**IBM Data Science Project: Select the best location for a new restaurant in Vancouver**

Background:

For the IBM Data Science Capstone Project, we are trying to answer to the following business problem.

A restaurant business owner with multiple locations opened across Canada decides to open a new restaurant in Vancouver. The new unit is going to be focused on Asian cuisine, which is the main specialization gastronomy area on which the restaurants chain is focusing.

Considering the high real-estate prices across Vancouver, intense competition and the high rates that the restaurant plans to apply, one of the variable to decide upon is the right location. The intention of the owner is to find an optimal location in an area which is close to sceneries eating delights, high frequency tourist sections of the city and easily accessible to wealthier inhabitants.

The analysis can be driven by using unsupervised machine learning to create clusters of district areas potentially candidates for the optimal location. The new restaurant will be situated closest to culinary centers and tourist attractions.

Data:

To conduct the analysis we need the following sets of data:

1. List of the main districts of Vancouver which is obtained through a csv file. List of districts to be obtained from Wikipedia website: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Vancouver
2. Geo-coordinates of districts selected at point 1 above to be retrieved in using geocoder tool
3. Venues from each district are collected using Foursquare API.

Problem solve:

After obtaining the complete data in the desired formatting, we apply the k-means methodology in order to create cluster of districts and determine areas where the restaurant should be located.

Analysis begins by uploading wiki data through csv. file and creating a list of districts of Vancouver, together with the geo-coordinates of each district. Basically, the imported list of districts is used in geocode python library to get the latitude and longitude of each district in the list. Districts and their coordinates are stored in a pandas data frame format. When done, this includes the following details: District, Name, Latitude and Longitude.

| | District | Name | Latitude | Longitude |
|---|---|---|---|---|
| 0 | Vancouver-Downtown | Downtown | 49.283393 | -123.117456 |
| 1 | Vancouver-False Creek | False Creek | 49.274751 | -123.106131 |
| 2 | Vancouver-Fraserview | Fraserview | 49.218416 | -123.073287 |
| 3 | Vancouver-Hastings | Hastings | 49.280673 | -123.032600 |
| 4 | Vancouver-Arbutus ridge | Arbutus ridge | 49.240968 | -123.167001 |
| 5 | Vancouver-Kingsway | Kingsway | 49.256732 | -123.089712 |
| 6 | Vancouver-Langara | Langara | 49.219437 | -123.118026 |
| 7 | Vancouver-Mount Pleasant | Mount Pleasant | 49.263330 | -123.096588 |
| 8 | Vancouver-Point Grey | Point Grey | 49.264019 | -123.195022 |
| 9 | Vancouver-Quilchena | Quilchena | 49.243838 | -123.149094 |
| 10 | Vancouver-West End | West End | 49.284131 | -123.131795 |
| 11 | North Vancouver-Lonsdale | North Lonsdale | 49.343624 | -123.072751 |
| 12 | North Vancouver-Seymour | North Seymour | 49.556758 | -123.045975 |
| 13 | North Vancouver-Capilano | North Capilano | 49.342203 | -123.111585 |
| 14 | North Vancouver-Garibaldi | North Garibaldi | 49.740507 | -123.083205 |
| 15 | Vancouver-Burrard | Burrard | 49.285636 | -123.119815 |
| 16 | Vancouver-Little Mountain | Little Mountain | 49.241853 | -123.113496 |

The next step is to retrieve the venues of each district. This is completed with the help of Foursquare.com credentials via API. Data is retrieved in json format. We setup a limit of 100 venues for each district and a radius of 1000 meter from the coordinates of district center. In addition to that, we determine which venues are the most common within each district. First 10 venues retrieved are shown below:

| | Neighborhood | Neighborhood Latitudine | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Vancouver-Downtown | 49.283393 | -123.117456 | Rosewood Hotel Georgia | 49.283429 | -123.118911 | Hotel |
| 1 | Vancouver-Downtown | 49.283393 | -123.117456 | Gotham Steakhouse & Cocktail Bar | 49.282830 | -123.115865 | Steakhouse |
| 2 | Vancouver-Downtown | 49.283393 | -123.117456 | Hawksworth Restaurant | 49.283362 | -123.119462 | Lounge |
| 3 | Vancouver-Downtown | 49.283393 | -123.117456 | SEPHORA | 49.284092 | -123.117204 | Cosmetics Shop |
| 4 | Vancouver-Downtown | 49.283393 | -123.117456 | Abercrombie & Fitch | 49.282274 | -123.118685 | Clothing Store |
| 5 | Vancouver-Downtown | 49.283393 | -123.117456 | Hyatt Regency Vancouver | 49.284934 | -123.120407 | Hotel |
| 6 | Vancouver-Downtown | 49.283393 | -123.117456 | Vancouver Art Gallery | 49.282827 | -123.120457 | Art Gallery |
| 7 | Vancouver-Downtown | 49.283393 | -123.117456 | The Keg Steakhouse + Bar - Dunsmuir | 49.283438 | -123.116363 | Restaurant |
| 8 | Vancouver-Downtown | 49.283393 | -123.117456 | Mogu: Japanese Street Eats | 49.284118 | -123.117531 | Food Truck |
| 9 | Vancouver-Downtown | 49.283393 | -123.117456 | Disney store | 49.281689 | -123.119850 | Toy / Game Store |
| 10 | Vancouver-Downtown | 49.283393 | -123.117456 | Cartems Donuterie | 49.283833 | -123.113554 | Donut Shop |

After collecting this additional data, a new data frame includes the districts and separate columns for “n” most common venues of each districts. At this stage, the columns look like this: Neighborhood, 1st most common venue, 2nd common venue and so on up to the 10th most common venue.

| | Neighborhood | 1st Most common venue | 2nd Most common venue | 3rd Most common venue | 4 Most common venue | 5 Most common venue |
|---|---|---|---|---|---|---|
| 0 | North Vancouver-Capilano | Coffee Shop | Bank | Sandwich Place | Park | Convenience Store |
| 1 | North Vancouver-Lonsdale | Coffee Shop | Bar | Chinese Restaurant | Sandwich Place | Park |
| 2 | Vancouver-Arbutus ridge | Sushi Restaurant | Park | Bakery | Tea Room | Coffee Shop |
| 3 | Vancouver-Burrard | Hotel | Dessert Shop | Food Truck | Coffee Shop | Sandwich Place |
| 4 | Vancouver-Downtown | Hotel | Dessert Shop | Coffee Shop | Restaurant | Sandwich Place |

Unsupervised machine learning will be applied by using the K-means methodology. In order to do this, first we need to use one-hot encode to create dummy variables to transform the venue categories values and allow the machine learning process.

K-means requires an optimal number of clusters to be used. For determining the appropriate district clustering, the parameter for the optimal number of clusters will be identified by using silhouette score approach. We create a chart to show the silhouette scores for a range number of clusters. The highest score on the chart becomes the optimal number of clusters to initiate.

The number of clusters mentioned above is going to be used in the K-means process. The end result will have each district assigned with a cluster label into the data set.

Results:

The clustered data will let us know which cluster is the best for the solution of our problem. Most common venues and their frequency are a valuable indicator when considering the cluster to include the potential restaurant location. We will advise the owner to consider district from the cluster where most of the lively part of the city is present with a lot of gastronomy and tourist venues on site.

Conclusion:

By using various Python libraries we are able to analyse and provide the output and recommendation to support decisional process. As a result, business owner selects the most profitable location with the most benefits available to his customers.