Title: To analyze suitable place for opening New Shopping Mall

Work Done by: Pranjali Khadse

Introduction :-

Shopping malls are preferred by peoples these days because of multifunction availabilities. Peoples can have grocery shopping, accessories shopping and at the same time there are fun activity corners for children's also.

For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city nearby IT Hubs and many more are to be built. Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. So here, location plays a crucial role which will decide whether the opening of mall will be profitable or failure.

Business Problem:-

The objective of this capstone project is to analyze and select the best locations in the city of Pune, India to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city , if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

Data :-

To solve the problem, we will need the following data:

1. List of neighborhood's in Pune city.
2. Latitude and longitude coordinates of those neighborhood's. This is required in order to plot the map and to get the venue data.

3. Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them:-

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune) contains a list of neighborhood's in Pune, with a total of 46 neighborhood's which comes under Pune Municipal Corporation. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful soup packages. Then we will get the geographical coordinates of the neighborhood's using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology :-

- Firstly, we need to get the list of neighborhoods in the city of Pune which come under Pune Municipal Corporation.
- This list can be obtained by scraping Wikipedia page by using beautiful soup library for web scraping. However, this is just a list of names. In order to do detail analysis, we would also need coordinates for neighborhoods present in the list
- We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. This can be done with the help of geocoder library.
- After gathering the data, we will populate the data into a pandas Data Frame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a

sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Pune .

- Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. 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 neighborhoods 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 neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood 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 analyzing the "Shopping Mall" data, we will filter the "Shopping Mall" as venue category for the neighborhoods.
- Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.
- We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for "Shopping Mall".
- The results will allow us to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results:-

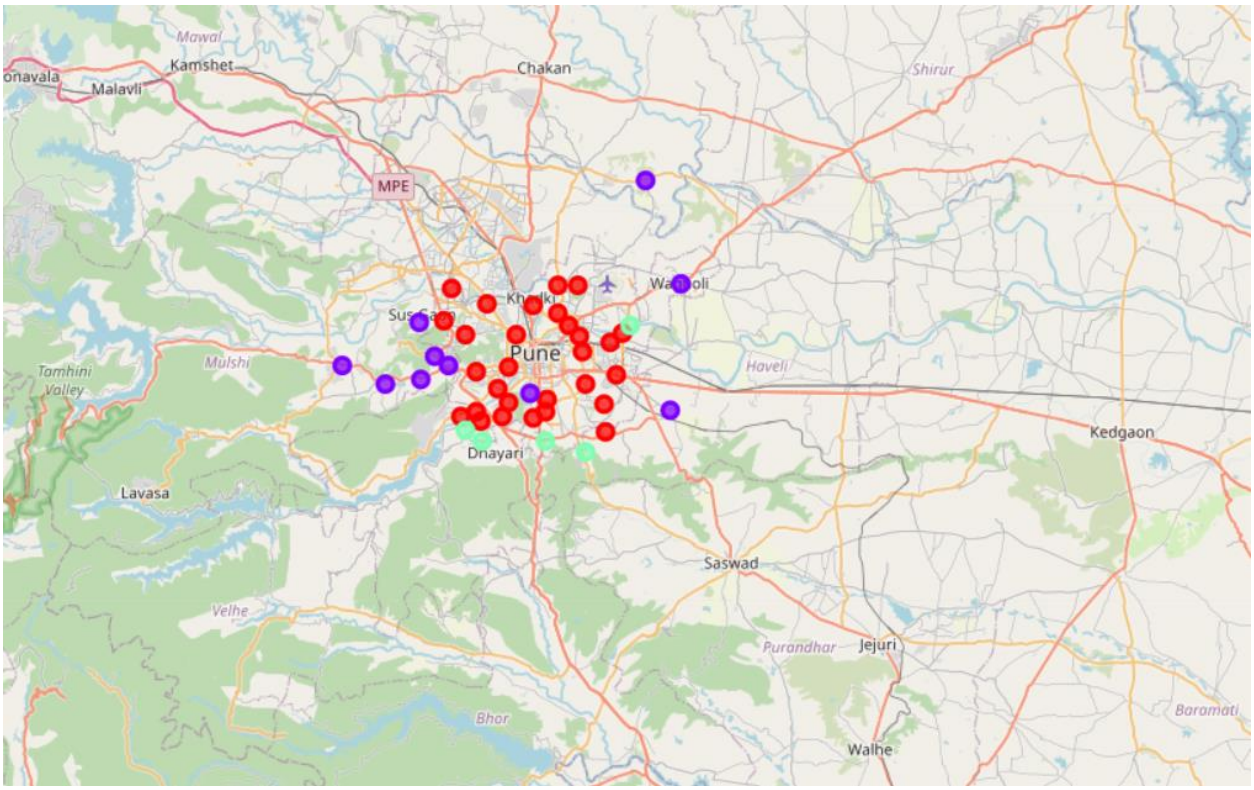
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Shopping Mall":

• Cluster 0: Neighborhoods with moderate number of shopping malls

• Cluster 1: Neighborhoods with no existence of shopping malls

• Cluster 2: Neighborhoods with high concentration of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.



Observation:-

- From analysis of clusters it is evident that cluster 2 has more shopping malls as compared to cluster 1 and cluster 0.
- Also, we can see that cluster 1 has no malls at all when compared to cluster 2 and 0
- So, from above two observations we have cluster 1 to open shopping malls in order to get more revenue, as there would be no competition for new mall opened

- There is one unique observation about cluster 2. The neighborhoods that are present in cluster 2 have renounced IT parks in the area nearby to shopping malls.
- So, there is also a chance for developers to propose new plan for IT parks in neighborhoods present in cluster 1 and 0. Although another analysis for IT Park proposal would be needed to select places, but still the idea would work out fine.
- Main takeaway from this analysis for developers is, no investment for new proposals should be made in neighborhoods of cluster 2 as there is already enough competition going on among those present.

Conclusion:-

- In this project analysis we just considered the areas which are under Municipal corporation. There are other factors which need to be considered.
- Factors like quality of housing area , any industries or IT parks in vicinity these are also the factors that would add quality to result. As in where the mall could be in more good business.
- Other API's like Zomato should also be considered in order to get all venues present in the areas of cluster.