Coursera Capstone Project

Hotel Location Prediction at Whitefield,Bengaluru

By: Dony George

INTRODUCTION

Whitefield in Bengaluru is a prominent place in India's Silicon Valley. The area is occupied by bachelor IT professional from all over India

A city this size has a lot restaurants and variety.

PROBLEM

- Success of a restaurant is often heavily influenced by location.
- My client 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.
- Building the wrong restaurant in a good location does not guarantee success.

SOLUTION

- Using Foursquare data, we can determine the current distribution of restaurants in an area to identify under-served restaurant types.
- Segmenting the current restaurants will allow us to identify the best locations for a specific under-served type of restaurant.

DATA SOURCES AND CLEANING

The core data source for this project is Foursquare data with "Food" as the top-level category.

Feature Selection

- Name (to identify uniqueness)
- Category
- Location (latitude, longitude)

Data Cleaning

API calls made to Foursquare return JSON data. It's necessary to strip out much of what is returned to isolate only the features mentioned in Feature Selection.

Tools

Python packages:

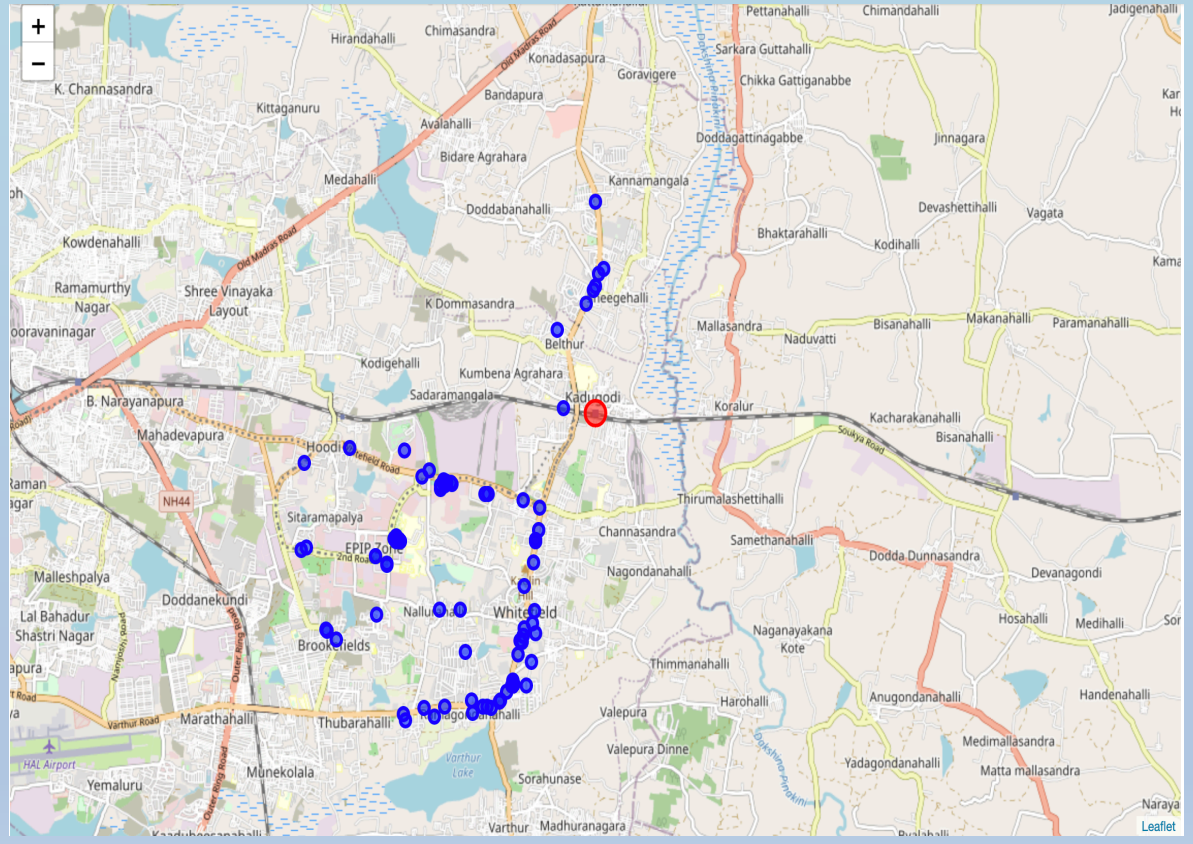
- pandas (for dataframes and analysis)
- numpy (to help handle the data)
- scikitlearn (for k-means clustering)
- matplotlib (to create visuals)

EXPLORATORY DATA ANALYSIS

Starting Location

We pull data from Foursquare based on a radius from a given location.

.



Visualize

With all our data into a simple dataframe, what does a map of the current restaurants around Whitefield look like?

This is a useful step to get an impression if clustering will be the right approach.

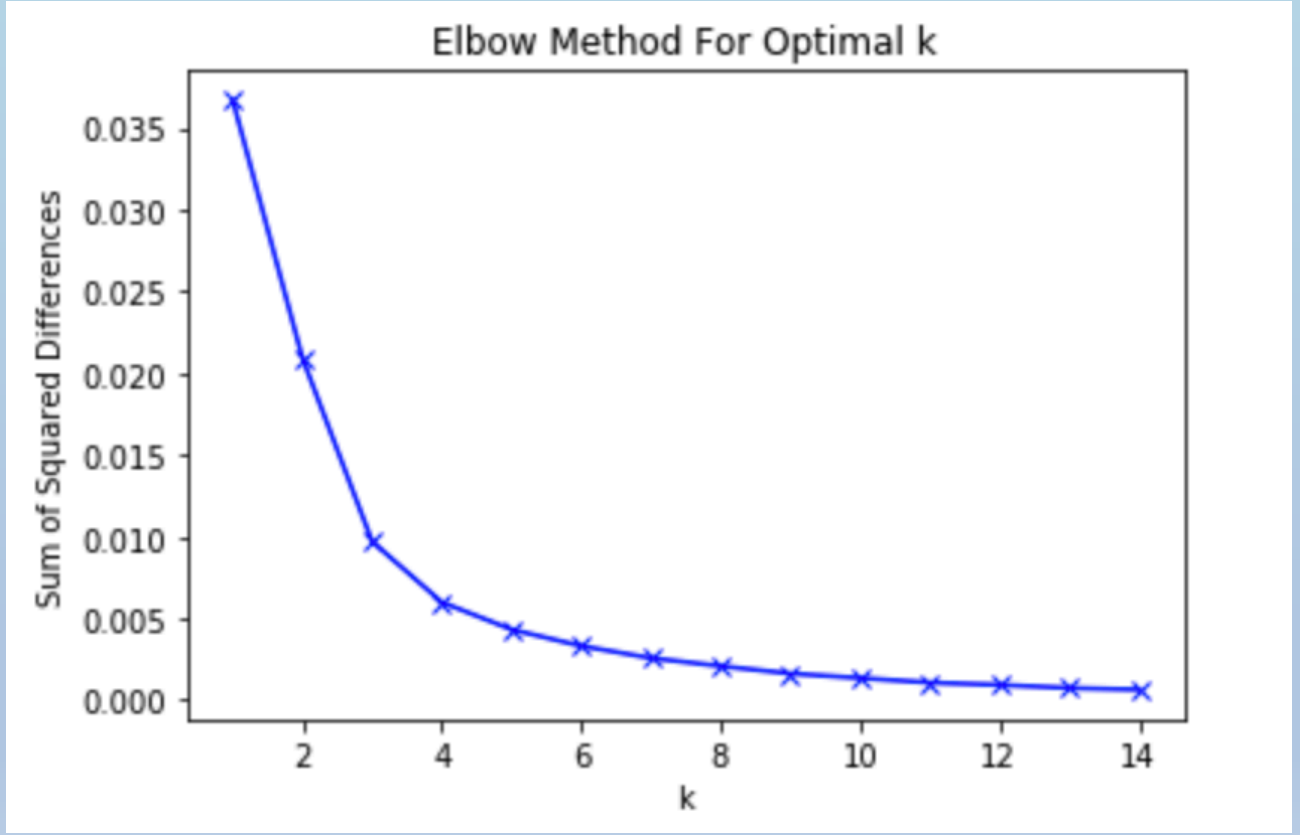
k-MEANS CLUSTERING

How Many Clusters

I did not want to pick an arbitrary value for *k*, so I utilized the Elbow Method to help determine the optimal *k*.

In this method, I plot the sum of squared differences (ssd) and determine the *k* where there is less value in adding *k* based on the reduction in sum of squared differences.

I determined for my data, the appropriate *k* was four. This means, I have 4 distinct groups of restaurants in the Macon area which I will use to compare.



PREDICTIVE MODELING

Setting up *k*-means clustering

k = 4

Using scikitlearn, we perform Kmeans clustering using 4 to start segmenting the current restaurant locations into 4 distinct clusters.

n_init = 15

Because we randomly drop in centroids and seek the optimal location, we can get different outcomes if we run the process multiple times.

Running this 15 times helps us truly come to an optimal location for a centroid within each cluster.

PREDICTIVE MODELING

Results of *k*-means clustering

- Performing k-means clustering on our dataset will assign each restaurant a label between 0 and 3.
- The label represents their cluster.
- This information is built back into the original dataframe so we have an easy to read table of the location name and the cluster it belongs to.

The Current Dataframe (head)

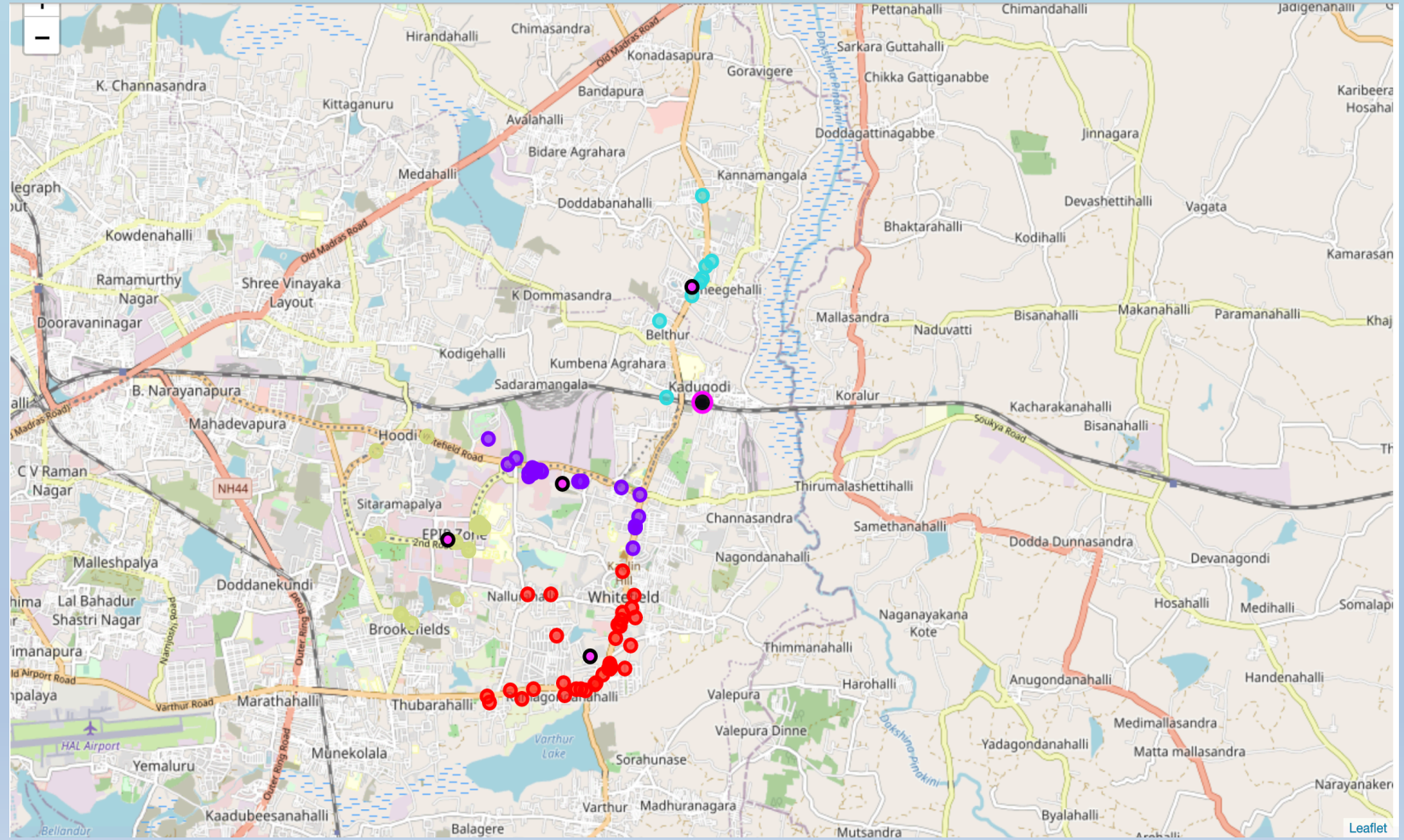
	name	categories	lat	lng	Cluster
0	Barbeque Nation	BBQ Joint	12.987283	77.736093	1
1	M Cafe (Marriott)	Restaurant	12.979192	77.728240	3
2	Latitude	Restaurant	12.986769	77.737480	1
3	Pizza Hut	Pizza Place	12.986671	77.736303	1
4	Herbs and Spices	Eastern European Restaurant	12.968145	77.750862	0
5	Adyar 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

VISUALIZE THE CLUSTERS

Cluster Information

- **Whitefield:** pink circle, black fill
- **Cluster Centroids:** black circle, pink fill

- Clusters are color coded dynamically.
  - **Cluster 0** – Red (north-northwest)
  - **Cluster 1** – Purple (west)
  - **Cluster 2** – Teal (south)
  - **Cluster 3** – Olive (northeast)



EXAMINING THE CLUSTERS

FREQUENCY

Rank the categories

- Need to determine where a category is under-represented.
- Ranking the categories by frequency within each cluster offers a way to see this.
- My approach looks at the top 5 most common types of restaurants within a given cluster.

ONE-HOT ENCODE

Categories to Dummies

- Need dummy variables instead of categorical variables to determine frequency.
- One-hot encoding of the variables effectively translates the categorical variables into dummy variables.
- From here we can look at the mean frequency of a category across the cluster.

GOAL

Under-Represented

- In this situation, under-represented will be determined by the following criteria:

1. Category is represented in the top 5 for at least 2 other clusters.
2. Category is not in the top 5 for a selected cluster.

- If those criteria are met, I will select the most represented category based on criteria 1 and the cluster which applies to criteria 2.

THE RESULTS

Indian Restaurant

This 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.egory in cluster 1, 2 and 3 (often by a large margin).

Conclusion

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 client to look into an Indian Restaurant within the area defined as Cluster 0 & 2.

.