Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/12376
DOI URL: http://dx.doi.org/10.21474/IJAR01/12376

ISSN NO. 2320-5407

*RESEARCH ARTICLE*

# CLUSTERING ANALYSIS FOR RESIDENTIAL AREAS BASED ON NEIGHBORHOOD AMENITIES

**Karan Bhowmick**

## *Manuscript Info*

***Manuscript History***
Received: 25 November 2020
Final Accepted: 28 December 2020
Published: January 2021

## *Abstract*

The use of urban land in cities can be improved and the poor execution of Urban planning is related to the problem of housing. The problem of housing has become acute because of the tremendous increase of urban population and unplanned growth of the cities. Mumbai has a population of 20,411,000 thus it is the target of our analysis project. Affordable housing in Mumbai has become an unfathomable challenge, it one of the most complex probes in this city. About 42% of Mumbai's housing comprises slums. With this in mind, our aim is to help the decision of buying houses, by recommending localities with basic amenities. We hope to make the process of scrutinizing residential buildings more streamlined. We also hope to underscore areas with housing potential in this study. We use K-Means Clustering to cluster the different neighborhoods of Mumbai, based on the availability of 31 amenities in the neighborhood. We have used Data from Wikipedia to get the list of neighborhoods in Mumbai, and we use Foursquare API to get a list of amenities in each area of the neighborhood. We then evaluate the model using silhouette score and plot a graph using folium to show the different clusters on the map of Mumbai.



## Introduction:-

**Objectives:**

To cluster the neighborhoods in Mumbai according to the prospect of growth.
Visualize the clusters on the map of Mumbai to streamline decision-making related to housing

**Research Background:**

In this section we discuss the variables we used to conduct clustering and the method with which we procured data.

**Understanding the Variables:**

31 amenities have been selected for analysis, and the measure of each is the division of the number of instances of that amenity by the total number of amenities to discern the level of availability.

**The amenities selected are:**

Airport, Scenic Lookout, Shopping Mall, Supermarket, Convenience Store, Department Store, Electronics Store, Pharmacy, Train Station, Bus Station, Chinese Restaurant, Fast Food Restaurant, Pizza Place, Indian Restaurant, Asian Restaurant, Café, Coffee Shop, Bar, Wine Shop, Bank, ATM, College Academic Building, Beach, Monument / Landmark, Golf Course, Park, Multiplex, Movie Theater, Stadium, Club House, Gym.

**Corresponding Author:- Karan Bhowmick**

**Data Collection:**
We have used Wikipedia to get the names of the 136 neighborhoods of Mumbai. BeautifulSoup (a library in Python) for scraping the Wikipedia Page of the Neighborhoods in Mumbai. Subsequently, we assigned latitude and longitude values to each neighborhood using geocoder (a library in Python).

Then, we proceed with Foursquare API to get all of the amenities available in these neighborhoods. We then use feature selection to trim the amenities based on a study by Strutt & Parker (a real estate agency).

## Research Methodology:-

In this section, we describe the procedure we followed for the clustering of Residential Areas in neighborhoods. We also delve into feature selection after assessing the quality of variables. We discuss K-Means Clustering and the evaluation metrics we used to evaluate our model.

**Data Preparation and Feature Selection:**
The dataset has neighborhood names with the latitude and longitude coordinates, along with the values of the amenities/total amenities. We then use a ranking system inspired by a study from Strutt & Parker real estate agency, to create a priority of variables. The ranking was based on the effect of these parameters on the cost of a Residential Building.Thus, the houses are in accordance with the findings, Airport and Scenic outlook make a neighborhood prone to development.
For other factors, the order of precedence is as follows:
1) Shopping Stores
2) Transport Hub
3) Restaurants
4) Green Spaces
5) Sports and Recreation

| Neighborhood | Airport | Scenic Lookout | Shopping Mall | Supermarket | Convenience Store | Department Store | Electronics Store | Pharmacy | Train Station |
|---|---|---|---|---|---|---|---|---|---|
| Khar Danda | 0.028571 | 0.000000 | 0.000000 | 0.000000 | 0.028571 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| Land's End, Bandra | 0.000000 | 0.083333 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| Powai | 0.000000 | 0.024390 | 0.000000 | 0.000000 | 0.000000 | 0.024390 | 0.000000 | 0.000000 | 0.000000 |
| Gokuldham | 0.000000 | 0.021739 | 0.043478 | 0.000000 | 0.021739 | 0.000000 | 0.021739 | 0.000000 | 0.000000 |
| Marine Lines | 0.000000 | 0.021739 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.043478 |
| Nariman Point | 0.000000 | 0.020833 | 0.020833 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |

**Figure 1:-** Aggregated values of amenities/total amenities to show availability ratio.

We decide on 8 categories to create a prioritized scoring system for all neighborhoods. We multiply the variables of each column depending on the priority, from a range of 1 to 8. We multiply 8 with the variables that affect prices the most and so on.

**Amenities with magnitude 8 in terms of availability:**
Airport, Scenic lookout.

**Amenities with magnitude 7 in terms of availability:**
Shopping Malls, Supermarket, Convenience Store, Department Store, Electronics Store.

**Amenities with magnitude 6 in terms of availability:**
Train Station, Bus station.

**Amenities with magnitude 5 in terms of availability:**
Chinese Restaurant, Fast Food Restaurant, Pizza Place, Indian Restaurant, Asian Restaurant.

**Amenities with magnitude 4 in terms of availability:**
Cafe, Coffee Shop, Bar, Wine Shop.

**Amenities with magnitude 3 in terms of availability:**
Bank, ATM, College Academic Building.

**Amenities with magnitude 2 in terms of availability:**
Beach, Golf Course, Park, Club House.

**Amenities with magnitude 1 in terms of availability:**
Monument, Multiplex, Stadium, Movie Theatre, Gymnasium.

**This is done to standardize ametric that can be used to cluster the neighborhoods, it is named 'Lifestyle Score':**

| | Lifestyle Score |
|---|---|
| 0 | 1.114286 |
| 1 | 2.291667 |
| 2 | 1.731707 |
| 3 | 2.500000 |
| 4 | 1.913043 |

**Figure 2:-** Lifestyle Score.

**Model Development:**
We develop a model using the unsupervised machine learning algorithm, K-Means Clustering. Unsupervised algorithms, unlike supervised learning algorithms, provide unlabeled data and form patterns and features of their own to make sense of the model. Clustering Analysis can be defined as a classification of data points to subgroups/clusters. The data points in each subgroup are similar, while the data points in other subgroups/clusters are different. The metrics used to discern similarity can be based on either correlation or euclidean distance between the data points.

We use K-Means Clustering algorithm, which is an iterative algorithm that divides data points into k distinct clusters, while making the intra-cluster data points as similar as possible while keeping the inter-cluster data points as different as possible. It works to minimize the distance of each data point in a cluster to the centroid of that cluster, the lesser the variation in a cluster the more homogeneous it is. It achieves this by calculating the sum of the squared distance between the data points and the cluster's centroid.

**The approach that K-Means follows is Expectation-Maximization method. The objective function used to get the best model is:**
**Equation 1:**-Expectation-Maximization

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} ||x^i - \mu_k||^2$$

Here, **wik** =1 if xi belongs to cluster k, **xi** is the data point, **μk** is the centroid of the cluster.

Here, the E-method is achieved by assigning data points to a particular cluster. This is done by differentiating J wrt wik and updating the cluster assignments made.

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m} \sum_{k=1}^{K} \|x^i - \mu_k\|^2$$

$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_j \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

**Figure 3:-** The E- Method.

Now, the M-method is achieved by computing the centroid of each cluster. Now we differentiate J wrt μk and recomputing the values of centroids after the E-step.

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m} \sum_{k=1}^{K} \|x^i - \mu_k\|^2$$

$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_j \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

**Figure 4:-** The M-method.

We assign k=5 for our experiment, we use Lifestyle Score as the discerning factor for cluster assignment.

**Cluster 1:**
We can see here that Cluster 1 groups regions with an okay LifeStyle Score in the range: 0.857 - 1.800.

| | Neighborhood | Lifestyle Score | Cluster |
|---|---|---|---|
| 88 | Dhobitalao | 1.800000 | 0 |
| 89 | Four Bungalows | 1.790323 | 0 |
| 90 | Walkeshwar | 1.769231 | 0 |
| 91 | D.N. Nagar | 1.760870 | 0 |
| 92 | Breach Candy | 1.758065 | 0 |
| 93 | Powai | 1.731707 | 0 |
| 94 | Versova, Mumbai | 1.727273 | 0 |
| 95 | Badhwar Park | 1.710000 | 0 |
| 96 | SEEPZ | 1.687500 | 0 |
| 97 | Ferry Wharf | 1.666667 | 0 |

**Figure 5:-** Cluster 1 Lifestyle Score.

**Cluster 2:**
We can see here that Cluster 1 groups regions with an average Lifestyle Score. Range : 1.880 - 2.359

| | Neighborhood | Lifestyle Score | Cluster |
|---|---|---|---|
| 37 | Kajuwadi | 2.358974 | 1 |
| 38 | Thakurdwar | 2.346939 | 1 |
| 39 | Mahalaxmi, Mumbai | 2.344828 | 1 |
| 40 | Tardeo | 2.344828 | 1 |
| 41 | Uttan | 2.333333 | 1 |
| 42 | Land's End, Bandra | 2.291667 | 1 |
| 43 | Mira Road | 2.272727 | 1 |
| 44 | Khotachiwadi | 2.271186 | 1 |
| 45 | Churchgate | 2.270000 | 1 |
| 46 | Chira Bazaar | 2.259259 | 1 |

**Figure 6:-** Cluster 2 Lifestyle Score.

**Cluster 3:**
We can see here that Cluster 2 groups regions with poor Lifestyle Score.
Range : 0.000 - 0.500.

| | Neighborhood | Lifestyle Score | Cluster |
|---|---|---|---|
| 118 | Madh Island | 0.5 | 2 |
| 119 | Malabar Hill | 0.0 | 2 |
| 120 | Guru Tegh Bahadur Nagar | 0.0 | 2 |
| 121 | Dadar | 0.0 | 2 |
| 122 | Fanas Wadi | 0.0 | 2 |
| 123 | Gorai | 0.0 | 2 |
| 124 | Jagruti Nagar | 0.0 | 2 |
| 125 | Kannamwar Nagar | 0.0 | 2 |
| 126 | Koliwada | 0.0 | 2 |

**Figure 7:-** Lifestyle score of Cluster 3.

**Cluster 4:**
We can see here that Cluster 3 groups regions with a good Lifestyle Score.
Range : 2.405 - 3.048

| | Neighborhood | Lifestyle Score | Cluster |
|---|---|---|---|
| 4 | Kamathipura | 3.047619 | 3 |
| 5 | Dongri | 3.000000 | 3 |
| 6 | Dedh galli | 2.944444 | 3 |
| 7 | Uran | 2.909091 | 3 |
| 8 | Yashodham | 2.869565 | 3 |
| 9 | Charni Road | 2.844828 | 3 |
| 10 | Princess Street (Mumbai) | 2.822581 | 3 |
| 11 | Matunga Road, Mumbai | 2.804124 | 3 |
| 12 | Virar | 2.772727 | 3 |
| 13 | I.C. Colony | 2.764706 | 3 |

**Figure 8:-** Lifestyle score of Cluster 4

**Cluster 5:**
We can see here that Cluster 4 groups regions with a good Lifestyle Score.
Range : 3.500 - 3.818.

| | Neighborhood | Lifestyle Score | Cluster |
|---|---|---|---|
| 0 | Anushakti Nagar | 3.818182 | 4 |
| 1 | Aarey Forest | 3.750000 | 4 |
| 2 | Thakkar Bappa Colony | 3.560000 | 4 |
| 3 | Pali Village | 3.500000 | 4 |

**Figure 9:-** Lifestyle score of Cluster 5

We then proceed to evaluate our model.

**Model Evaluation:**
We use the silhouette coefficient to evaluate our clustering model. Silhouette Coefficient's value ranges from -1 to +1. If the coefficient is closer to +1, it signifies that mean clusters are well apart from each other and clearly distinguished. If it is closer to -1, mean clusters overlap and have not been assigned properly. Finally, if it is closer to zero, it means the distance between the clusters is not significant.

Silhouette Coefficient is used to analyze the coherency of data points in the clusters.

**Equation 2:**
Silhouette Score
Silhouette Score = (b-a)/max(a,b)

Here, **b** is the average inter-cluster distance and **a** is the average intra-cluster distance or the distance between data points in a cluster.

For our model the Silhouette coefficient is **0.58354.**

## Results and Discussion:-

**Neighborhoods in South Mumbai:**

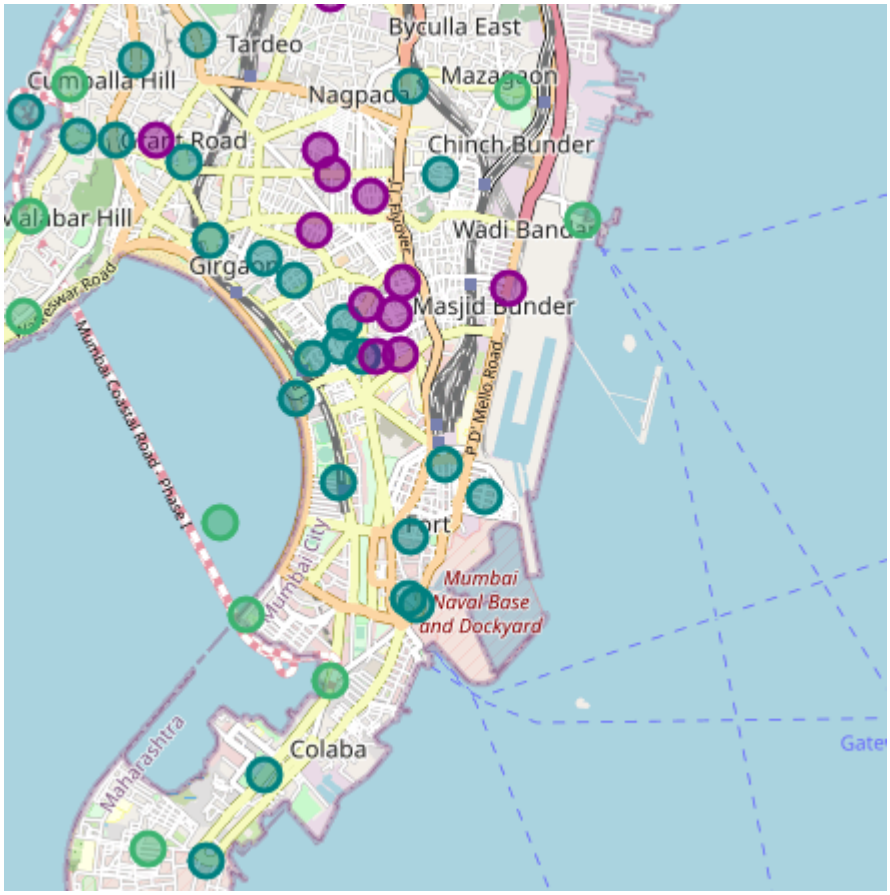


**Figure 10:-** Clusters formed in South Mumbai.

**Neighborhoods in Central Mumbai:**

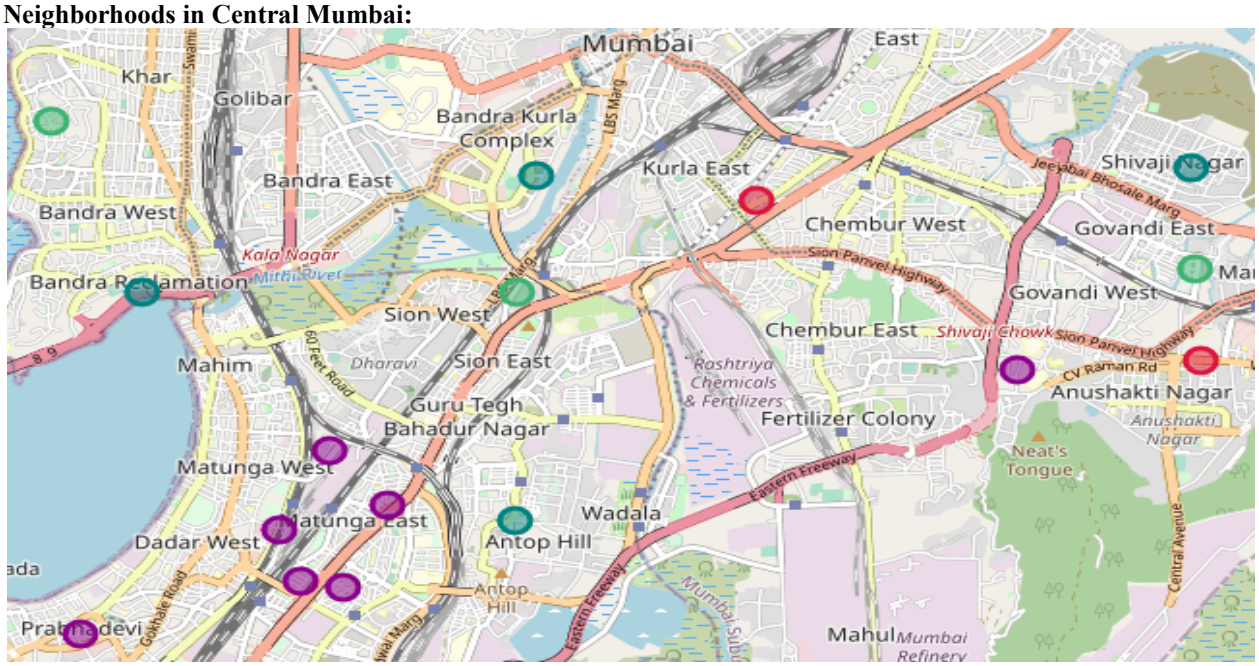


**Figure 11:-** Clusters formed in Central Mumbai.

**Neighborhoods in Upper-Central Mumbai:**

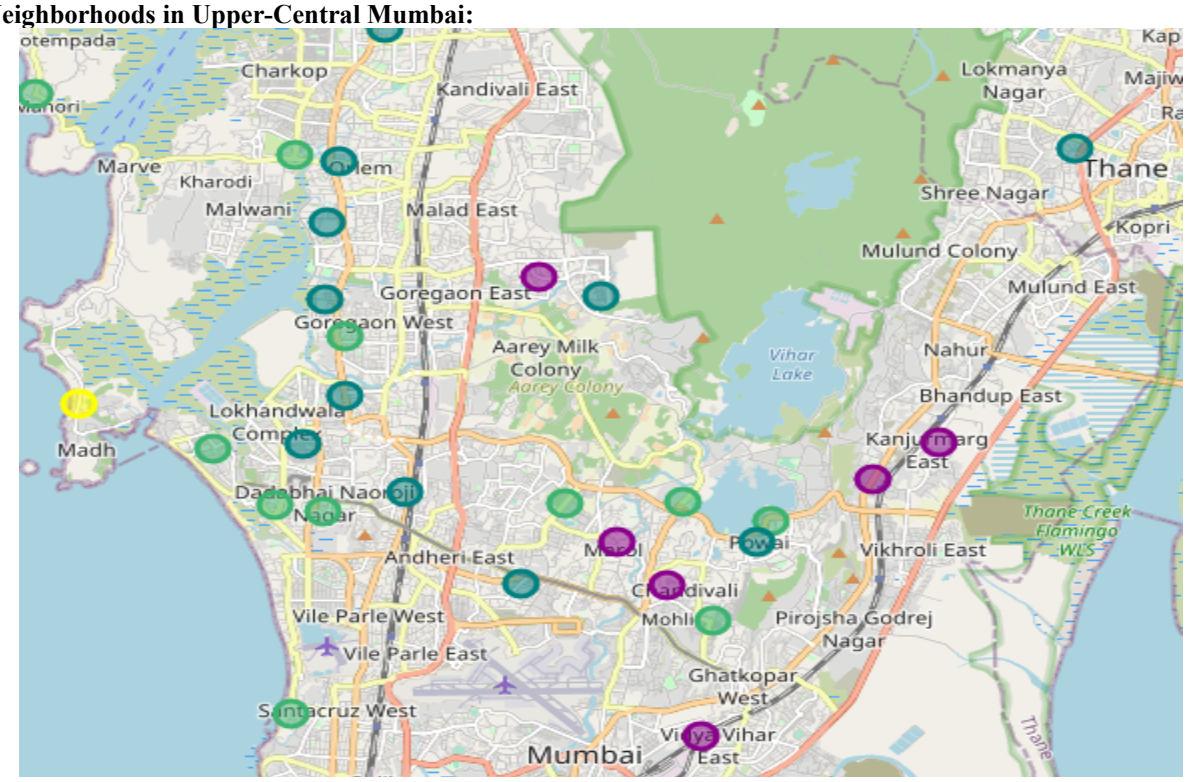


**Figure 12:-** Clusters formed in upper-central Mumbai.

**Neighborhood Clusters in North Mumbai:**

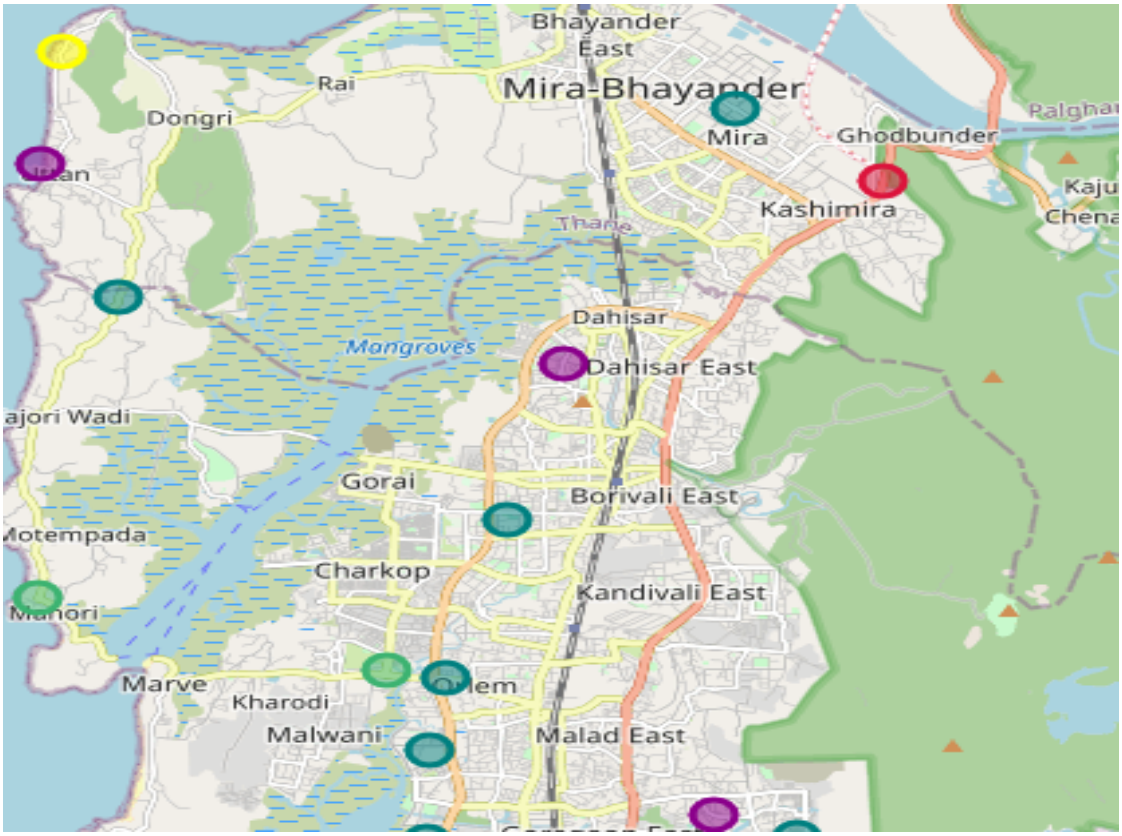


**Figure 13:-** Clusters formed in North Mumbai.

**Neighborhood Clustering all over Mumbai:**

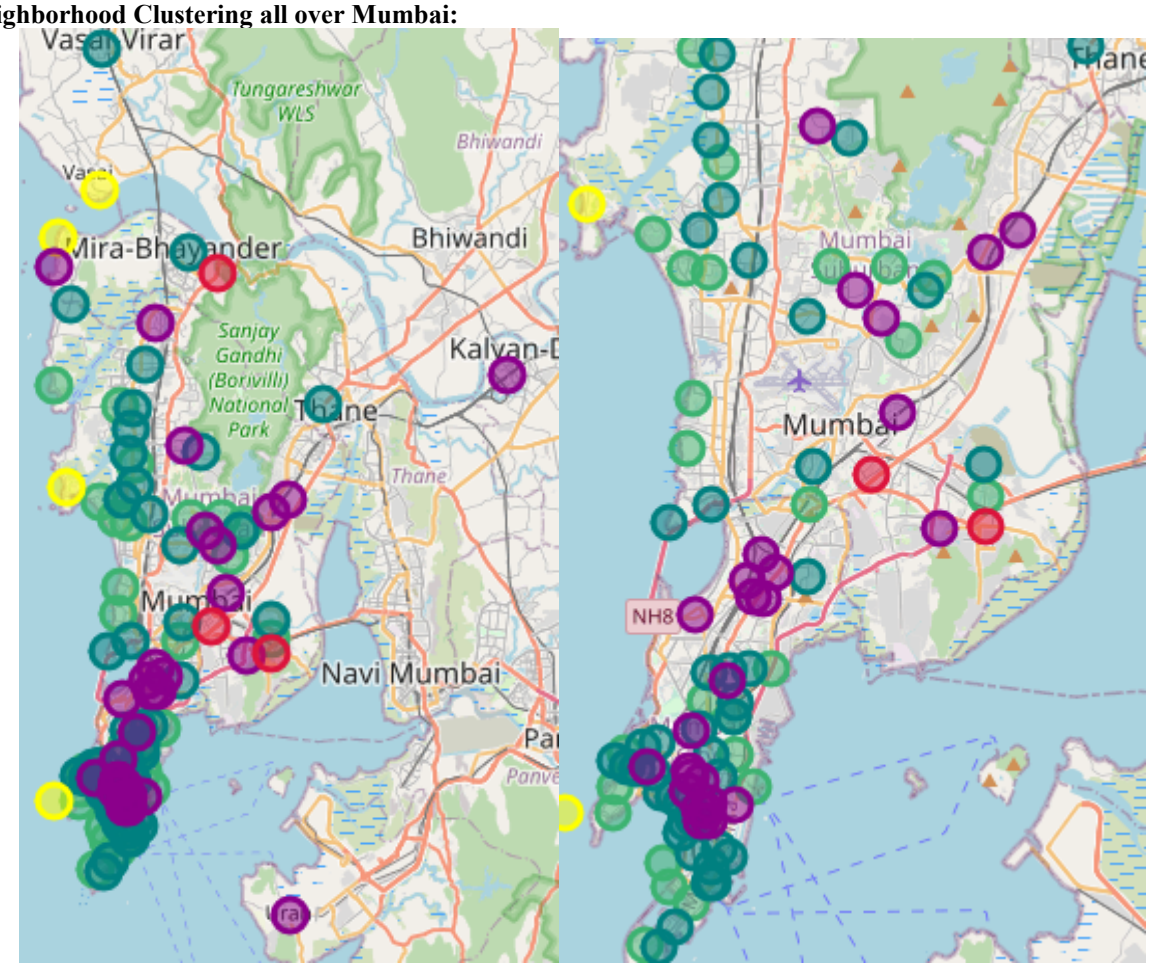


**Figure 14:-** Overview of cluster concentration throughout Mumba**i.**

**Color Index:**
Cluster 1 (poor): medium sea green
Cluster 2 (average): teal
Cluster 3 (the worst): yellow
Cluster 4 (good): dark magenta
Cluster 5 (the best): crimson
From these visualizations, we can see that Central Mumbai has some of the best neighborhoods based on availability of amenities, followed by South Mumbai, North Mumbai and Upper-Central Mumbai.

## Conclusion:-

The neighborhoods that are magenta or red show great prospect for growth and have a multitude of amenities in their vicinity. These neighborhoods should be the top picks for anyone looking to invest in property or buying a house. Thus, we have successfully concluded our clustering and helped discern the 136 neighborhoods of Mumbai.

## References:-

1. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
2. https://www.struttandparker.com/knowledge-and-research/which-local-amenities-add-most-value.