Niket V Patil

12/02/2021

Finding similar neighbourhoods in three indian cities

# Introduction

Identifying & Clustering regions in above three cities to find areas which share similarity based on number of popular venues (Types of Restaurants, Stores, Tourist Destinations etc.)

## Background

* In last decade, cities in India have developed at astonishing rate. With developing cities House pricing also goes up. If a resident in one city wants to relocate to different city or neighborhood, which city/neighborhood should he choose that is most similar to his current neighborhood. This project will try to answer this question. We will base our similarity on a kind of Blueprint for neighborhoods based on popular venues such as Restaurants, Malls, Natural Areas, Tourist Destinations etc.
* Audience of this project might include Home Buyers or Renters, Real Estate Agents. Also, this project might be helpful to those who are looking to open new Businesses to find appropriate locations to open their news business.

# DatA

To complete this project, we will require data including names of neighbourhoods in Mumbai, Chennai & Bengluru. With popular venues for each neighbourhood, along with location data which should include Latitude & Longitude.

To get neighbourhoods names we'll have to scrape Wikipedia pages. To find venues we'll use Foursquare API (Venue Recommendation), & for Location data we'll use Mapquest API for cities where location data is not available from Wikipedia Page.

# Preparation

After getting names of Areas from Wikipedia I created a Dataframe. For two of the three cities location data was provided on Wikipedia page. So after scraping for Mumbai & Chennai from Wikipedia I have got neighbourhood names as well as latitude and longitude values. For Bengluru Geocoding needed to done. With help of Mapquest API that was achieved.

Next step was to use prepared dataset to run through Foursquare API to query venues in each city.

# Methodology

Unsupervised clustering method called k-means clustering was used for this study to cluster our observations into six different clusters. Correct number for clusters was derived by using Elbow method.

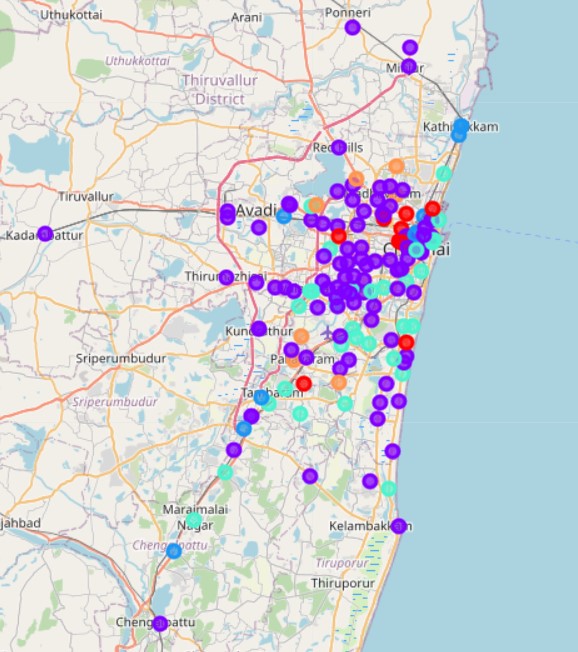
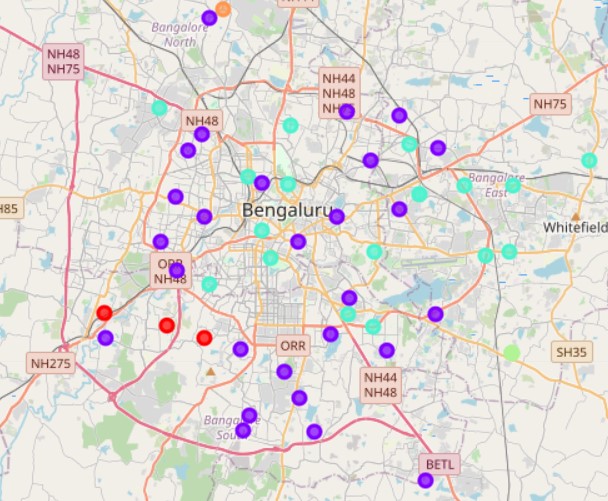
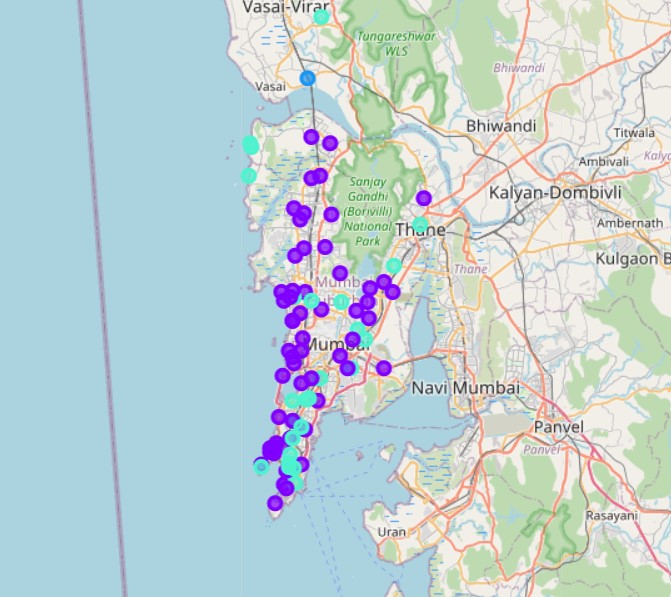
Top five venue categories were created based on frequency of occurrence in dataset. Using One-hot encoding each venue was converted to yes/no column. Then grouping rows by area mean of frequency for each venue category. By this way five most common venues were found for each neighbourhood.

# Result & Discussion

To visualize clusters and render the clusters folium library was used. Clusters are mapped onto the map of respective cities. There are two large clusters & four smaller ones.

# figures

Three cities Mumbai, Bengluru & Chennai.



# Clusters

* Cluster 1: Shopping Area Cluster, General Store dominate this cluster. So these are maybe highly residential/shopping areas.
* Cluster 2: Developed Urban Cluster, this is our biggest cluster with lots of different venues. Most likely large urban centres with lots of Pubs, Cafes, Hotels & Restaurants.
* Cluster 3: Train Cluster, this seems Peaceful places with Art galleries, Yoga studios. Food places are not as prevalent here as some other clusters.
* Cluster 4: Restaurant (Indian) Cluster is the second big cluster along with cluster 2. They both hold almost all values among them. In addition to Indian Restaurants being the prevalent venue type, the cities in this cluster are characterized by a diverse set of amenities, indicative of larger urban centres.
* Cluster 5: Italian Restaurant (Outlier) Cluster 5, In this cluster either places are very close by and small, or we might have made some mistake. If not, then we can say Italian Restaurants define this cluster along with Yoga Studio.
* Cluster 6: Old/Small Areas Cluster 6: is fourth small cluster. Where most common venue is ATM. Flower Shops seem abundant in this cluster. And as with other clusters Food venues are famous here too.

# Conclusion

Neighbourhoods are three cities were scraped from Wikipedia & using Foresquare API we got venues for those neighbourhoods. We have total 333 Areas from three cities & for them total 2682 venues. We used K-means algorithm to group six different clustering groups. Out of which two clusters had most values.

With such analysis new home buyers, renters, real estate agents & entrepreneur could find places in new cities which are similar. Though due to time constraints we could not analyse and work on data in depth. More through study could help find answers we were trying to solve.

Throughout the process of this study we uncovered limitations in comprehensively addressing the business problem at hand. Nevertheless, we did find some interesting patterns among our refined dataset of three cities. Next steps in the process might be to supplement the data used to cluster cities with additional sources, such as the average home price, population size & crime rate.