# Coursera Capstone

IBM Applied Data Science Capstone

**Opening a new Restaurant in Calgary, Alberta, Canada**

*By Jimmy Zhu*

*August 2020*

# Table of Contents

# Introduction

Keeping a restaurant running requires a continuous stream of eaters placing orders for the food that the restaurant offers. The quality of the food, marketing plan, and competitive pricing are all factors that will drive that traffic but location is also a key decision for the restaurant owner to make which can allow the restaurant to flourish. The more restaurants in a given neighborhood, the greater the odds that restaurant traffic may diminish. Hence, the business question is what neighborhoods are underserved by restaurants and what are some of the restaurants that already have a presence in that area. This question can apply for brand new restaurant owners who are looking to enter a market or for restaurant owners looking to expand into regions with similar competitive landscape.

# Data

Location data will be webscraped from Wikipedia identifying neighborhoods belonging to the same postal code within Calgary, Alberta, Canada. Each neighborhood will be passed into the Foursquare API to retrieve data of the venues within that neighborhood. There will be a maximum of 100 venues within 500m of the latitude and longitude of each respective neighborhood. The data retrieved from Foursquare includes the venue name, venue latitude, venue longitude, and venue category. The venue category describes the category that a venue belongs to, such as Sports Bar for 'Shark Club Sports Bar & Grill'. Venues returned belong to a large variety of venues from banks to hotels to restaurants so an additional column titled 'General Venue Categories' will be added. This new column will generalize the venue category into their appropriate business sectors such as Banks, Food, Entertainment, Health and Fitness, Hospitality, Retail, and Services. Now each venue has a General Venue Category assigned to it so we can compare businesses which are more related to one another. The excel file 'Venues.xlsx' will be created to store the General Venue Categories for each venue category given by Foursquare. The excel file will be read then the data reinserted into the pandas dataframe for further processing to review the relevant dataset for the project.

In this project, the venues belonging to 'FOOD' will be isolated and grouped together by the neighborhood. Within a pandas dataframe, each unique postal code and the corresponding neighborhood(s) belonging to that postal code will have a single row with the quantity of restaurants within those neighborhood(s) and all the names of each of those restaurants. K-means clustering will be applied to the dataset based on the quantity of restaurants in each neighborhood(s) so that five clusters are defined to distinguish areas that have similar quantities of restaurants. An interactive map showing Calgary will include markers color coded by the cluster they belong to with a popup which identifies the name of the neighborhood(s) belonging to a unique postal code, the cluster label, and the restaurants that are in that area. Using this, restaurant owners can make decisions of where to start up their restaurant.

# Methodology

Knowing that the restaurant wants to open up in Calgary, the code begins by obtaining a dataset of the neighborhoods within Calgary based on their unique postal code. This is available on the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T) but the table on this Wikipedia page includes information about boroughs across Alberta. Webscraping will be performed to extract the data from the table on the website by using the Python requests and beautifulsoup packages. The table headers are also webscraped and passed into the dataframe so that the entire pandas dataframe is populated with the appropriate headers and data to match the table provided on Wikipedia. The table includes the Postal Code, Borough, Neighborhood, Latitude, and Longitude. Therefore, all the geodata necessary to locate each neighborhood is now available if the original table was complete.

The pandas dataframe is checked to ensure completeness of data to avoid errors while data wrangling. Tests are run to check if there are any neighborhoods not assigned. If there are unassigned neighborhoods, then the corresponding borough will be set as the neighborhood. It appears that boroughs without assigned neighborhoods are smaller boroughs hence the absence of neighborhood names. Any boroughs that are missing latitude and longitude data is passed into the Nominatim module from the geopy.geocoders library in order to obtain the appropriate latitudes and longitudes.

As this business problem pertains to restaurant owners in Calgary, AB, the location dataframe is filtered to only include the Calgary borough. Each row has the latitude and longitude coordinates necessary to locate each neighborhood via Foursquare and obtain a list of information about venues within each neighborhood. Since this Jupyter notebook is being shared online publicly, environmental variables were created and placed on local storage. These environmental variables are then accessed as the variables CLIENT_ID and CLIENT_SECRET within the Jupyter notebook whenever calls to the Foursquare API are made. The function getNearbyVenues receives a series of arguments for the name of the neighborhood(s), neighborhood latitude, and neighbor longitdues then sends out an explore query to the Foursquare API for each neighborhood. The API returns a json file with up to 100 venues within 500m of each neighborhood latitude and longitude coordinates. The function takes the relevant data from the JSON and saves it into a pandas dataframe with the headers; Neighborhood, Neighborhood Latitude, Neighborhood Longitude, Venue, Venue Latitude, Venue Longitude, and Venue Category.

The venue category describes the category that a venue belongs to, such as Sports Bar for ‘Shark Club Sports Bar & Grill’. Venues returned belong to a large variety of venues from banks to hotels to restaurants so the venue category column is exported to an excel file called ‘Venues.xlsx’. Using the excel file, a General Venue Category will be assigned to each venue category returned by Foursquare. These General Venue Categories will be Banks, Food, Entertainment, Health and Fitness, Hospitality, Retail, and Services. While tedious, this excel can be read back into the code and attached to the venues pandas dataframe for further

processing of the data. Should this code be repeated for another city, the code will check the 'Venues.xlsx' spreadsheet to identify the General Venue Category of all the venue categories provided by Foursquare. If there are new venue categories, then those new venue categories will be appended to the bottom of the 'Venues.xlsx' spreadsheet for additional user input for identification of the General Venue Category. Upon further updating of this spreadsheet, less user input will be required. Please note the code in this Jupyter Notebook is setup so that it is run cell by cell following the instructions provided in the markups.

In this project, the venues belonging to 'FOOD' will be isolated and grouped together by the neighborhood. Within a pandas dataframe, each unique postal code and the corresponding neighborhood(s) belonging to that postal code will have a single row with the quantity of restaurants within those neighborhood(s) and all the names of each of those restaurants. Now all the data for the relevant industry and the relevant area is present and ready to be clustered together. K-means clustering divides the data up into a user defined k-number of non-overlapping clusters. No internal structure is provided so this is an unsupervised algorithm that clusters objects (in this project, the quantity of restaurants) across the different clusters. Five clusters are selected for this project so that each neighborhood is placed into a cluster with other neighborhoods which have similar quantities of restaurants. These results in addition to the location information for each neighborhood are entered into a Folium map with markers at the longitude and latitude coordinate of each neighborhood. The markers are color coded to the cluster that each neighborhood belongs to and will provide a popup with the following information; the neighborhood name(s), the cluster label and the names of the restaurants within each neighborhood. This map allows restaurant owners to identify neighborhoods that are ripe for market entry or for market expansion into a competitive environment that is similar to their existing market.

# Results

The k-means clustering clustered neighborhoods with the following traits:

- Cluster 0: 1<= Quantity of Restaurants <= 3
- Cluster 1: 23<= Quantity of Restaurants <= 30
- Cluster 2: 38<= Quantity of Restaurants <= 38
- Cluster 3: 11<= Quantity of Restaurants <= 18
- Cluster 4: 4<= Quantity of Restaurants <= 5

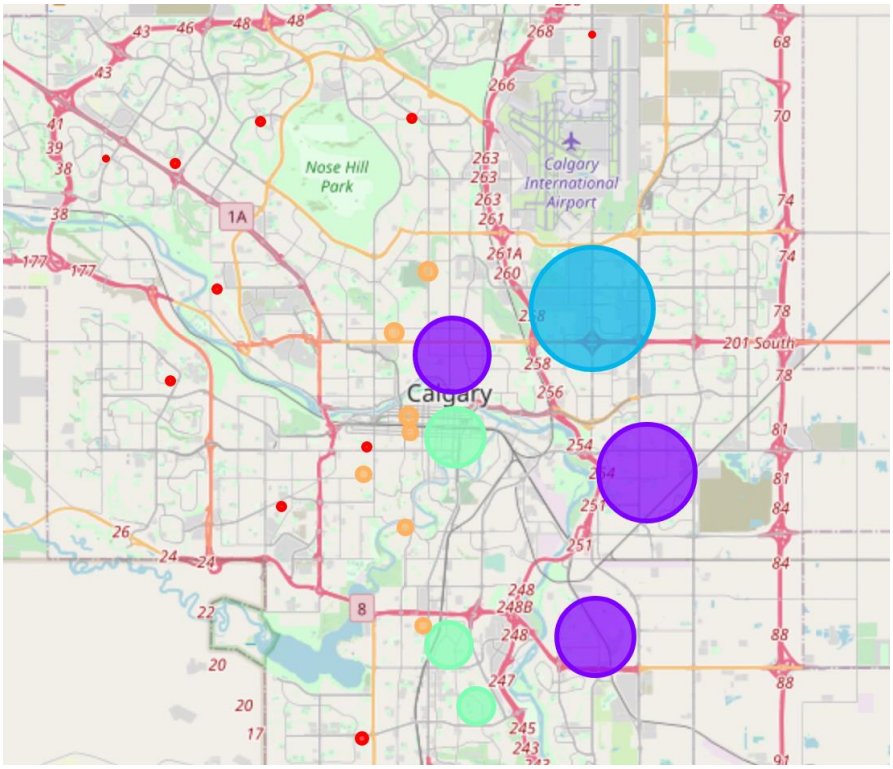


Figure 1: Folium Map with Markers for each Neighborhood segmented by the quantity of restaurants present

Table 1: Information about Restaurants in each Neighborhood (df_NumRestaurants)

| | Cluster Labels | Neighborhood | # of Restaurants | Venues |
|---|---|---|---|---|
| 0 | 2 | Inglewood, Burnsland, Chinatown, East Victoria... | 38 | Grumans Catering and Delicatessen, 5 Vines Win... |
| 1 | 1 | Connaught, West Victoria Park | 30 | Holy Grill, Posto Pizzeria and Bar, The Casbah... |
| 2 | 1 | Oak Ridge, Haysboro, Kingsland, Kelvin Grove, ... | 24 | The Keg Steakhouse + Bar - Macleod Trail, Bole... |
| 3 | 1 | City Centre, Calgary Tower | 23 | Gyu-Kaku Japanese BBQ, Holy Grill, Posto Pizze... |
| 4 | 3 | Rosscarrock, Westgate, Wildwood, Shaganappi, S... | 18 | Moti Mahal Restaurant, Shawarma Knight, Cluck ... |

# Discussion

The markers on Figure 1 are scaled based on the quantity of restaurants so just looking at the map tells you that the blue marker area has much more restaurants than any of the areas with a red marker. Patterns can be extrapolated for areas that have less restaurants and can provide franchise owners with information about finding the best location for restaurants. The location can be selected to avoid highly served areas, or to ensure that multiple restaurants are not cannibalizing sales. For restaurants that are looking to satisfy consumers that currently do not have as many options then it appears that neighborhoods on the far West and far north neighborhoods of Calgary are currently underserved in terms of the quantity of restaurants available. Perhaps for franchise owners, opening up multiple restaurants along the West side ring road would allow for strong market penetration and optimize transportation routes for improved delivery costs and freshness of ingredients.

These recommendations must be scrutinized based on the fact that the map provided in figure 1 only considers the quantity of restaurants. With additional tags identifying the type of food provided (such as bars, Asian, middle eastern, western European, etc), the k-means clustering could consider that secondary information (though this data would have to undergo one hot encoding before it can be clustered via k-means) and better cluster the neighborhoods based on the quantity of restaurants competing for the same consumer food palette. That is another element that restaurants may use when deciding whether or not to enter a particular neighborhood. Furthermore, some neighborhoods are likely more densely populated so the number of consumers in each neighborhood varies which limits the importance of the quantity of restaurants in each neighborhood. This is yet another variable that could have been considered when undergoing k-means clustering which better assists owners with deciding on the most lucrative neighborhood to enter. Further research would have to look at procuring these additional datasets for processing and review.

One important limitation may be the data set available from Foursquare for the city. It is possible that information is outdated and incomplete so there must be verification on the completeness of the Foursquare dataset retrieved.

# Conclusion

To answer the question of which neighborhoods would be ideal locations to open up a restaurant in Calgary, AB, the quantity of restaurants within each neighborhood was used. This began with finding all of the neighborhoods within Calgary Alberta. 122 postal codes with one or more neighborhoods within Calgary were identified along with their corresponding longitudes and latitudes. These were locations were used as center points for the Foursquare API to identify nearby venues within each of these neighborhoods. After processing the venues data that was returned, cluster 0 or the red markers in Figure 1 identify locations where consumers are underserved by the quantity of restaurants present. Now that users have a visual aid in identifying potential markets, those users can click on the marker which opens a popup that identifies the names of all the restaurants in those neighborhoods. Users can identify their future competition and make decisions on whether their restaurant can thrive in those environments.