IBM Data Science Capstone Project

Introduction:

A stakeholder is investing in a new Fast Food restaurant in Toronto, CA. As an investor, they want to make sure that the owner of this restaurant is considering many variables when deciding which location to choose for their new restaurant. The end goal, of course, for the stakeholder and the owner is profitability of this new restaurant. Many factors can affect profitability of a restaurant, so I am going to do an analysis of some of those variables so that the owner can make a more informed decision regarding the location of this new restaurant.

The questions and thus variables that I will be trying to address here are limited, but powerful. Through the data described in the next section, I will be able to answer the following questions, which will be very useful in the decision-making process for this owner and stakeholder.

1) How many neighborhoods are there in this postcode area?
    a. This will be a proxy for population of the postcode since more neighborhoods implies higher population.
2) How many restaurants are there overall in each postcode?
    a. This will indicate how crowded the industry is overall in each postcode, which will help the owner decide if there is room in a certain postcode for new competition.
3) How many similar restaurants are there in each postcode?
    a. This will allow the owner to see how many restaurants there are that are in a similar category, which will provide information about direct competition to his new restaurant.

With these questions answered, I will cluster the postcodes using the k-means algorithm, which will allow us to more easily balance total number of restaurants with the number of direct competitors in fast food restaurants.

Description of the data:

I will be using two main data sources. First, I scraped the website https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M to create a pandas data frame that includes the postcode, borough name, and neighborhood names within each postcode. I also used the geopy geocoder library in python to get the latitude and longitude of each postcode as well. This dataset allowed me to answer the first question above. Second, I used the foursquare API to input the geographical information from the first dataset to gather information about restaurants near each postcode. This dataset allowed me to answer questions two and three above.