# Selecting apartment in Bengaluru based on locality around that apartment

Shambu Nandish

July 28, 2020

# Business Problem

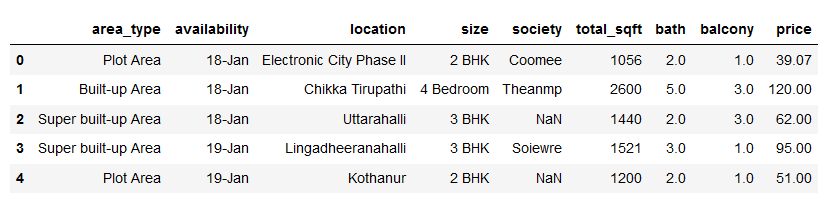
Bengaluru is home for IT companies mainly. Along with IT companies, Bengaluru is also a major contributor to economy. Since a lot of people migrate to this place, it’s difficult for people to find apartments on their own without knowing the neighborhood. Every have their own choice for locality. Some want a lot of restaurants near their locality. Like people who don't know to cook require restaurants nearby.

Choosing a suitable locality depends on varied factors including age and family status. In this project, we aim to characterize Bengaluru's neighborhoods based on the types of businesses and amenities that are situated in those neighborhoods. We will then use this data to perform K-means clustering on the locality and come up with two useful widgets that identify similar neighborhoods with price of the apartment which we have as a part of dataset.

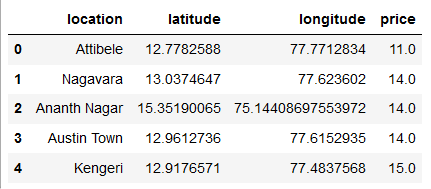
This analysis will be especially useful for non-local workers who have migrated and who are trying to find a home in the city. It will also be useful for local Bengaluru residents who, despite their familiarity with the city, can still benefit from the data.

# Data used for analysis

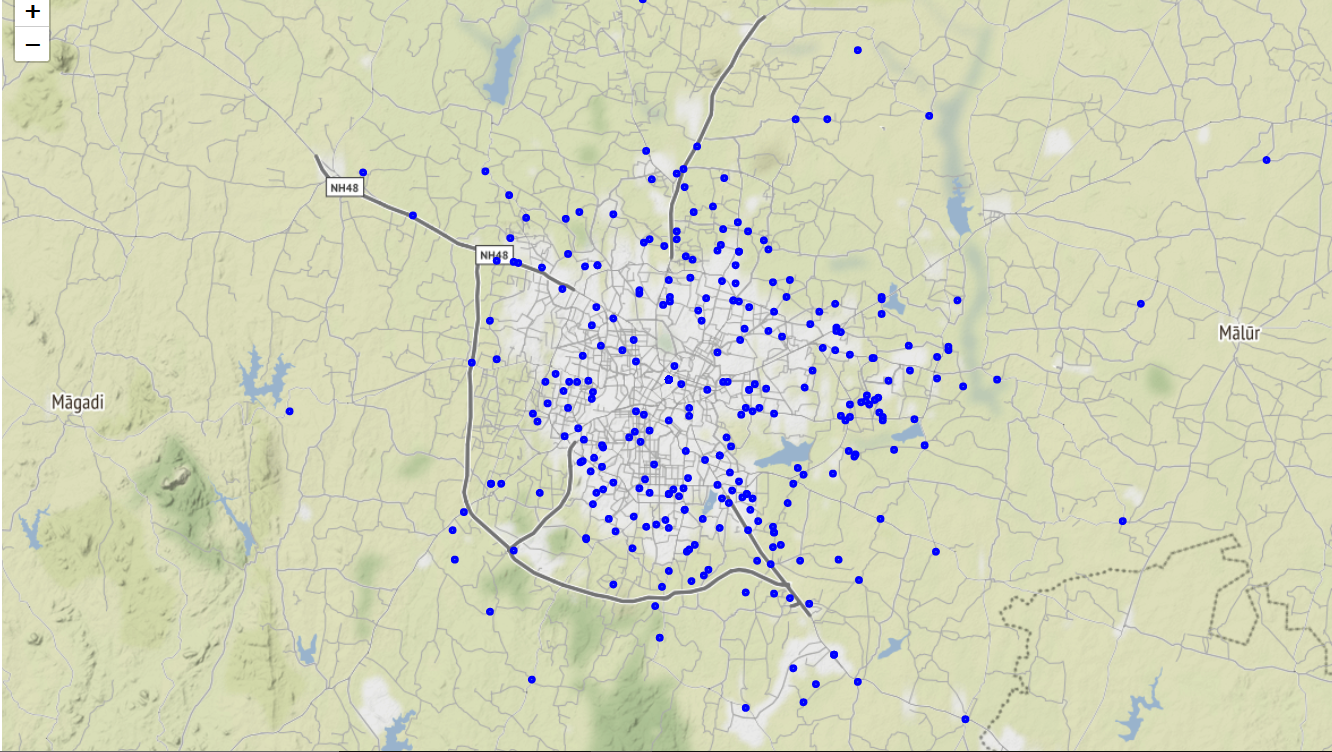
We will use data from Kaggle which was earlier used for prediction of prices in Bengaluru.



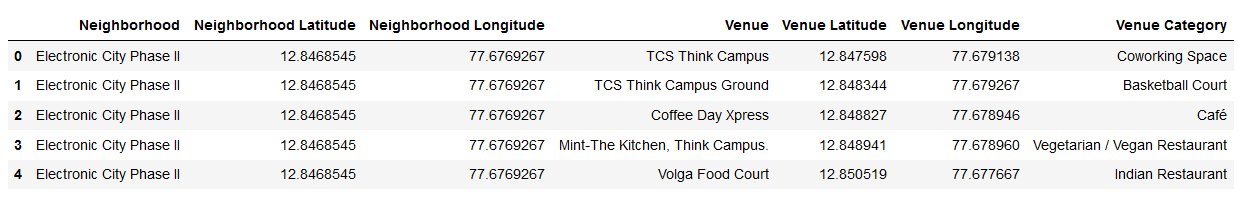
This dataset has the Type of area, availability, location, size (1/2/3 BHK), society name, total square foot of that apartment, number of bathrooms, number of balconies, price of that for 400 neighborhoods located within Bengaluru. We choose only two columns (location and price). We will append this data with the coordinates of each of these neighborhoods that we retrieved using the Python GeoPy module.



After plotting data on map using folium

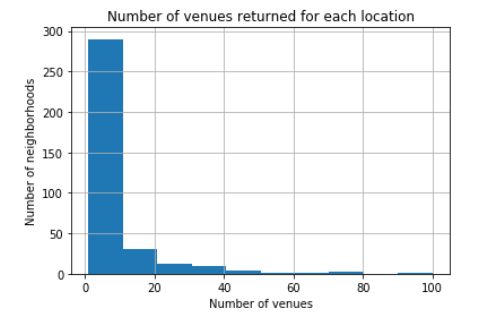


In addition to the above data, we will use data from Foursquare API which is a local search-and-discovery service that provides recommendations of places to go near a user’s current location. We will utilize the Foursquare API to collect information on the venues within each neighborhood in order to characterize and profile each neighborhood. This neighborhood characterization can be achieved in a multitude of ways depending on the project goal and availability of data. For this project, we will characterize each neighborhood based on the entertainment, food, and shopping options located in that neighborhood. We made this choice because the intended target of this project is for people who want to buy apartments in Bengaluru.



# Methodology*:*

In the histogram below, we can see that the Foursquare data returned less than 15 venues for around 20 neighborhoods. While this probably indicates that these neighborhoods are secluded and residential, it could also suggest that more data is needed. This can be focused on in future iterations of this project. Also, there are 2 neighborhoods for which no results (i.e. venues) were returned from Foursquare. These neighborhoods have accordingly not been included in the following analysis.



We start to transform our data to get it ready for the K-means clustering algorithm. We create indicator variables for all of our categories, group the rows by neighborhood and then standardize them to get, for every neighborhood, the proportion of venues in each category. These steps help us get the data in a form that’s easier to understand and enable us to develop a rough profile of each neighborhood.

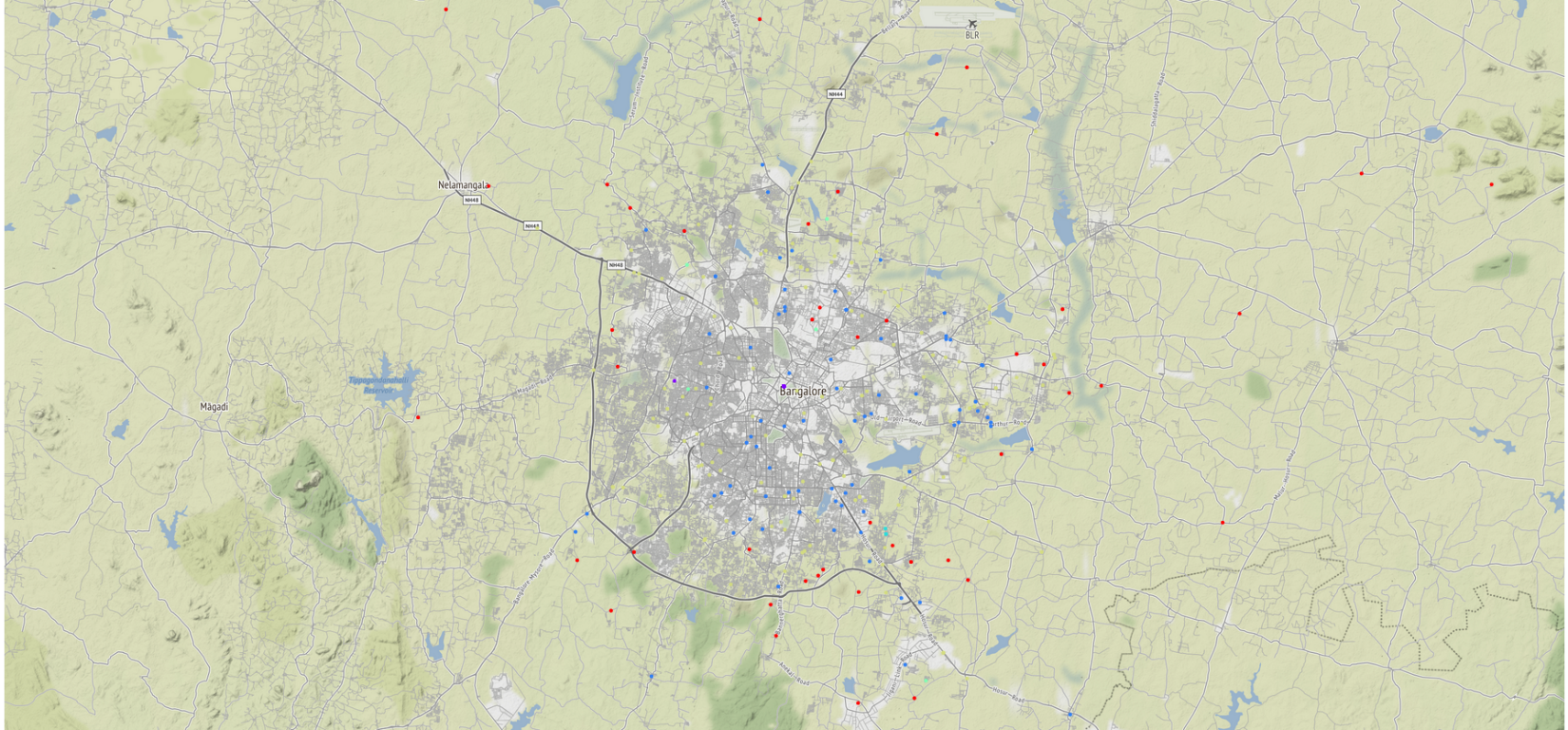


We have successfully characterized each neighborhood by using the data from Foursquare. This means we can get to the most interesting part (at least for me) of the project, data modeling.

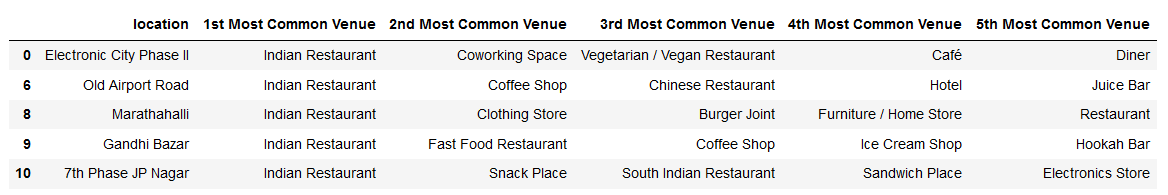
After running the K-means algorithm once more using K=6, we assign cluster labels 0 to 5 to 353 neighborhoods. We dropped five neighborhoods earlier because Foursquare didn't return any data on venues in those neighborhoods. These neighborhoods are assigned with the label 7 as a placeholder until better/more data becomes available.

# Results

We will now plot markers for all the neighborhoods on a map of Bengaluru with each marker colored according to its assigned cluster. Clicking on a marker will show you the location’s name, price of the apartment and cluster label.



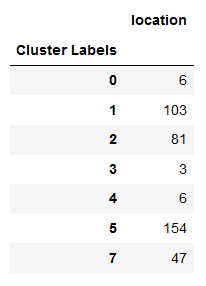
We can now examine any cluster that interests us and determine the discriminating venue categories that distinguish that cluster. Based on the defining categories, you can come up with a general profile of the cluster. For example, neighborhoods with cluster label 0 have a lot of Indian restaurants and cluster 5 have all the types of the places available which most of the common man needs.



Based on the cluster, we can select the location and find price of the apartment.

# Discussion

In the table below, we can see that clusters 1,2 and 5 are the most common. Some clusters have only one neighborhood which could indicate overfitting. Real-world knowledge of the dataset can assist in such a situation.



Our analysis can be improved over time as the quality and quantity of data increases. Other variables that characterize a neighborhood such as population demographics, population density, crime statistics, etc., could be incorporated in a future analysis of the subject.