Similarities among Neighbourhoods in Bangalore and Mumbai

Applied Data Science Capstone by IBM on Coursera

Likhit J Jain

March 11, 2020

1. Introduction

1.1 Background

Bangalore is the capital of the Indian state of Karnataka. It has a population of over ten million, making it a megacity and the third-most populous city and fifth-most populous urban agglomeration in India. Bangalore is sometimes referred to as the "Silicon Valley of India" (or "IT capital of India") because of its role as the nation's leading information technology (IT) exporter. A demographically diverse city, Bangalore is the second fastest-growing major metropolis in India.

Mumbai, also known as Bombay is the capital city of the Indian state of Maharashtra. According to UN, as of 2018, Mumbai was the second most populous city in India and the seventh most populous city in the world with a population of 19.98 million. As per Indian government population census of 2011, Mumbai was the most populous city in India with an estimated city proper population of 12.5 million. Mumbai is the financial, commercial and entertainment capital of India. It is also one of the world's top ten centres of commerce in terms of global financial flow, generating 6.16% of India's GDP and accounting for 25% of industrial output, 70% of maritime trade in India and 70% of capital transactions to India's economy. Mumbai's business opportunities, as well as its potential to offer a higher standard of living, attract migrants from all over India, making the city a melting pot of many communities and cultures.

1.2 Problem

Now, let us assume a person wants to move from Bangalore to Mumbai for whatever reason. This person has absolutely no idea about Mumbai. But this person wants to move to a locality in Mumbai that is very similar to that of Bangalore. My proposal is to implement such a system where similar neighbourhoods in Bangalore are mapped to similar neighbourhoods in Mumbai so that the person who is moving has a better idea of the neighbourhoods of Mumbai. The person would not have to make a blind decision as to where to purchase/rent an apartment.

1.3 Interest

Obviously, anyone who is looking to move from Bangalore to Mumbai or vice versa would be interested in knowing which neighbourhoods are similar between the two cities. This will help the individual make informed choices.

2. Data Acquisition and Cleaning

2.1 Data Sources

The list of neighbourhoods was scraped from the web. A list of neighbourhoods in Bangalore can be found here and a list of neighbourhoods in Mumbai can be found here. The geographical coordinates of Bangalore and Mumbai and their respective neighbourhoods were fetched using the Python Geocoder Package. The location data of each of the neighbourhood were fetched using the FourSquare API.

2.2 Data Cleaning

Web Scraping was done using the BeautifulSoap package in Python. Data scraped from these sources were merged into one Pandas data frame. There was not much cleaning required other than correcting the spelling of some of the neighbourhoods and removing redundant data. After cleaning the data there were 95 rows and 3 columns i.e. neighbourhood, latitude and longitude.

3. Methodology

3.1 Feature Extraction

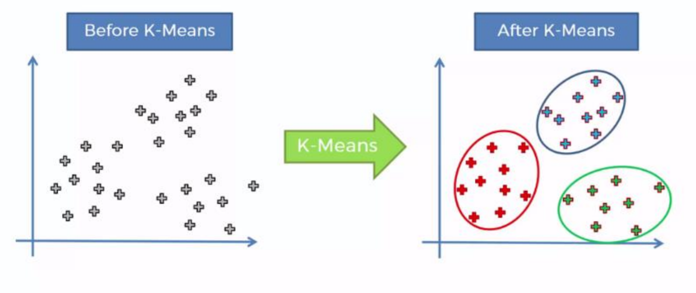
For feature extraction One Hot Encoding is used in terms of categories. Therefore, each feature is a category that belongs to a venue. Each feature becomes binary, this means that 1 means this category is found in the venue and 0 means the opposite. Then, all the venues are grouped by the neighbourhoods, computing at the same time the mean. This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category.

3.2 Unsupervised Learning

For the purpose of doing unsupervised learning to found similarities between neighborhoods, a clustering algorithm is implemented. In this case K-Means is used due to its simplicity and its similiraty approach to found patterns.

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the multi-dimensional features.

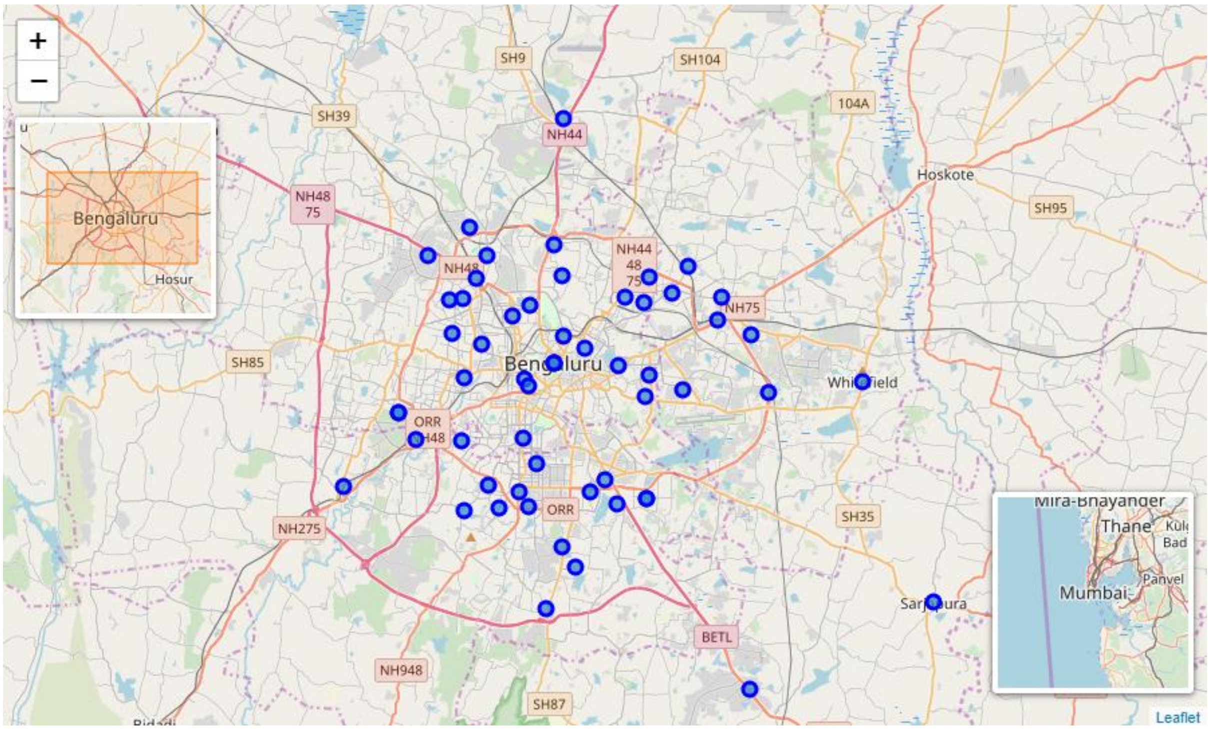
In the following figure there is a graphical example of how a K-Means algorithm works. As it is possible to see, dispersion is minimized by representing all clustered data into one group or cluster.

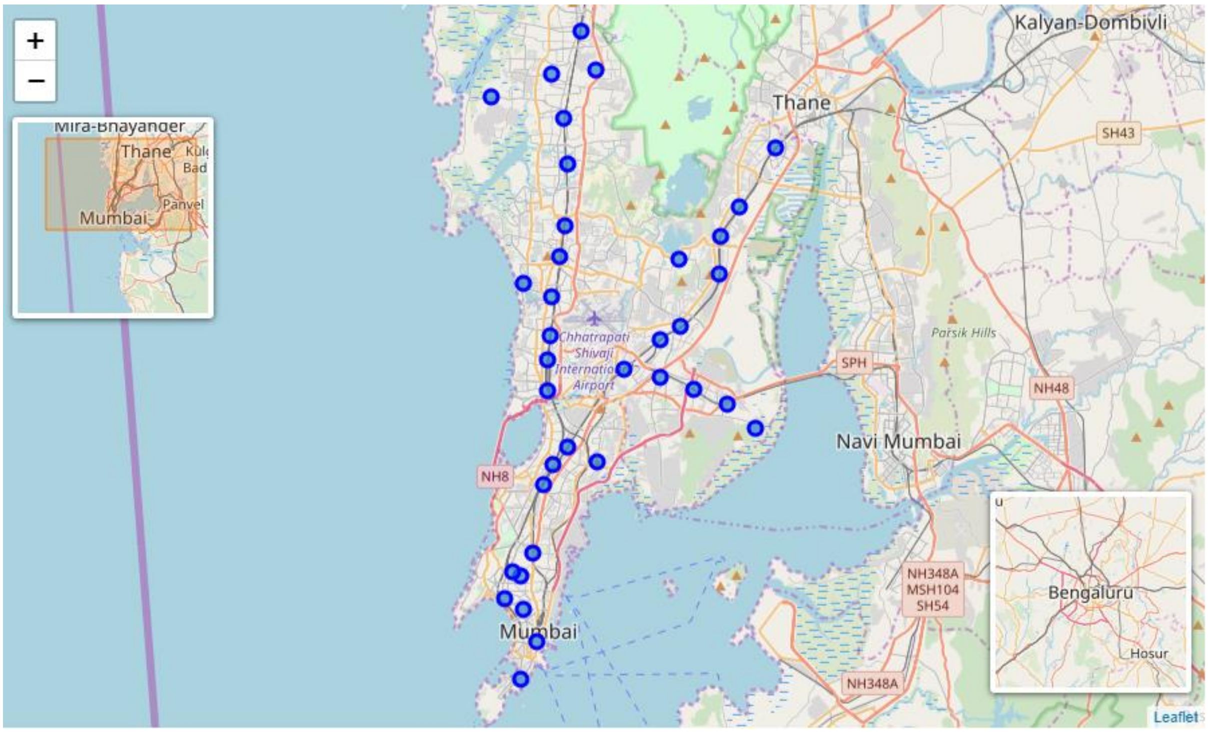


It is necessary for this algorithm to have a prior idea about the number of clusters since it is considered an input of this algorithm. For this reason, the elbow method is implemented. A chart that compares error vs number of clusters is done and the elbow is selected. Then, further analysis of each cluster is done.

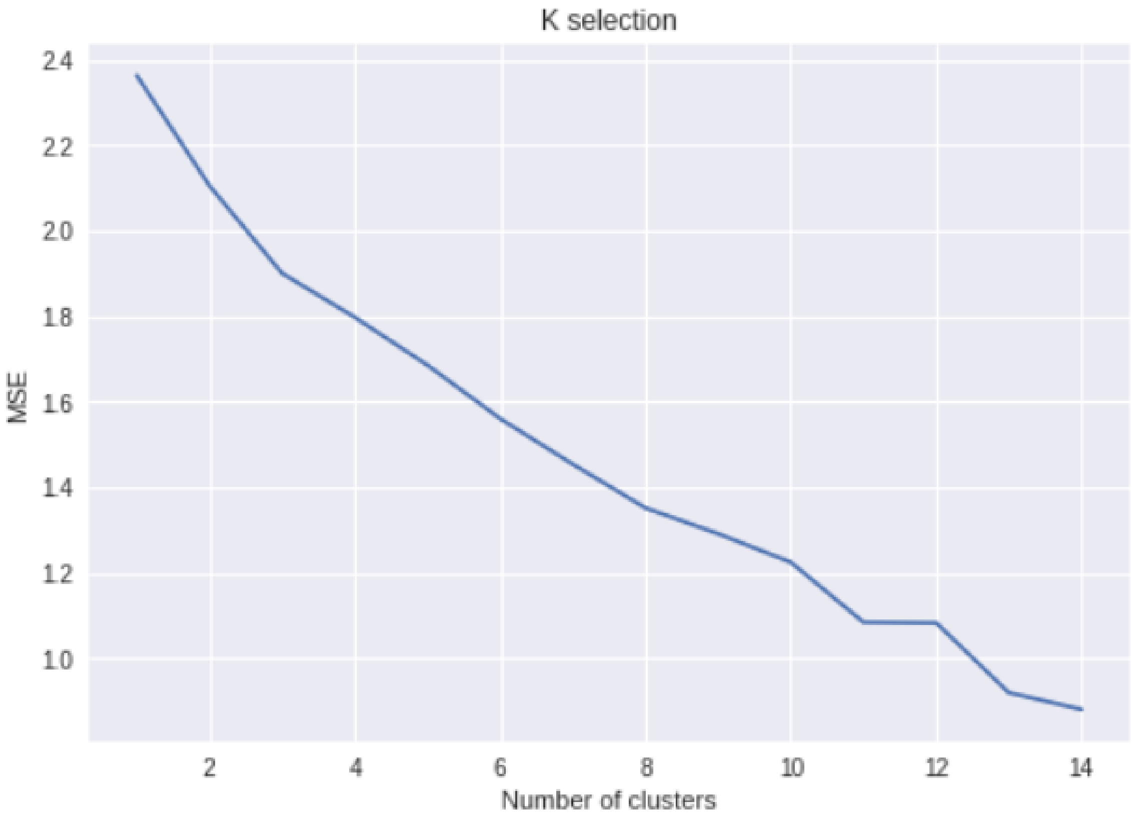
4. Results

Firstly, data is plotted in a geographical map to get a notion of the world location. In the two following images are shown the neighbourhoods in Bangalore and Mumbai respectively.



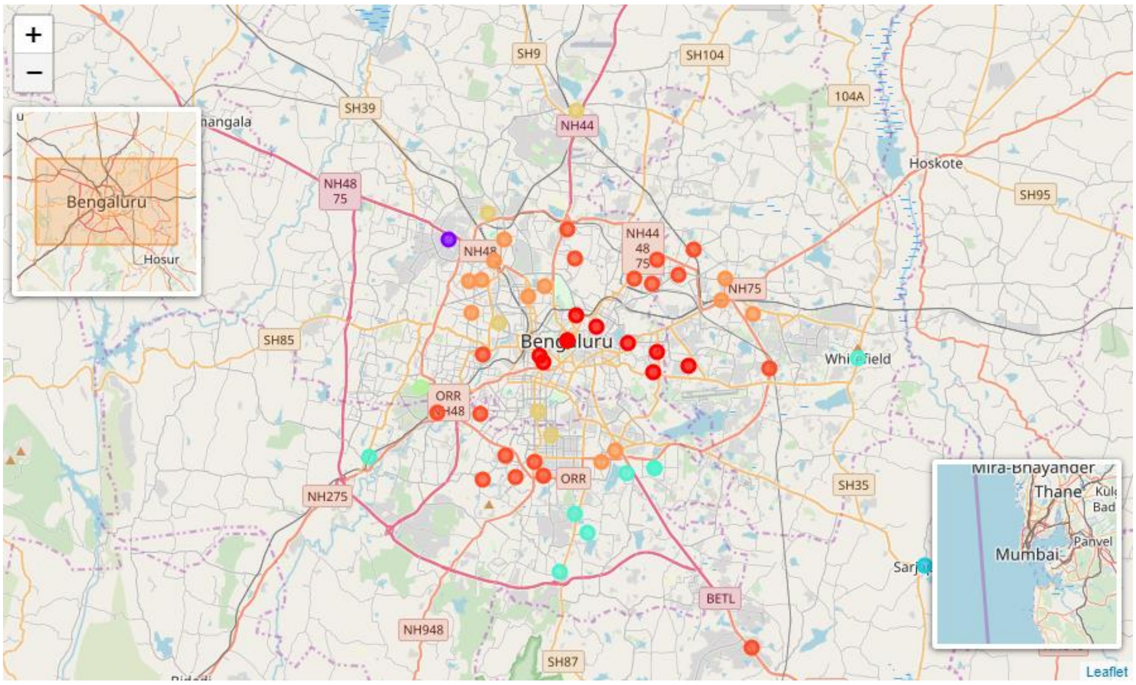


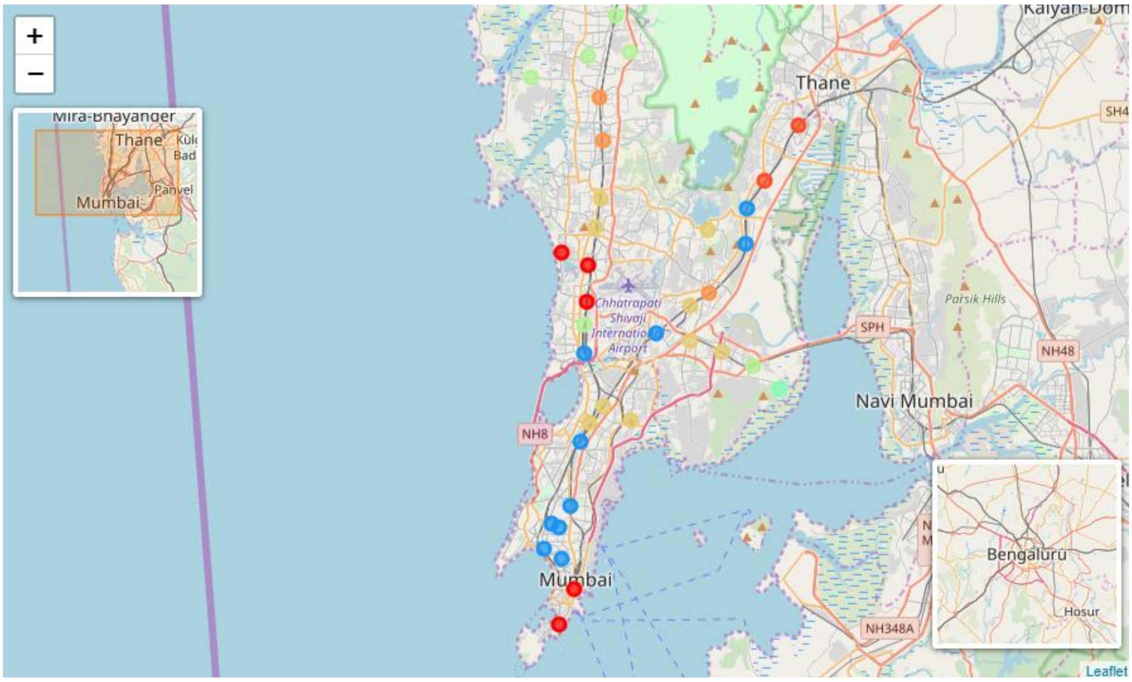
Secondly, the cluster algorithm is implemented. For this purpose, it is necessary to have a prior idea about the number of clusters. Therefore, the mean squared error (MSE) is plotted vs the number of clusters. The number of clusters start with a value of 1 increasing until a value of 15. This chart is shown in the image below.



As it is expected, the MSE decreases over the number of clusters. The elbow method here is implemented in order to select the appropriate number of groups. In this case, it is possible to see that the elbow is found more or less around 11. The MSE found below this number shows little changes rather than big ones. Finally, once the number of clusters is fixed, the clustering algorithm is repeated through samples and each neighbourhood is labelled according to the clusters found.

For visualization purposes, the geographical data is again plotted but with different colours. Each colour represents the cluster for which that neighbourhood belongs. This image is shown below.





In this image it is evident that cluster algorithm is not segmenting the neighbourhoods for location areas. This means that it is not true that geolocation of neighbourhoods is correlated with the categories of the venues around each neighbourhood. Yet, it is possible to see which neighbourhoods within Bangalore are more similar to the neighbourhoods within Mumbai. Those neighbourhoods that are similar among them belong to the same cluster. Hence, they have the same colour in the image above.

5. Discussion

It is worth to note that this work is useful only for those who live in Bangalore or in Mumbai. The reason is because there is a limited amount of data, we can request using the FourSquare API. Consequently, it will have a greater cost than the Lite version.

Moreover, there are clusters with just one neighbourhood. In the results we found out that this cluster has a frequency of 1 in garden places. This means the cluster is not segmenting correctly data and the centroid is located in the exact position of that neighbourhood. This neighbourhood has a high frequency of garden places around. Hence, we can say the algorithm is doing great since there is no other cluster with similar venues around.

6. Conclusion

In this work a segmentation between two different countries is done. This segmentation involves the neighbourhoods in Bangalore and the neighbourhoods in Mumbai. The data is downloaded and the venues around the neighbourhoods is acquired using the Foursquare API. One Hot Encoding is used for converting the categories of the venues into a feature matrix. Then, all venues are grouped by neighbourhoods and at the same time the mean is calculated. Hence, the resulting features used are the frequency of occurrence from each category in a neighbourhood.

The K-Means clustering algorithm is used for finding similarities between all the neighbourhoods listed in the feature matrix. The elbow method is used for selecting the appropriate number of clusters. Hence, the K selected is 11. Results show that there are 11 groups due of 11 clusters.

Finally, any user who wants to move from Bangalore to Mumbai and vice versa can use this system to get a notion or idea about what is the best suitable place for him.