Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Bakery in Vancouver, BC

Parisa Shiri

August 2019

Introduction

Vancouver is a beautiful city which attracts many tourists from all over the world each year. Trying out new food places such as restaurants, cafes and bakeries are one of the favorite activities that tourists do. Among these places, bakeries with special treats and sweets bring many people in the store. From tourists to locals, no one can resist the smell of freshly baked pastries. Therefore, there are many bakeries around Vancouver, with possibility of building more of them. This can be particularly a good opportunity to collaborate with big artisan bakeries from other countries and bring them to Vancouver as a new business.

Business Problem

When deciding to open a new bakery, there are many questions that need to be answered to guarantee a successful and profitable business. One of these questions is about the location of the bakery. Which neighborhood is the most suitable one? How are different bakeries distributed in the area? The objective of this report is to explore the neighborhoods in Vancouver to find out the best possible locations to open a new bakery by using data science methodology and machine learning techniques such as clustering.

Data

The following data are used to find the best possible locations for opening a new bakery in Vancouver:

- List of neighborhoods in Vancouver with their coordinates. This list can be found and downloaded from:
  https://data.vancouver.ca/datacatalogue/localAreaBoundary.htm
  The file formats available on this website are KML, SHP, Google Map, XLS and CSV. In order to read the data in Jupyter Notebook, an online file converter was used to first convert KML to GeoJSON, which is a form of JSON file that can be easily loaded and represented as dataframe. The file contains name of neighborhoods in Vancouver and their geometry coordinates. In order to represent each neighborhood with only one latitude and longitude rather that its polygon shape, the data can be manipulated to find the average of latitudes and longitudes for each neighborhood (discussed in detail later).

- List of bakeries in Vancouver. This data was obtained using Foursquare developer account. For more information, please refer to https://developer.foursquare.com

Methodology

First, the information of the neighborhoods in Vancouver are obtained from the city of Vancouver website at:

https://data.vancouver.ca/datacatalogue/localAreaBoundary.htm

The file is converted into geojson file using an online file converter. The geojson file then is loaded in a jupyter notebook. When the data is populated in a pandas dataframe, it can be seen that each neighborhood has a series of latitudes and longitudes referring to each corner of the neighborhood. On top of that, this information is in the form of a list with extra unnecessary brackets. In order to use forsquare API later, we need to represent each neighborhood with one latitude and longitude rather than its geometrical shape. Therefore, first the list needs to get flattened to eliminate extra brackets. Then the latitudes of all corners of the neighborhood are averaged to find the latitude of the neighborhood. Same thing is done for longitudes as well. Now we have neighborhood names and their coordinates which can be easily visualized using folium library in Python.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 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 neighborhoods 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 neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each

neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the data for “bakery”, we will filter venue category with “bakery” for the neighborhoods. 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 will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Bakery”. The results will allow us to identify which neighbourhoods have higher concentration of bakeries while which neighbourhoods have fewer number of bakeries. Based on the occurrence of bakeries in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new bakeries.

Results

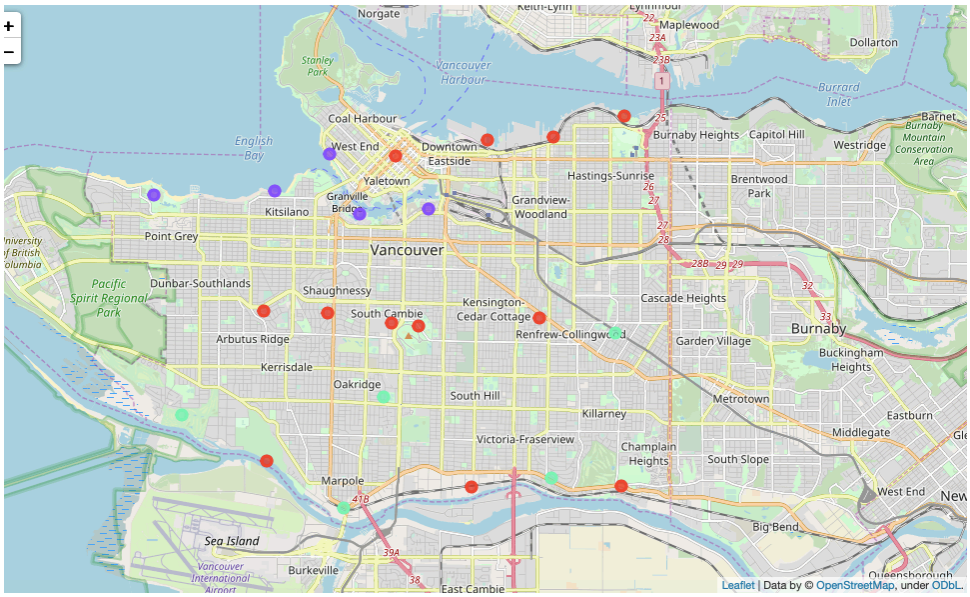
The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Bakery”:

• Cluster 0(red): Neighbourhoods with moderate number of bakeries

• Cluster1(purple): Neighbourhoods with high number of bakeries

• Cluster 2(mint): Neighbourhoods with low concentration of bakeries.

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

As observations noted from the map in the Results section, most of the bakeries are concentrated in the northern and downtown areas of Vancouver, belonging to cluster 0 and 1. On the other hand, cluster 2 has very low number to no bakeries in the neighbourhoods. This represents a great opportunity and high potential areas to open new bakeries as there is very little to no competition from bakeries. Meanwhile, bakeries in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of bakeries. From another perspective, the results also show that the oversupply of bakeries mostly happened in the downtown area of the city, with the southern areas with very few bakeries. Therefore, this project recommends property developers to capitalize on these findings to open new bakeries in neighbourhoods in cluster 2 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new bakeries in neighbourhoods in

cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 1 which already have high concentration of bakeries and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of bakeries, there are other factors such as population and income of residents that could influence the location decision of a new bakeries. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new bakery. In addition, this project made use of the free developer account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results. Another suggestion for future is to consider the classification bakeries based on specialty as well, as some bakeries are more popular for their bread than artisan bakery and this can affect the final decision for the location of the new bakery.

Conclusion

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 and investors regarding the best locations to open a new bakery in Vancouver. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 2 are the most preferred locations to open a new bakery. 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 bakery.

References

- Neighborhoods in Vancouver. Retrieved from: https://data.vancouver.ca/datacatalogue/localAreaBoundary.htm

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

Appendix

Cluster 0 (moderate)	Cluster 1 (high)	Cluster 2 (low)
Arbutus-Ridge	West Point Grey	Renfrew-Collingwood
Downtown	Mount Pleasant	Dunbar-Southlands
Sunset	West End	Victoria-Fraserview
Strathcona	Kitsilano	Oakridge
Grandview-Woodland	Fairview	Marpole
Hastings-Sunrise		
Kensington-CedarCottage		
Kerrisdale		
Killarney		
South Cambie		
Shaughnessy		
Riley Park		