Coursera Capstone

**IBM Applied Data Science Capstone**

***Segmenting And Clustering Localities In Mumbai For a Potential Residential Buyer***

By: Devbrat Mahapatra

**Business Understanding**

A client is planning to buy a residential property in Mumbai, India. The client wishes to explore the different areas of Mumbai before buying a residential property. This project will aid in exploring the area as well as show the average price/sqft in the particular area. Clusters of similar localities will also be shown to the client.

**The Location**: Mumbai is the dream city of India. It houses the film industry and is also the financial hub of India. As a result, every year people shift to Mumbai from various parts of India and this project will cater to many clients.

**Foursquare API**: This project would use Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

**Work Flow**: Using credentials of Foursquare API features of nearby places of the neighborhoods would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

**Clustering Approach**: To compare the similarities of neighborhoods or areas, we decided to explore neighborhoods, segment them, and group them into clusters. To be able to do that, we need to cluster data which is a form of unsupervised machine learning.

**Data**

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

* List of localities in Mumbai with the price range (per sqft). This defines the scope of this project which is confined to the city of Mumbai, the financial capital of India.
* Coordinates and Venue data of those localities. This is required in order to plot the map and also to get the venue data.

## Sources of data and methods to extract them

* This website <https://www.99acres.com/property-rates-and-price-trends-in-mumbai>) contains a list of localities in Mumbai, with a total of 244 localities. For convenience, we copy the data straight to an excel file for easier data wrangling. Then we will get the geographical coordinates of the locality using Tomtom Geocoder package which will give us the latitude and longitude coordinates of the localities.
* After that, we will use Foursquare API to get the venue data for those localities. Foursquare has one of the largest database 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.

**Methodology**

* Firstly, we need to get the list of localities in the city of Mumbai. Fortunately, the list is available in the 99acres (<https://www.99acres.com/property-rates-and-price-trends-in-mumbai>). We will do data wrangling for proper formatting of the data. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful TomTom Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. We need an API key to retrieve the data. After gathering the data, we will populate the data into a pandas Data Frame and then visualize the localities 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 Mumbai.
* Next, we will use Foursquare API to get the top 10 venues that are within a radius of 500 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 localities 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 locality and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each locality by grouping the rows by locality and taking the mean of the frequency of occurrence of each venue category.
* 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 use the elbow method to find the optimal k value. We find the optimal value of k=4, therefore the number of clusters are four. The results will allow us to determine the concentration of the venues.

**Results**

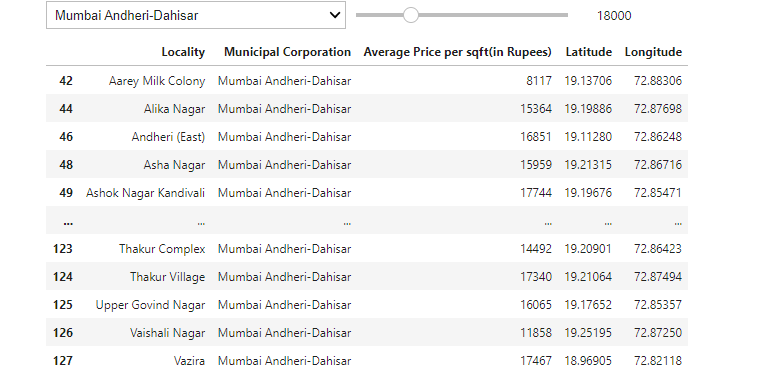


Fig: A dashboard created for the client wherein there is a dropdown menu to choose the municipal corporation and a slider for choosing the price.

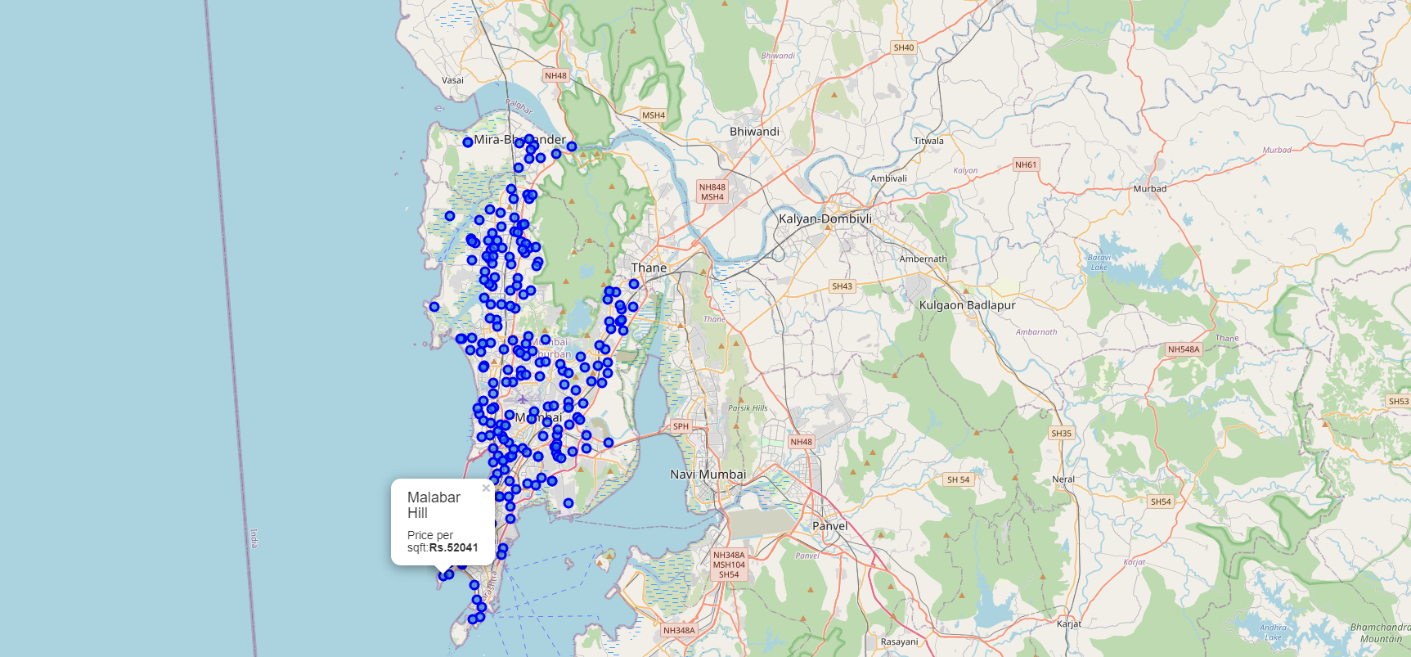


Fig: An interactive map created for the client with a popup label which has the following details:

1. Locality
2. Price per sqft.

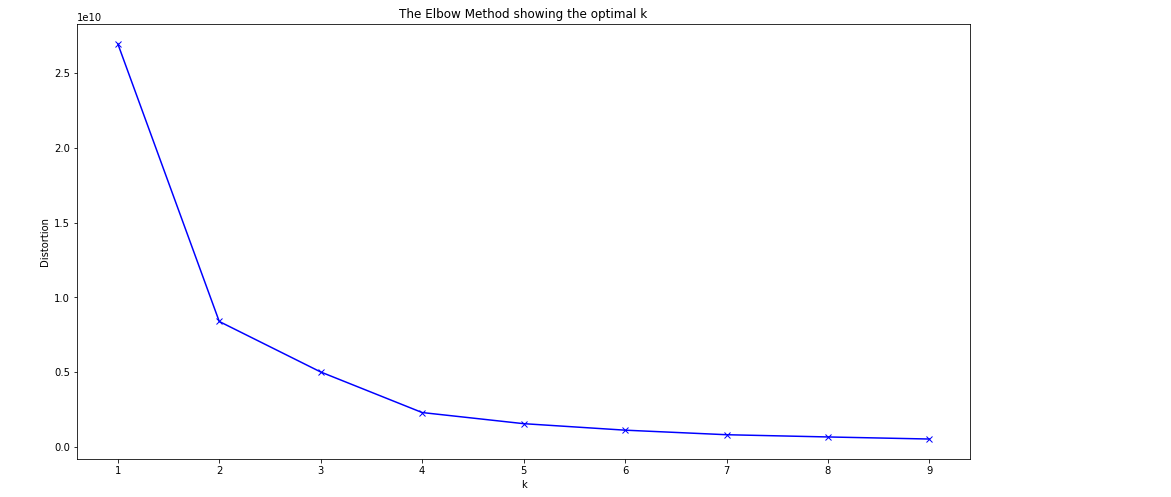
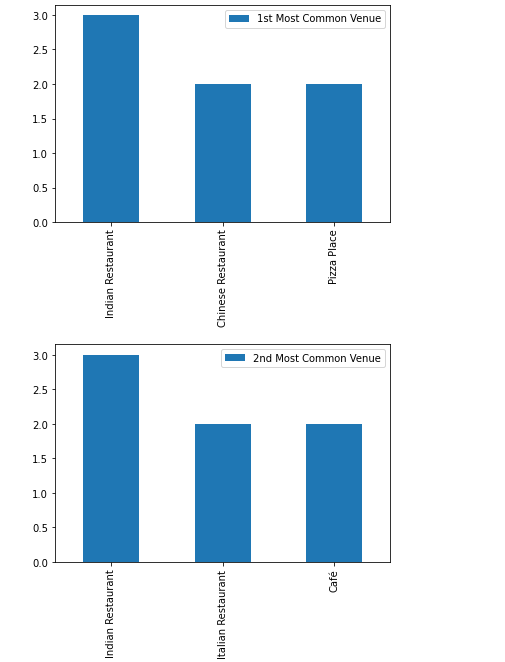
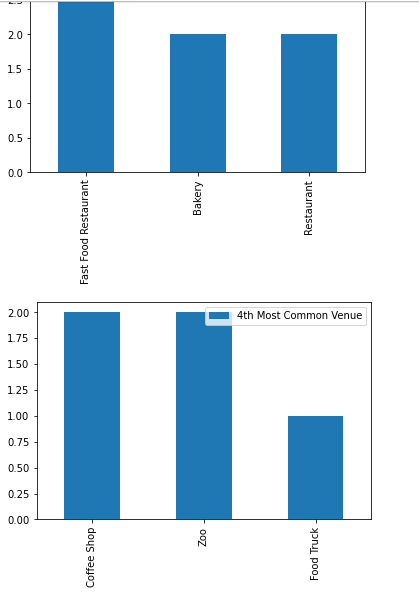


Fig: Elbow method graph for finding optimal k

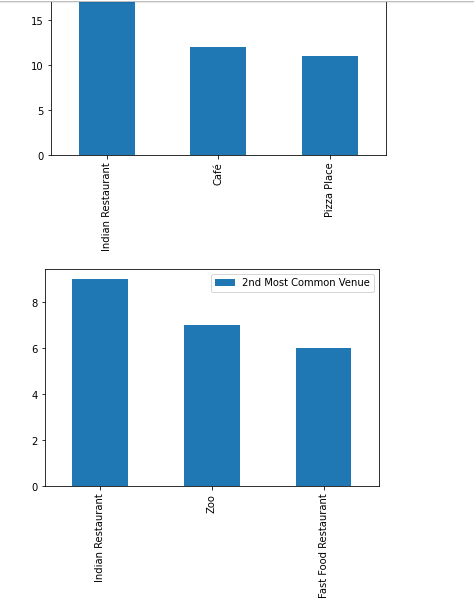
**Examining the clusters**

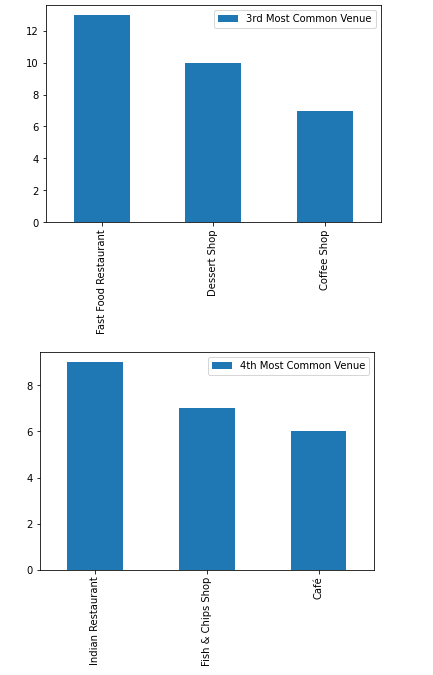
**Cluster 1**

****

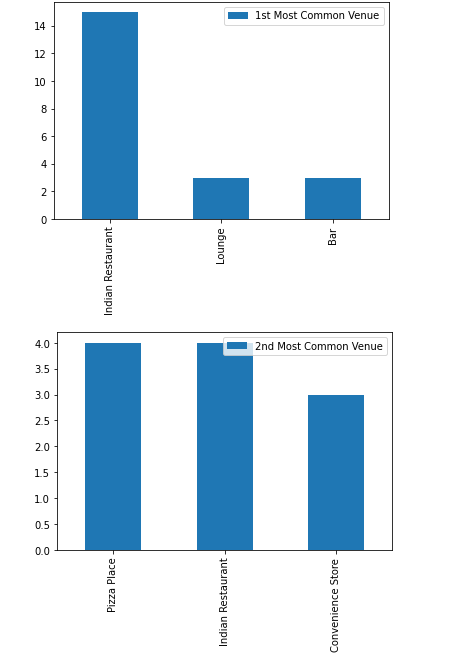


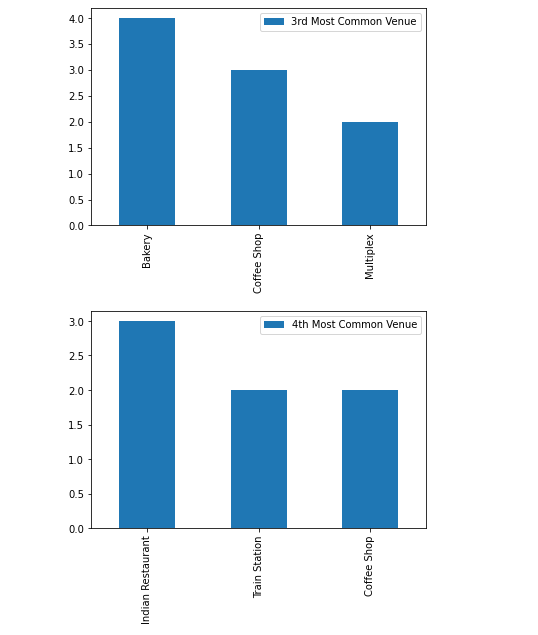
**Cluster 2**

****

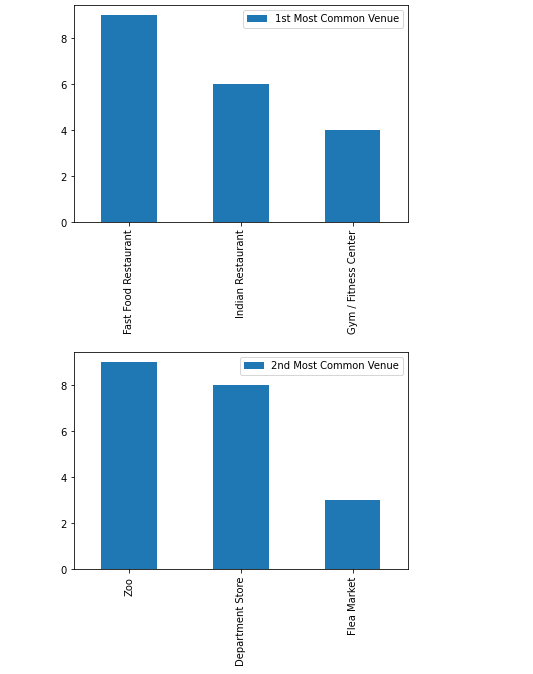


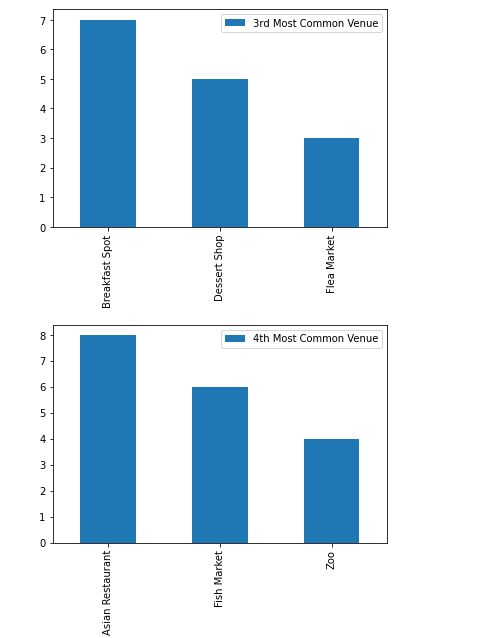
**Cluster 3**

****

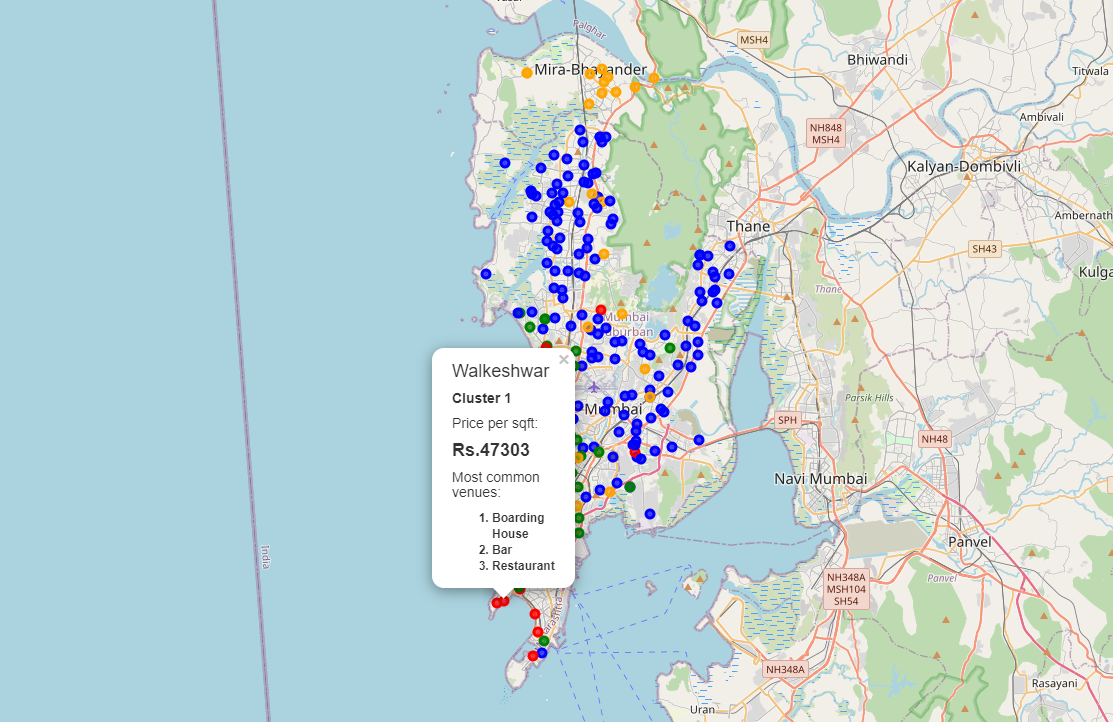


**Cluster 4**





**Cluster map**

****

An interactive map created for the client with a popup label which has the following details:

1. Locality
2. Cluster
3. Price per sqft.
4. Most common venues around

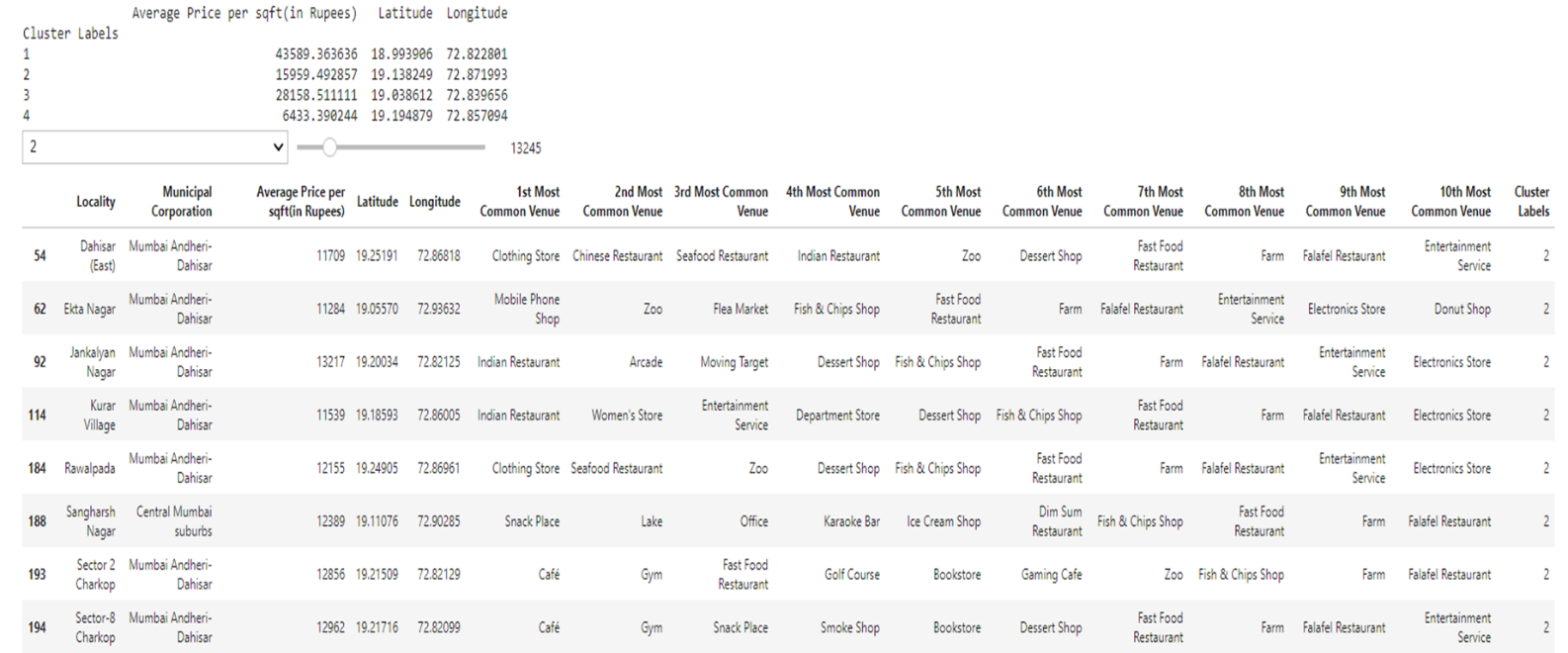


Fig: A dashboard with a dropdown menu to choose the cluster label and a slider for choosing the price with the average price per sqft. for every cluster

# Suggestions for Future Research

* Better location accuracy.
* Better cluster model accuracy.
* The price per sqft. data is only pertaining to residential apartments in Mumbai and would like to include price for standalone houses.
* Data regarding the crime rates in the areas can be included to determine which areas to safer to live.
* This project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned.

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The localities in cluster 1 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

Category:Suburbs in Kuala Lumpur. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs\_in\_Kuala\_Lumpur

Foursquare Developers Documentation. *Foursquare*. Retrieved from <https://developer.foursquare.com/docs>

Malay Mail. (2018, March 14). Malls facing meltdown as glut continues. *Malay Mail*. Retrieved from <https://www.malaymail.com/s/1597735/malls-facing-meltdown-as-glut-continues>

Tan, H. H. (2018, April 5). An oversupply of retail space in Malaysia. *StarProperty.my*. Retrieved from [http://www.starproperty.my/index.php/articles/property-news/an-oversupply-of-retail-space-in-](http://www.starproperty.my/index.php/articles/property-news/an-oversupply-of-retail-space-in-malaysia/) [malaysia/](http://www.starproperty.my/index.php/articles/property-news/an-oversupply-of-retail-space-in-malaysia/)

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster 0**   * Bangsar South * Bukit Bintang * Bukit Nanas * Bukit Tunku * Chow Kit * Damansara Heights | * Damansara Town Centre * Damansara, Kuala Lumpur * Dang Wangi * Jalan Cochrane, Kuala Lumpur | * Jalan Duta * Kampung Baru, Kuala Lumpur * Medan Tuanku * Mont Kiara * Segambut | * Setiawangsa * Shamelin * Taman Desa * Taman Tun Dr Ismail |
| **Cluster 1**   * Alam Damai * Ampang, Kuala Lumpur * Bandar Menjalara * Bandar Sri Permaisuri * Bandar Tasik Selatan * Bandar Tun Razak * Batu 11 Cheras * Batu, Kuala Lumpur * Bukit Jalil * Bukit Kiara * Bukit Petaling * Cheras, Kuala Lumpur | * Desa Petaling * Federal Hill, Kuala Lumpur * Happy Garden * Jinjang * Kampung Datuk Keramat * Kepong * Kuchai Lama * Maluri * Miharja * Pantai Dalam * Putrajaya | * Salak South * Semarak * Sentul Raya * Setapak * Sri Hartamas * Sri Petaling * Sungai Besi * Taman Bukit Maluri * Taman Cheras Hartamas * Taman Connaught * Taman Ibukota | * Taman Len Seng * Taman Melati * Taman Midah * Taman OUG * Taman P. Ramlee * Taman Sri Sinar * Taman Taynton View * Taman Wahyu * Titiwangsa * Wangsa Maju |
| **Cluster 2**   * Bangsar * Bangsar Park | * Brickfields * KL Eco City | * Lembah Pantai * Pudu, Kuala Lumpur | * Taman U-Thant |