Coursera Capstone

IBM Applied Data Science Capstone

Analyzing Restaurants in Bangalore

Introduction

Bangalore is one of the few metropolitan cities in India which has a lot to offer in various perspectives such as food, entertainment, job opportunities, tourism and whatnot. I'll be taking a small subset of this inexhaustive list and try to solve a problem that could be possibly solved or made easier with the help of Data Science and Machine Learning tools at our disposal.
For a city this big a size, it is always going to be difficult for a visitor or even a resident to decide upon an place to eat in which should be both affordable and well-rated. With so many options available at hand, making such a decision is always a difficult ask because not all of us can afford to visit all the luxurious places and while others don't have the patience to have a try them all one by one.

Problem

According to the problem already mentioned above we'll try formulating it as such so that it could be answered feasibly with the help of Data Science. The problem can be classified as differentiating out the varying varieties of restaurants available in Bangalore and whether we can group similar restaurants together which have a similar rating and a relatively affordable price for various ranges so as that it could fit into certain categories and the decision of visiting these restaurants could be made easier.

Data

To solve the following problem we'll be making use of the following data:

- Fetching the data from Geopy for getting the coordinates of Bangalore (since Geopy seems to be down we'll just do a Google Search)

- Data from the Foursquare API making appropriate calls to get a list of Trending venues across a radius of about 9kms.
  This data would contain information about venues which are not restaurants as well so we'll have to perform a data cleaning process and get rid of such entries.

- Data from the Zomato API as well, which would contain the restaurants (which will be used to drop the non restaurant venues from the Foursquare data) and contain the ratings and average price of these restaurants as well which form the backbone of the analysis.

Methodology

The first step to approach this problem was to collect the data which has already been described above. Then the next most important thing was to analyze what the data from from the respective APIs actually contained and how could they be used in a proper way.

The “explore” endpoint for the Foursquare API was pretty useful to get the venues according to the parameters we set and similarly the fetched data in turn helped us to get the corresponding data from the Zomato API as well.

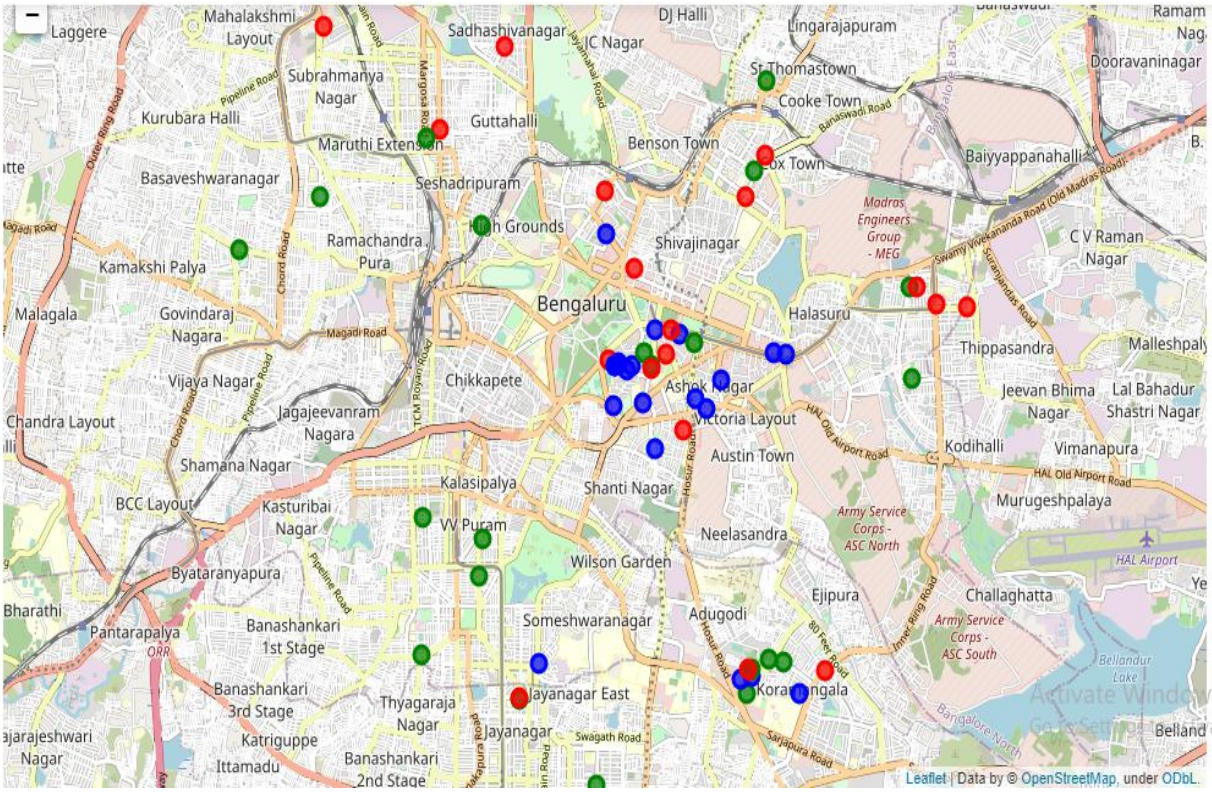
The latitude and longitude data from both of these datasets were pretty helpful to accurately cluster the restaurants according to their correct geographic location as well. This was easily represented visually on a map through the Folium library.

The average price and ratings of the restaurants from the Zomato API formed the features of to cluster the restaurants together with the help of k-means algorithm from scikit-learn library.

Results

The following results were obtained after the running the k-means clustering algorithm:
We have 3 clusters formed in the following map:

1. Cluster 0 (colour - green) : These are the restaurants who have a relatively low price and an highly rated (average > 4.5)
2. Cluster 1 (colour - red): These are the restaurants which are cheap but are not rated as well as the ones above so these are the ones to avoid (average rating < 4)
3. Cluster 3 (colour - blue): These are the restaurants which are rated pretty high but are also quite expensive (Average price > 2000₹)



Discussion

First and foremost the choice of 3 clusters were made as suggested by the Elbow curve technique which made 3 groups of such restaurants as described above. The grouping could have been slightly more discreet with 1 or 2 more labels used which may have grouped expensive restaurants having similar ratings together. The general idea remains the same and results obtained from the analysis of the machine learning model above suggest us to select restaurants according to our own choices but made easier with the plots obtained above since we can avoid the restaurants which have received a relatively low rating from the users.
Further, we can suitably choose these restaurants according to the their price range and how affordable it is.

Conclusion

A simple clustering machine learning algorithm was used in this analysis which was sufficient to separate out necessary information about the various restaurants in Bangalore which could prove to be useful for anyone who is having a hard time making such a decision themselves. Not to mention that this can be further improved further using various techniques.
To mention a few, we can retrieve some additional info about the locality of restaurants so that some of them located in a smaller area could be grouped together accordingly.
We could engineer the features even more or add some more features to obtain a more accurate model.
Lastly, data science techniques are almost always helpful to any problem in an ingenious way.