Capstone Project – IBM Coursera

**Clustering Neighborhoods of Bangalore, IN based on Venue Information**

Jayadev K A – 2020

1. **Introduction**

India is a land of flavors, since independence the country has undergone a drastic transformation both in terms of Economic development and the very fabric of Indian lifestyle. With tremendous influence from the west, arising due to the rapid influx of technological jobs and the deep rooted family values the evolution has been remarkable. As the younger generation gets exposed to new trends, concepts such as Pubs, Café, Fast food, Mall’s etc. once believed to be only accessible on other countries are quickly becoming a reality and constitute a majority of the business in several metro cities. Bangalore, being the IT capital of India has numerous multinational companies employing thousands. This ultimately has led in the rise of Pubs, Café, Fast food, Mall’s etc. As these venue’s become more and more common there is an issue of selecting the right locality to maximize the sales. Most often than not Entrepreneurs are in a dilemma about the location to setup their new ventures either a Restaurant, Fast food or shop etc.

This study tries to collect information on various venues located in the city of Bangalore and assist new entrepreneurs in deciding on the location for their business ideas. Once, all the information has been collected based on the Foursquare location data for the various geo locations in Bangalore, the study classifies different locations as major Food, sports, shopping hub etc. The final result will help guide new business owners in selecting the neighborhood respectively based on the type of business i.e. Restaurants, Fast food etc.

1. **Data Description**

The study derives information from secondary data that was scrapped from sites mentioned below:

1. **Data Source:**
   1. Bangalore Location Data : <https://finkode.com/ka/bangalore.html>
   2. Bangalore Venue Data: <https://foursquare.com/>
2. **Data Pre-Processing:**

Locations data was derived based on the postal information from site “a”. Once the location information was scrapped using the “BeautifulSoup” library in Python, the “geopy” library was used to retrieve the latitude and longitudes of every location. As postal data comprises of location having the same geo-locations but different PIN codes, the data was cleaned to remove any duplicate geo-locations. It was also noticed that couple of stray geo-locations were populated, therefore using the quartile outlier method using the latitude metric these locations were successfully eliminated from the final list of neighborhoods. The final list contained 90 neighborhood data for which venue information would be collected.

Using the Foursquare location data, with radius set to 1000 meters and limit of venues as 100 Venue details for various locations were collected. The resulting dataset spanning 900 plus rows as shown in the sample table (Table 2.1).



Table 2.1 – Sample Data Set

The Foursquare resulted in 200 plus different venue category.

As we would be performing clustering, the above data set needs to be formatted into the right format that can be readily plugged into the clustering model to check for similar characteristics among neighborhoods. Firstly, neighborhoods with less than 10 venue details were excluded to maintain consistency. For the remaining neighborhoods, the top 10 venue category was identified and placed under the respective columns as shown below.

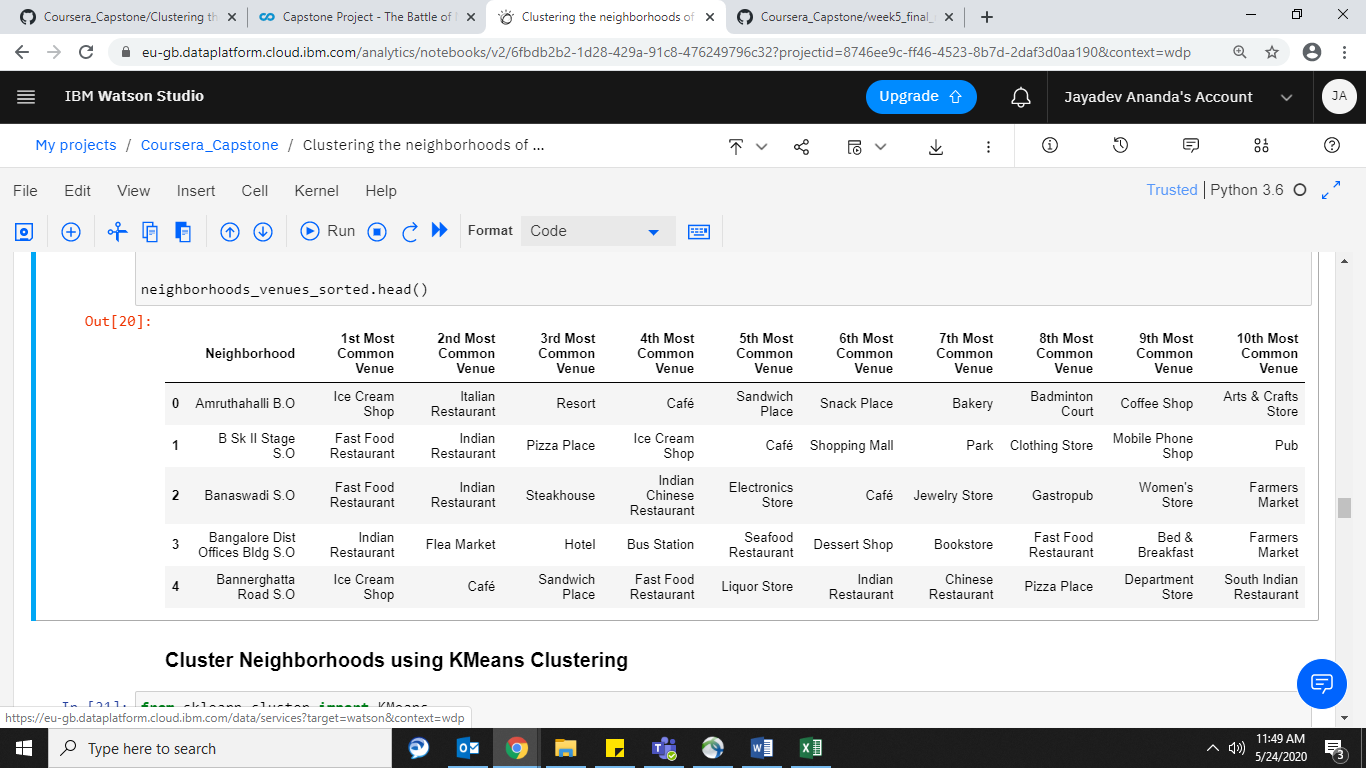


Table 2.2 – Transformed Data Set

1. **Methodology**

For clustering purposes, the K- Means clustering technique was selected. The K-Means clustering clusters the observations into K clusters based on the distance of each observation to the centroid. The major principle behind the algorithm is large distance of separation between observations in inter-clusters and small distance of separation between observations in intra-clusters. This makes K-Means relatively simple and suited to solve this problem.

To identify the optimal K value, we would use the elbow method which measures the sum of squared distances within the clusters. More the clusters the sum of squared distances approaches to Zero. From the Elbow graph, we find that the major change in squared distances occur with two clusters. Therefore, for our analysis only 2 cluster model will be selected.

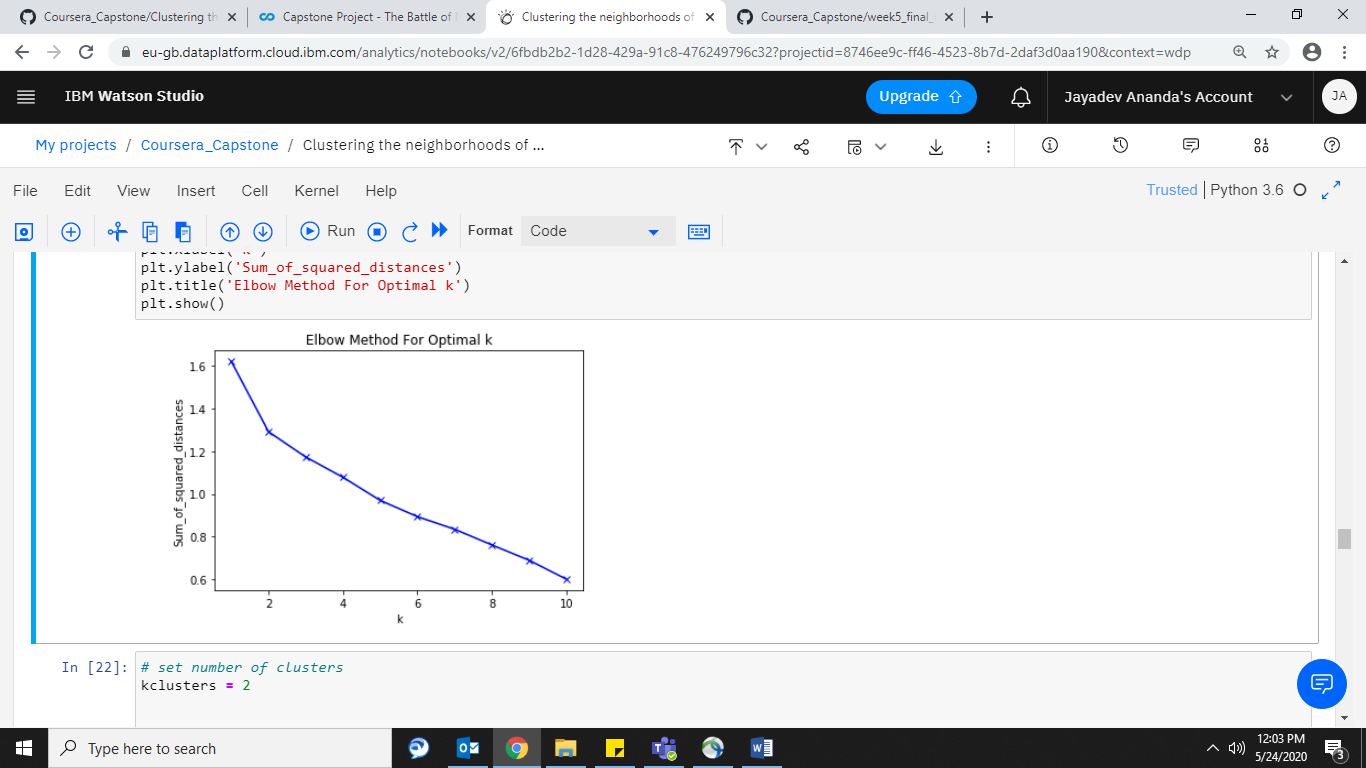


Figure 3.1: Elbow diagram.

Using 2 cluster for our model, we will be able to identify which neighborhoods will be suitable or provide favorable conditions for the right business.

1. **Results**

The results from the K-Means cluster with K selected as 2, yields the following observations:

1. **Cluster 0**: These neighborhoods are characterized by high number of Ice-cream Shops, Café & Tea Rooms, Hotels and Resorts, Pub’s, Vegan and Italian Restaurants, Fast Food Joins, Clothing stores & Farmers Market.

Cluster 0 comprises of neighborhoods like Amruthahalli, Bannerghatta Road, Chickpet,

CMP Centre and School, Electronics City, H.A.L II Stage, Hoodi, Indiranagar, Jayanagar,

Malleswaram, Nagarbhavi ,Seshadripuram ,Sivan Chetty Gardens.

1. **Cluster 1:** These neighborhoods are characterized by high number of Indian, Seafood and Chinese Restaurants, Departmental stores, Electronic stores, Jewelry stores & Gym’s.

Cluster 0 comprises of neighborhoods like B Sk II Stage, Banaswadi, Basavanagudi, Basaveshwaranagar, C.V.Raman Nagar, Chikkalasandra, Domlur, Immedihalli, Koramangala VI Bk, Mahalakshmipuram Layout.

The above two clusters have been visualized in the map below, cluster 0 is represented in “Red” and cluster 1 has been represented in “Violet”.

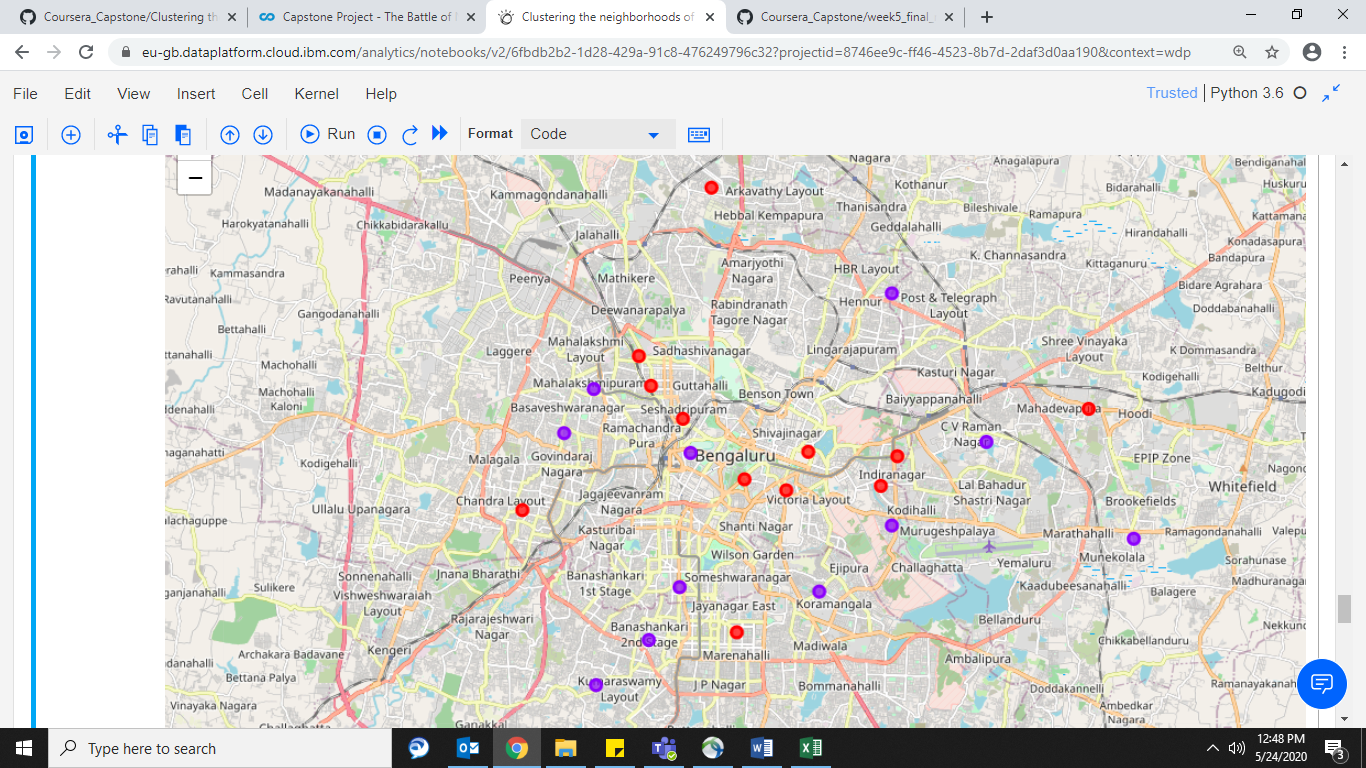


Figure 4.1 Bangalore Venue Cluster Map

1. **Discussion**

Having a successful business is down to many factors such as a rock solid business plan, location of the business and most importantly catering to the customer needs. Most often than not, business fail to match the business objective with these factors. For example, while deciding to open a Fast food joint or a Juice junction, it is wise to select an area that is populated by young customers that drive majority of the sales. Similarly, while deciding the location for a restaurant it is apt to select locations having family concentration. The results of this study provides a backdrop of current business concentration in various major locations throughout Bangalore. Cluster 0 as mentioned on the results page, seems to have high numbers of Ice cream parlors, café, pubs etc. Therefore, any individuals looking to open any business establishments of similar genres should concentrate their business in cluster 0 locations due to the presence of high demand. In similar fashion, business which are family oriented like Indian or Asian restaurants, Departmental stores, jewelry stores etc. can concentrate their stores in cluster 1 locations.

This study also provides business owners with another dimension of insights, consider a scenario where in an entrepreneur has understood there is lot of potential in the Koramangala region. However, is not sure of what business to set up. Since, Koramangala is part of cluster 1 one can assume business that have seen a success in other locations part of cluster 1 can also find good demand in Koramangala. On comparisons with Domlur another location part of Cluster 1, we find that Koramangala has very less Breakfast spots. Hence, providing the necessary inputs for business idea generation.

1. **Limitations & Suggestions for the Future**

The current data set seems to be limited with venue information available only from Foursquare.com. This does not comprise of the overall venue information in Bangalore. The future study should also look at gathering information from local service and delivery partners like Zomato or Swiggy. Also, Google houses information on all the business across Bangalore. These, online delivery applications along with Google contain comprehensive data of all the restaurants & other local business and serve as an ultimate sample for this analysis.

Another factor that influences any business owners in making decisions is the cost or rent details by location. Adding these information would make this analysis even more versatile and an end to end guide in helping Entrepreneurs select a location given a business idea or vice versa.

1. **Conclusion**

The study looked at finding solutions to an age old problem that has lot of potential entrepreneurs spend sleepless nights right from understanding what business they are going to set up, understand customers and ultimately finding the right location for the business. As mentioned in the results, the study classified several locations into groups of 2. Cluster 0, had prominent signs of young customers preferring business catering their needs and tastes. Similarly, cluster 1 grouped locations where business meet the needs of fairly elder customers, families etc. Using this as a reference, future business owners can pin point the right locations based on their business ideas or scope out the right business ideas based on the need currently not being catered to.