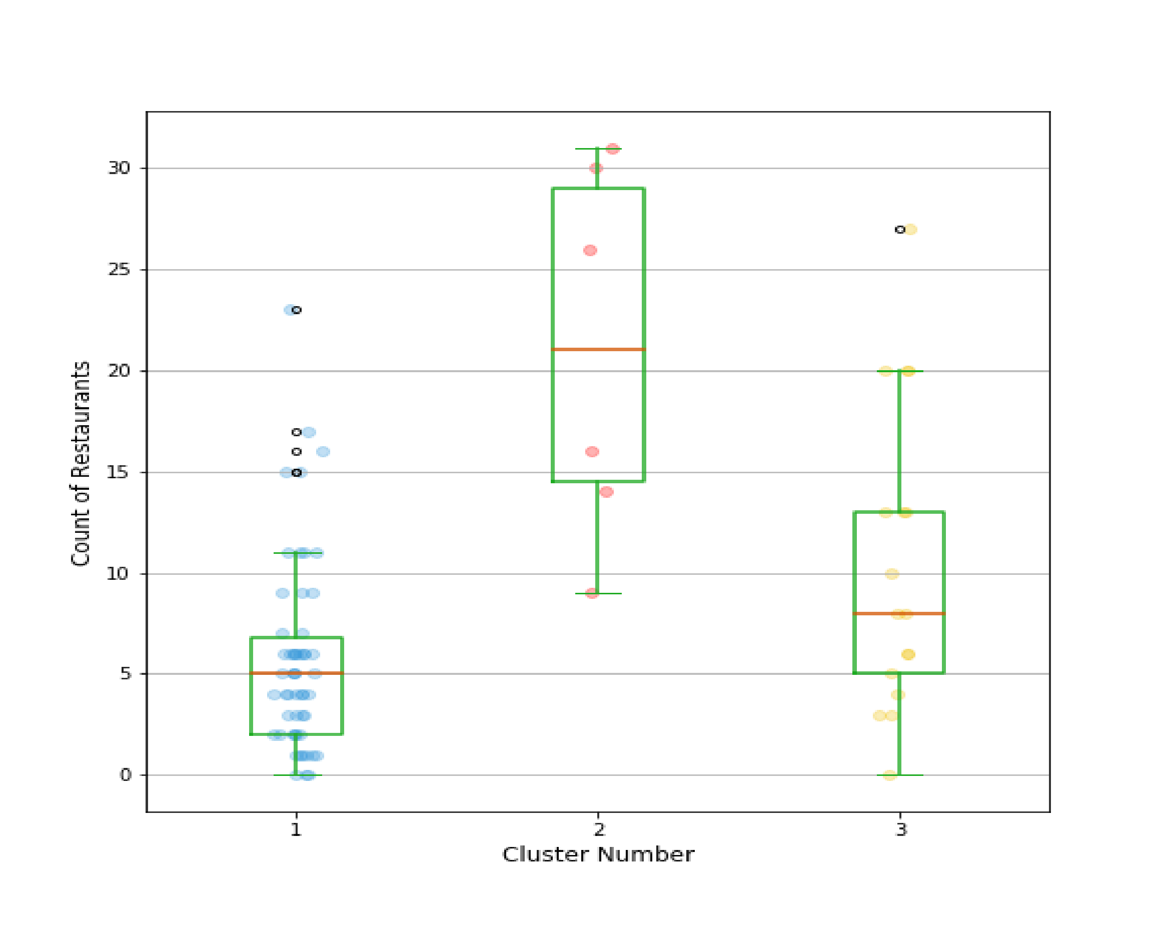
- Community areas in cluster 1 has generally lower number of restaurants than cluster 2 and 3 barred for some of the outliers in cluster 1. This indicates lesser competition of restaurants in community areas in cluster 1 compared to cluster 2 and 3.



Result – Box Plots

Count of Restaurants

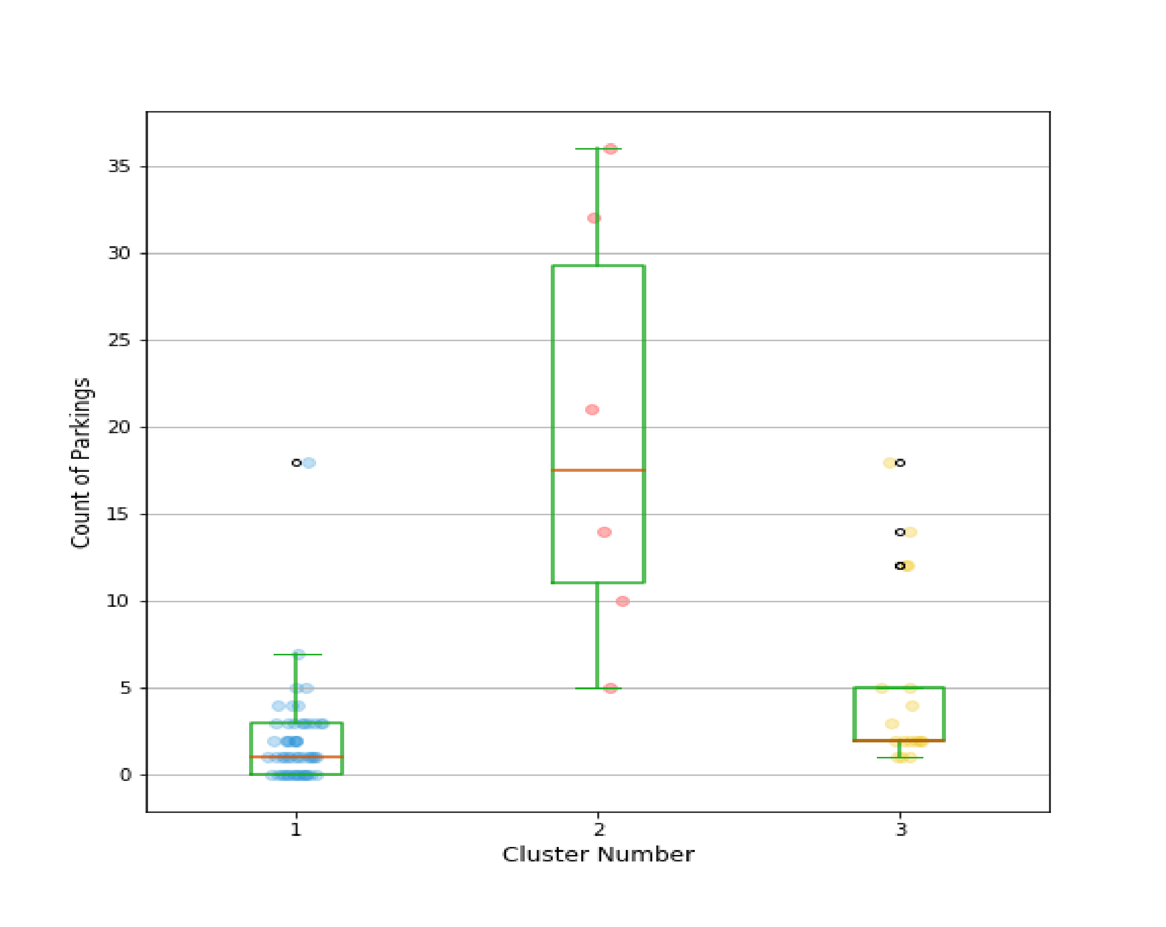
- Cluster 2 consists of areas with highest number of restaurants, hence greater competition while cluster 3 falls below somewhere between cluster 1 and 3 in terms of competition between restaurants in the community areas.



Result – Box Plots

Count of Car Parkings

- The box plots summarize distribution of community areas, based on number of car parkings, within each cluster and across the three clusters.
- Community areas in cluster 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



Result - Summary

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

Recommendations

Cluster 1

- Community areas in cluster 1 are more suitable for restaurants with cheaper menu due to lower per capita income of the community areas.

- Significantly higher prices will deter customers due to low Per Capita Income in this cluster.

- 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.

Recommendations

Cluster 1

- 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. Cost associated with renting or building new car parking.

- The most appropriate strategy here will be to differentiate in terms of cost rather than quality.

Recommendations

Cluster 2

- Cluster with highest score on each feature.

- Highest Per Capita Income in this cluster means customers in these areas can comfortably afford higher priced menu.

- Highest number of restaurants in these areas mean highest competition compared to cluster 1 and 3.
- There are greater chances for higher level of demand and growth given the highest Population Density per Square Mile in these areas.

Recommendations

Cluster 2

- 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. This may be explained by the wealthiness of the region and higher population density.

- 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.

- The appropriate strategy here will be differentiation in terms of quality rather than price. Higher prices will signify status quo matching the needs of customers in these community areas.

Recommendations

Cluster 3

- 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. So, the level of demand and growth will depend on where a restaurant is located in cluster 3.

Recommendations

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.

Conclusion

- The objective of the report was to segment and cluster community areas in Chicago based on five features i.e. Population Density Per Square Mile, Per Capita Income, Count of Crimes, Count of Restaurants and Count of Parkings.
- The data was obtained and processed to determine the relevant features of each community area.
- Haversine formula was used to assign a venue to a nearest community area.
- More accurate measures of distance 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.
- The data was standardized to remove the effects of difference in unit of measurement and size.

Conclusion

- K-Means unsupervised machine learning algorithm was used to cluster the unlabeled data that reduce the squared distance of data points from a cluster mean.
- Optimal number of clusters were determined using inertia graph of the algorithm.
- This distribution of community areas within each cluster was visualized using Chicago's map.
- The distribution of community areas within each cluster and comparison of clusters was made using box plots.
- Cluster 1 generally had lower population density, per capita income, count of restaurants, count of car parkings and average count of crimes. The appropriate strategy here would be to differentiate on price rather than quality and status quo.

Conclusion

- Cluster 2, on average, had scored highest on each feature. The appropriate strategy here would be to differentiate on quality of restaurant by charging higher price. Higher price would signify status of customers.
- Cluster 3 had generally average score on each features compared to the other two cluster except count of crime that was lowest in this cluster. The appropriate strategy here would be to offer quality restaurant service with mid price range menu.
- The above strategies are based on broad characteristics of the clusters.
- Further insight may be gained by performing additional clustering algorithm on each cluster to identify community areas with most attractive features.
- Rating data of restaurants may be obtained to gain further insight at the level of competition but this data may come with cost.

Conclusion

- Crime data may be refined to include only relevant crimes.
- Latest data on the features may be available that would enhance the quality of data and clustering.
- Further on the ground research would be needed make more informed decision about opening a restaurant, for example, to check preference and taste of consumers, availability of area and other inputs, etc.

Thank You!