Capstone Project - The Battle of Neighborhoods (Week 5)

K-mean clustering Algorithm

Author: Avnish Omprakash Yadav

Date: 20th august 2020

Clustering localities in Bangalore, India based on Restaurants (K-means Clustering)

- Why clustering of localities in Bangalore required?

Answer: We are performing clustering of localities to assist new person to choose best place to live if he/she is visiting first time to Bangalore.

- How can clustering localities help person visiting to Bangalore?

Answer: We will provide clustered map of Bangalore City where restaurant will be pointed based on price for two person, and rating.

Example: Restaurant having low price and average rating around 4.

Data Collection

- Four Square Database used to collect venues detail with 35km of Bangalore City. We will take latitude and longitude from response return from Four Square Database of every venue. We have found 100 venues.

Below table contain sample data collected from Four Square Database.

Name	Latitude	Longitude	Category
UB City	12.97170898	77.59590528	Shopping Mall
Truffles - Ice & Spice	12.97180162	77.6010306	Burger Joint
Toscano	12.97198038	77.59606565	Italian Restaurant
Smoke House Deli	12.97165618	77.59825418	Deli / Bodega

Data Collection

- Zomato Api is used to collect data of near by restaurant around venues return from Four Square Database.

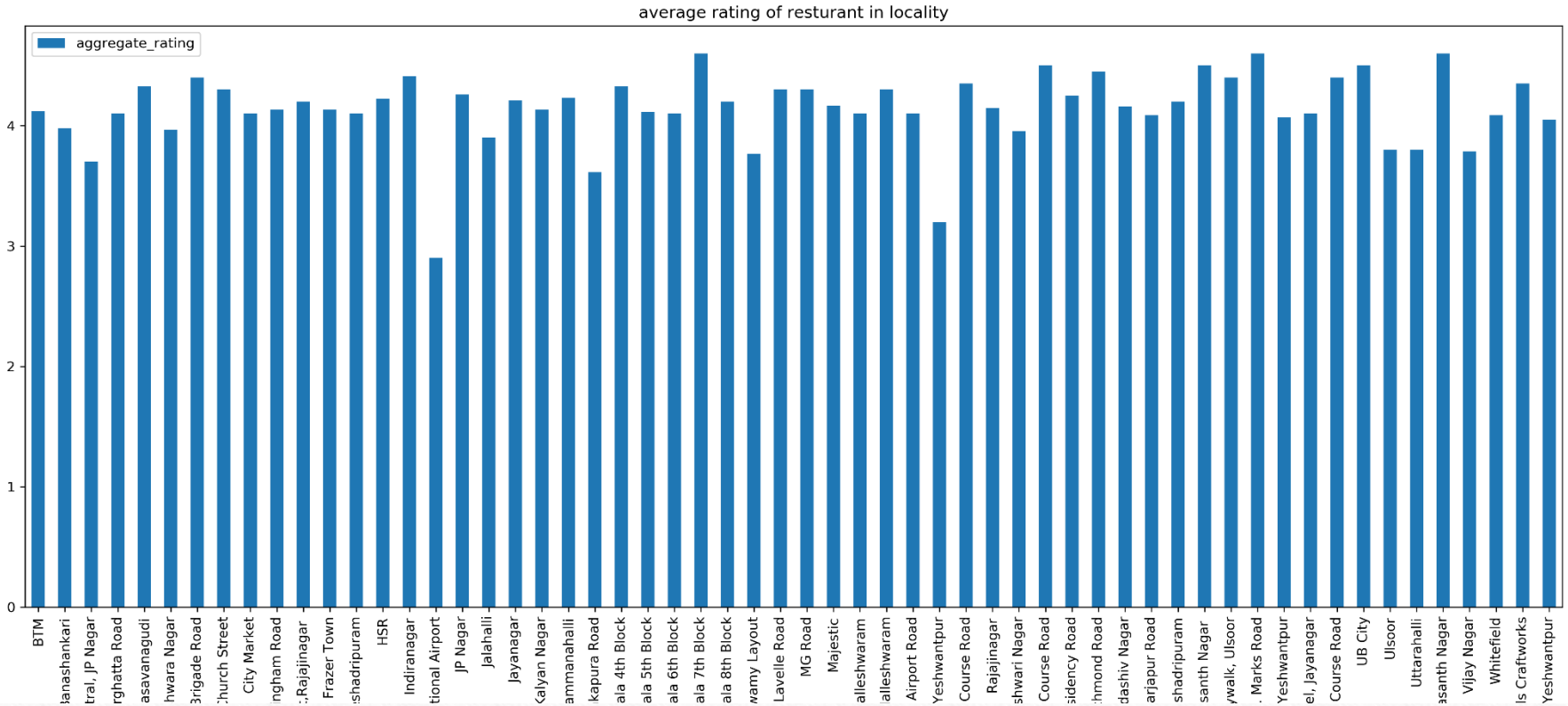
venue_name	venue_latitude	venue_longitude	locality	average_price_for_two	price_range	aggregate_rating	votes
Green Theory	12.968645	77.602743	Residency Road	950	2	4.1	3327
Communiti	12.97221919	77.60836903	Residency Road	1500	3	4.7	7155
Hard Rock Cafe	12.97603394	77.60156728	St. Marks Road	2500	4	4.8	5920
Cafe Azzure	12.97495904	77.60760259	MG Road	900	2	4.3	3839
Olive Bar And Kitchen	12.96688576	77.60817122	Richmond Road	1800	3	4.6	2448

Data Wrangling

- Collected data from Four Square database and Zomato api is not in structured format. We have create to function which will convert receive detail from Four Square database into structure format.
- "**get_banglore_venue_detail**" function will convert Four Square database result into structured format as we have show in Data collection section. We have collected 100 venues from Four Square database.
- "**get_banglore_resturant_detail**" function will convert Zomato api result into required structured format as we have already shown in data Collection. We have collected **249 restaurant** detail from Zomato Api.

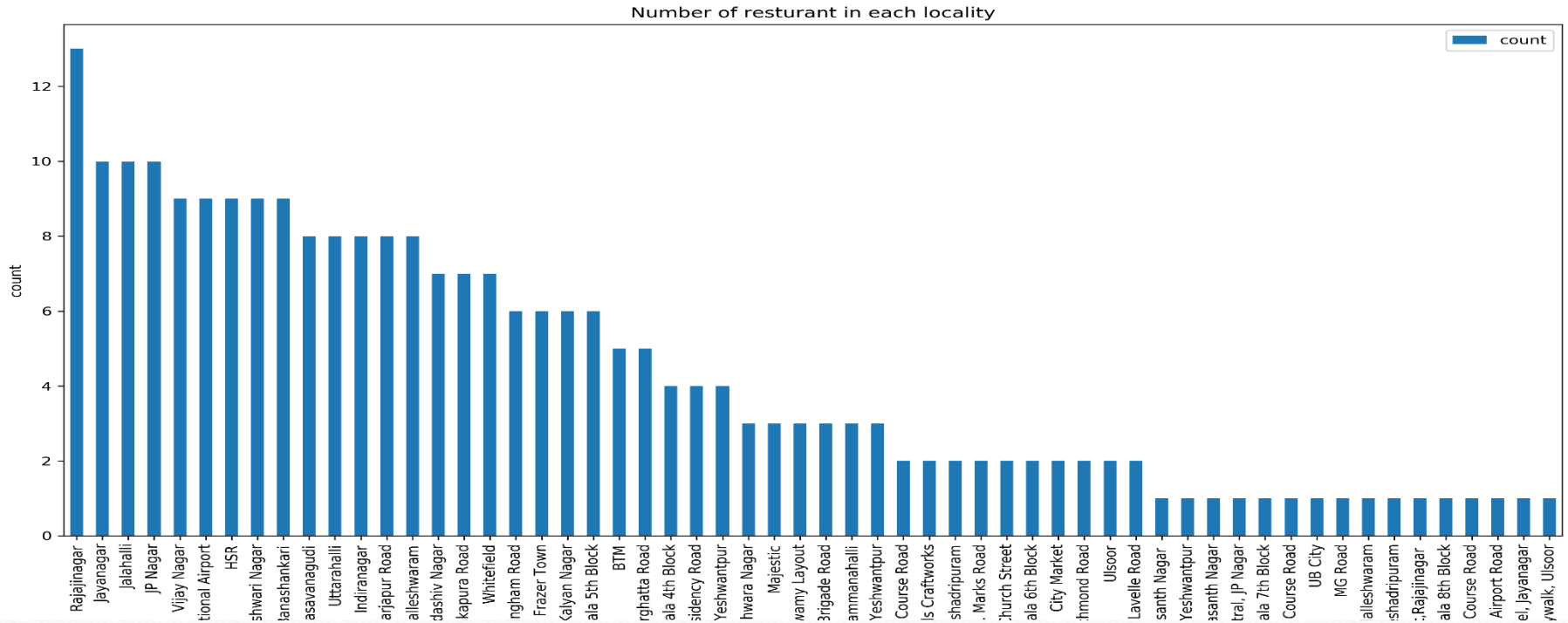
Data Analysis

Based upon graph represented below we can see that most of restaurant having rating around 4 some having rating more than 4.

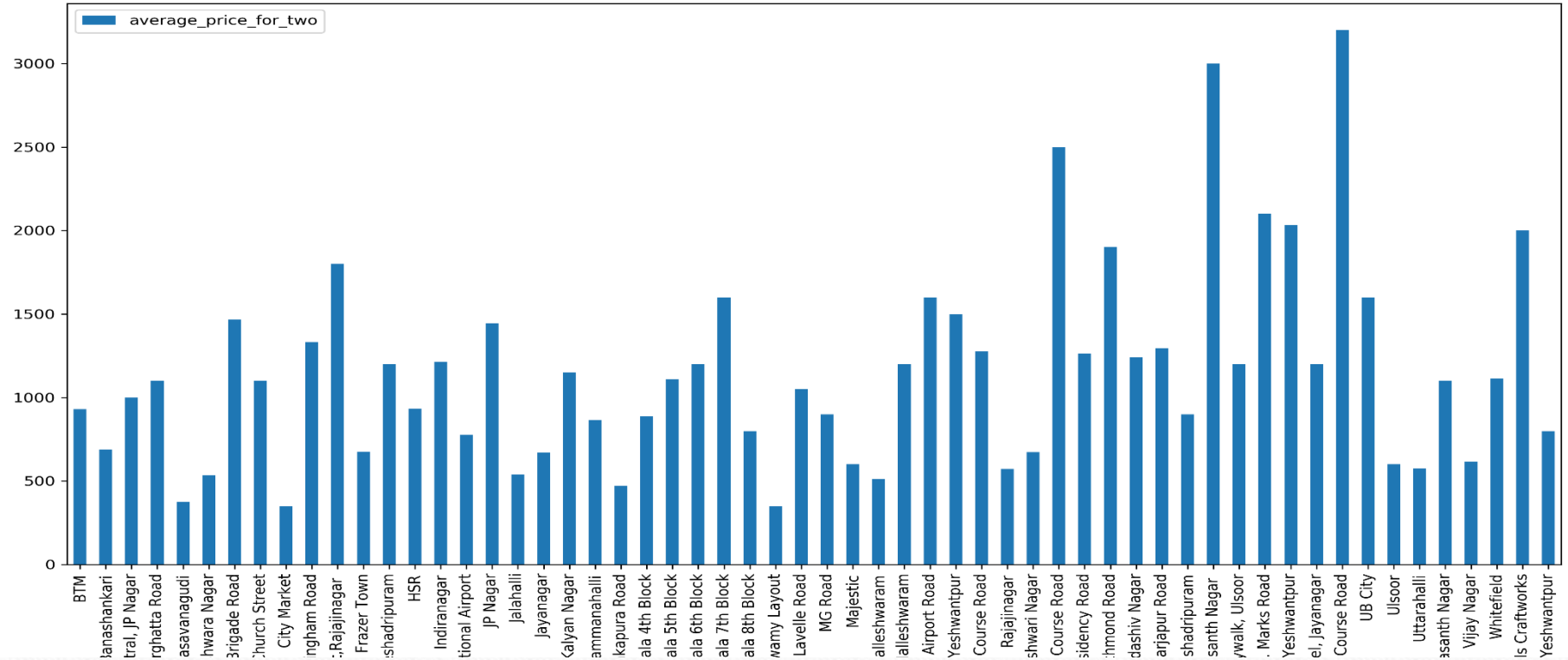


Data Analysis

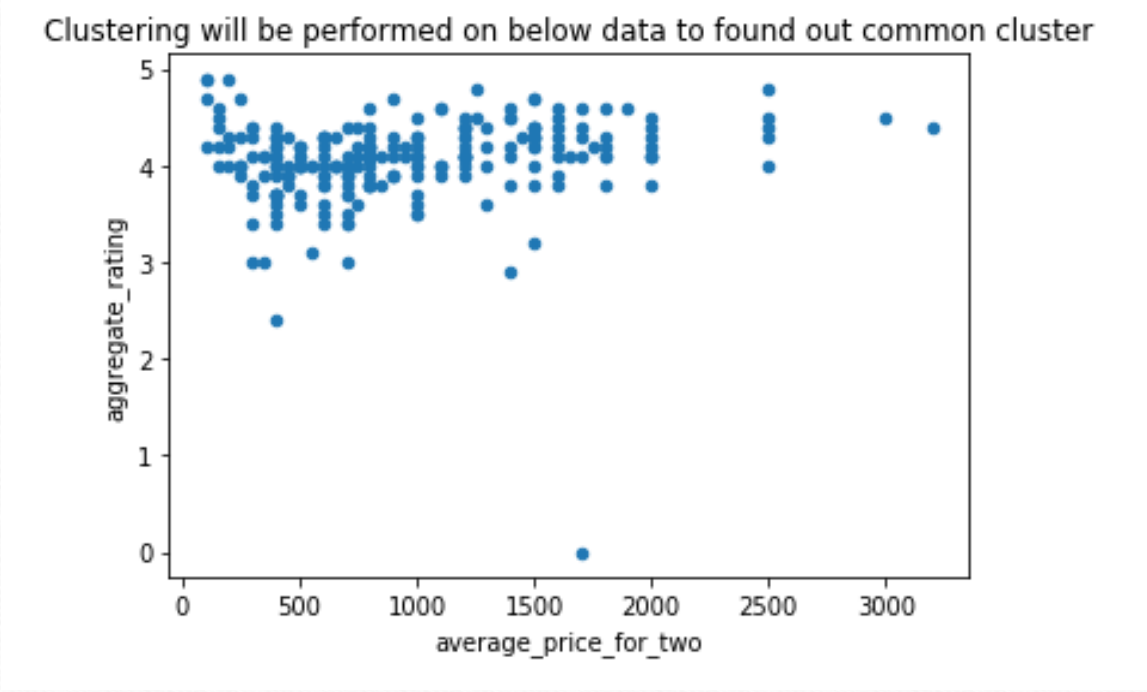
Below graph represent number of restaurant available in each locality. From below graph also we can come to some conclusion to choose locality to live in Bangalore.



Data Analysis



Data Analysis



We can see that there only two or three cluster can form. This graph indicate that once cluster will have more number of restaurant. Restaurant have low price and high ratings will come into below categories:

- Low price and high rating
- High price and high rating
- Average price and average rating

K-means (Unsupervised learning)

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into **k** clusters in which each observation belongs to the cluster with the nearest **mean** (cluster centers or cluster centroid), serving as a prototype of the cluster.

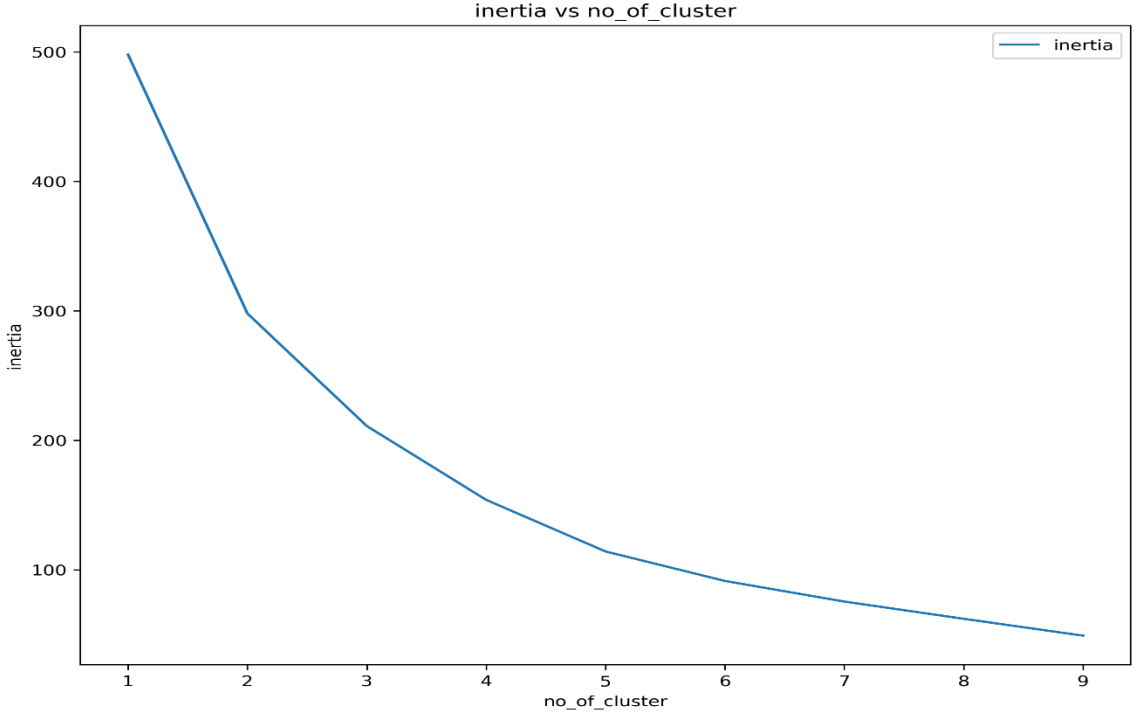
How to choose value of k (No. of cluster.)?

We will use elbow method to decide value of k which will nest fit our data.

The **elbow method** runs **k-means** clustering on the dataset for a range of values for **k** (say from 1-9) and then for each value of **k** computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

From this graph we can see that how inertia decrease when we have increase the value of cluster. The **K-means** algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the **inertia** or within-cluster sum-of-squares (see below).

We will choose the value of k after which inertia decrease slowly. As per my observation I have picked 3 as value of k.



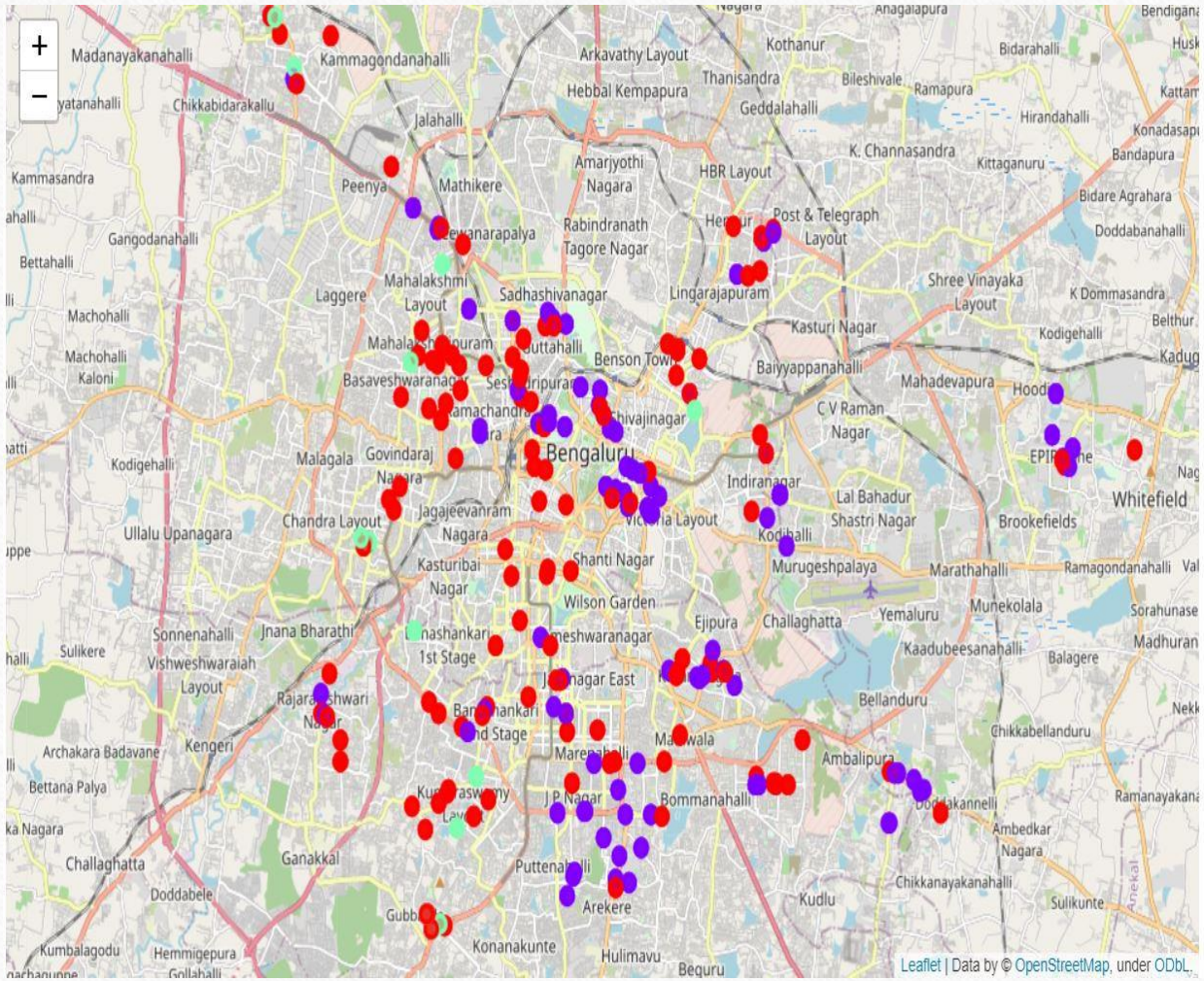
Conclusion

From this graph we can see that there were three cluster we have created. A person visiting to Bangalore can choose a desired place to live in Bangalore by considering where he can get reasonable restaurant.

Blue: Cluster 0 has average price and it has rating around between 3.5 and 4.

Red: Cluster 1 is medium rating around 4. And restaurant is not enough expensive

Aqua: cluster 2 has highest rating more than 4 and average price.



Thank You!