Applied Data Science Capstone

Final Report

Neal T. Murphy

Introduction Section:

Foreign investors are looking at various cities in the United States to move and open a restaurant. They are wanting to learn more about prospective cities and in which neighborhoods there is a concentration of restaurants. While they want their new restaurant to be near other popular eating destinations, they also don't want to be crowded in with too many of the same style.

Based on input from our customer as to which city they are considering, and what type of restaurant they want to open, this program will gather data from Foursquare, process the information, and return a map, with suggested neighborhoods in the specified city. These neighborhoods will be selected by seeing where there is a high concentration of restaurants, but a lower number of restaurants of the particular style of our investors. This will be useful to our stakeholders as they make their choice of where to move once they come to the country, and maximize their investment once their own restaurant is operational.

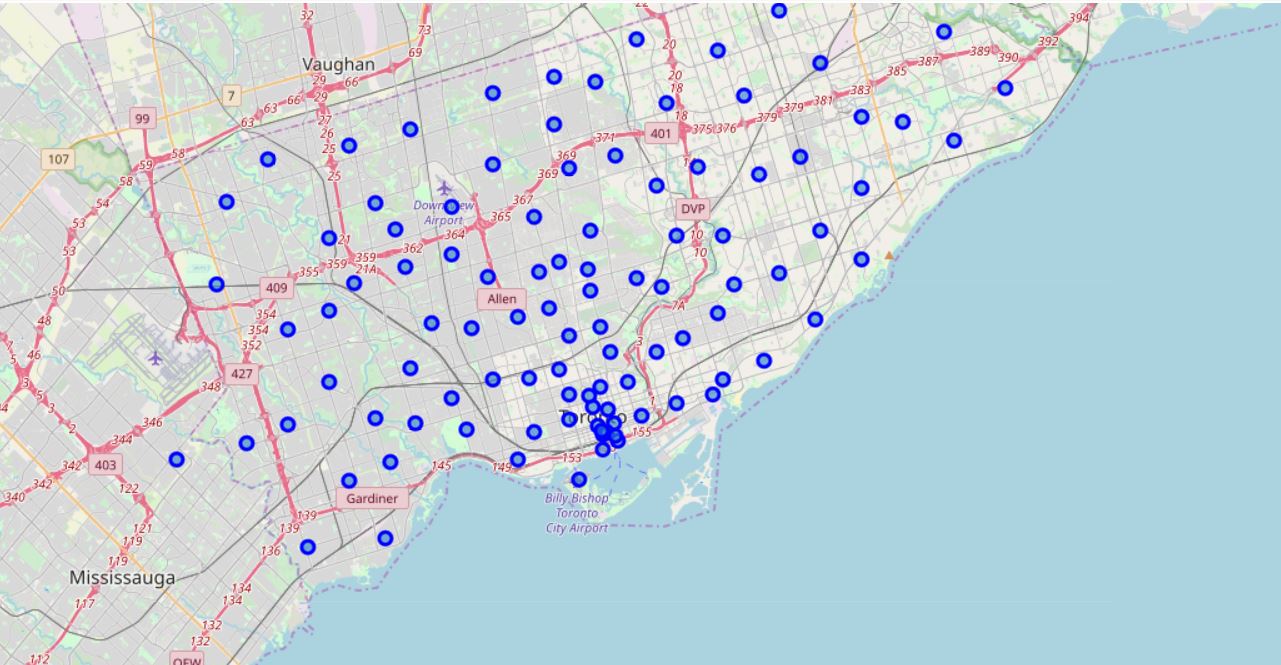
Data Section:

We will be building off of our existing code that analyzed the various neighborhoods around Toronto, Canada. As such, we will be using Folium, Foursquare, and all of the other functions we utilized in the week 3 assignment. We will still be using Foursquare location data to find the various venues in all of the different neighborhoods. Through the use of the Geocode function, this code could be applied to any desired city, not just Toronto like in this example. We would also need to gather some basic data from our investors, such as which city they want to build a new restaurant (in this example, Toronto), and what type of restaurant they want to start, (a Chinese restaurant in this example). Our resulting maps will not just be a map of all of the different neighborhoods. It will filter the results to show areas that already attract customers, but do not have many of the style our investors want to open.

Methodology Section:

Python’s read.html function is used to collect a list of neighborhoods and postal codes from Wikipedia. After confirming which table we want to use, we trim null values out of the Boroughs column, and group the lines based on their postal code. To clean up the data a bit further, we check to find any null values in the Neighbourhood column, and set them equal to the Borough name. Now that our list is complete, we need to use the geocoder function to lookup each of these neighborhoods and assign them a latitude and longitude. In the event that our geocoder function is not working, we can find these same values through a web-search and exporting a csv file, which we would then import to its own Python dataframe. When we have two dataframes with a common column between them, we can use the join command to make one master list.

Now that we have a workable set of data, we can create our first map. Using Folium and the geolocator function, we pass the function a string to search, and get a longitude and latitude. Then we can run through the rows of our dataset and add each data point to the map as its own neighbourhood. The resulting map is shown below.



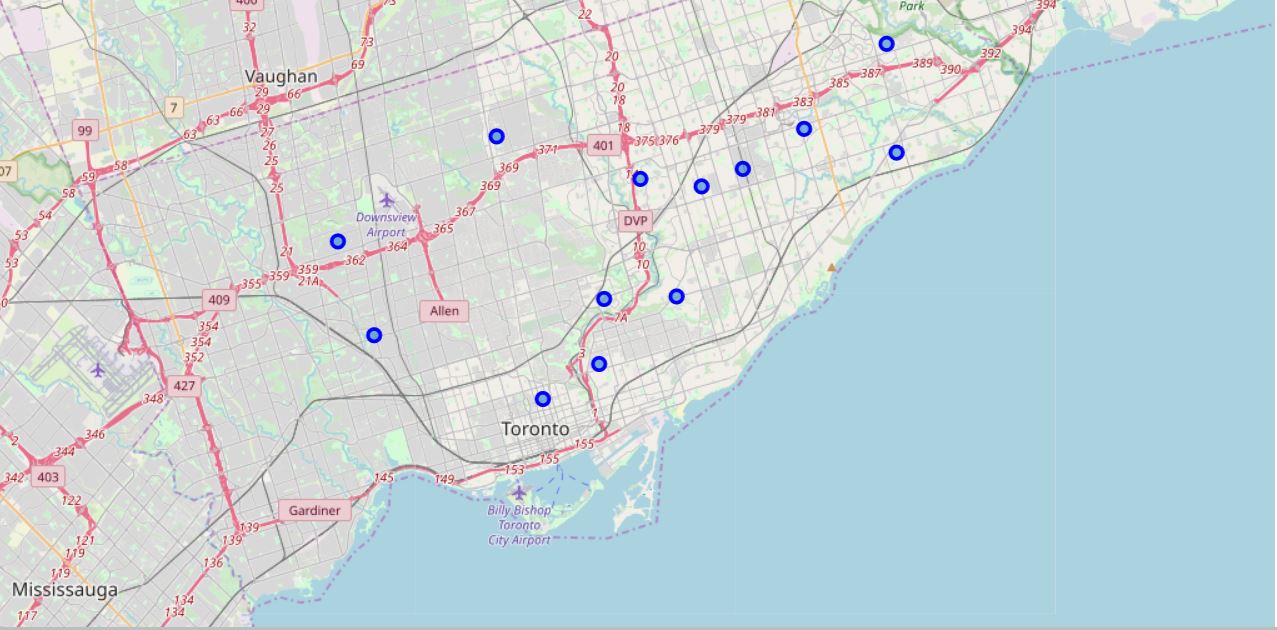
From here, we will use Foursquare to lookup the various venues in each neighborhood. We want to see all venues, not just the restaurants, so we can find out which neighbourhoods have a higher concentration of restaurants. Once we have a list of venues in each neighborhood, we will group them together for further processing.

We can now see what venues are in each neighbourhood, and in what category they fall. Now we will normalize the data to find out how much weight each venue category has in its neighborhood. Using this, we will be able to determine which neighbourhoods are more popular for restaurants. We now organize the data in such a way that gives us the most popular category of venue in each area.

Since we are trying to find where is best to put a restaurant, we create a new dataframe of only the lines that have any type of restaurant as the most common type of venue. At the same time, we want our investors style of restaurant to stand out from the others. So we will ask our investors what type of restaurant they are going to open, Chinese in this example, and remove any neighborhood where a Chinese restaurant is already the most common venue type.

Results Section:

Now that we have filtered through all the data, we have cut down the entire city from a list of over one hundred possible boroughs, to just thirteen. Furthermore, we have plenty of additional data for our investor to make the best decision as to where to put their restaurant. They can see which neighbourhood is still attracting their desired clientele, and where in the city that would be.



Discussion Section:

As it stands, the program requires that the coder pass the information regarding which city to search, and for what style of restaurant a potential investor is planning. For future development of this program, I would incorporate the ability to request input from a user as the program runs as to which city they want to search, and what type of venue they are looking to open. This would expand the usability of the program beyond those familiar with coding, allowing it to be used by the investors themselves, instead of having to go through a 3rd party. It would also allow the expansion beyond just restaurants into any style of venue where there is opportunity. I chose to look at restaurants because that is what my family does.

Conclusion Section:

In this study, we were able to look at all neighbourhoods of a city, see what venues are most popular in each. We looked at where restaurants are most likely to be located and see what style is lacking in these areas to be remedied by our investors. Through this, we discovered where potential opportunity lies for investors to come in and build a new shop to serve an under utilized client base. I will attempt to continue to develop the program to expand its capabilities and attempt to find a truly optimized location for a potential new venue.