IBM Applied Data Science Capstone

Opening a New Mall in Vancouver, British Columbia

By: Andrew Shrestha

Date: February 6th, 2021

Introduction:

Vancouver has been ranked consistently as one of the best cities to live in the world. This is due to its beautiful and accessible nature environments with mountains, ski resorts, hiking trails, beaches, parks, and shopping experience. These factors tend to be the driving factor for tourism in Vancouver since its ranking is number 2 most popular destination in Canada with over 15 million tourists per year.

Vancouver is highly successful in the retail industry, as it hosts 4 of the top 10 malls (in terms of sales per sqft.) in Canada. These numbers are a great indicator of the general population, where the public/tourists are very inclined to visit and spend time in malls to meet their growing demand for greater shopping experience. Malls are successful at generating great revenue for both property developers/ owners, as well as the city itself, and so it would be of great interest to look into pursuing and expanding this avenue.

Business Problem:

The objective of this Capstone Project will be to utilize data and select the best locations in the city of Vancouver to open a new Shopping Mall. Through the use of techniques such as foursquare analysis, clustering, and segmenting, we would like to solve the following business problem: “Where would be the optimum location to construct a new mall for property developers/owners in the city of Vancouver?”

Data:

- A list of neighborhoods in Vancouver. These will be the locational data that we will confine our analysis to
- Both the Latitude and Longitude coordinates of each of the neighborhood data points so that we may plot it on a map
- Data on Shopping Malls or related so that we may use this in order to perform clustering on the neighborhoods

Data source:

- Obtain a list of 45 neighborhoods from Wikipedia (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Vancouver)
- Web scraping techniques to grab the data from Wikipedia (python and Beautifulsoup)
- Geocoder to extract longitude and latitude of the neighborhoods
- Foursquare API to obtain shopping mall venue data of each neighborhood
- Folium for map visualizations

Methodology:

The first step we need to take in this project is to obtain the number of neighbourhoods in the selected location of Vancouver through our Wikipedia source: (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Vancouver). This is done by using a combination of web scraping via python and beautifulsoup packages to grab the list and prepare it for further analysis.

Once this is accomplished, we will be extracting the geographical points/coordinates of longitude and latitude from each location using our geocoder.

After the longitude and latitude data is gathered, we will be converting the data to a more visual representation with the use of pandas dataframe in combination with the Folium package. This will create a map visual of the data and neighborhoods in question.

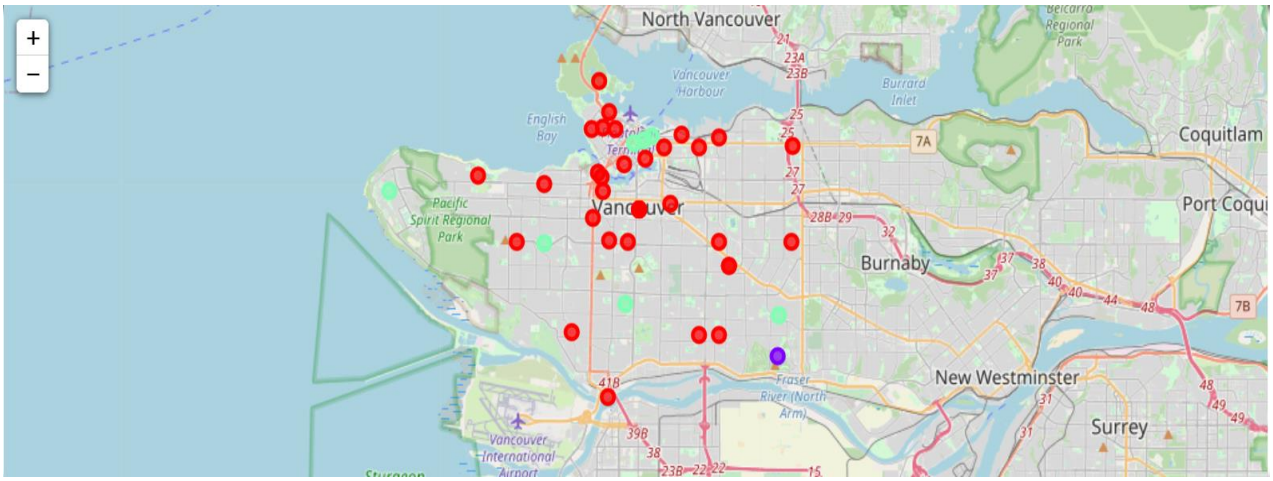
Using the Foursquare API, we can run the coordinates through calls and grab the venue longitude and latitude, category, and name. This means that we are able to check the unique values and categories that are returned from all venues. We will also be prepping the data for our next clustering step through the use of grouping neighborhoods as well as the frequency mean of each category. (the category of "Shopping Mall" will be used as a filter category for each neighborhood)

Finally, we will cluster the data through the unsupervised machine learning technique of k-means. Clustering into 3 different groups with "frequency" of shopping malls being the filter, we will be able to figure out the concentration of these malls in different areas and areas of opportunity, thus helping us in solving the question of which location would be best to open a new shopping mall.

Results:

We have obtained results on 3 clusters of neighborhoods given the frequency of the "Shopping Malls" they have in the vicinity. The results are as follows:

- Cluster 0 = Neighborhoods that have no presence of shopping malls
- Cluster 1 = Neighborhoods that have the most significant presence of shopping malls
- Cluster 2 = Neighborhoods that have a moderate presence of shopping malls



The results of the clustering are shown in the above map where the red points are Cluster 0, purple are Cluster 1, and light green are Cluster 2

Discussion:

Continuing the conversation noted in the observations of results within the project, We see that Cluster 0 has no shopping mall presence, where as Clusters 1 and 2 have high and moderate mall presence respectively. Based on this information, there is an opportunity to utilize the fact that Cluster 0 has no mall presence as a way to have successful interest and demand as it will be the first one available and also this cluster will see little competition with other malls due to its locational advantage of being in Cluster 0.

We also see that Cluster 0 has the most points and is more wide spread geographically, which means that the population living in these areas will not have to travel as far or use the substitute of online shopping, thus ensuring a decent demand and attraction towards this new mall.

Therefore, it will be advised to any prospective project managers/ developers to build a mall in the area of Cluster 0 and avoid Cluster 1 as it has the highest concentration of shopping mall presence.

Conclusion:

Through our use of various data preparation and machine learning techniques, we have analyzed and successfully identified the solution to the business problem of where we could build a new shopping mall with the goal of maximizing profits. With the help of our visualizations and data clustering, we have identified that it would be most beneficial to recommend prospective developers to build a new mall in the area of Cluster 0 as it shows the most opportunity for generating demand and success. We can also derive from this that we should stay away from Cluster 1 when allocating a location due to the fact that this cluster has the highest presence of shopping malls in the area.