APPLIED DATA SCIENCE CAPSTONE – Week 05

Beno Wajsfeld 02-NOV-2020

INTRODUCTION

The problem I would like to solve is that I want to invest but not exactly where, in Manhattan or Toronto, which neighborhood and not exactly in what kind of business or market. It can be a coffee shop or a gym, or a restaurant. To make a decision in what kind of business I want to invest and in which neighborhood of Manhattan or Toronto, I need to have an initial market analysis using customers' feedback given to venues in their neighborhoods.

The main idea is to know which are the most visited venues in each neighbor of New York and venues in each neighbor in Toronto.
In order to simply this operation, I took only Manhattan neighbors from New York and in Toronto I took the neighbors with Toronto in its name to have neighbors and venues analyzed.
So let's go to and find Manhattan and Toronto neighbors.
First New York and its neighbor Manhattan. After that, Toronto…

NEW YORK (Manhattan)

DOWNLOAD AND EXPLORE DATASET - MANHATTAN

The first thing I did is to download the newyork_data.json file and transformed into a panda_dataframe from features key because all relevant data are in it. The dataframe contains Borough, Neighborhood, Latitude and Longitude.
New Your has 5 boroughs and 306 neighborhoods. After that you can see the map showing all the 306 neighborhoods.
But Manhattan is our goal so Manhattan was obtained from NewYork initial dataframe created. The result is a dataframe with 40 neighborhoods of Manhattan. You can check on the map and visualize its 40 neighborhoods.
We are ready to get maximum top 100 venues from each neighborhood.

EXPLORING NEIGHBORHOODS - MANHATTAN

Initially a function was defined to bring maximum 100 venue in a 500 of Marble Hill for testing. Using Foursquare Credentials and Marble Hill Latitude and Longitude, a url was created to bring maximum 100 venues in Marble Hill in a 500 meter radius. 22 venues were found. We are also interested in the category of each venue. Another url was created to bring the categories.
Now cleaning the Jason we created a dataframe with venue name, it category, its latitude and its longitude.
Now we applied the same sequence of program done with Marble Hill to all neighborhoods in Manhattan.
The Manhattan dataframe result is 3210 lines and columns as follows: Neighborhood, Latitude, Longitude, VENUE, Venue Latitude, Venue Longitude and Venue Category. After we grouped venues by neighborhood and we realized that there are 321 unique categories

ANALYZING EACH NEIGHBORHOOD – MANHATTAN

The last dataframe was converted into a dataframe of 3211 lines and 322 colums. 3211 venues in 40 neighborhoods distributes in 322 categories.

After we grouped rows by neighborhood and took the mean of the frequency of occurrence of each category. The new dataframe has 40 lines (neighborhoods) and 322 (categories).

We got the 10 most visited category venue of each neighborhood and created a new dataframe with 40 lines (neighborhood) and 10 columns (10 most visited venue category).

CLUSTER NEIGHBORHOODS – MANHATTAN

In order to find a pattern between the 10 most venue category visited, we executed K-means to cluster the neighborhood into 7 clusters. The cluster will group neighborhoods using the 10 most venue visited following a certain pattern determined by the K-means. The clusters were created and mapped indicating the different clusters by different dot colors.

EXAMINE CLUSTERS - MANHATTAN

After determining the 7 cluster, I analyzed their 10 venues most visited and then I classified each cluster into venue category in importance order: See the map with cluster/neighborhood locations.

RESULTS:

CLUSTER 0:

- FIRST: Chinese, Italian, Japanese, Sushi and American restaurants
- THIRD: Coffee

CLUSTER 1:

- MOST: Theater, SPA, Park, Hotel, Civic Center, PLAZA, Gym…commercial places!
- FEW: Restaurants (mainly in Central Harlem an in Civic Center), coffee shops, clothing store

CLUSTER 2:

- MOST: coffee shops and coffee
- FEW: restaurants (Mexican, Chinese)

CLUSTER 3:

- MOST: amusement (park, baseball field), Transportation (Harbor/Marina, Heliport, Boat or Ferry)

CLUSTER 4:

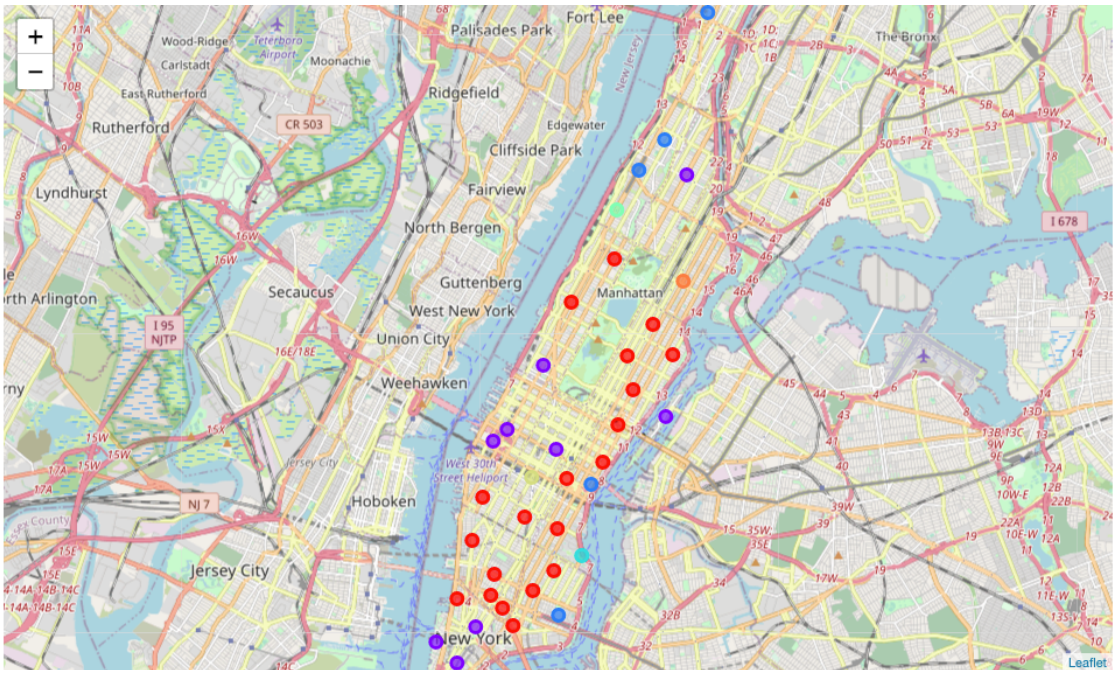
- MOST: Park
- FEW: Restaurants(American, Burger Joint, Ethiopian)

CLUSTER 5:

- MOST: Korean Restaurant
- FEW: Hotel, Japanese and American Restaurant

CLUSTER 6:
- MOST: Mexican restaurant
- FEW: Latin American Thai and French restaurants

Out[41]:



TORONTO

DOWNLOAD AND EXPLORE DATASET – TORONTO
In this case I didn't have a Jason file with Toronto data. The solution was to use a table in Wikipedia and use BeautifulSoup to extract the data inside one of its table that has PostCOde, Borough and Neighborhood.
But taking a look in the dataframe, a lot of cleaning was necessary:
- Exclude \n from PostCode .
- Exclude all lines that has "NOT ASSIGNED" in its Borough.
- Replace all neighborhood with "NOT ASSIGNED" by the Borough.
- Lines with the same ZipCode, consider the first one the master and add on the master neighborhoods the other neighborhood names separated by comma.

We found 103 neighborhood in Toronto. Now let's find their latitude and longitude. For that we crossed this dataframe with a dataframe of Postal Code/Latitude/Longitude.
The new dataframe now has 10 boroughs and 103 neigborhoods.
Like in New York Manhattan was chosen, here in Toronto only boroughs that have Toronto in its name.
The new dataframe has 39 neighborhoods.

EXPLORE NEIGHBORHOODS – TORONTO

As it was done with Marble Hill, The Beaches at Toronto was chosen.

- Its latitude and longitude were determined
- The top 100 venues in 500 meters were downloaded using Foursquare
- Follow the same procedure done with MANHATTAN.

In the case of The Beaches, only top 4 venues were found.

Then the same procedure done with The Beaches to find the top 100 venues in 500 meters radius was used with all neighborhood of Toronto: total of 1624 venues in 39 neighborhoods of Toronto and classified in 237 unique categories.

ANALYZE EACH NEIGHBORHOOD – TORONTO

Again, the neighborhoods were grouped and the mean of the frequency of occurrence of each category calculated. This dataframe has 39 neighborhoods (lines) and 237 categories (columns). Then for each neighborhood we took the highest 10 means of the frequency of occurrence of category and created and a new dataframe. It has 39 neighborhoods (lines) and 11 columns (Neighborhood plus 10 most categories).

CLUSTER NEIGHBORHOODS - TORONTO

In order to find a pattern between the 10 most venue category visited, we executed K-means to cluster the neighborhood into 5 clusters. The cluster will group neighborhoods using the 10 most venue visited following a certain pattern determined by the K-means. The clusters were created and mapped indicating the different clusters by different dot colors.

EXAMINE CLUSTERS - TORONTO

After determining the 5 clusters, I analyzed their most 10 venues categories and then I classified each cluster into venue category in importance order: See the map with cluster/neighborhood locations.

RESULTS:

CLUSTER 0:

- MOST: Nature: Park, Trail, playground
- FEW: Stores and restaurants

CLUSTER 1:

- MOST: Café, coffee shop, Restaurants (Thai, Italian , Mexican, Vietnamese)
- FEW: Bar.

CLUSTER 2:

- MOST: Sandwich café, café, bakery (Quick meals)
- FEW but on all neighborhoods: Stores, shops and restaurants

CLUSTER 3:

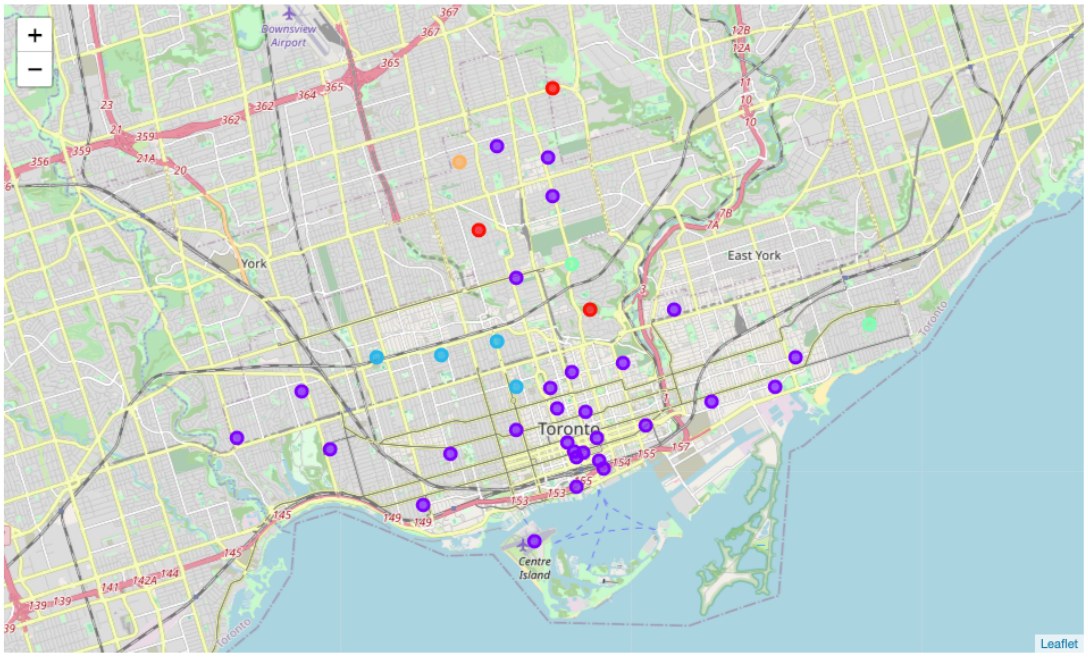
- MOST: Nature and heathy food activities (yoga)

- FEW: store and restaurants

CLUSTER 4:

- MOST: Nature (garden) and cultural activity (Music Venue)

Out[126]:



DISCUSSION

Now let's discuss the results and try to choose where I would like to invest and open a business.

Looking the cluster from both places I can see that restaurants and coffee related activities where the most visited by customers.

Maybe this is a good business.

But I have a question: Is it possible that with only the results of the clustering I can really choose everything? In my opinion it is a good beginning to understand the places and what king of business are most visited.

Take a look in those tables:

MANHATTAN

CLUSTER	Population	Area (mile2)	Density (people/mile2)	AVG income US$
0	749300	8.31	90179.66	118676.79
1	211730	15.79	13408.75	78938.21
2	373280	4.03	92702.50	50010.86
3	16540	0.16	104360.00	96200.00
4	40940	0.60	68240.00	82300.00
5	17110	0.70	24350.00	126800.00
6	118860	1.56	76060.00	37500.00
TOTAL:	1527760	31.15	49045.47	88949.49
Total of venues collected:	3210	40 neighborhoods	Total of venues considered in clusters:	1212

TOTONTO

CLUSTER	Population	Area (mile2)	Density (people/mile2)	AVG income CAD
0	53842	9.87	5455.20	127554.18
1	354784	45.69	7764.96	65372.71
2	84096	8.58	9802.50	63982.77
3	47590	8.86	5371.54	101943.09
4	29249	2.06	14210.00	85991.00
TOTAL:	569561	75.06	7588.35	75160.13
Total of venues collected:	1624	39* neighborhoods	Total of venues considered in clusters:	814

From the notebook I collected all data related to all clusters and their 10 most venues visited.

I visited the links below:

Manhattan:
https://statisticalatlas.com/county-subdivision/New-York/New-York-County/Manhattan/Household-Income#figure/neighborhood

Toronto:
https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/

https://torontocentral.evrealestate.com/ListingDetails/983-Roselawn-Ave-Toronto-ON-M6B-4M9/W4542612

From the links I would discover important data related to each neighborhood like population, area, median income of the population. Of course I didn't find all data from all neighborhood but it gave an idea of the characteristics of each one.
I also calculated how many of each venue were found on each neighborhood and the total of each venue in the cluster.
I am sure that those data are also important to make an important decision…all I want is to minimize the risks, bad ones! It took me quite big amount of time…only 40 neighborhoods each place.

Check the links bellow to see Manhattan and Toronto table with all data I collected:

Manhattan: https://github.com/bwajsfeld/IBM_CAPSTONE/blob/main/Research_NY.pdf

Toronto :
https://github.com/bwajsfeld/IBM_CAPSTONE/blob/main/Research_TO_OFICIAL.pdf

If you want to see the notebook:
https://nbviewer.jupyter.org/github/bwajsfeld/IBM_CAPSTONE/blob/main/C9_W5_BATTLE_NY_TO.ipynb

(I use nbviewer to be possible to see the map!)

With the data collected I can answer some queations, for example:

- Which neighborhood has the highest median income?
- Which one has the highest area? I could calculate venue per miles2 an check which neighborhood could accept another venue without saturating the neighborhood. Which neighborhood is lacking a kind of busisness?.
- Which neighborhood has the highest people density people/miles2? It means more customers.

- I can create new indicators and compare businesses and neighborhood that could help to check if the business I have chosen as café/coffee shops (I love coffee) and restaurants bring me the best indicators.

Now choosing between Manhattan and Toronto…

Well, Toronto has more area and population but Manhattan has more median income like salary. Maybe Manhattan is much more expensive to open a business but maybe you will more revenue.

CONCLUSION

The nest step of this research is to really analyze all data I found and try to minimize the risks of opening a business where ever the result place would be.

The clustering itself is a powerful tool to understand a certain pattern in the distribution of venues in our case but to make a decision only on them in our case is to risky.

I exercised to collect more data and have more information to make a good decision but I will try to make it happen for café/coffee shop or any of those restaurants listed or after a good market research bring a new kind of restaurant.