IBM Data science professional.

Capstone Project – opening a new Multiplex in Pune, India.

Business problem.

- Huge number of movies are released in India every year, with people being movie enthusiasts, multiplexes are great sources of revenue.
- Location of a multiplex plays an important role in determining footfalls and revenue.
- Objective – To select best locations in the city of Pune for opening a multiplex.

Data

- Data required.
  - List of neighbourhoods in Pune.
  - Latitude and Longitude coordinates of these neighbourhoods.
  - Venue data collected using the Foursquare API.

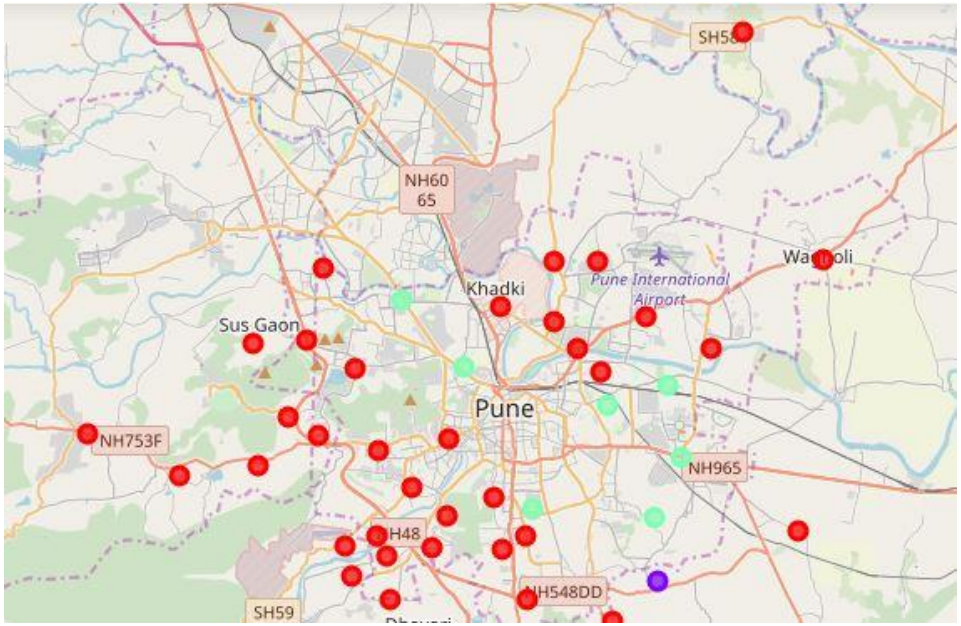
- Sources of Data
  - Wikipedia page for neighbourhoods list - https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune
  - Foursquare API for venue data.
  - Geocoder package for getting latitude and longitude.

Methodology

- Web scraping wikipedia page to get the list of neighbourhoods and creating a pandas dataframe.
- Get latitude and longitude coordinates using the geocoder package and add these to the dataframe.
- Group data by neighborhood and take the mean frequency of occurrence of each venue category.
- Filter venue category by shopping mall and perform K-means clustering.
- Plot the clutser on map of Pune city using Folium library.

Results

- The neighbourhoods are categorised into 3 categories.
  - Cluster 0 – almost no multiplexes.
  - Cluster 1 – few multiplexes.
  - Cluster 2 – majority of multiplexes.



Discussion.

- Cluster 0 has no multiplexes, this offers a good opportunity to open multiplexes.
- Cluster 1 has very multiplexes, this again is a good opportunity to open multiplexes.
- Cluster 2 has majority of multiplexes, multiplex owners should refrain from opening multiplexes in these areas.

Conclusion

- Multiplex owners should open new multiplexes in cluster 0 and 1.