Restaurant Location Prediction Bengaluru, India

By : Dony George

April 26, 2020

1. Business Problem

1.1 Background

Whitefield in Bengaluru is a prominent place in India's Silicon Valley. The area is occupied by bachelor IT professional from all over India. The area has a lot of food joints and I have a friend who wants to start a business. He is set on starting a restaurant and has asked for my help to figure out what kind and where to put it. He doesn't care what type of restaurant, but he wants it to be successful.

1.2 Solution

I have decided to help him by using Foursquare data to cluster the current restaurants around Whitefield and identify restaurant types that are well established in one segment, but not in another. When complete, I hope to be able to offer him a location and type of restaurant suggestion based on current data, which he can then take and start doing more research on his own.

2. The Data

2.1 Data sources

This project relies entirely on Foursquare data. I have used the “Food” category within the Foursquare data with starting point for data as Whitefield.

2.2 Feature selection

There are only 4 fields within the results which we need to start our analysis. The name of the venue, the category of the venue, and the latitude and longitude (location) of the venue. The location is critical because we will be doing k-means clustering based on the locations of each restaurant.

2.3 Data cleaning

We make API calls to collect Foursquare data and get the results in JSON format. These results have a lot of meta-data which needs to be removed before we can effectively use the data. Additionally, we want to get that information into a dataframe to leverage different python packages in the analysis. The final result of this data cleaning process is a dataframe with the four desired columns ready for to explore.

3. Exploratory Data Analysis

3.1 Understand the data

Since I approached this project with an intentionally broad scope regarding the outcome, I found it very valuable to visualize the data as I went through the process. Additionally, I collected some different statistics regarding the data. I spent a lot of time slicing and viewing this data in an effort to ‘gut-check’ each step as I went.

3.1.1 Can the data give us the answer we are looking for

At a high level, I was able to see that the area around Whitefield has 95 different restaurants that fell into 32 categories. This was an important insight because it allowed me to see that there would be some good groupings among the restaurants.

3.1.2 Visualize the data

The next step in my exploration was to visualize the 95 restaurants on a map. I performed this step to see if clustering still felt like the right approach. Though the data is pretty spread out, as I reviewed the map, I was able to see that there were definite areas where I could see the data clustering. In Figure 1, the red circle is the location of Whitefield which we used as the starting point. The blue circles are the locations of each restaurant returned by Foursquare.

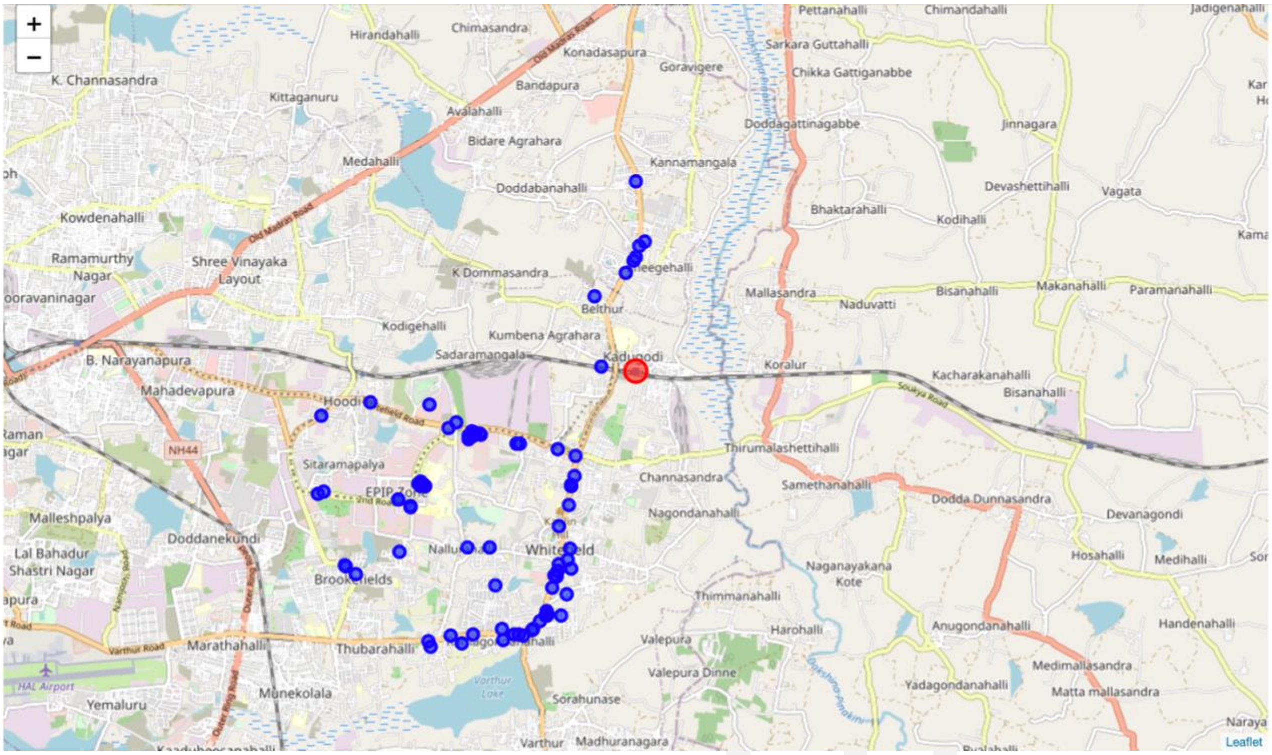


Figure 1 – Restaurants around Whitefield (red).

3.2 How many clusters

I have established that the data seems suitable for clustering, but one challenge with k-means clustering is determining how many clusters to use. One approach here is called the “Elbow Method”. In this method, I visualize the sum of squared differences within the latitude and longitude over different values of k. The resulting chart should look like a bent arm and the location of the ‘elbow’ would represent the optimal k. The principle here is that the steeper

portion of the graph has significant reductions in the sum of squared differences, but after the
'elbow' the improvements in reductions becomes less and less for each additional k. The best result is a very distinct elbow, but this was not the case in my data. (This can happen with data that does not have very distinct clusters.) Figure 2 shows the results from performing the Elbow Method.

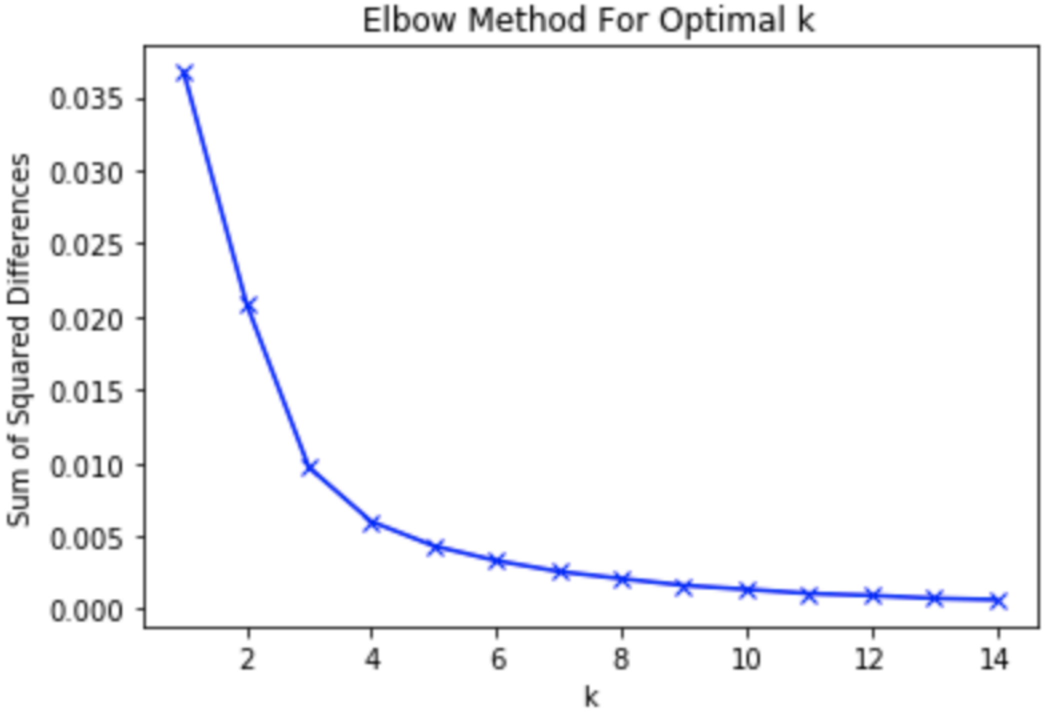


Figure 2 – Elbow Method to determine optimal k.

I was not satisfied with a distinct break within this chart. I see a break at k=2, but I also still see pretty significant reductions in the sum of squared differences between k = 2 and 4. So rather than looking at the raw data, I looked at the rate of change and determined that after k=4 there was a lot less improvement, and so I settled on using k=4 for further analysis.

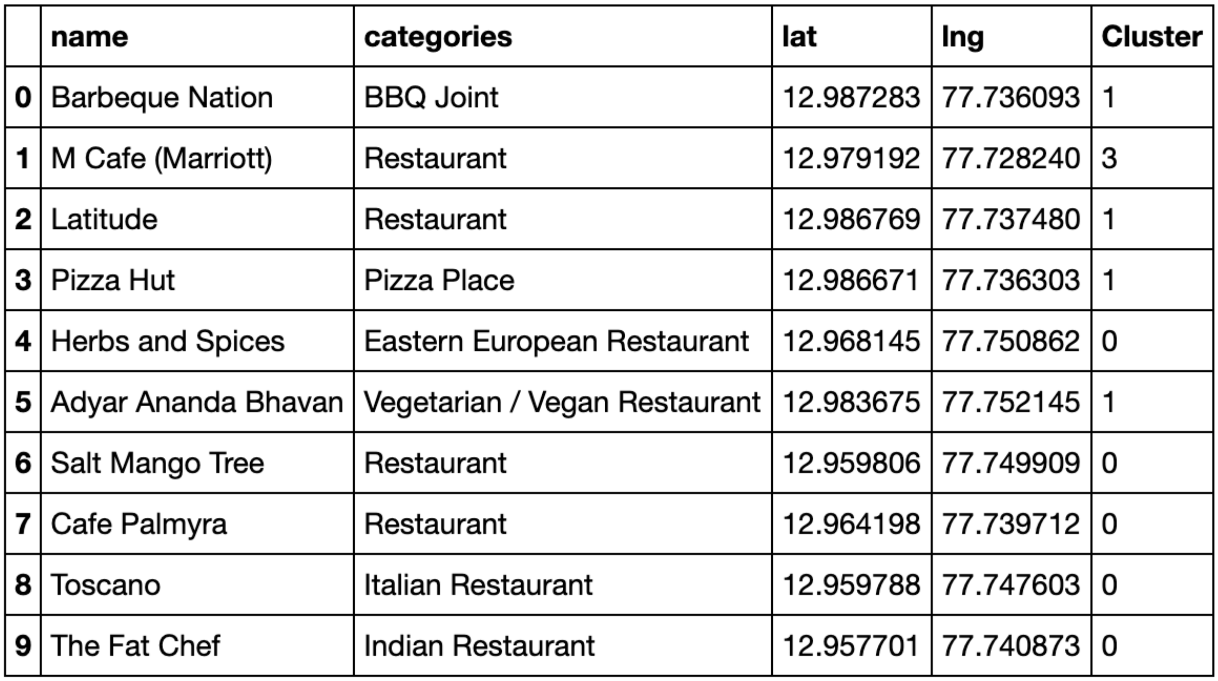
4. Predictive Modeling

4.1 Classification models

4.1.1 Setting up k-means clustering

Given the optimal k, I used the scikitlearn python package to perform 15 iterations of k-means clustering in an attempt to find the optimal centroids for each of my four clusters.

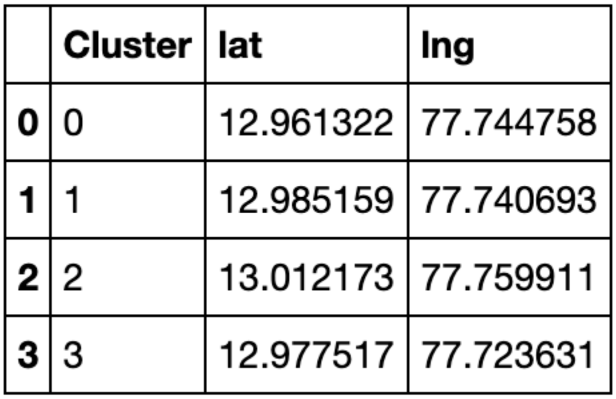
To perform this step, I isolated only the latitude and longitude of each restaurant and ran k-means clustering with k=4 and n_init = 15. The array that is returned by this process labels each of the restaurants with a category 0 through 3. From there, I add these labels back to my original dataframe to see which cluster each restaurant belongs in. Figure 3 shows my current dataframe with this information added.



	name	categories	lat	lng	Cluster
0	Barbeque Nation	BBQ Joint	12.987283	77.736093	1
1	M Cafe (Marriott)	Restaurant	12.979192	77.728240	3
2	Latitude	Restaurant	12.986769	77.737480	1
3	Pizza Hut	Pizza Place	12.986671	77.736303	1
4	Herbs and Spices	Eastern European Restaurant	12.968145	77.750862	0
5	Adyar Ananda Bhavan	Vegetarian / Vegan Restaurant	12.983675	77.752145	1
6	Salt Mango Tree	Restaurant	12.959806	77.749909	0
7	Cafe Palmyra	Restaurant	12.964198	77.739712	0
8	Toscano	Italian Restaurant	12.959788	77.747603	0
9	The Fat Chef	Indian Restaurant	12.957701	77.740873	0

Figure 3 – nearby_venues dataframe with the addition of "Cluster"

From here, I need to identify the centroids themselves. This process sets the centroid for each cluster at the center of each cluster, so the easiest way to identify the location of the centroids is just to look at the mean latitude and longitude values for each of the clusters. Figure 4 shows the results of the means of each cluster.



	Cluster	lat	lng
0	0	12.961322	77.744758
1	1	12.985159	77.740693
2	2	13.012173	77.759911
3	3	12.977517	77.723631

Figure 4 – Centroid Locations

4.1.2 Visualizing the clusters

At this point, it's important to validate that the steps I've taken up to this point make sense. The easiest way to do this is to visualize the previous map we produced and color-code by cluster. Figure 5 is the updated map. Whitefield and the centroids are not part of my data from Foursquare and have been coded black and pink to offset them from the rest of the data. Whitefield is pink circle, black fill. Centroids are black circle, pink fill. The venues themselves are rainbow colored, assigned dynamically in the plot. Cluster 0 is red, cluster 1 is purple, cluster 2 is teal, and cluster 3 is olive

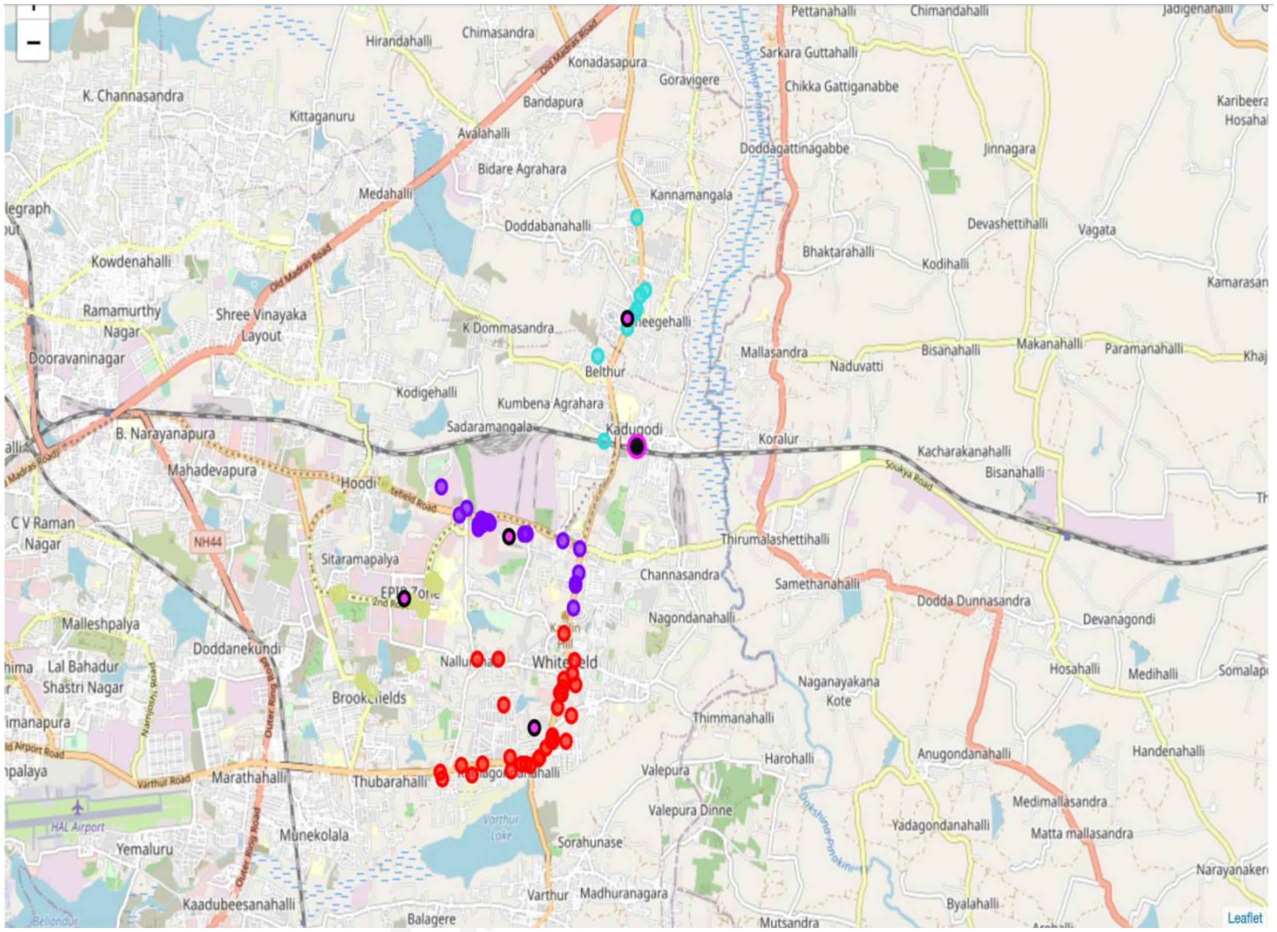
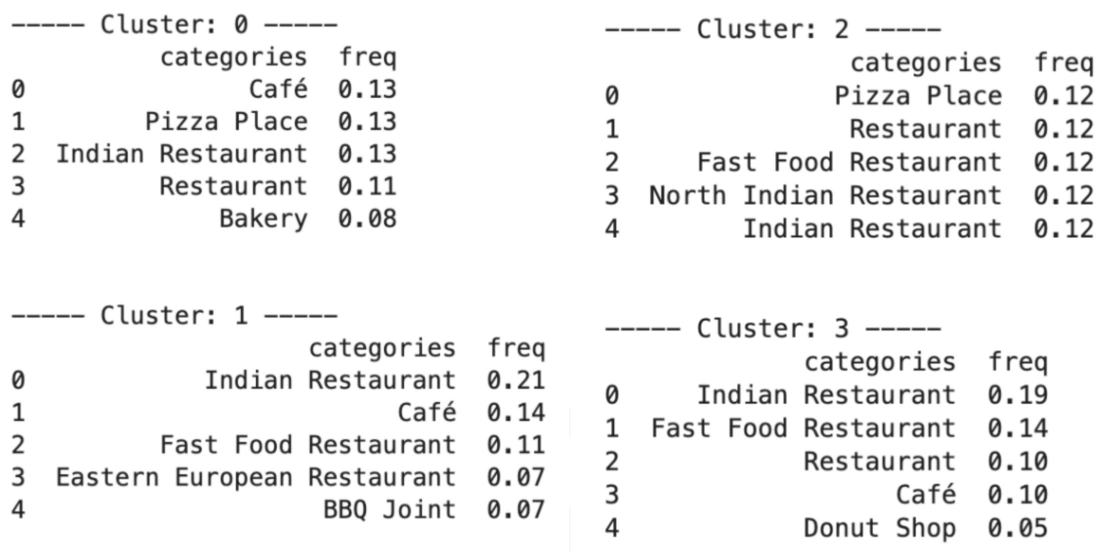


Figure 5 – Four clusters of restaurants in Whitefield, Bengaluru.

4.1.3 Examining the clusters

The overall goal is to determine where a restaurant category is under-represented in one cluster, but highly represented in other clusters. From there, we can establish that X type of restaurant should be located in Y cluster. I need to examine the clusters more closely to properly determine the correct X and Y in this situation.

To do this, I need to ‘one-hot’ encode all the categories into dummy variables so I can look at the mean representation of each category across each of the clusters. From there, I am able to report on the top 5 categories for each of the clusters



```
----- Cluster: 0 -----
         categories  freq
0              Café  0.13
1       Pizza Place  0.13
2  Indian Restaurant  0.13
3        Restaurant  0.11
4            Bakery  0.08
```

```
----- Cluster: 1 -----
                  categories  freq
0          Indian Restaurant  0.21
1                       Café  0.14
2       Fast Food Restaurant  0.11
3  Eastern European Restaurant  0.07
4                   BBQ Joint  0.07
```

```
----- Cluster: 2 -----
                 categories  freq
0               Pizza Place  0.12
1                Restaurant  0.12
2      Fast Food Restaurant  0.12
3  North Indian Restaurant  0.12
4         Indian Restaurant  0.12
```

```
----- Cluster: 3 -----
              categories  freq
0      Indian Restaurant  0.19
1   Fast Food Restaurant  0.14
2             Restaurant  0.10
3                   Café  0.10
4             Donut Shop  0.05
```

Reviewing the frequencies of each category between the clusters reveals that Indian Restaurants are the #1 most common venue in clusters 1 & 3 with high freq, but in clusters 0 & 2, Indian Restaurants are yet to show their prominence in their respective clusters.

At this point, I can recommend my friend to look into an Indian Restaurants within the area defined as Cluster 0 & 2.

5. Conclusions

In this study, I used Foursquare data to determine the restaurant landscape for Whitefield, Bengaluru, India. With this information I was able to tell my friend to look into an Indian Restaurant within the area defined as Cluster 0 & 2.

6. Full-Disclosure

I had help from many of the Data Science forums to get to the final point in this projects. Thanks to them all