IBM Applied Data Science Capstone

Recommendation of site location to build Shopping Mall

By : Sanskriti Tiwari

Aug, 2020

- **Introduction :**

In recent years, our lives have become busier and due to this we don't get time to relax ourselves with family and friends. In a recent survey, shopping was considered as a great refreshment activity. Many people find that visiting a shopping mall is a great way to relax themselves during holidays and weekends.

Nowadays, if we look around, we find numerous people in shopping malls. This is because Shopping malls are a one-stop destination for people. They are very convenient as people can watch movies, dine at restaurants, do grocery shopping, play games and perform many such leisure activities all at one place.

In fact, some people just visit malls to hang out and don't shop at all. It is a place where one can find everything they want in one place which makes it easier for people to visit one mall and finish up with their shopping.

Property builders take advantage of this fashion and build more shopping malls to satisfy the demands of people. As a result of this, there are numerous shopping malls in the city of Kuala Lumpur, Malaysia and numbers increase continuously. Shopping malls provide consistent rental incomes to property builders.

However, building a shopping mall requires many factors to be taken into consideration. One of the most important factor is the location of shopping mall. Location is the most important decision as it determines whether the mall will be success or a failure.

- **Business Problem :**

The main objective of this capstone project is to analyse and select the best location to open a shopping mall in the city of Kuala Lumpur, Malaysia.

We will use data science methodology and machine learning technique like clustering and segmentation to achieve our goal. This project aims to answer the business question : In the city of Kuala Lumpur, if a builder wants to open a shopping mall, where would you suggest him to open it?

- **Data :**

  To provide a solution to the problem, we will need the data:

  1. List of neighborhoods in the city of Kuala Lumpur, capital of Malaysia, a country in South-East Asia.
  2. Geographical coordinates (Latitude and Longitude) of these neighborhoods in order to plot map and get venue data associated to these neighborhoods
  3. Venue data with respect to each neighborhood, particularly of shopping malls. This data will be used to cluster neighborhoods

- **Source of Data and how to extract it :**

  We will use web scraping technique to collect the list of neighborhoods in Kuala Lumpur. Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur) contains all the neighborhoods present in Kuala Lumpur. We extracted this page using Python packages and BeautifulSoup library. Then we used geocoder library of Python to get latitude and longitude associated with each neighborhood.
  Then, we used Foursquare API (https://foursquare.com/developers/apps) to get the venues present in every neighborhood. Foursquare has a database of 150+ million places and is used by 130,000+ developers. Foursquare API provided every category of venue like coffee shops, theatres, parks, museums, etc. But we were particularly interested in shopping malls. So we extracted shopping malls from the data provided by Foursquare API.

- **Approach :**

  Firstly, we will get list of neighborhoods in the city of Kuala Lumpur, Malaysia. This was available at https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur . We will then use web scraping using Python packages and BeautifulSoup library to extract neighborhood list. Now we need to get geographical coordinates of these neighborhoods. Then we used geocoder library of Python to get latitude and longitude associated with each neighborhood.

Then, we used Foursquare API (https://foursquare.com/developers/apps) to get the top 100 venues present within the radius of 4 km for every neighborhood. We created a Foursquare developer's account for this purpose. Using our Foursquare Client ID and secret key, Foursquare API provided every category of venue like coffee shops, theatres, parks, museums, etc. But we were particularly interested in shopping malls. So, we extracted shopping malls from the data provided by Foursquare API. API returned the venue data into JSON format. We extracted the Venue name, category, latitude and longitude from it. We then populated all of this data into pandas dataframe.

Fetching Venue data using Foursqaure API

In [ ]:
```
venue_info = []

for lat, lng, neighborhood in zip(data['Latitude'], data['Longitude'], data['Neighborhood']):

    # create the API request URL
    html = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}".format(
        client_id,
        client_secret,
        version,
        lat,
        lng,
        radius,
        limit)

    result = requests.get(html).json()["response"]['groups'][0]['items']

    for venue in result:
        venue_info.append((
            neighborhood,
            lat,
            lng,
            venue['venue']['name'],
            venue['venue']['location']['lat'],
            venue['venue']['location']['lng'],
            venue['venue']['categories'][0]['name']))
```

Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking mean of the frequency of occurrence of each venue category. By doing so, we are preparing the data so that in can used for clustering. Since we are analyzing the "Shopping Mall" data, we will filter the "Shopping Mall" as venue category for the neighborhoods.

Then, we will perform clustering on the shopping mall data using K-Means Clustering algorithm. **K-means 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 most popular clustering algorithm. We clustered our data into 4 clusters. The results of these clusters

helped us to identify the neighborhoods having higher concentration of shopping malls and those with fewer shopping malls. The clusters formed will help us to answer our business problem.

Applying K Means Clustering

```
train = df_mall.drop('Neighborhood', axis = 1)

kmeans = KMeans(n_clusters = 4, random_state = 2)
kmeans.fit(train)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=4, n_init=10, n_jobs=None, precompute_distances='auto',
    random_state=2, tol=0.0001, verbose=0)
```

```
kmeans.labels_[0:20]
```

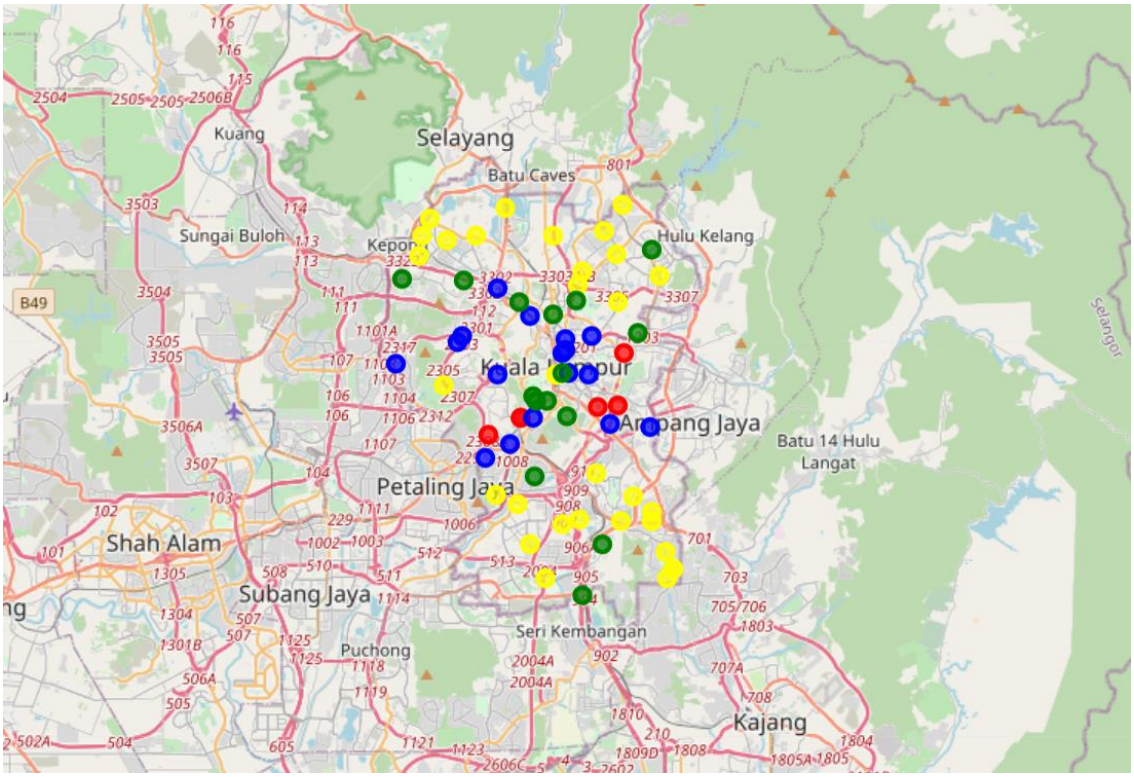
```
array([0, 3, 3, 0, 3, 0, 1, 1, 2, 0, 0, 2, 2, 0, 0, 2, 3, 2, 0, 2],
      dtype=int32)
```

- **Results :**

  The 4 clusters obtained as a result of k-means clustering categorized neighborhoods based on the frequency of occurrence of shopping malls:

  - Cluster 0: Neighborhoods with least number of shopping malls
  - Cluster 1 : Neighborhoods having highest number of shopping malls
  - Cluster 2 : Neighborhoods densely populated with shopping malls
  - Cluster 3 : Neighborhoods having moderate number of shopping mall

The results of clustering is visualized on the map below with Cluster 0 in red, Cluster 1 in blue, Cluster 2 in green and Cluster 3 in yellow.



- **Discussion :**
  From the above analysis, it seems that neighborhoods in Cluster 1 and 2 are densely populated with shopping malls and therefore, the owners have to face an intense competition. On the other hand, Cluster 0 has a very few numbers of shopping malls. Therefore, neighborhoods in Cluster 0 are suitable for the construction of a new shopping mall because they have a little or no competition and there are high chances that the shopping mall will produce more profit for the client. However, Cluster 3 can also be taken into consideration as there are only few shopping malls there. Lastly, the client should be suggested to avoid cluster 1 and 2 as they are densely populated with shopping malls.

- **Limitations and Suggestions for Future Reseaches :**
  In this solution, we only considered one factor i.e. the frequency of occurrence of shopping malls. However, there are many other factors that affect the location. Some of these are density of population, average income of

population in the neighborhood, type of occupation of population, etc. In future researches, all the above-mentioned factors can also be taken into consideration.

- **Conclusion :**

  From the analysis done in our project, we conclude that the best location for opening a shopping mall are the neighborhoods present in Cluster 0.
  Also, it seems that neighborhoods in Cluster 1 and 2 are densely populated with shopping malls and therefore, the owners have to face an intense competition. Property dealers should avoid neighborhoods present in Cluster 1 and 2.
  On the other hand, Cluster 0 has a very few numbers of shopping malls. Therefore, neighborhoods in Cluster 0 are suitable for the construction of a new shopping mall because they have a little or no competition and there are high chances that the shopping mall will produce more profit for the client. However, Cluster 3 can also be taken into consideration as there are only few shopping malls there.

- **References :**

  - https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur
  - https://foursquare.com/developers/apps
  - https://python-visualization.github.io/folium/
  - https://pypi.org/project/geocoder/
  - https://pypi.org/project/beautifulsoup4/
  - https://brainly.in/question/14669042