Battle of The Neighbourhood's

Finding best neighbourhood in the city of Bangalore

This project aims to utilize all the concepts learned through IBM Data Science Professional Course. We define a domain problem, the data that will be used and using this we are able to analyse it by applying various ML tools. Analysis of data is step by step procedure, involves Data Gathering, Data Cleaning, Exploratory Data analysis, Data Modelling and interpreting (final output). In this project we will go through all these process and provide a conclusion that can be leveraged by the business stakeholders to make their decision.

Background:

Bangalore the Silicon Valley of India, Karnataka's capital city. It is famous for traffic 24/7. With a population of over 15 million, Bengaluru is the third largest city in India and 27th largest city in the world. Bangalore is one of the most ethnically diverse cities in the country, with over 51% of the city's population being migrants from other parts of India. Historically a multicultural city, Bengaluru has experienced a dramatic social and cultural change with the advent of the liberalization and expansion of the information technology and business process outsourcing industries in India. IT companies in Bengaluru employ over 35% of India's pool of 1million IT professionals. Being a diversity in nature and the hub of interactions between ethnicities brings many opportunities for entrepreneurs to start or grow their business. It is a place where people can try the best of each culture, either while they work or just passing through.

Business Problem:

The city's 21st century dining culture has been shaped by locals, expats and scores of working professional from across India who have moved to the city during the past decade or two. This makes food businesses to come up with new start up ideas by offering numerous cuisines across the world class dishes and get productive. The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighbourhood in Bangalore to open a restaurant. Being a land of multi-diverse, people wishes to taste new dishes and there are numerous opportunities to open a new restaurants like Italian. Through this project, we will find the most suitable location for an entrepreneur to open a new Italian restaurant in Bangalore, India.

Target Audience:

This project is aimed towards Entrepreneurs or Business owners who want to open a new Italian Restaurant or grow their current business. The analysis will provide vital information that can be used by the target audience

Data Overview:

The data that will be required will be a combination of CSV files that have been prepared for the purposes of the analysis from multiple sources which will provide the list of neighbourhoods in Bangalore (via Kaggle), the Geographical location of the neighbourhoods (via Kaggle & Geocoder package) and Venue data pertaining to Italian restaurants (via Foursquare). The Venue data will help find which neighbourhood is best suitable to open an Italian restaurant

Data acquisition:

Source 1: Neighborhoods of Bangalore and Geographical Location via Kaggle

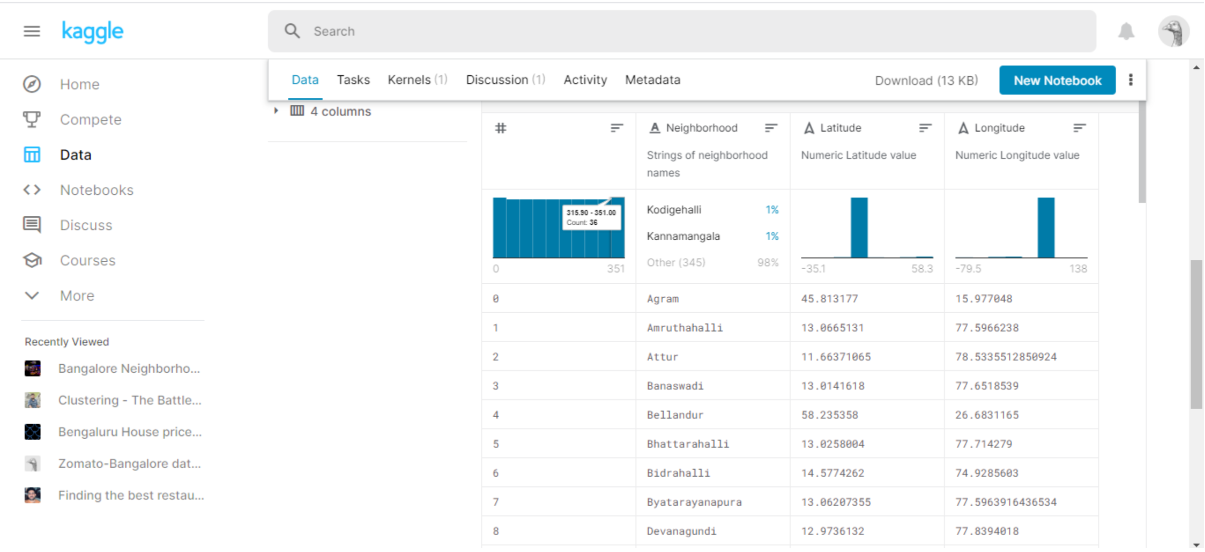


Figure - 1

1. https://www.kaggle.com/rmenon1998/bangalore-neighborhoods?select=blr_neighborhoods.csv

The Kaggle site shown above (Figure -1) provided all the information about the neighbourhood's, latitude's and longitude's present in Bangalore. The file was in the CSV format, so we had to attach to a pandas data frame (Figure – 2).

	City	Neighborhood	Latitude	Longitude
0	Bangalore	Agram	45.813177	15.977048
1	Bangalore	Amruthahalli	13.066513	77.596624
2	Bangalore	Attur	11.663711	78.533551
3	Bangalore	Banaswadi	13.014162	77.651854
4	Bangalore	Bellandur	58.235358	26.683116

Figure -2

Source 2: Venue Data using Foursquare:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	venue	venue_categories	venue Latitude	venue longitude
0	Agram	45.813177	15.977048	Amélie	Dessert Shop	45.813842	15.979011
1	Agram	45.813177	15.977048	Corner bar	Bar	45.812930	15.979440
2	Agram	45.813177	15.977048	Trg bana Josipa Jelačića	Plaza	45.813032	15.976868
3	Agram	45.813177	15.977048	Tržnica Dolac	Farmers Market	45.814070	15.977261
4	Agram	45.813177	15.977048	Cremme Zagreb	Dessert Shop	45.814987	15.976296

Figure – 3

We performed a bit of data cleansing. It is seen through above figure (Figure -3) that the neighbourhoods are grouped by the name of the neighbourhood, so data clustering is made easier later on.

In order to analyse the data we must require type of restaurants that contain a particular venue for a neighbourhood. After an exploratory data analysis we get the mean(how often the Italian restaurants located) of Italian restaurants to their respective neighbourhood which we will further use to cluster them and make a decision for best neighbourhood in the city of Bangalore.

	Neighborhood	ATM	Accessories Store	Afghan Restaurant	Andhra Restaurant	Arcade	Arts & Entertainment	Asian Restaurant	Athletics & Sports	Auto Garage	...	Theme Park Ride / Attraction	Toy / Game Store	Train Station	T
0	Achitnagar	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	
1	Adugodi	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.166667	0.0	...	0.000000	0.000000	0.0	
2	Agram	0.0	0.0	0.0	0.0	0.0	0.0	0.033333	0.000000	0.0	...	0.033333	0.000000	0.0	
3	Akkur	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.066667	0.0	
4	Alahalli	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
97	Virgonagar	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.5	...	0.000000	0.000000	0.0	
98	Whitefield	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	
99	Yelachenahalli	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	

Frequency of Categories per neighbourhood

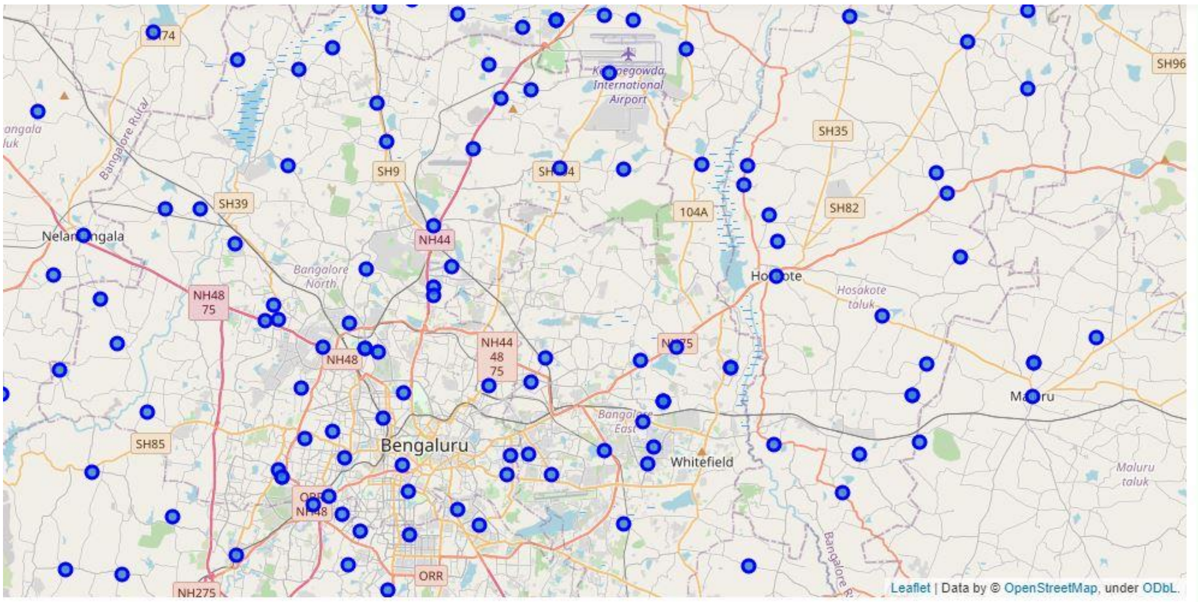
Methodology:

Data Cleaning:

After understood a business domain problem and collected the data required to start the analysis process using panda's data frames cleaning plays a crucial role, the data being acquired from kaggle dataset it is cleaned.

Exploratory Data Analysis:

Using the data we gathered then created a map using Folium and displaying all the neighbourhood’s it was located in.



Next, we used the Foursquare API to get a list of all the venues in Bangalore which include cafe shops, post offices; railway stations Spanish Restaurants etc..,

Getting this data was crucial to analysing the number of Italian Restaurants all over the Bangalore. There were 5 Italian restaurants in Bangalore. We then merged the foursquare venue data with the neighbourhood data which gave us the nearest venue for each of the neighbourhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	venue	venue_categories	venue Latitude	venue longitude
0	Agram	45.813177	15.977048	Amélie	Dessert Shop	45.813842	15.979011
1	Agram	45.813177	15.977048	Corner bar	Bar	45.812930	15.979440
2	Agram	45.813177	15.977048	Trg bana Josipa Jelačića	Plaza	45.813032	15.976868
3	Agram	45.813177	15.977048	Tržnica Dolac	Farmers Market	45.814070	15.977261
4	Agram	45.813177	15.977048	Cremme Zagreb	Dessert Shop	45.814987	15.976296

Venue table merged with Neighborhood data

Modelling :

In order to analyse the data categorical data must converted to a numerical using a technique of one hot encoding . For each of the neighborhoods, individual venues were turned into the frequency at how many of those venues were located in each neighbourhood. If the neighbourhood contains venue it indicate with ‘1’ else with ‘0’.

	Neighborhood	venue_categories_ATM	venue_categories_Accessories Store	venue_categories_Andhra Restaurant	venue_categories_Arcade	venue_categories_Arts & Entertainment	venue_cate
0	Agram	0	0	0	0	0	
1	Agram	0	0	0	0	0	
2	Agram	0	0	0	0	0	
3	Agram	0	0	0	0	0	
4	Agram	0	0	0	0	0	

One hot encoding of venues corresponding to Neighborhoods

Then we grouped those rows by Neighborhood and calculated the mean of frequency of occurrences of each venue category.

	Neighborhood	ATM	Accessories Store	Afghan Restaurant	Andhra Restaurant	Arcade	Arts & Entertainment	Asian Restaurant	Athletics & Sports	Auto Garage	...	Theme Park Ride / Attraction	Toy / Game Store	Train Station	T
0	Achitnagar	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	
1	Adugodi	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.166667	0.0	...	0.000000	0.000000	0.0	
2	Agram	0.0	0.0	0.0	0.0	0.0	0.0	0.033333	0.000000	0.0	...	0.033333	0.000000	0.0	
3	Akkur	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.066667	0.0	
4	Alahalli	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
97	Virgonagar	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.5	...	0.000000	0.000000	0.0	
98	Whitefield	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	
99	Yelachenahalli	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	

Grouped Neighborhood by the mean of frequency of each venue

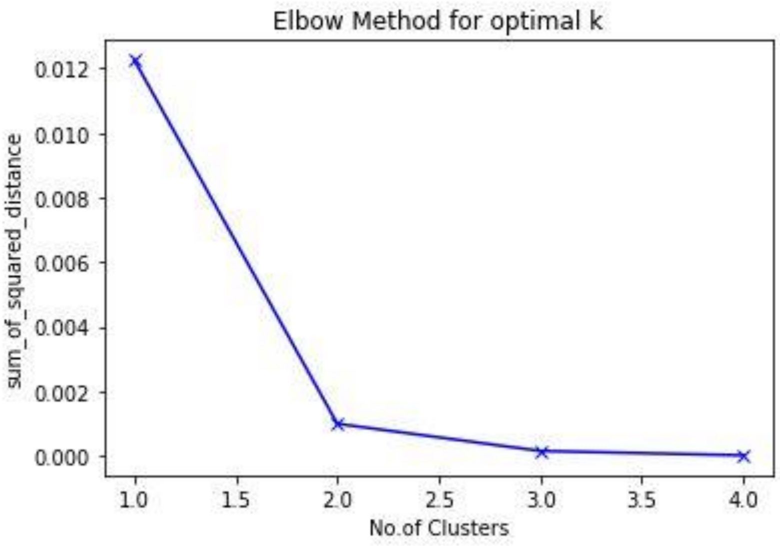
After, we created a new data frame that only stored the Neighbourhood names as well as the average frequency of Italian Restaurants in that Neighbourhood. This allowed the data to be summarized based on each individual Neighbourhood and made the data much simpler to analyse.

	Neighborhood	Italian Restaurant
0	Achitnagar	0.000000
1	Adugodi	0.000000
2	Agram	0.066667
3	Akkur	0.000000
4	Alahalli	0.000000
...	...	...
97	Virgonagar	0.000000
98	Whitefield	0.000000
99	Yelachenahalli	0.000000
100	Yelahanka	0.000000
101	Yeliyur	0.000000

New data frame storing Neighborhoods and the mean of Italian Restaurant in that Neighborhood

K-Means Clustering:

To make the analysis more interesting, we wanted to cluster the neighbourhoods based on the neighbourhoods that had similar averages of Italian Restaurants in that Neighbourhood. To do this we used **K-Means** clustering. To get our optimum K value that was neither over fitting nor under fitting the model, we used the **Elbow Point** Technique. In this technique, we ran a test with different number of K values and measured the accuracy and then choose the best K value. The best K value is chosen at the point in which the line has the sharpest turn. In our case, we had the Elbow Point at K = 2. That means we will have a total of 2 clusters.



Finding optimal K-value (Elbow Method)

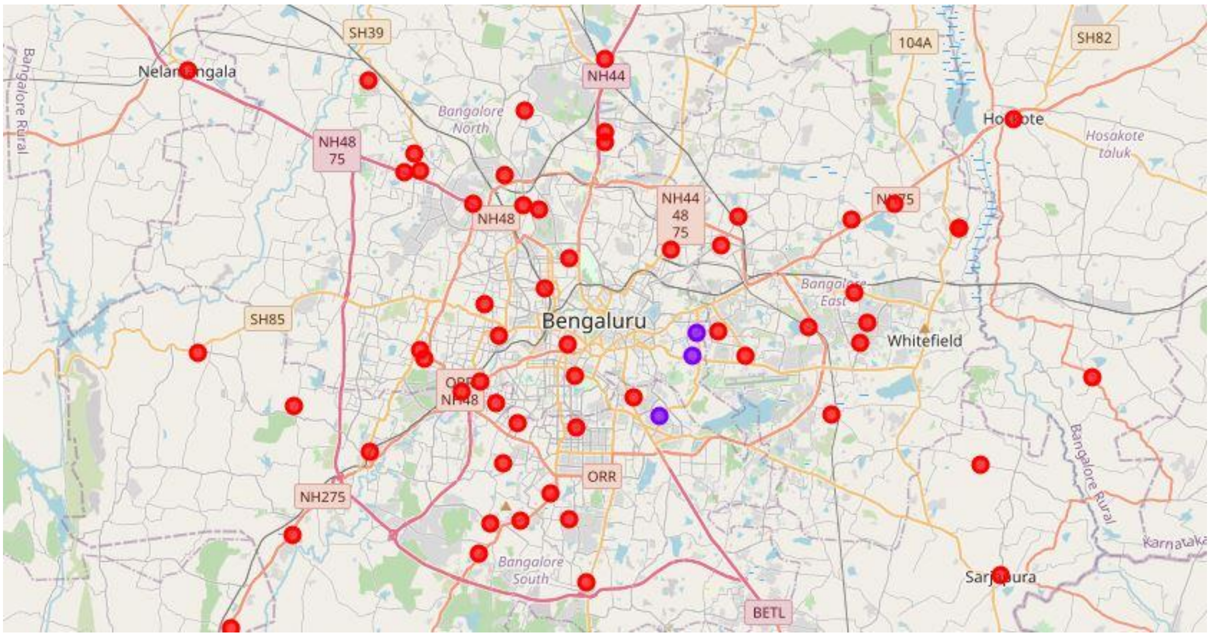
In k-Means clustering it segmented & cluster the data based on similarities of data. Neighbourhoods that had a similar mean frequency of Italian Restaurants were divided into 2 clusters. Each of these clusters labelled as 0 and 1.

	labels	Italian Restaurant	Neighborhood
0	0	0.000000	Achitnagar
1	0	0.000000	Adugodi
2	1	0.066667	Agram
3	0	0.000000	Akkur
4	0	0.000000	Alahalli

Corresponding cluster labels were added

After, we merged the venue data with the table above, creating a new table which would be the basis for analysing new opportunities for opening a new Italian Restaurant in Bangalore. Then we created a map using the Folium package in Python and each neighbourhood was coloured based on the cluster label.

- Cluster 1 – Red
- Cluster 2 – Blue

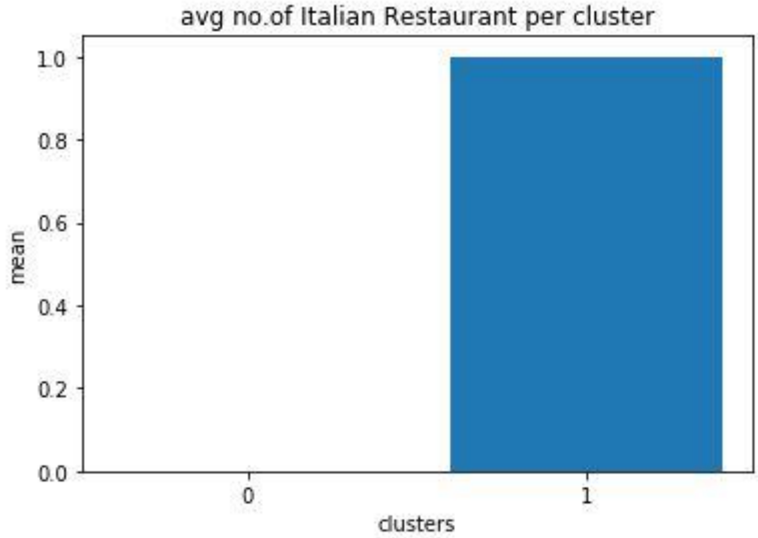


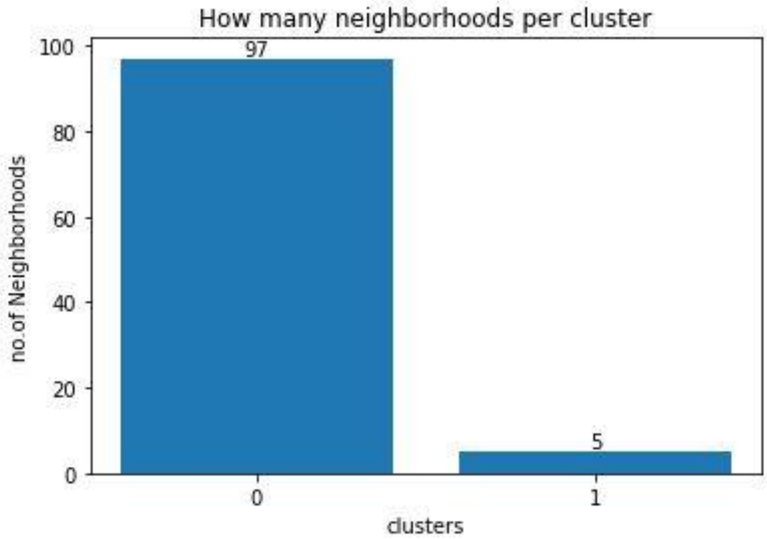
Map with clusters

The above map shows the two clusters that had a similar mean frequency of Italian Restaurants.

Data Analysis:

We have a total of 2 clusters (0,1). Before we analyse them one by one let's check the total amount of neighbourhood's in each clusters and the average Italian Restaurants in that cluster. From the bar graph that was made using Matplotlib , we can compare the number of Neighborhoods per Cluster. We see that Cluster 1 has the no neighborhoods that contain Italian restaurants while cluster 2 has the least (5). Then we compared the average Italian Restaurants





no.of Neighborhoods per clusters

Cluster Analysis:

This information is vital from the bar graph of average number of Italian Restaurants per cluster, we can see that cluster 1 has no Italian restaurants while cluster 2 contains a few Italian restaurants. From map we can see that neighbourhoods in cluster1 are more widely populated but cluster2 are a little. Now let's analyse the each cluster individually.

Cluster 1 (Red):

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	venue	venue_categories	venue Latitude	venue longitude	City	Latitude	Longitude	labels	Italian Restaurant
30	Amruthahalli	13.066513	77.596624	Reliance Fresh	Convenience Store	13.066264	77.596961	Bangalore	13.066513	77.596624	0	0.0
31	Amruthahalli	13.066513	77.596624	State Bank of India ATM	ATM	13.066130	77.598380	Bangalore	13.066513	77.596624	0	0.0
32	Amruthahalli	13.066513	77.596624	Shoba Hypermarket	Department Store	13.066882	77.595574	Bangalore	13.066513	77.596624	0	0.0
33	Amruthahalli	13.066513	77.596624	Sip N Crunch	Juice Bar	13.066003	77.597030	Bangalore	13.066513	77.596624	0	0.0
34	Amruthahalli	13.066513	77.596624	Bata	Shoe Store	13.065699	77.599402	Bangalore	13.066513	77.596624	0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
592	Vidyanagara	13.018496	76.108893	Sunrise Cafe and Restaurant	Café	13.018644	76.109861	Bangalore	13.018496	76.108893	0	0.0
593	Vidyanagara	13.018496	76.108893	Apollo Pharmacy	Pharmacy	13.017034	76.107567	Bangalore	13.018496	76.108893	0	0.0

Cluster 1 contain no Italian Restaurants but had 124 unique venue categories with 386 venues and having a least average of Italian Restaurants. Since it doesn’t contain any Italian Restaurants.

Cluster 2 (Blue):

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	venue	venue_categories	venue Latitude	venue longitude	City	Latitude	Longitude	labels	Italian Restaurant
23	Agram	45.813177	15.977048	Boban	Italian Restaurant	45.811613	15.976522	Bangalore	45.813177	15.977048	1	0.066667
29	Agram	45.813177	15.977048	Carpaccio	Italian Restaurant	45.811286	15.974799	Bangalore	45.813177	15.977048	1	0.066667
81	Domlur	12.962467	77.638196	Spaghetti Kitchen	Italian Restaurant	12.964539	77.641657	Bangalore	12.962467	77.638196	1	0.047619
132	Indiranagar S.O (Bangalore)	12.973291	77.640467	Chianti Ristorante & Wine Bar	Italian Restaurant	12.970167	77.640346	Bangalore	12.973291	77.640467	1	0.033333
220	Begur	52.480709	13.451829	Pipaso	Italian Restaurant	52.478816	13.445998	Bangalore	52.480709	13.451829	1	0.062500
298	Koramangala	12.934011	77.622230	Chianti	Italian Restaurant	12.933537	77.621862	Bangalore	12.934011	77.622230	1	0.033333
469	Begur	52.480709	13.451829	Pipaso	Italian Restaurant	52.478816	13.445998	Bangalore	52.480709	13.451829	1	0.062500

Cluster 2 contains only 5 unique neighborhoods which contain only Italian Restaurants with highest mean of 0.06667. In the map we can see that a few blue dots spread in map.

Distance Measure:

After cluster analysis the most interesting part comes that is recommending places in order to that we must calculate the distance s of the neighborhoods from the city centre.

	Neighborhood	Dist_frm_center-neigh
20	Jakkur	11.96 km
21	Kadugodi	19.40 km
22	Kalkunte	206.74 km
23	Kannamangala	28.19 km
24	Kodigehalli	14.80 km
25	Kothanur	51.92 km
26	Krishnarajapuram	92.49 km
27	Kundalahalli	13.46 km
28	Lingarajapuram	5.74 km
29	Mahadevapura	199.45 km
30	Medimallasandra	25.61 km

Distance Data Frame from city centre

Before, we recommend the best places. Let's explore the cluster 2, to add distance corresponding to each neighbourhood merge the distance dataframe with new cluster data frame.

	Neighborhood	Dist_frm_center-neigh
0	Agram	6790.05 km
1	Domlur	4.83 km
2	Indiranagar S.O (Bangalore)	4.97 km
3	Begur	7127.08 km
4	Koramangala	5.14 km

From above, we can see that distance of Agram and Begur are much higher, from the initial data we acquired from kaggle having a wrong latitude and longitudes for those neighbourhood's, data cleaning not performed. So, we are going to remove those Neighborhoods.

	Neighborhood	Dist_frm_center-neigh	N
0	Domlur	4.83 km	
1	Indiranagar S.O (Bangalore)	4.97 km	
2	Koramangala	5.14 km	

After cleaning the data, now explore the venues that contain Italian Restaurants and distance from each neighbourhood.

	Neighborhood	Dist_frm_center-neigh	venue	Dist_from_neigh-venue
0	Domlur	4.83 km	Spaghetti Kitchen	0.44 km
1	Indiranagar S.O (Bangalore)	4.97 km	Chianti Ristorante & Wine Bar	0.35 km
2	Koramangala	5.14 km	Chianti	0.07 km

From above we can see that distances of neighbourhood from city centre and neighbourhood to venues. The Italian Restaurants are available less than 6km from city centre.

To recommend the places, it must be near to the centre of Bangalore and there will be less or no any type of restaurants. For analysis, we will merge the cluster 2 data with distance data frame which we got as a new data frame.

	Neighborhood	Dist_frm_center-neigh	Neighborhood Latitude	Neighborhood Longitude	venue
0	Jalahalli H.O	9.71 km	13.046453	77.548380	Kadamba Gardenia
1	Jalahalli H.O	9.71 km	13.046453	77.548380	Kanti Sweets (Jalahalli)
2	Jalahalli H.O	9.71 km	13.046453	77.548380	D Needs
3	Jalahalli H.O	9.71 km	13.046453	77.548380	New Arya Bhavan
4	Konanakunte	9.66 km	12.886018	77.579141	Kabab Mahal

To know which neighbourhood contain how many type of any restaurants, we calculated no.of any type restaurants from above data frame.

	Neighborhood	No.of any restaurants
0	Jalahalli H.O	3
1	Konanakunte	1
2	Peenya Dasarahalli	2
3	Horamavu	2
4	Nagarbhavi	1
5	Mallathahalli	6

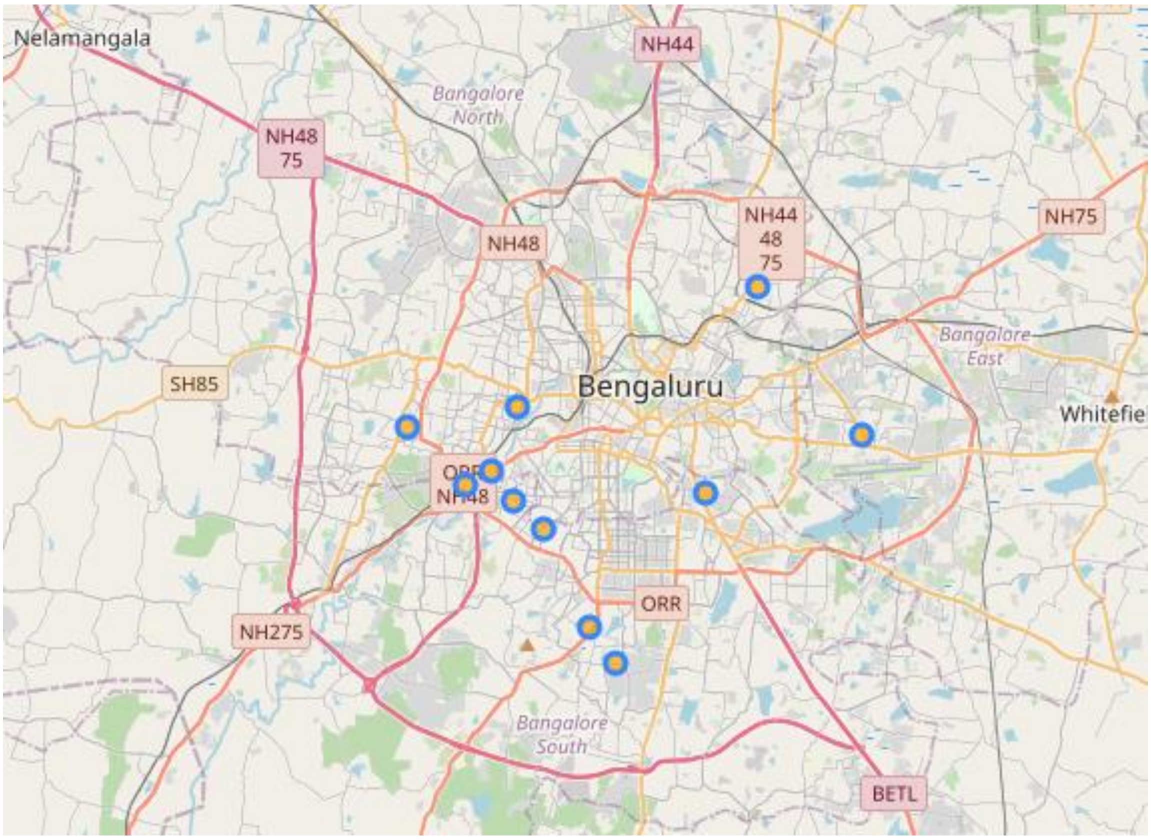
From above figure we can see that neighbourhood 'Jalahalli H.O' contains 3 restaurants which we may consider as a competitors from other type similarly to all the neighborhoods. By further analysis we created a new data frame which illustrates the count of no.of any restaurants.

	Neighborhood
No.of any restaurants	
1	9
2	4
3	4
5	3
6	1
7	1
8	1
13	1

This data frame shows that number of neighbourhood that contain different type of any restaurants, we can see that there are 9 Neighbourhoods having a single restaurant which we may consider as a recommended places because there are least competitors and also there are two places namely 'Deepanjalinagar' and 'Girinagar S.O' not having any type of restaurants.

Result:

Let's visualize map of recommended places:



And finally the recommended places are:

```
Adugodi                         => 3.63 km
Vijayanagar S.O (Bangalore)     => 5.29 km
Lingarajapuram                  => 5.74 km
Kathriguppe                     => 6.23 km
Girinagar S.O (Bangalore)       => 6.46 km
Deepanjalinagar                 => 6.71 km
Vimanapura                      => 7.54 km
Nayandahalli                    => 7.80 km
Yelachenahalli                  => 8.61 km
Nagarbhavi                      => 9.43 km
Konanakunte                     => 9.66 km
```

Discussion Section:

There are a few Italian restaurants available in cluster 2 which represented with blue color. After analyses of cluster 1 distance measure the optimum place to pu t a new Italian Restaurant in Bangalore as there are many neighborhoods in the area but little to no Italian Restaurants, therefore eliminating any competition th e first best place that have a great opportunity would be Deepanjalinagar ang se cond best place will be vijay nagar. Some of the drawbacks of this analysis are — the clustering is completely based on data obtained from the Foursquare API. Also, the analysis does not take into consideration of the Italian population acros s neighbourhoods as this can play a huge factor while choosing which place to o pen a new Italian restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Italian restaurant in these locations with little to no competition.

Conclusion:

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in a way that it was similar to how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information, control the content and break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighbourhoods of Bangalore, get a great measure of data from Kaggle. We visualized utilizing different plots present in Matplotlib libraries. Similarly, we applied AI strategy to anticipate the error given the information and utilized Folium to picture it on a map.

Places that have room for improvement or certain drawbacks give us that this project can be additionally improved with the assistance of more information and distinctive Machine Learning strategies. Additionally, we can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Jewellery shop and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data science.