Name: Kaemon Derrick

Date: 4/12/2019

Course: Advanced Data Science with IBM Specialization

Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods (Week 1)

Description

Problem: Charlie, a data scientist, is looking to invest money into a new business. His business partner and long-time friend is a chef and had decided to open a new Chinese restaurant in Toronto. Charlie is has decided to leverage his data-science skills to identify the perfect location (postal code) for the new restaurant.

Charlie must consider the market conditions including the current competition and population of the location to make the best choice when choosing the restaurant location.

The location of the restaurant must be such that there is constant foot traffic.

Determining the venues around the proposed location will be key to choosing the right location. The venues will be ranked using a point system, each venue contributing to the point count for a location (postal code).

Venues such as schools, movie theatres etc. will be weighted more heavily than a location such as a retirement home.

Charlie will create a comprehensive report on his finding to present to other project stakeholders as part of the business analysis.

Restaurant:

City: Toronto, ON, Canada

Type of Restaurant: Chinese

Tier: Mid

Target Audience: Middle-Upper class

Preferred location: In or near an area with a large Chinese population with little competition. Close to a mall or in a mall is preferred.

Data

Foursquare API - Location and Venue Data

1. Foursquare will be used to identify all the postal codes in Toronto. Once identified, they will be ranked with a points system based on population and surrounding venues.
2. Foursquare API will be again used to gather all venues within a postal code. These venues will be analyzed and given a point value based on characteristics (to be determined later).

Stats Canada – Population Data

1. Stats Canada provides valuable statistics on the Canadian population. This information will be used to identify the most and least populated postal codes in the city.

Geo-location Data

1. Geo-location information for the postal codes provided by the course in a previous week's assignment. Will be used as input into the foursquare API.

Methodology

1. To start the analysis on city of Toronto, the city is needs to be divided up into manageable locations – postal codes.

   The postal code information is pulled from Wikipedia:
   https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

   Those postal codes with multiple neighborhoods have been combined into one entry in the table (Shown highlighted in yellow).

   The postal codes are identified in the city as shown in the snippet below:

	Postal_Code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront (Toronto), Regent Park
3	M6A	North York	Lawrence Heights, Lawrence Manor
4	M7A	Queen's Park (Toronto)	Queen's Park (Toronto)

2. The geo-location data for each of the postal codes is in the file: Geospatial_Coordinates.csv. This file was provided as part of the course material for this course.

   The co-ordinates are added to the data frame shown below:

	Postal_Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront (Toronto), Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park (Toronto)	Queen's Park (Toronto)	43.662301	-79.389494

3. The population data for all postal codes in Canada was downloaded from the statistics Canada website at: 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.

   The extraneous data has been removed as after some analysis, Toronto only has postal codes with 'M' as the beginning letter.

   The below screenshot shows the new data frame.

	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Population
0	M3A	North York	Parkwoods	43.753259	-79.329656	34615
1	M4A	North York	Victoria Village	43.725882	-79.315572	14443
2	M5A	Downtown Toronto	Harbourfront (Toronto), Regent Park	43.654260	-79.360636	41078
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763	21048
4	M7A	Queen's Park (Toronto)	Queen's Park (Toronto)	43.662301	-79.389494	10

4. Since the restaurant will rely on persons that are in close-proximity to the restaurant location, only postal codes with greater than 25,000 population will be considered. Those postal codes with less than 25,000 population are dropped from consideration.

   Removing those postal codes with a population that does not meet the clip leaves only 51 rows in the data frame.

5. Next, nearby venues are identified for each postal code and a new data frame is created. The data frame will be used to record the venues and point value of each venue.

	Postal_Code	Postal_Latitude	Postal_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	M3A	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	M3A	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	M3A	43.753259	-79.329656	DVP at York Mills	43.758899	-79.334099	Road
3	M3A	43.753259	-79.329656	TTC Stop #09083	43.759655	-79.332223	Bus Stop
4	M5A	43.654260	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery

6. By examining the data frame, we can see that there are 247 unique categories. These categories are ranked based on what they can contribute to or detract from the location. For example, a school is ranked high (more points) because it can increase the foot traffic and therefore the potential for increased sales.

   A point value of one (1) will be given to venues that do not fit in categories identified.

7. A point value is assigned to each venue in the Toronto_venues dataframe as shown below.

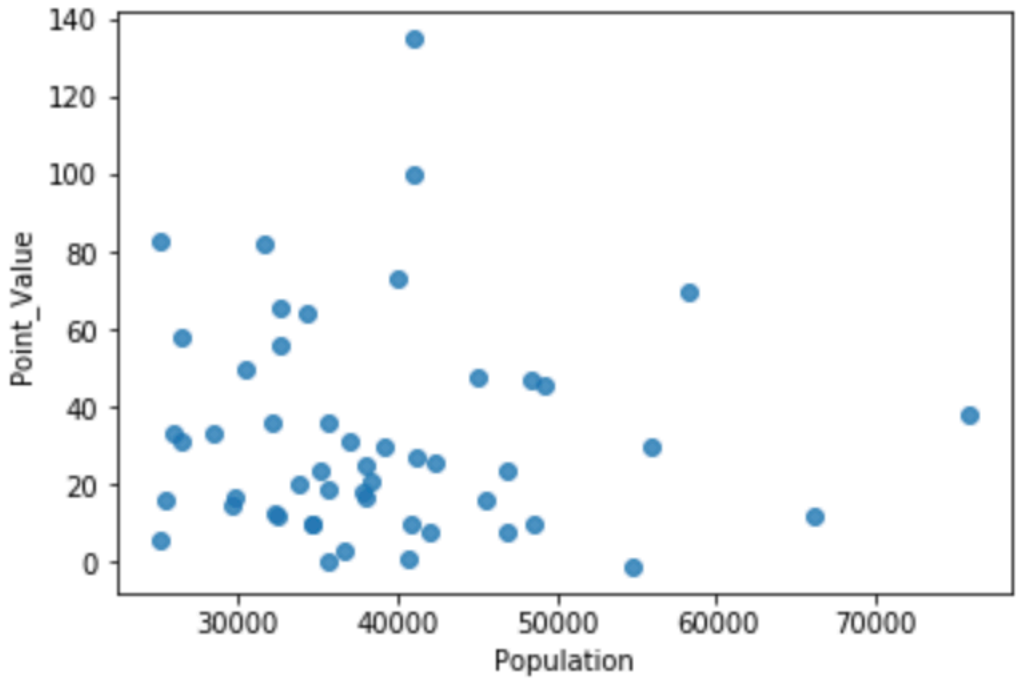
	Postal_Code	Postal_Latitude	Postal_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category	Point_Value
0	M3A	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park	3
1	M3A	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop	4
2	M3A	43.753259	-79.329656	DVP at York Mills	43.758899	-79.334099	Road	1
3	M3A	43.753259	-79.329656	TTC Stop #09083	43.759655	-79.332223	Bus Stop	2
4	M5A	43.654260	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery	1

8. Next, the values for the venues are summed up to give a total for the postal codes. This value is added to the toronto_df data frame.

	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Population	Point_Value
0	M3A	North York	Parkwoods	43.753259	-79.329656	34615	10
1	M5A	Downtown Toronto	Harbourfront (Toronto), Regent Park	43.654260	-79.360636	41078	135
2	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242	35594	19
3	M1B	Scarborough, Toronto	Rouge, Toronto, Malvern, Toronto	43.806686	-79.194353	66108	12
4	M6B	North York	Glencairn	43.709577	-79.445073	28522	33

9. Next, I sort the dataframe and analyze the points compared to the population. This is done to see how the points are affected when population increases. A larger population could bring more venues and thus more points.

	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Population	Point_Value
1	M5A	Downtown Toronto	Harbourfront (Toronto), Regent Park	43.654260	-79.360636	41078	135
24	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191	40957	100
11	M4E	East Toronto	The Beaches	43.676357	-79.293031	25044	83
23	M4K	East Toronto	The Danforth West, Riverdale, Toronto	43.679557	-79.352188	31583	82
35	M6P	West Toronto	High Park, The Junction South	43.661608	-79.464763	40035	73

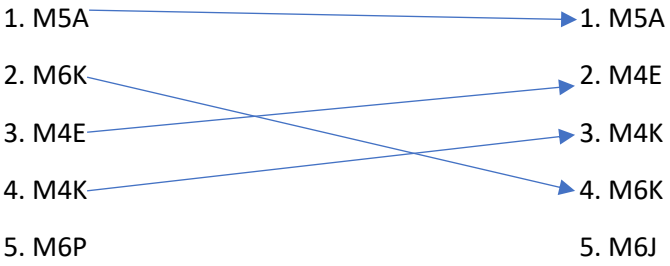


10. Population does not seem to have a large impact on the points for the postal code. Though, this can also be the case that there are many venues (e.g restaurants – competition) in the area that is bringing the point value down.

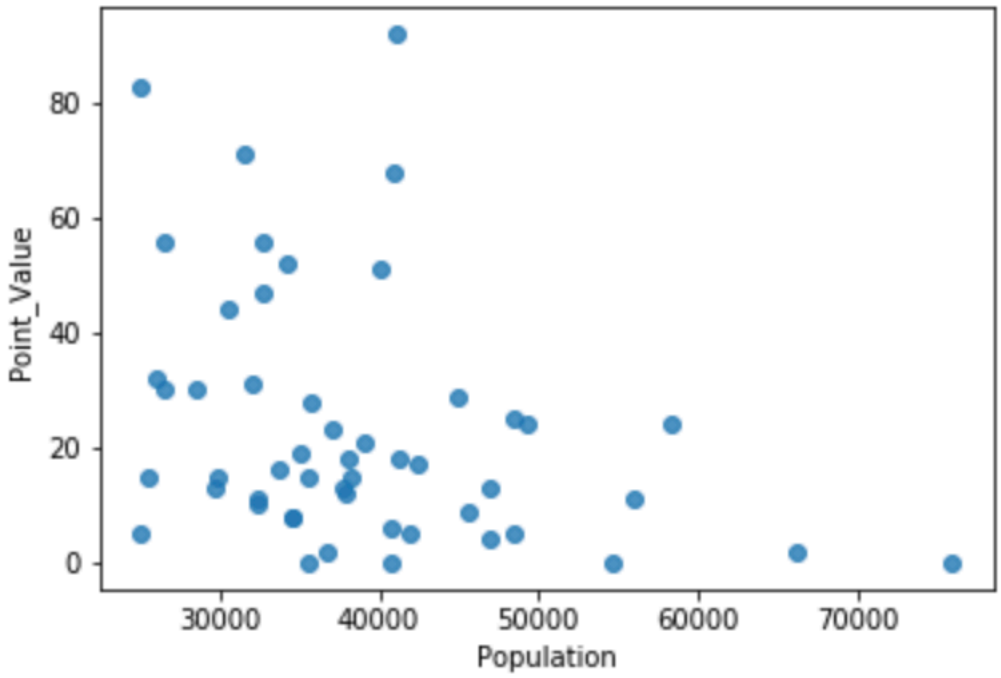
    To account for the population skewing the data one way or another, the population will be normalized to a value between 0 and 1. Then, the point value will be multiplied by 1 minus the normalized population to fairly determine the point value.

	Postal_Code	Borough	Neighborhood	Latitude	Longitude	Population	Point_Value
0	M5A	Downtown Toronto	Harbourfront (Toronto), Regent Park	43.654260	-79.360636	41078	92
1	M4E	East Toronto	The Beaches	43.676357	-79.293031	25044	83
2	M4K	East Toronto	The Danforth West, Riverdale, Toronto	43.679557	-79.352188	31583	71
3	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191	40957	68
4	M6J	West Toronto	Little Portugal, Toronto, Trinity–Bellwoods	43.647927	-79.419750	32684	56

We can see that the point values are changed, and the top 5 postal codes are different. Though 'M5K' is still ranked number one, other postal codes now rank differently. The changes to the top 5 are shown below:



11. The new scatter plot can be seen here



12. From the analysis above, the top three postal codes to place a new Chinese restaurant are as follows:

    1. Postal Code: M5A
       Borough: Downtown Toronto
       Neighborhood: Harbourfront (Toronto), Regent Park
       Latitude: 43.6542599, Longitude: -79.3606359
       Population: 41078

    2. Postal Code: M4E
       Borough: East Toronto
       Neighborhood: The Beaches
       Latitude: 43.67635739999999, Longitude: -79.2930312
       Population: 25044

    3. Postal Code: M4K
       Borough: East Toronto
       Neighborhood: The Danforth West, Riverdale, Toronto
       Latitude: 43.6795571, Longitude: -79.352188
       Population: 31583

13. The postal codes determined to be the top three are displayed on the map below. Either of these postal codes will be a good choice to open a new Chinese restaurant in. They provide the best environment for a restaurant opening with many attractors and little competition (comparatively).