2. DATA

2.1 Data Requirement

Since this project is aimed to group localities in Bangalore as mutually exclusive clusters based on their similarity, there should be a dataset of the type of buildings/venues on every single locality. List of Locality names and their geographical information is the next requirement. Also we need average rents in these localities, so a dataset on rent prices is also required.

2.2 Data sources

1. Foursqaure.com:- Foursquare.com provides location based data on major places around the world. It has abundance of crowd sourced data on places, venues and even reviews(tips) on them.
2. Wikipedia:- List of Localities in Bangalore is obtained from this Wikipedia list.
3. 99acres.com:- It is one of India’s leading real-estate website in which people can buy, sell, rent properties.
4. Geocode by Awesome Table:- It is an addon in Google sheets which finds the Latitude and Longitude of an Address.

2.3 Data Collection

To obtain data on rent of houses in Bangalore, I scraped the website 99acres.com using Beautiful Soup. Also I used the wikipedia list to get the names of Localities in Bangalore. Then I Geocoded this data to obtain respective latitudes and longitudes of localities. Foursquare provides an API for developers to extract information from their database. In this project we use the explore feature of Foursquare API which returns an user defined number of nearby venues with in a radius from a geographical location. For this project I set the limit for number of venues as 75 and a radius of 1.5 km.

2.3 Data Cleaning and Data Preparation

The scraped rent data had 4060 records of houses available for rent. Attributes like price and area was in object datatype. For example Price –‘15,000’, Area-‘9,000’. These issues were fixed. In this dataset there was about 957 localities, obviously the localities mentioned in the website was very much detailed . However for the sake of simplicity I only took the records where the locality name was matching with the Wikipedia list. This truncated the rent dataset to 656 rows with 46 unique localities. The attributes of the data set were:-

Title	**Locality**	**Building/ Street**	**Price**	**Area**	**Number_of_Bedrooms**	**Number_of_Bathrooms**	**Description**

Since we are only displaying average price of a locality, localities which has under 5 records were dropped. The resultant rent data had 34 unique localities. Since this is not a predictive modelling project this should not cause us any issues. If the need to improve the accuracy of average price occurs , all the dropped rows can be used in the future.

Now we needed the venues data from Foursquare. So every Locality was geocoded, i.e. respective latitude and longitude was added. And this details were passed to the

Foursquare API. The resultant data had top 75 venues in a radius of 1.5 Km, and its attributes are :-

	Locality	**Locality Latitude**	**Locality Longitude**	**Venue**	**Venue Latitude**	**Venue Longitude**	**Venue Category**

There was 156 unique venue categories like ATM, Bus Station, Park, Lake etc. Frequency of each category for every Locality was calculated. Clustering algorithm will cluster localities based on this frequency data. And corresponding clusters will be obtained. Merging these clusters with average prices in the rent data for each locality is the required output.