IBM Data Science Professional – Capstone Project.

Introduction

For many people living in Pune and the whole of India, watching a movie is one the most preferred ways of spending some quality leisure time with their friends, families and loved ones. A larger number of Bollywood and Hollywood movies are released in India every year given the huge movie watching population of the country. Pune being a metropolitan city possesses a variety of multiplexes catering to the diverse population. However, taking into account the growing population of the city and the increased infrastructure development of areas, opening new multiplexes would prove to be a good source of revenue for multiplex owners.

Business Problem.

If a multiplex owner is looking to open a new multiplex, which neighbourhoods should be targeted ?

The objective of this capstone is to determine neighbourhoods where there are no or very few multiplexes .

Target Audience of this project

The target audience of this project is multiplex owners and other stakeholders looking to open a new multiplex catering to the growing population of the city.

Data

To solve this problem, we need the below data –

- List of neighbourhoods in Pune – this will be used as the reference for selecting the areas.
- Latitude and longitude coordinates of the neighbourhoods – this is required to plot maps and also to get the nearby venues.
- Nearby venues data for all the neighbourhoods to get a sense of how many multiplexes are there in each neighbourhood.

Sources of data and methods to extract them

We get the above data from different sources. To get a list of neighbourhood, we utilise Wikipedia which has a great list of neighbourhoods in Pune - https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune.

To get latitude and longitude of these areas, we use the geocoder library in Python.

Further, to the acquire the nearby venues of each neighbourhood, we use the Foursquare API with with client ID and secret as authentication.

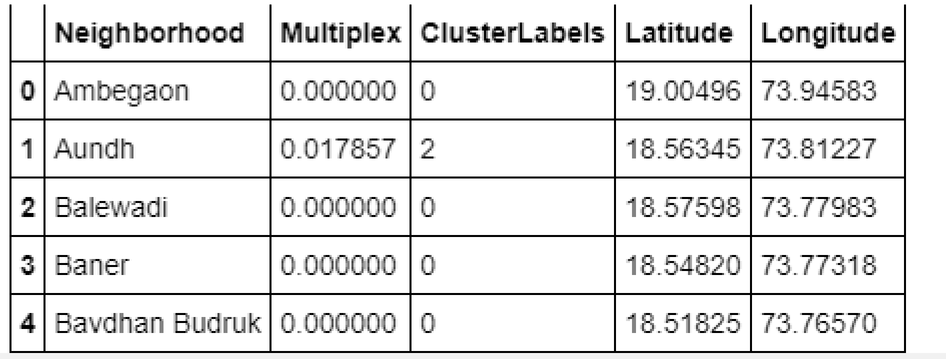
Methodology

First we need to get the list of neighbourhoods in Pune. We use the Wikipedia page to get this information. To make it available inside our Jupyter notebook environment, we use the requests package to get the raw html and parse it using the popular web-scraping package Beautiful Soup. We see the that the neighbourhoods' list is under the "ul" html tag, we find the ul tag from raw html, iterate over the items and append them to a python list. We finally create a dataframe using this list.

To get the latitude and longitude coordinates of these neighbourhoods, we use the geocoder.arcgis() method of the geocoder package, to iterate over all the neighbourhoods, find their latitude and longitude coordinates . We then concatenate these coordinates to the original neighbourhoods dataframe. We visualize the neighbourhoods in a map using Folium package

To get the venues data for these neighbourhoods, we use the Foursquare API. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Multiplex" data, we will filter the "Multiplex" as venue category for the neighbourhoods.

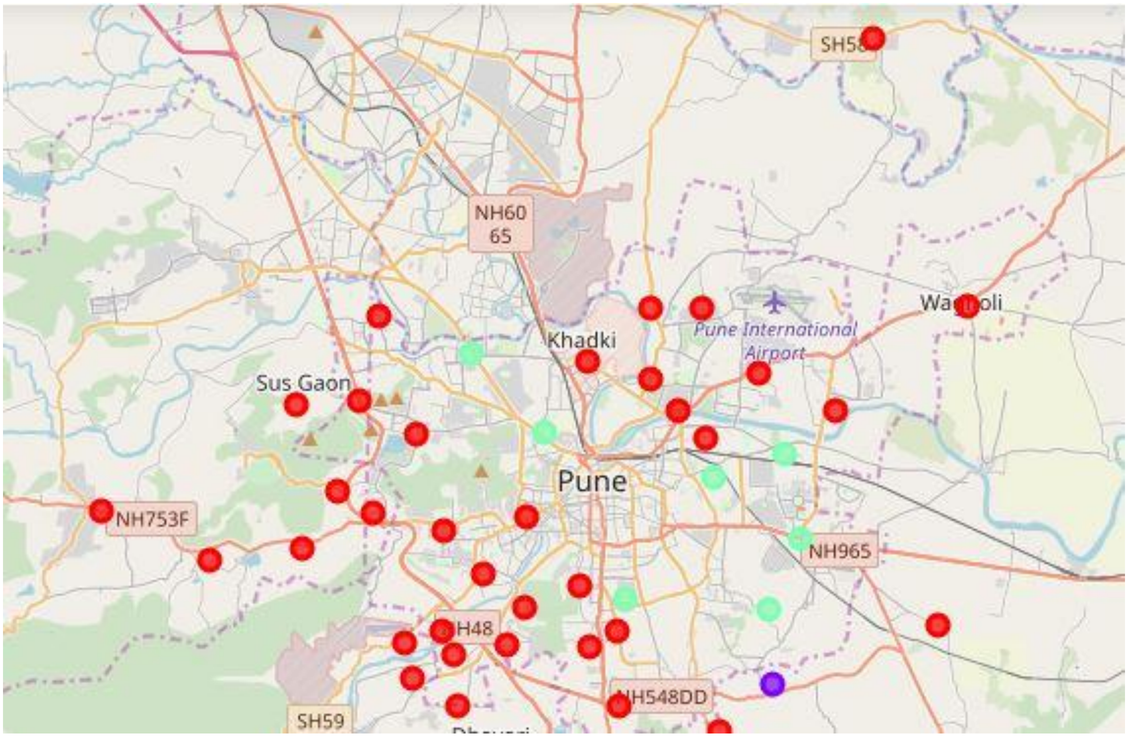
We then perform clustering of the data using the popular K-means clustering algorithm, which is a partition based clustering algorithm. We decide the number of clusters to be 3. After fitting the model, the cluster labels are returned which are then merged with the dataframe.



	Neighborhood	Multiplex	ClusterLabels	Latitude	Longitude
0	Ambegaon	0.000000	0	19.00496	73.94583
1	Aundh	0.017857	2	18.56345	73.81227
2	Balewadi	0.000000	0	18.57598	73.77983
3	Baner	0.000000	0	18.54820	73.77318
4	Bavdhan Budruk	0.000000	0	18.51825	73.76570

Results

We plot the clusters on a map using Folium.



The three clusters are denoted by three different coloured dots in the above map.

Red – Cluster 0

Blue – Cluster 1

Green – Cluster 2

Cluster 0 has 37 data points, these neighbourhoods has almost 0 multiplexes.

Cluster 1 has just 1 datapoint, there is presence of multiplex in this neighbourhood.

Cluster 2 has 7 neighbourhoods, majority of multiplexes are concentrated in these areas.

Discussion

From the clustering results and statistical analysis, we can conclude that clusters 0 and 1 have the least number of multiplexes and thus, offer a great business opportunity for the multiplex developers. Cluster 2 has the majority of multiplexes which offer tough competition to not just any planned multiplex but to one another as well.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of multiplexes, there are other factors such as population and income of residents that could influence the location decision of a new multiplex.

Conclusion

In this project, we have identified a business problem, determined which data are required and collected them from various sources, performed data wrangling, cleaning, and feature creation to transform the data in the form required by the machine learning algorithm. We then a applied an unsupervised machine learning algorithm called K-means clustering with to divide our data into 3 clusters depending on the frequency of occurrence of multiplexes in areas. From the clustering results we can conclude that clusters 0 and 1 are conducive to the opening of new multiplexes while areas in cluster 2 offer tough competition due to high concentration of multiplexes.