**Available Dataset and its Usage**

**Datasets available** – There will be 3 dataset that will help me resolve the business problem in question.

The ***first dataset*** will be from Foursquare API. This dataset will contain many columns but the ones which will solve my purpose are the following:

**venue id, name, address, geographical location (latitude and longitude), postal codes, category**.

This dataset will form the base of the project. This main purpose of this dataset will be to provide me with the list of postal code and geographical locations of all popular places in Mumbai. It will be very difficult to extract the locality from the address column in the dataset as there are various ways to write a locality name in Mumbai. Hence to proceed further, I will be using the postal codes to determine the localities with the help of subsequent dataset.

The ***second dataset*** will be from a website (<https://mumbai7.com/postal-codes-in-mumbai/>) having the most up-to-date list of all **postal codes** along with their respective **locality name** in Mumbai. I will be scrapping the dataset directly from the website. It will help to standardize the locality names and further help in merging the above dataset with the third dataset. In other words, it will help connect the first and third dataset on a common column i.e. locality and subsequently lead the formation of final dataset.

However, there are issues that needs to be resolved here. There will be some postal codes and localities that need to be revisited. Reason being, real estate prices are not as per the postal codes in Mumbai. Couple of localities are lugged under one locality in real estate world. The other reason to revisit them is few postal codes are incorrect or the fact that there are plenty of ways to write a locality name, which doesn’t help with the normalization. These few records need to be corrected one by one.

The ***third dataset*** will be from another website (www.99acres.com). This website will provide with the real estate prices of all localities in and around Mumbai. For all the localities, we will get following:

**Locality name, min retail price per sq ft, max retail price per sq ft**, quarter change, rental rates for 1, 2, & 3 BHK.

Out of all, the first 3 will be required as of now and merged with the first dataset. Now the dataset is complete for further processing. Then another column will be introduced, which will be the mean of Minimum Retail Price per square feet and Maximum Retail Price per square feet. This column along with the latitude and longitude of the popular venues will be used in clustering model to find the apt coordinates of centre of each cluster, the model defines.