**Exploratory analysis of Mumbai**

**Business problem**

Mumbai is the commercial capital of India. Every year thousands of professionals move into Mumbai due to transfers, better job prospects etc. Mumbai is a large city and be quite unnerving to someone who is new to this place. The attempt of this project is to segment the city into areas with similar characteristics based on K-Means clustering. This helps the professionals new to Mumbai to narrow down their search as to where to look for a place of stay. They get an idea about the nature of a location based on the kind of most common venues in the locality.

**Data**

The primary data needed for the exercise is the locality wise zip code along with corresponding longitude and latitude. This is extracted from the website [www.geonames.org](http://www.geonames.org). This website has the zip code wise longitude and latitude data for several countries. The data for India is first downloaded and from this data a csv file containing the zip code wise longitude and latitude for Mumbai is culled out for the purpose of this analysis.

The sample data of this newly created Mumbai dataset was as follows:

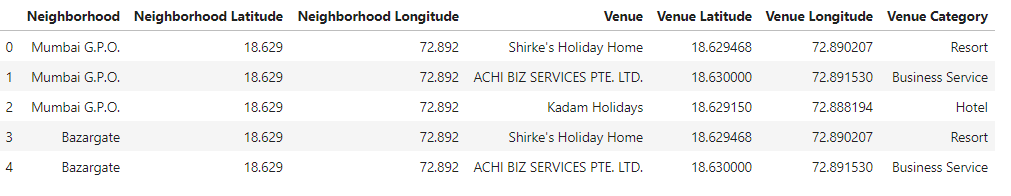
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Pincode | Place | State | City | Latitude | Longitude |
| 400001 | Mumbai G.P.O. | Maharashtra | Mumbai | 18.629 | 72.892 |
| 400001 | Bazargate | Maharashtra | Mumbai | 18.629 | 72.892 |
| 400001 | Town Hall (Mumbai) | Maharashtra | Mumbai | 18.629 | 72.892 |
| 400001 | Tajmahal | Maharashtra | Mumbai | 18.629 | 72.892 |
| 400001 | Stock Exchange | Maharashtra | Mumbai | 18.629 | 72.892 |

The pin code goes up to 400104. Multiple locations within same pin code is available. There are a total of 239 locations captured in the dataset used for the exploratory analysis.

Foursquare API is then used to provide information about venues close to each location.

**Methodology**

* Using foursquare API the top 100 venues within a radius of 500 m from each location is captured along with their latitude and longitude
* The resultant data is converted into a pandas data frame, a sample of which which looks like this



* The top 10 most common venues for each neighborhood is identified and this is converted into a pandas data frame which leads to an output that gives the top ten most common venue for each neighborhood. The sample of resultant output looks like this



* Based on the common venue characteristic the neighborhoods are segmented into clusters using k-means clustering. A total of 8 clusters was used. Summary of the output is provided in the appended MS-EXCEL file

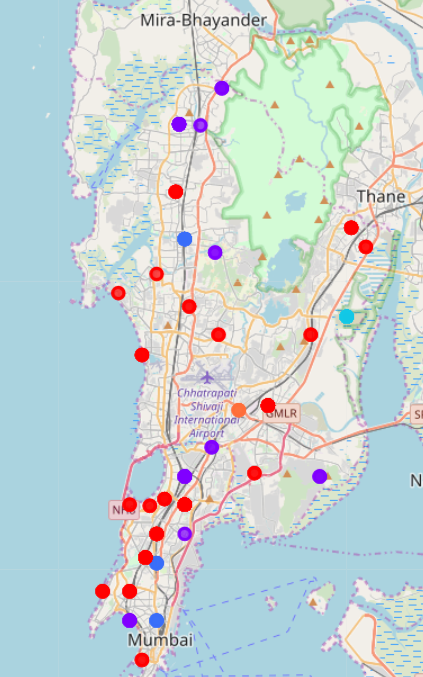


**Results**

The output gives an indication of the kind of area each location is and what are the nearby areas that share similar characteristic. This will help the professionals new to Mumbai to get a quick idea about each location. A sample output of a cluster is given below



The cluster map looks as follows



**Discussion**

There are limitations to this analysis. The extent to which foursquare has captured data in Mumbai may not be as much as it has been done in major cities in US or UK. The output may hence be very different if the extent of mapping by foursquare is a more improved one. Also pincode wise latitude and longitude data available in geonames also has some issues associated with it in terms of accuracy. Authentic Government of India data is not available. Over a period of time as more data is available the output quality will improve.

**Conclusion**

Knowing the top 10 most common places in a locality gives a good idea about the nature of a locality. Clustering it gives an idea about locations that have a similar characteristic. This project helps user identify locations which share similar characteristics in Mumbai.