COURSERA CAPSTONE PROJECT

Problem Statement :

Opening a new shopping mall in Toronto

By : Amrita Ghosh

Introduction :

Shopping malls are always a very trendy thing and always get good attention from residents as well as tourists. Now, there are shops in shopping malls which cater to different budgets. Shoppers prefer shopping centres/malls to stand-alone shops for various reasons. They have their own parking facility. There is a wide variety of products available. There are products from competing producers available under one roof. So, they can compare and make purchases. They have facilities such as restrooms. They have gaming zones. There are food courts with a wide variety of cuisine. There are movie theatres in shopping centres. All these features making shopping a fun-filled and satisfying experience. Since shopping centres are the most sought-after shopping destinations, it is beneficial for a businessman to set up a store in a shopping mall. Generally, retail store owners rent shop space in a mall. Renting store space benefits the businessman in many ways. Shopping malls are usually located in prime locations which are easily accessible. So the location of the mall is a very important decision to be taken.

Problem Statement :

The objective of this Capstone Project is to help opening a new shopping mall in Toronto. Using the location data obtained and the various clusters of venues in the neighbourhoods of Toronto we can find a prime location for this shopping mall ensuring that this becomes a successful project.

Target audience :

This project is useful to the property developers and investors looking to open or investing new shopping malls in the financial city of Toronto. According to research, many of the top shopping malls of Canada are located in Toronto being led by Yorkdale Shopping Centre ,Toronto ON, having productivity nearly $1,905.00.Thus the shopping mall business is beneficial in this area .All this makes me think that this will be a highly prospectus project for the present times as well as for the future. The shopping mall is a good business idea due to the country's continued obsession with building more shopping space despite oversupply.

Data section :

The various data sources to be used are :

1. We use the Foursquare location API data to explore the venues around any neighbourhood .This helps us to get the information regarding the popularity of that neighbourhood .

2. We need to find the geographical location like latitudes and longitudes of the venues to create a map. This ensures that we can visualize our data properly so that we can determine the best location in the neighbourhood for our shopping mall.

3. After clustering the neighbourhoods we need to find out the more prominent clusters. As more prominent the cluster will be more people would visit the shopping mall resulting in good money.

Sources of data :

We get the data about the postal codes ,boroughs and neighbourhoods of Toronto from the provided Wikipedia site :

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

We can use web scraping to extract the data regarding the borough and the neighbourhoods in Toronto. Using data wrangling or data cleaning methods we can easily clean the data.

To get the latitudes and longitudes of all the latitudes and longitudes of the neighbourhoods we use the csv file :

```
http://cocl.us/Geospatial_data
```

Then we will get the latitudes and longitudes of the coordinates using the geocoder package in Python.

```
address = 'Toronto'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto City are {}, {}.'.format(l
atitude, longitude))
```

We will use the Foursquare location API call to explore the venues around these neighbourhoods.

```
url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}".format(
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        long,
        radius,
        LIMIT)
```

We will mainly focus on the shopping mall category of the venue to solve our problem.

This project involves use of many data science skills like data analysing, powerful data visualizing tools like folium and machine learning tools like K- means clustering.

Methodology :

First we go the Wikipedia site and extract the names of all the boroughs and neighbourhoods of Toronto using web scraping. This uses the python packages pandas and lxml parser. After getting all the data we convert that into a pandas dataframe having columns of postal code, borough and neighbourhoods.

Now we have to begin the exploratory analysis and data wrangling. For this we remove those rows which do not have a borough and group the neighbourhoods having the same borough. The rows which do not have a neighbourhood are given the same names as that of the borough.

Then we get the csv file containing the latitudes and longitudes of all the neighbourhoods of Toronto and finally merge both the data frames to produce a data frame containing the postal codes ,borough, neighbourhoods, latitudes and longitudes.

Then we use the geocoder package to get the latitudes and longitudes of Toronto and the folium package to create a map of Toronto to show all the neighbourhoods.

Using the foursquare API we explore the top 100 venues in the neighbourhoods within a radius of 500 m. Not only did we get the venue name but also its latitude, longitude and venue category.

Then we can group the boroughs and neighbourhoods in the neighbourhood and perform one hot encoding on that data. This produces a data frame with dummy variables showcasing various venue categories like airports , shopping malls.

After this we can use the unsupervised machine learning tool K-means Clustering to divide all the neighbourhoods in Toronto. In this module we divide it into 5 clusters to visualize which clusters have more interesting venues .

Finally after examining individual clusters we decide to examine the presence of shopping malls in Toronto as the main problem statement is to find a good location for our shopping mall.

We create a new data frame containing the neighbourhoods and shopping malls columns only by slicing .

Then we merge this data frame with the one containing the postal codes and borough.

Finally to examine the spread of shopping malls we use K-means clustering again (this time number of clusters =3)to get the shopping malls data of various clusters.

Also when we found out the number of shopping malls in Toronto, there are about 39 .

Then we separately examine the shopping malls in each of the three clusters.

Results :

Thus after getting the result from the K-means clustering we see that we can cluster the neighbourhoods into 3 regions based on the frequency of occurrence of shopping malls.

Cluster 1 : Cluster with a very high concentration of shopping malls.

Cluster 2 : Cluster with a moderate concentration of shopping malls.

Cluster 3 : Cluster with no or very low concentration of shopping malls.

Discussions :

From the results obtained the investors and property developers should try investing in a shopping mall in either cluster 2 or cluster 3.Cluster 1 has more than 30 shopping malls which are already in stiff competition and oversupply. Thus opening a mall would be highly favourable in Cluster 3 as it has only one shopping mall till date thus providing the investors with almost zero competition. Cluster 2 faces only a moderate competition as there are few shopping malls over here but it is also a favourable location for establishing a new mall.

Conclusion :

This project is solely based on the idea of opening a shopping mall on the frequency of the number of the shopping malls in the neighbourhoods but opening a mall requires taking more factors. Additional factors would be the population in the nearby areas, land spaces that is the availability of free space ,commercial or residential area ,the income of the people living there. So the location data needs to be refined with the data for all these factors so that we can get the best location for establishing the shopping mall.

But apart from that based on the frequency we concluded that areas in cluster 2 and 3 are highly favourable for opening a shopping mall. We used the concepts of data wrangling, exploratory analysis, data visualising and machine learning tools like K-means clustering to complete this project.