IBM Data Science Capstone Project – Week 1

The Problem

Opening a restaurant, or any business for that matter, is one of the bigger decisions that a person can make in their lives. It is a huge decision, and decisions made during set up can make or break the business. An example of such decisions includes:

- Restaurant theme
- Restaurant location
- Timing of opening

For the purpose of this study, we will focus on location, namely within which of the postal code areas (i.e.: T2H vs T1Y), and various other parameters (population and income measures) in an attempt to create clusters of postal code areas to assist in making this decision a little easier.

According to the following CNBC article, approximately 60% of all restaurants fail within the first year, and 80% fail prior to their 5th (https://www.cnbc.com/2016/01/20/heres-the-real-reason-why-most-restaurants-fail.html). The number one reason cited is choice of location – in this case, that choice really does seem to be 'do or die'.

In the city of Calgary, Alberta, a shift in overall person and business traffic flow since the oil industry downturn (beginning in late 2014) has further complicated this decision. Historically, the downtown core was the hot spot for openings of new restaurants, with revenue fed by a seemingly never-ending flow of clients with large expense accounts. That was when office vacancy rates were approximately 3%. Today they are hovering around 30%, with at least 100,000 fewer people making the daily trip to the downtown core.

Additionally, there have other forces at play that complicate the overall picture. Some have been sparked by the COVID-19 pandemic, others by general changes to the City of Calgary's overall community planning philosophy. These considerations include:

- Community Population/Density: Calgary is historically a city of suburb communities, full of single-family homes. In recent years, due to the ever expanding 'urban sprawl', there has been an increasing push for multi-family densification. These demographic changes should be considered in any analysis
- Walkability/ride-ability: another push for increased walkability of communities, and an increasing number of cycling lanes on community roads (with the corresponding reduction in the ease of driving and finding parking) should be considered when evaluating communities to open in, especially when considering serving alcohol

It should be noted that cost of operation is certainly a big consideration. However, by community area, there is no repository for property tax and rental rates published within Calgary. For this reason, the analysis will focus solely on the potential for revenue generation, with the assumption that a savvy business owner will be able to control costs in order to achieve profitability.

The Data

To create this evaluation/model, the following data will be used to cluster neighborhoods:

- Calgary Postal code (and contained neighborhood) data, complete with the geographic center of each postal code area. Note there are 34 postal code areas located within the City of Calgary proper (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T)
- Population by postal code area: taken from the 2016 census analysis (StatsCan) https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&SR=1&S=22&O=A&RPP=9999&PR=0
- Income data by community area: although I was unable to get income data per postal code area, I was able to get the data from the Calgary Mortgage and Housing corporation (CMHC). Since it was by area, I had to assign each community a postal code and group by postal codes to average the income per community. Any postal codes with no reported income metrics were assigned the median of all the Calgary communities/postal code areas

  For the analysis, I chose to focus on after tax income, so kept only median household income after tax. A link to the data follows: https://www03.cmhc-schl.gc.ca/hmip-pimh/en/TableMapChart/TableMatchingCriteria?GeographyType=MetropolitanMajorArea&GeographyId=0140&CategoryLevel1=Population%2C%20Households%20and%20Housing%20Stock&CategoryLevel2=Household%20Income&ColumnField=HouseholdIncomeRange&RowField=Neighbourhood&SearchTags%5B0%5D.Key=Households&SearchTags%5B0%5D.Value=Number&SearchTags%5B1%5D.Key=Statistics&SearchTags%5B1%5D.Value=AverageAndMedian
- Community Walkability score: since Calgary publishes this index per community, communities will need to be grouped into postal code areas, like what was done for the income analysis, and a median walkability score per postal code area will need to be stored. A link follows: https://www.walkscore.com/CA-AB/Calgary
- Foursquare API: the Foursquare API will be utilized to provide information on the number and type of restaurants that are within a specified distance from the center of the postal code area. This will be used to determine the number and type of restaurants already present, and as part of the information used to cluster the communities by their potential suitability for a new restaurant

The data will be grouped (after many calculations, merges, and joins) into a single data frame for analysis. Multiple tests will be run with different numbers of clusters (6, 8, and 10 are being thought of for the beginning analysis) to determine the model that best delineates the suitability of communities/postal code areas for analysis. Finally, the clusters will be classified (i.e.: rich, dense, walkable vs medium income, spread out, poor walkability) to give areas a profile for suitability of business set up