Capstone Project -The Battle of Neighborhoods (Week 5)

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

Author: Avnish Omprakash Yadav
Date: 20thaugust 2020

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

Index

1. Introduction:

People visiting new cities would be highly interested in the localities with the best restaurants in the city. People might want to know how good a given restaurant is based on the ratings the restaurant has received and would like to know the price range the Restaurant falls under so that they can make informed budget decisions. Also, they would like to know the best localities where they could find these restaurants. The information of ratings and price range of various restaurants in the city and their localities in form of graphs, charts and maps would help people decide which restaurant to choose among the many restaurants in the city. And also which locality to visit. Also combining the location of the restaurants in the city with their price and rating information would help visitors make easy decisions about the locations they should visit. A map of the restaurants and another map of the localities with specific color attributes will be plotted to highlight their position. Further, we will classify the various locations into different clusters using a Machine Learning Algorithm, the K-means clustering Algorithm. This enables any visitor to take a quick glance and decide what place to visit.

2. Data Collection:

To get location and other information about various venues in Bangalore, two APIs were used. The Foursquare API and the Zomato API. The Foursquare's explore API was used to fetch venues up to a range of 35 kilometers from the center of Bangalore. The names, categories and locations (latitude and longitude) of these venues were collected. Using the name, latitude and longitude values obtained from the Foursquare API, we used the Zomato search API to fetch data from its database. The Zomato API allows to find only restaurants based on a search criteria using the name, latitude, longitude, etc. The data from the two APIs do not match completely because Foursquare API retrieves all venues in Bangalore and the Zomato API retrieves only restaurants in Bangalore. So, we combine the two datasets to get only Restaurants from the Foursquare API and the corresponding ratings and price information from the Zomato API.

We use various techniques of Data cleaning to get the final dataset.

We have extracted information of following field/ attributes from Four Square Database.

Four Square Database	
Name	The name of the venue.
Category	The category type as defined by the API
Latitude	The latitude value of the venue.
Longitude	The longitude value of the venue.

Sample data extracted from Four Square Database. (we have extracted 100 venues.)

Name	Latitude	Longitude	Category
UB City	12.97171	77.59591	Shopping Mall
Truffles - Ice & Spice	12.9718	77.60103	Burger Joint
Toscano	12.97198	77.59607	Italian Restaurant
Smoke House Deli	12.97166	77.59825	Deli / Bodega
JW Marriott Hotel Bengaluru	12.97236	77.59505	Hotel
Cubbon Park	12.97704	77.59528	Park
Corner House	12.97298	77.59997	Ice Cream Shop
Harima	12.96775	77.60007	Sushi Restaurant
M.G Road Boulevard	12.97577	77.60398	Plaza
Infinitea	12.98716	77.59483	Tea Room
The Oberoi	12.97346	77.61829	Hotel

Data extracted from “Four square” database contains latitude and longitude. We have used latitude and longitude of Venue to request Zomato Api to return restaurant around venues.

We have extracted information of following field/ attributes from Zomato Api.

Zomato Api	
Name	The name of the restaurant.
Locality	The locality of the restaurant.
Rating	The average rating of the restaurant given be users.
Price range	The price ranges the restaurant belongs to as defined by Zomato
Price for two	The average cost for two people dining at the restaurant.
Latitude	The latitude value of the restaurant.
Longitude	The longitude value of the restaurant.

Sample data extracted from Zomato Api. (we have extracted 249 restaurant.)

venue_name	venue_latitude	venue_longitude	locality	average_price_for_two	price_range	aggregate_rating	votes
Green Theory	12.96865	77.60274	Residency Road	950	2	4.1	3327
Communiti	12.97222	77.60837	Residency Road	1500	3	4.7	7155
Hard Rock Cafe	12.97603	77.60157	St. Marks Road	2500	4	4.8	5920
Cafe Azzure	12.97496	77.6076	MG Road	900	2	4.3	3839
Olive Bar And Kitchen	12.96689	77.60817	Richmond Road	1800	3	4.6	2448
Arbor Brewing Company	12.97006	77.61081	Brigade Road	2000	4	4.2	8947
Church Street Social	12.97155	77.59851	Lavelle Road	1500	3	4.3	8266
Farzi Cafe	12.97206	77.5959	UB City	1600	3	4.5	2817
MISU	12.97075	77.60079	St. Marks Road	1700	3	4.4	2225
Brik Oven	12.97468	77.60543	Church Street	1200	3	4.4	3044

Data extracted from Zomato Api contains two most import information price and rating.

We also extracted number of feedback each restaurant received from customer which is valuable information. If restaurant has feedback from most of customer means that rating about restaurant is valid.

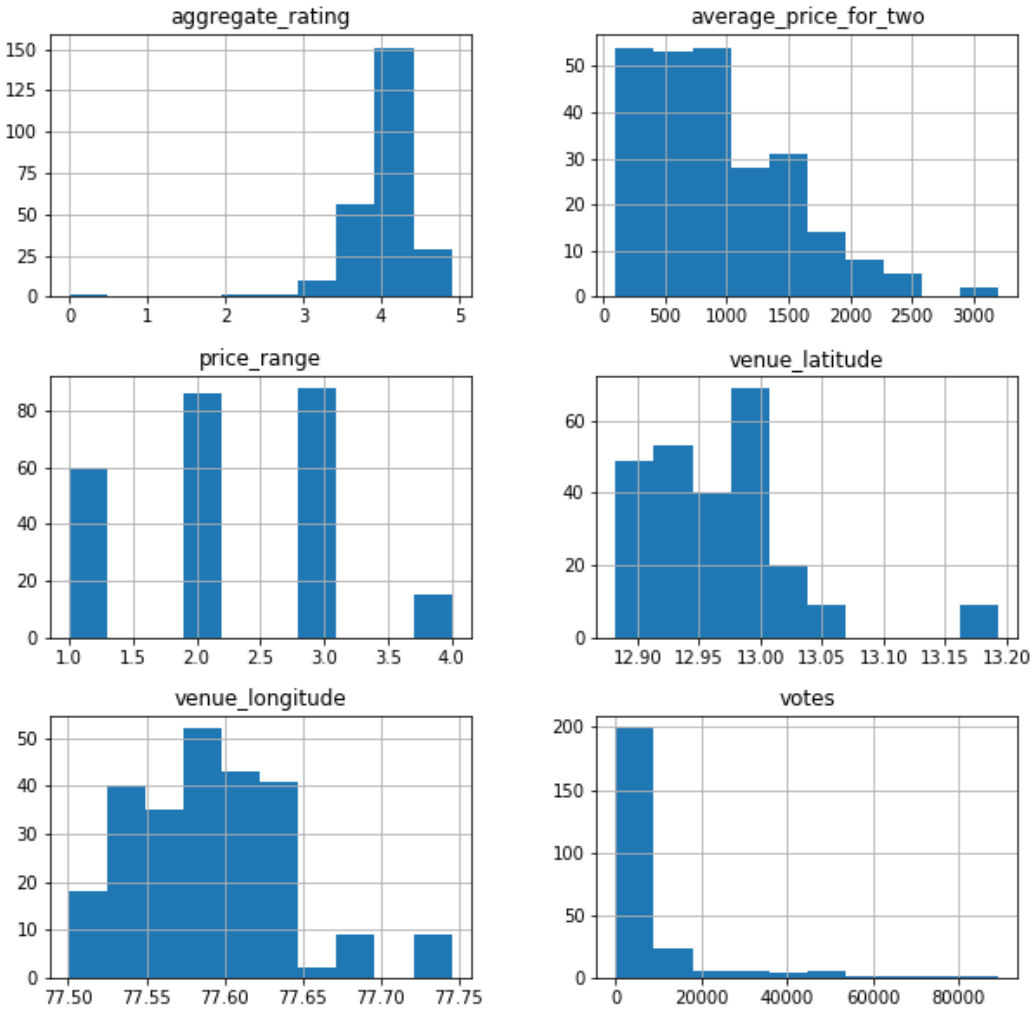
Price and rating can help many person to determine which restaurant can best fit there expectation in terms of cost and quality.

3. Methodology:

We will describe what we have analyzed from data that we have collected from Zomato Api.

Data Analysis: We have plotted histogram to see the distribution of numeric field in our dataset.

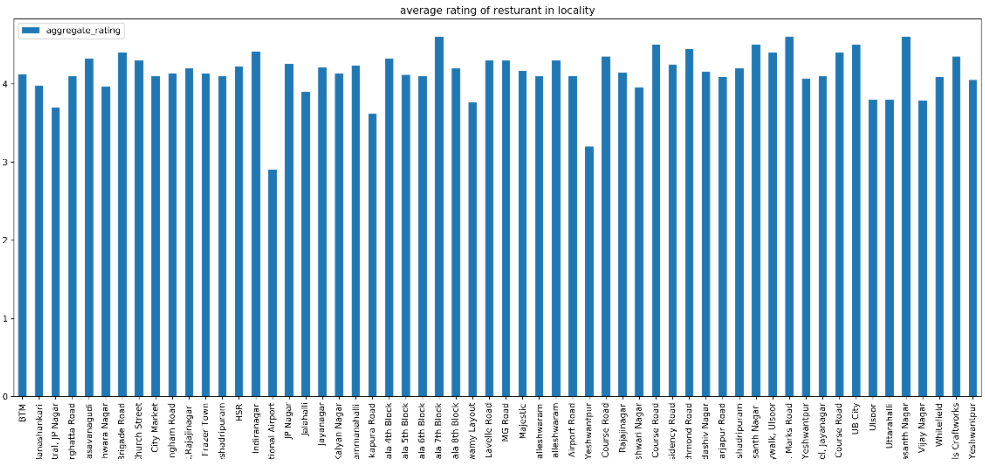
Below graph represent various data distribution of various field. We will explore each one of them one by one.



Aggregate Rating:

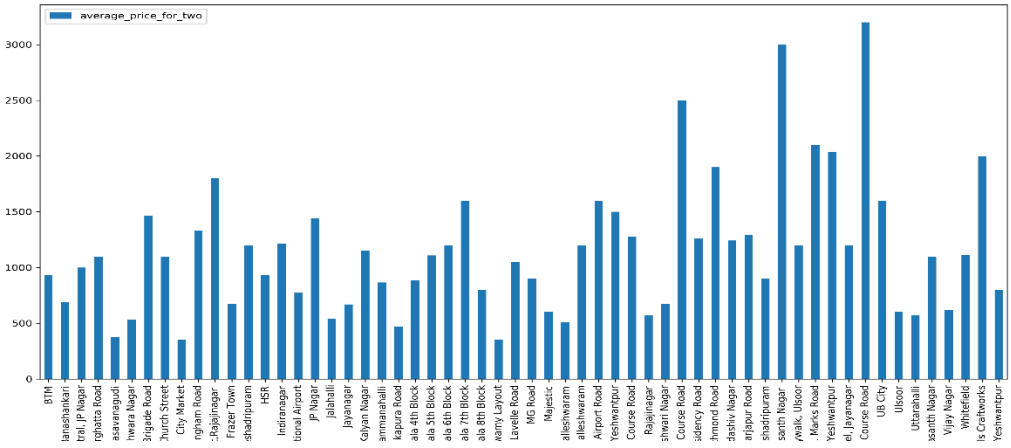
1. Most of restaurant having rating more than 4 (around more than 150 restaurant out of 249).
2. None of restaurant having rating exact 5.
3. There are few restaurant which has rating less than 3 (approximately 10)
4. More than 50 restaurant have rating between 3.5 and 4.

   Just analyzing one single histogram of one attribute we have received very crucial detail about restaurant of Bangalore locality.



Average price of two:

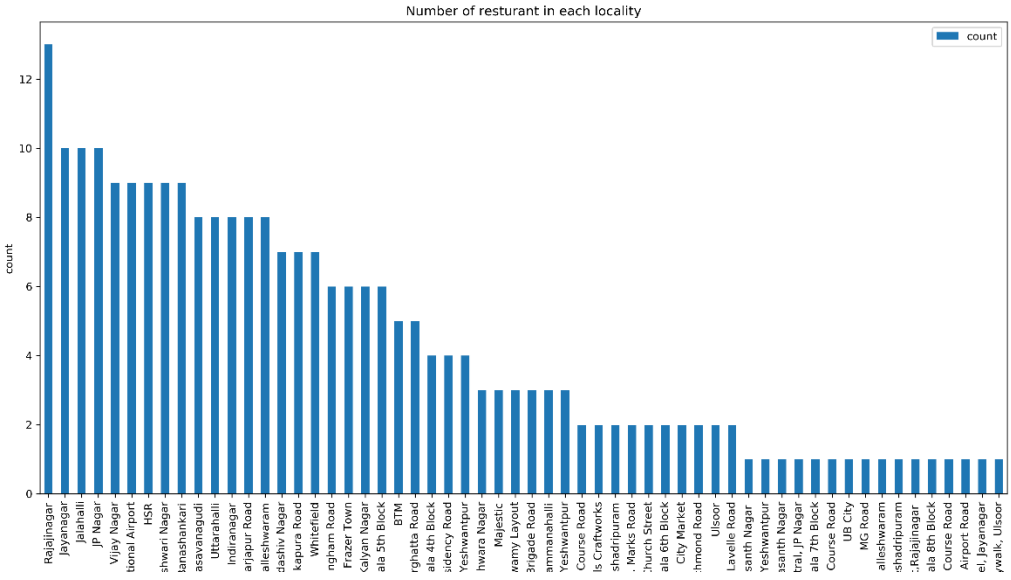
1. There are hardly two restaurant which cost around 3000.
2. More than 50% of restaurant cost comes below 1000.
3. There are approximately 17 restaurant which has price more than 2000.



Number of restaurant in each locality:

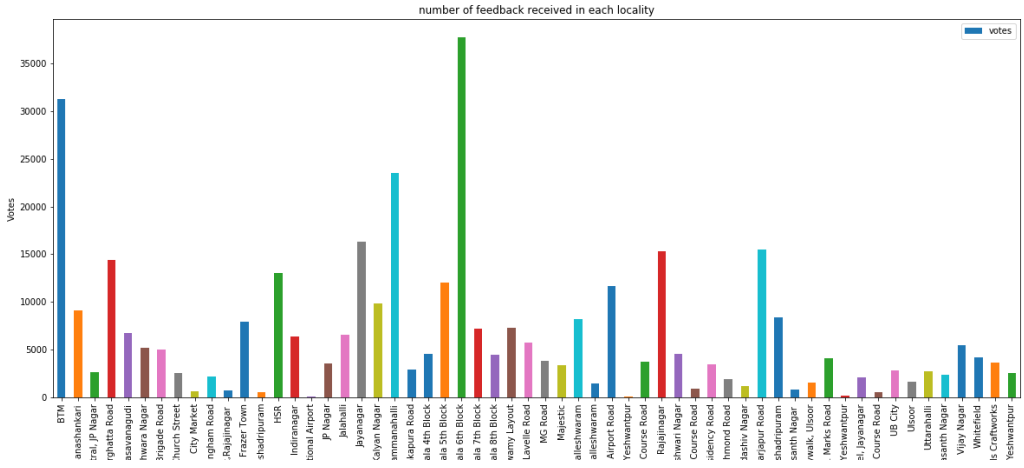
Graph depicts us that locality “Rajajinagar ” has 13 restaurant. Rajajinagar has highest number of restaurant availabe as compare to other locality. Some locality has only “1” restaurant available.

Below bar chart can help people to decide which locality would be best for them.



Votes: Based on number of feedback available for each customer one can understant which restaurant is popular. locality “Koramangala 6th Block” has huge number of feedback. “Koramangala 6th Block” may have famous restaurant where people usually visit most of the time.

I hope below bar chart graph will be usefult to new people visiting to Banglore.



We have used unsupervised machine learning K-means clustering algorithm to find our different cluster of restaurant of Bangalore localities.

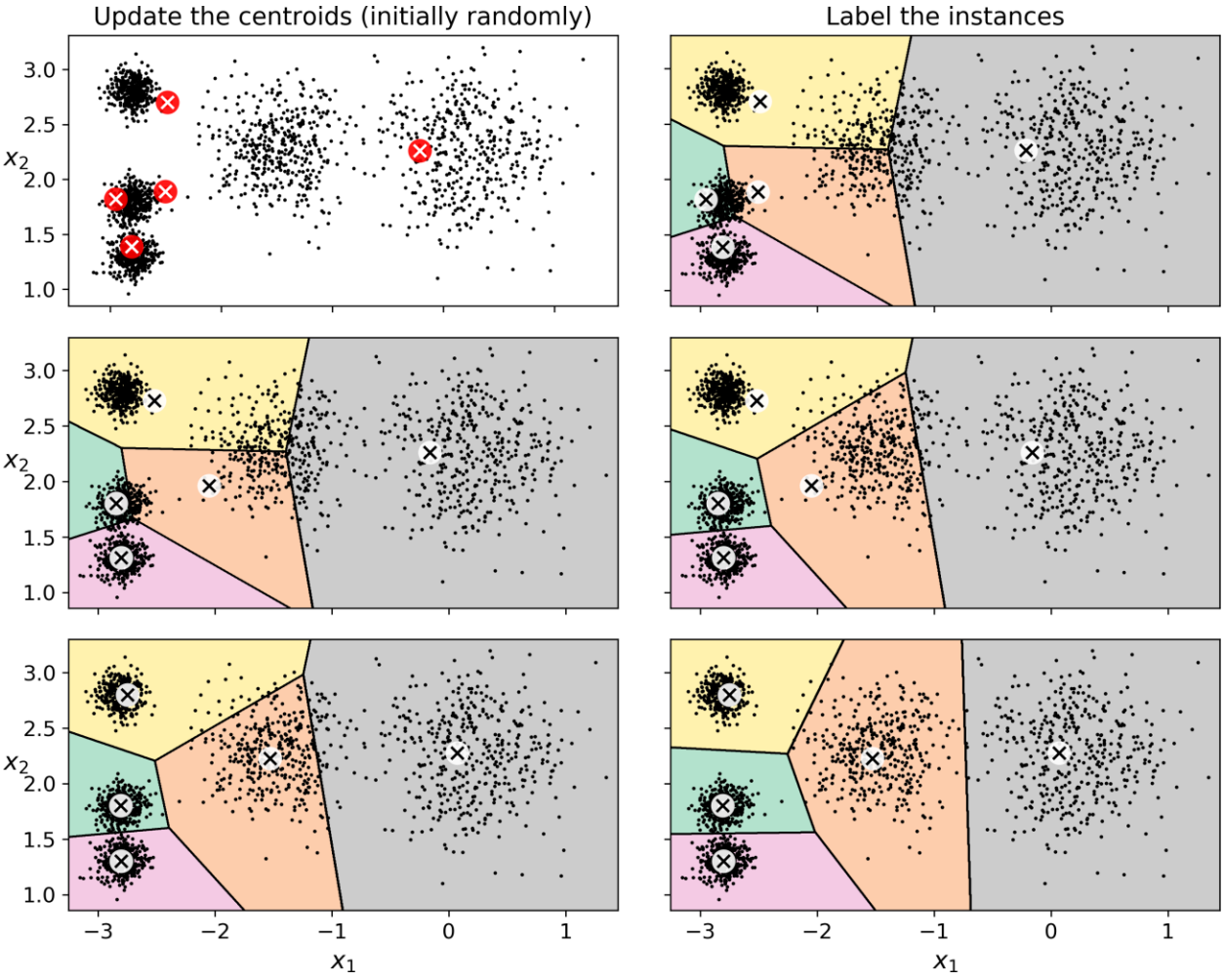
What is K-Means clustering algorithm?

The K-Means algorithm is a simple algorithm capable of clustering dataset very quickly and efficiently, often in just few iteration. It was proposed by Stuart Lloyd at the Bell Labs in 1957 as a technique for pulse –code modulation, but it was only published outside of the company in 1982, in a paper titled “Least square quantization in PCM”. By then, in 1965, Edward W. Forgy had published virtually the same algorithm, so K-Means is sometimes referred to as Lloyd-Frgy.

How does the K-Means Algorithm works.

Well it is really quite simple. Suppose you were given the centroids: you could easily label all the instances in the dataset by assigning each of them to the cluster whose centroid is closest. Conversely, if you were given all the instance labels, you could easily locate all the centroids by computing the mean of the instances for each cluster. But you are given neither the labels nor the centroids, so how can you proceed? Well, just start by placing the centroids randomly (e.g., by picking k instances at random and using their locations as centroids). Then label the instances, update the centroids, label the instances, update the centroids, and so on until the centroids stop moving. The algorithm is guaranteed to converge in a finite number of steps (usually quite small), it will not oscillate forever2. You can see the algorithm in action in Figure 9-4: the centroids are initialized randomly (top left), then the instances are labeled (top right), then the

centroids are updated (center left), the instances are relabeled (center right), and so on. As you can see, in just 3 iterations the algorithm has reached a clustering that seems close to optimal.



Finding the Optimal Number of Clusters.

To find optimal number of cluster we can train our model with different value of K (K={1…9}) cluster. Then we can plot K vs ineteria. We should choose where inertia drastically decreases.

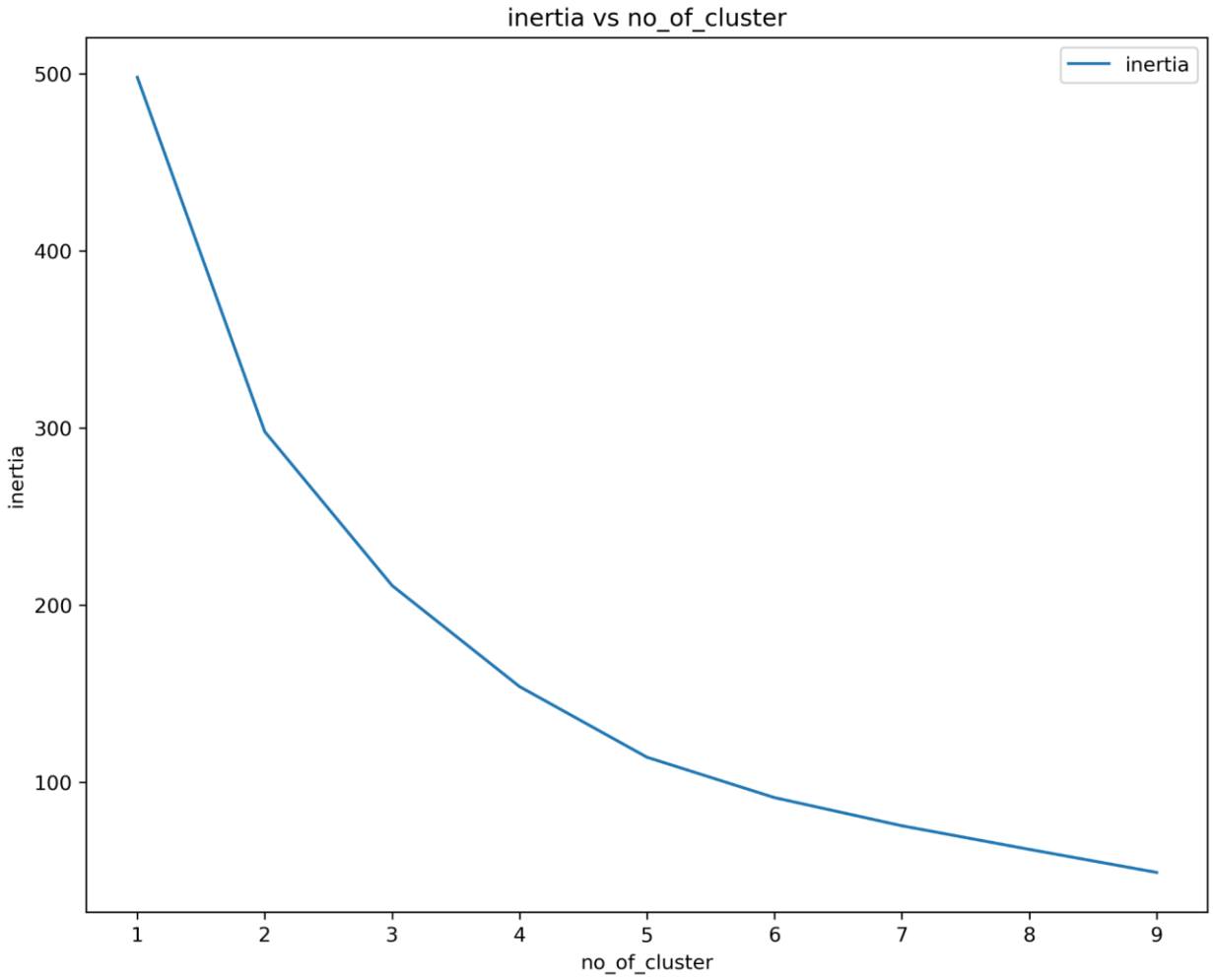
Below figure represent that value of k can be 2,3,4 so we have choosed 3 for our algorithm.

Why we should choose value of k where intertia decreses drastically?

What is Intertia: Average sum of square distance between point and it's centroid.

If we increases value of K every time then it will always decrease interia because if we have more number of cluster then we definitely get less interia "Average sum of square distance between point and it's centroid". If we choose k same as number of record available to us then we might get interia as "Zero". But it will not find out common cluster. Better to choose value of K where inertia decreases rapidly.

,



Why we have used K-Means Algorithm?

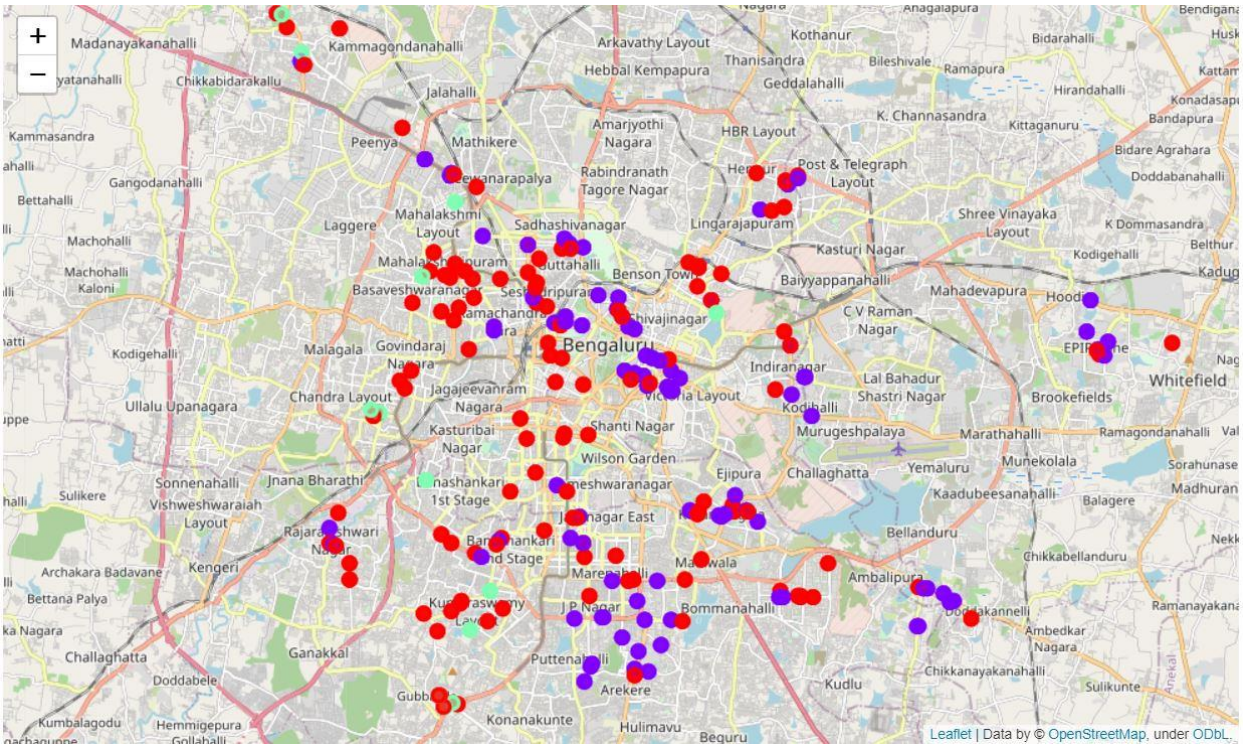
We have used K-Means algorithm because we wanted to find out different type of restaurant available in Bangalore localities. Example: We wanted to cluster in such manner so that any new person visiting to Bangalore can decide which place can best suit his need.

We will create cluster of restaurant in Bangalore city based on price and rating.

4. Result:

Figure show different cluster detected by K-Means clustering algorithm

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.

```
In [18]: banglore_resturant_details[banglore_resturant_details['label']==0]
```

Out[18]:

	venue_name	venue_latitude	venue_longitude	locality	average_price_for_two	price_range	aggregate_rating	votes	label
94	Moto Store & Café	12.987530	77.620858	Ulsoor	500	2	3.6	1595	0
123	Pizza Hut	12.997268	77.540406	Basaveshwara Nagar	600	2	3.4	4475	0
133	The Biryani Cafe	12.943092	77.541629	Banashankari	600	2	3.6	4025	0
185	Ji Hazoor	12.976187	77.726905	Whitefield	1300	3	3.6	6416	0
192	Cibo Esca	12.960728	77.528630	Vijay Nagar	1000	3	3.5	889	0
195	The Charcoal	12.960912	77.528437	Vijay Nagar	1400	3	2.9	575	0
198	Laziz Pizza	12.961651	77.526468	Vijay Nagar	600	2	3.5	4206	0
203	Cafe Casavista	13.066924	77.502156	Jalahalli	400	1	3.4	5198	0
204	Hotel Elite Restaurant	13.056744	77.507664	Jalahalli	750	2	3.6	5273	0
215	Adiga's	12.882698	77.546640	Kanakapura Road	350	1	3.0	5295	0
216	A2B - Adyar Ananda Bhavan	12.883607	77.548705	Kanakapura Road	550	2	3.1	244	0
220	Ovenstory Pizza	12.885713	77.544962	Kanakapura Road	400	1	3.5	1238	0

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

```
In [19]: banglore_resturant_details[banglore_resturant_details['label']==1]
```

Out[19]:

	venue_name	venue_latitude	venue_longitude	locality	average_price_for_two	price_range	aggregate_rating	votes	label
0	Green Theory	12.968645	77.602743	Residency Road	950	2	4.1	3327	1
3	Cafe Azzure	12.974959	77.607603	MG Road	900	2	4.3	3839	1
15	Dolci Desserts	12.986407	77.594977	Cunningham Road	750	2	4.2	1003	1
17	Mudpipe Cafe	12.988290	77.593878	Cunningham Road	750	2	4.0	997	1
21	The Green Path - Forgotten Food	12.990298	77.571834	Malleshwaram	600	2	3.9	1383	1
24	South Ruchis Square	12.984025	77.578110	Race Course Road	850	2	4.1	2126	1
28	Empire Restaurant	12.989005	77.574320	Seshadripuram	800	2	4.1	15690	1
29	CTR Shri Sagar	12.998270	77.569455	Malleshwaram	150	1	4.6	16857	1
31	Al-Bek	12.994248	77.571517	Malleshwaram	400	1	4.2	24562	1
32	Halli Mane	12.995624	77.571802	Malleshwaram	250	1	4.0	4133	1

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

```
In [20]: banglore_resturant_details[banglore_resturant_details['label']==2]
```

Out[20]:

	venue_name	venue_latitude	venue_longitude	locality	average_price_for_two	price_range	aggregate_rating	votes	label
1	Communiti	12.972219	77.608369	Residency Road	1500	3	4.7	7155	2
2	Hard Rock Cafe	12.976034	77.601567	St. Marks Road	2500	4	4.8	5920	2
4	Olive Bar And Kitchen	12.966886	77.608171	Richmond Road	1800	3	4.6	2448	2
5	Arbor Brewing Company	12.970062	77.610813	Brigade Road	2000	4	4.2	8947	2
6	Church Street Social	12.971549	77.598507	Lavelle Road	1500	3	4.3	8266	2
7	Farzi Cafe	12.972060	77.595901	UB City	1600	3	4.5	2817	2
8	MISU	12.970748	77.600788	St. Marks Road	1700	3	4.4	2225	2
9	Brik Oven	12.974678	77.605425	Church Street	1200	3	4.4	3044	2
10	Hammered	12.986476	77.594997	Cunningham Road	1400	3	4.1	6305	2
11	Millers 46	12.991611	77.594050	Vasanth Nagar	1100	3	4.6	2390	2
12	1Q1 Kitchen & Bar	12.983703	77.596924	Cunningham Road	2500	4	4.3	767	2

5. Discussion:

I have found very few restaurant has rating more than 4.5. Which comes under cluster 2nd.

Few restaurant has huge number of customer definitely those restaurant are famous.

If you are interested please check Data Analysis where we have shown a graph which depicts number of votes having each restaurant.

6. Conclusion:

We may collect data from other API apart from Zomato. Then we can combine.On new data we can perform our K-Means clustering. Just depend on Zomato api is not good enough. We should collect data from other api also who can provide detail of restaurant.

We have discussed cluster obtain by K-Mean algorithm in result section.