FINAL REPORT

CAPSTONE PROJECT

Jabin Choi

Coursera – IBM Data Science

12/30/2019

INTRO – PROBLEM DESCRIPTION

It is no doubt that Canada is one of many countries with immigrants. Scarborough, in particular, is one of the cities that recently attracted a lot of attention among immigrants. This project is useful for people planning to move to Scarborough by showing the results of comparing and analyzing dataset related to their average housing price and school rates among the venues registered in the app called Foursquare. For more details on Foursquare, please visit the following url:

** Foursquare URL: https:// foursquare.com

WHAT IS FOURSQUARE API?

In this project, dataset related to the average of housing price and school ratings among the venues registered in the app called Foursquare.

Foursquare is a social location service that provides users to explore the location around the world. The app enables users to download it through iPhone, Blackberry or Android devices and then connects their account to other social media accounts.

Foursquare API itself is a RESTful set of addresses to which developers can send requests. Users has nothing to download onto their server and it just allows developers to interact with the Foursquare platform.

The major purpose of using Foursquare API is to gather data source as it has a database of millions of places and the API provides the ability to perform location search and sharing as well as details on a business.

DATA APPROACH - METHDOLOGY

First, data will be loaded from Wiki url:

' https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M '

The data table looks like the followings:

Out[4]:

	Postalcode	Borough	Neighbourhood
0	0	0	0
1	M1A	Not assigned	Not assigned
2	M2A	Not assigned	Not assigned
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village

Data cleaning:

- Drop meaningless contents in the table – meaning handle 'Not assigned' and None values
- Group and re-index the table
- Add two columns called 'Latitude' and 'Longitude'

Once data cleaning is completed, the final data table before analyzing might look as follows:

Out[10]:

	Postalcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.811525	-79.195517
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.785665	-79.158725
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.765815	-79.175193
3	M1G	Scarborough	Woburn	43.768369	-79.217590
4	M1H	Scarborough	Cedarbrae	43.769688	-79.239440

Data from Foursquare API:

To gain datasets of various venues in different neighborhoods of that specific borough – Scarborough, we need to use 'Foursquare' locational information. The information contains venues name, locations, menus and photos.

- Neighbourhoods
- Neighbourhoods Latitude / Longitude
- Venues
- Venue names such as restaurants or stores
- Venue Latitude / Longitude
- Venue Category

In this project, we will extract information that we need such as venue names, categories and latitude / longitude. Once the columns are obtained from through the API, then it might look as follows if having chosen the radius to be 100 meters:

Out[25]:

	name	categories	lat	lng
0	SEPHORA	Cosmetics Shop	43.775017	-79.258109
1	Disney Store	Toy / Game Store	43.775537	-79.256833
2	American Eagle Outfitters	Clothing Store	43.776012	-79.258334
3	DAVIDsTEA	Tea Room	43.776613	-79.258516
4	Hot Topic	Clothing Store	43.775450	-79.257929

Methodology – K Means Clustering:

In this project, K-Means clustering that is a form of unsupervised machine learning is used for data analysis. By grouping similar data points (here, 'Neighbourhoods' column) together, we will discover underlying patterns and analyze dataset related to the average housing price and school ratings among the venues.

From the methodology, we are able to see top 5 most common venues per neighbourhood in Scarborough by setting 5 clusters (k=5).

SORTED DATA & RESULTS OF K-MEANS

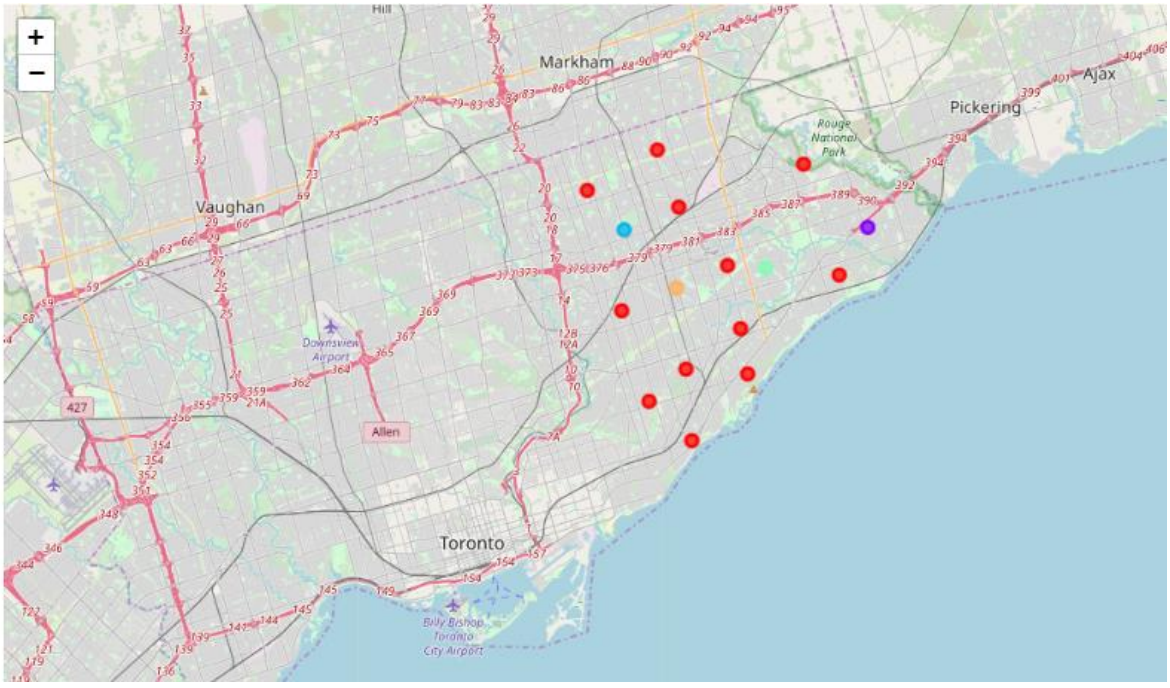
[Sorted data table]

Out[38]:

	Neighbourhood	1st Most common venue	2nd Most common venue	3rd Most common venue	4th Most common venue	5th Most common venue	6th Most common venue	7th Most common venue	8th Most common venue	9th Most common venue	10th Most common venue
0	Agincourt	Shopping Mall	Chinese Restaurant	Discount Store	Vietnamese Restaurant	Bakery	Grocery Store	Sushi Restaurant	Supermarket	Hong Kong Restaurant	Bubble Tea Shop
1	Agincourt North, L'Amoreaux East, Milliken, St...	Pharmacy	Zoo Exhibit	Gym	Golf Course	General Entertainment	Fried Chicken Joint	Flower Shop	Fast Food Restaurant	Electronics Store	Discount Store
2	Birch Cliff, Cliffside West	Gym	College Stadium	General Entertainment	Gym Pool	Skating Rink	Park	Fried Chicken Joint	Flower Shop	Fast Food Restaurant	Electronics Store
3	Cedarbrae	Playground	Zoo Exhibit	College Stadium	Golf Course	General Entertainment	Fried Chicken Joint	Flower Shop	Fast Food Restaurant	Electronics Store	Discount Store
4	Clairlea, Golden Mile, Oakridge	Bakery	Bus Line	Intersection	Coffee Shop	Soccer Field	Bus Station	Metro Station	Convenience Store	General Entertainment	Fried Chicken Joint

[Result of K-Mean Clustering, k=5]

Out[47]:



DISCOVERY THROUGH DATA ANALYSIS

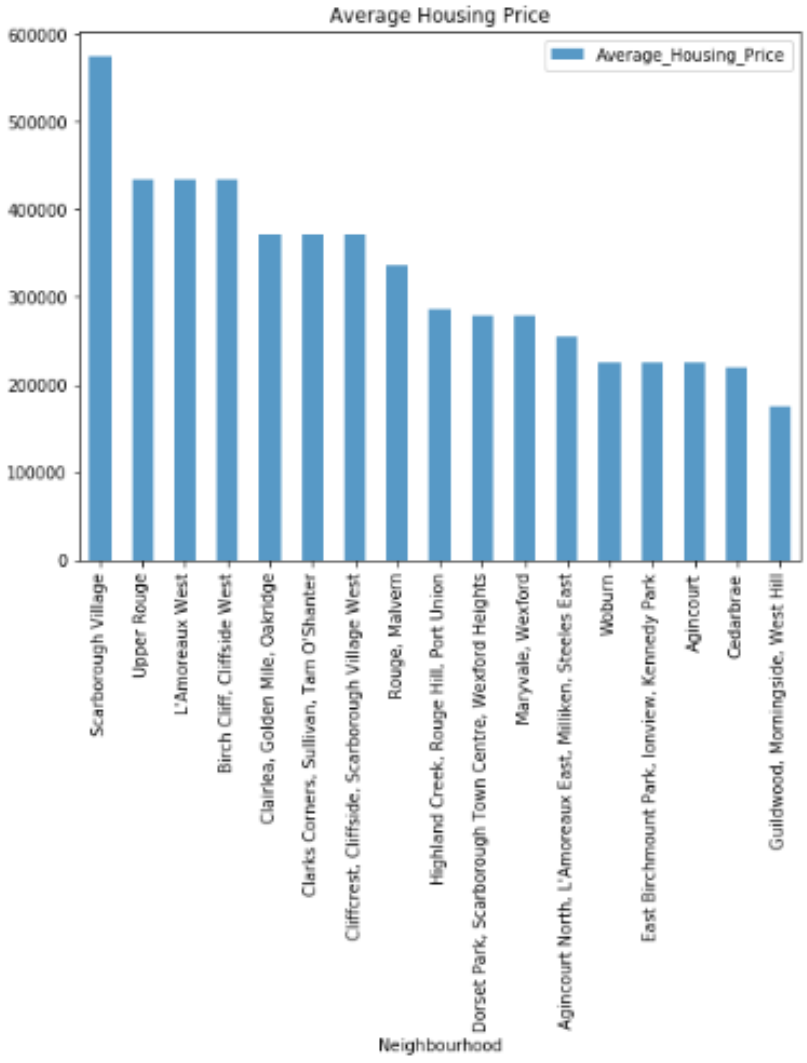
Once combine Scarborough neighbourhoods dataset with average of housing price, the result of the table is as follows:

Out[33]:

	Neighbourhood	Average_Housing_Price
0	Rouge, Malvern	335000.0
1	Highland Creek, Rouge Hill, Port Union	286600.0
2	Guildwood, Morningside, West Hill	175000.0
3	Woburn	225900.0
4	Cedarbrae	219400.0

The result of bar plot after sorting values by the average of housing price is as follows:

Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x1539c09a308>



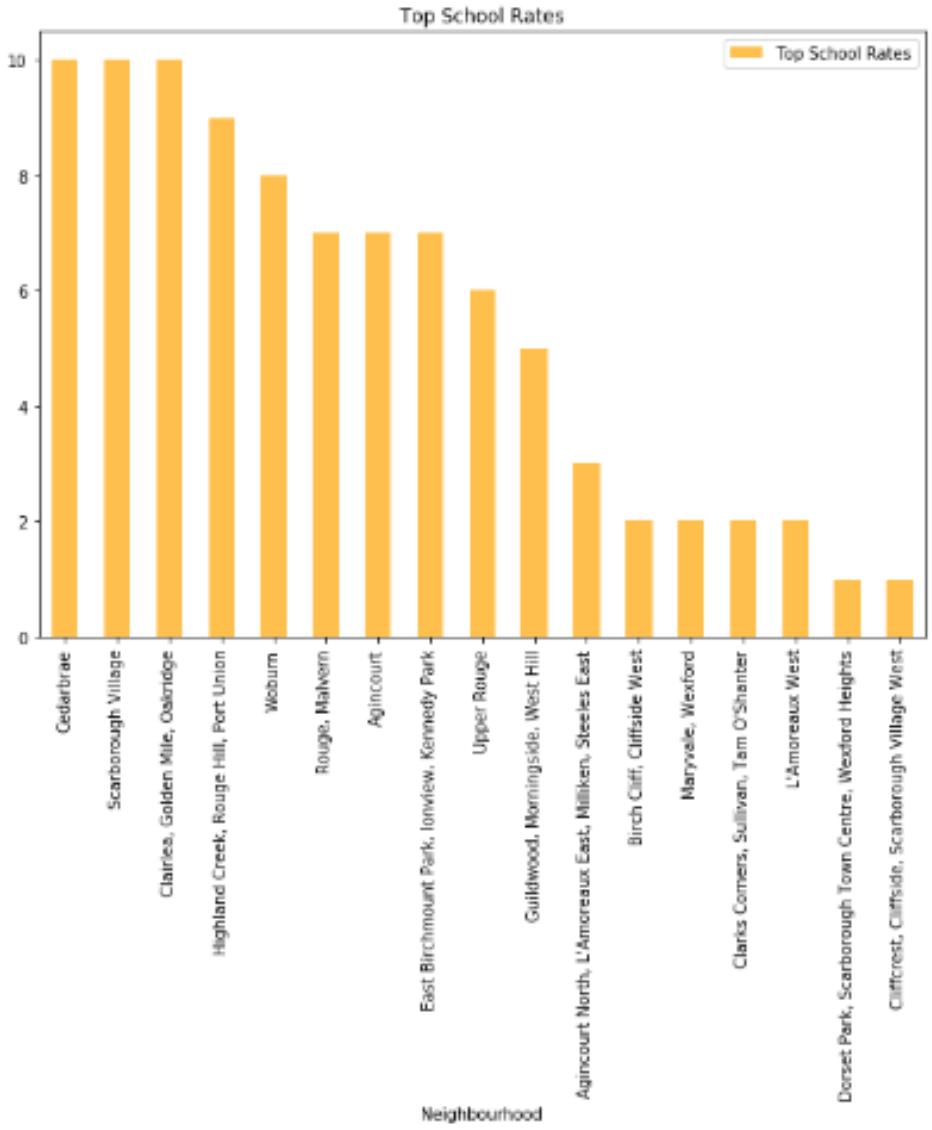
This time, we will combine Scarborough neighbourhoods data with top school rates. The table looks like the followings:

Out[30]:

	Neighbourhood	Top School Rates
0	Rouge, Malvern	7
1	Highland Creek, Rouge Hill, Port Union	9
2	Guildwood, Morningside, West Hill	5
3	Woburn	8
4	Cedarbrae	10

If we sort values by top school rates and then plot the table, then it looks like the following:

Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x1539be17a88>



CONCLUSION

[Result of Barplot 1]

From a bar plot named ‘Average Housing Price’, we are notified that ‘Scarborough Village’ is the neighbourhood that has the highest average of housing price and ‘Upper Rounge’ / ‘L'Amoreaux West’ are going after.

[Result of Barplot 2]

From a bar plot named ‘Top School Rates’, we are notified that ‘Cedarbrae’ is the neighbourhood that got the highest school rates and ‘Scarborough Village’ / ‘Clairlea, Golden Mile, Oakridge’ are going after.