Capstone Project - Relocation from New York City to Toronto

For Applied Data Science Capstone Course by IBM/Coursera

Table of contents

Introduction: Business Problem Description

This project is designed to resolve the following assumed business case:

A US business manager who is working for a large financial company and also living in Manhatten, New York city. Recently his company has asked him to relocate to Toronto Canada to lead the efforts to set up a branch business there. He needs to decide where to set up the business office in Toronto. He also needs to select a new residential location for his family, preferrably in a similar favorable location to his home in Manhatten in New York.

This business manager approached a specialized service company to help him to evaluate the location options in Toronto for both business office and his family home in Toronto. After signing the service contract, this service company starts to work on this project. This project could involve the following aspects:

1. First to evaluate and to understand the preferred characteristics from this US business manager on his current business office and family home in Manhatten in New York.
2. Second to collect and to evaluate information related to Toronto.
3. Compare the similarities and differences between New York and Toronto.
4. Based on US manager's preferred selection criterias to recommend location options on new business office and residential home in Toronto.

For each of above project aspects, we will conduct relevent data and analyze the data to provide quantitative asscessment. We will work the following data science project steps:

1. Collect neighborhood information between New York and Toronto.
2. Collect relevant venue information that could be required to compare between New York and Toronto.
3. Analyze the collected data to compare the similarities and differences between New York and Toronto.
4. Based on US business manager's proposed location selection preference criterias to recommend location options for both busines office and family home in Toronto.

Data Collection

Based on the above busienss problem descriptinon, we will collect the following data:

1. Neighborhood and venue information for New York city, including:
   (1) Neighborhood information for New York city.
   (2) Neighborhood information for Manhatten, New York.
   (3) Venues around existing business office and family home locations in Manhatten, New York city.

2. Neighborhood and venue information for Toronto, including:
   (1) Neighborhood information for Toronto.
   (2) Neighborhood information for Toronto financial central location.
   (3) Venue information for Toronto financial central location.

1. Collect neighborhood and venue Information for New York city

(1) Collect neighborhood Information for New York city

To get New York neighborhood information, we first download existing dataset newyork_data online. We convert the online dataset into pandas dataframe. The neighborhood information is then visualized on New York map.

(2) Collect neighborhood Information for Manhatten, New York

The Manhatten neighborhood data is a subset from New York neighborhood dataset. The data is visualized on a map.

(3) Collect venues around existing business office and family home

locations in Manhattan, New York city

We assume that the business office address is 100 Wall St New York, near New York Stock Exchange location.

Collect venues around existing business office in Manhattan, New York city

We are using Foursqure API to collect venue information

Explore Neighborhoods in Manhattan

Collect neighborhood and venue information for Toronto

We will collect the following neighborhood and venue information for Toronto:
(1) Neighborhood information for Toronto.
(2) Neighborhood information for Toronto financial central location. (3) Venue information for Toronto financial central location.

(1) Collect neighborhood information for Toronto

The Toronto neighborhood data is scraped from a Wikipedia page. The data was structured into pandas dataframe

(2) Collect neighborhood information for Central Toronto, the financial central location

The Toronto financial center is located in Central Toronto borough. The Central Toronto neighborhood data is a subset from Toronto neighborhood dataset. It is visually shown on a map.

(3) Collect venue information for Central Toronto, the financial center

Let's create a function to repeat the same process to all the neighborhoods in Central Toronto

Now write the code to run the above function on each neighborhood and create a new dataframe called *manhattan_venues*.

Methodology

In this project, we will use the following methods to analyze the data:

1. First, we will compare the total number of venues between Manhattan and Central Toronto. Such total venue number comparison could show the similarities and differences in terms of overall venue availability and distribution.
2. Second, for each neighborhood, we will rank top venues based on venue grouping. Such ranking information could show what venues are more available for a specific neighborhood. This local venue availability and concentration information will be valuable to match preferred venues with neighborhood options.
3. Third, we will apply k-means clustering to analyze the distribution of venue categories within different neighborhoods for both Manhattan and Central Toronto. This statistic significance based results could give good venue clustering distribution information for location selection.

To further refine the location selection criteria, we will specify the following US business manager's selection preferences for potential location selection options:

1. For new business office in Toronto, prefer location options at central financial district that exhibits similar venue clustering and vailability as those in Manhattan finncial district.
2. For new family home in Toronto, prefer location options that have good availability in recreation and social venues, such as park, musium and restaurants.

After proposing potential location options, we will also identify areas for further study to refine the location search by incorporating feedbacks and improvement suggestions from US business manager and his company decison makers after presenting this project report.

Analysis and Results

Analysis and Results

Analyze Manhatten Neighborhood and Results

Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category

Let's create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.

Run *k*-means to cluster the Manhattan neighborhood into 5 clusters

In [87]:
```python
# add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

manhattan_merged = manhattan_data

# merge toronto_grouped with toronto_data to add latitude/longitude for
manhattan_merged = manhattan_merged.join(neighborhoods_venues_sorted.set

#manhattan_merged = manhattan_merged.drop(['Unnamed:0'], axis=1)
manhattan_merged.head(30) # check the last columns!
```

Out[87]:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	4	Sandwich Place	Discount Store	Coffee Shop
1	Manhattan	Chinatown	40.715618	-73.994279	1	Chinese Restaurant	Cocktail Bar	American Restaurant
2	Manhattan	Washington Heights	40.851903	-73.936900	2	Café	Bakery	Mobile Phone Shop
3	Manhattan	Inwood	40.867684	-73.921210	2	Mexican Restaurant	Lounge	Café
4	Manhattan	Hamilton Heights	40.823604	-73.949688	0	Pizza Place	Coffee Shop	Mexican Restaurant
5	Manhattan	Manhattanville	40.816934	-73.957385	2	Italian Restaurant	Mexican Restaurant	Seafood Restaurant
6	Manhattan	Central Harlem	40.815976	-73.943211	1	African Restaurant	French Restaurant	Chinese Restaurant

7	Manhattan	East Harlem	40.792249	-73.944182	0	Mexican Restaurant	Bakery	Deli / Bodega
8	Manhattan	Upper East Side	40.775639	-73.960508	1	Italian Restaurant	Exhibit	Art Gallery
9	Manhattan	Yorkville	40.775930	-73.947118	1	Italian Restaurant	Gym	Coffee Shop
10	Manhattan	Lenox Hill	40.768113	-73.958860	1	Coffee Shop	Italian Restaurant	Pizza Place
11	Manhattan	Roosevelt Island	40.762160	-73.949168	1	Coffee Shop	Sandwich Place	Park
12	Manhattan	Upper West Side	40.787658	-73.977059	1	Italian Restaurant	Wine Bar	Bar
13	Manhattan	Lincoln Square	40.773529	-73.985338	1	Gym / Fitness Center	Theater	Café
14	Manhattan	Clinton	40.759101	-73.996119	1	Theater	Gym / Fitness Center	Hotel
15	Manhattan	Midtown	40.754691	-73.981669	1	Hotel	Coffee Shop	Cocktail Bar
16	Manhattan	Murray Hill	40.748303	-73.978332	1	Coffee Shop	Hotel	Sandwich Place
17	Manhattan	Chelsea	40.744035	-74.003116	1	Coffee Shop	Italian Restaurant	Ice Cream Shop
18	Manhattan	Greenwich Village	40.726933	-73.999914	1	Italian Restaurant	Clothing Store	Sushi Restaurant
19	Manhattan	East Village	40.727847	-73.982226	1	Bar	Wine Bar	Chinese Restaurant
20	Manhattan	Lower East Side	40.717807	-73.980890	2	Coffee Shop	Ramen Restaurant	Pizza Place
21	Manhattan	Tribeca	40.721522	-74.010683	1	Italian Restaurant	Spa	Park
22	Manhattan	Little Italy	40.719324	-73.997305	1	Bakery	Café	Bubble Tea Shop
23	Manhattan	Soho	40.722184	-74.000657	1	Clothing Store	Boutique	Art Gallery
24	Manhattan	West Village	40.734434	-74.006180	1	Italian Restaurant	New American Restaurant	American Restaurant

25	Manhattan	Manhattan Valley	40.797307	-73.964286	0	Indian Restaurant	Pizza Place	Coffee Shop
26	Manhattan	Morningside Heights	40.808000	-73.963896	2	Coffee Shop	American Restaurant	Park
27	Manhattan	Gramercy	40.737210	-73.981376	1	Bar	Italian Restaurant	Cocktail Bar
28	Manhattan	Battery Park City	40.711932	-74.016869	1	Park	Coffee Shop	Hotel
29	Manhattan	Financial District	40.707107	-74.010665	1	Coffee Shop	Hotel	Wine Shop

Finally, let's visualize the resulting clusters

Examine Manhattan venue clusters

Examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, you can then assign a name to each cluster. I will leave this exercise to you.

Cluster 1

In [63]: `manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 0, manhattan_`

Out[63]:

	Neighborhood	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
4	Hamilton Heights	Pizza Place	Coffee Shop	Mexican Restaurant	Café	Yoga Studio	Indian Restaurant	Sushi Restaurant
7	East Harlem	Mexican Restaurant	Bakery	Deli / Bodega	Pizza Place	Latin American Restaurant	Thai Restaurant	Spa
25	Manhattan Valley	Indian Restaurant	Pizza Place	Coffee Shop	Yoga Studio	Playground	Bar	Café

Cluster 2

In [64]: `manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 1, manhattan_`

Out[64]:

	1st Most	2nd Most	3rd Most	4th Most	5th Most	6th Most

	Neighborhood	Common Venue	Common Venue	Common Venue	Common Venue	Common Venue	Common Venue	
1	Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Salon / Barbershop	Vietnamese Restaurant	Bubble Tea Shop	
6	Central Harlem	African Restaurant	French Restaurant	Chinese Restaurant	Public Art	Cosmetics Shop	Seafood Restaurant	
8	Upper East Side	Italian Restaurant	Exhibit	Art Gallery	Bakery	Gym / Fitness Center	Juice Bar	C
9	Yorkville	Italian Restaurant	Gym	Coffee Shop	Bar	Pizza Place	Sushi Restaurant	
10	Lenox Hill	Coffee Shop	Italian Restaurant	Pizza Place	Sushi Restaurant	Gym / Fitness Center	Sporting Goods Shop	B
11	Roosevelt Island	Coffee Shop	Sandwich Place	Park	Indie Theater	Dry Cleaner	Bus Stop	
12	Upper West Side	Italian Restaurant	Wine Bar	Bar	Vegetarian / Vegan Restaurant	Mediterranean Restaurant	Bakery	
13	Lincoln Square	Gym / Fitness Center	Theater	Café	Concert Hall	Plaza	Italian Restaurant	
14	Clinton	Theater	Gym / Fitness Center	Hotel	American Restaurant	Italian Restaurant	Wine Shop	
15	Midtown	Hotel	Coffee Shop	Cocktail Bar	Theater	Clothing Store	American Restaurant	
16	Murray Hill	Coffee Shop	Hotel	Sandwich Place	Japanese Restaurant	Italian Restaurant	Gym / Fitness Center	
17	Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bakery	Nightclub	American Restaurant	
18	Greenwich Village	Italian Restaurant	Clothing Store	Sushi Restaurant	Café	Seafood Restaurant	Indian Restaurant	
19	East Village	Bar	Wine Bar	Chinese Restaurant	Mexican Restaurant	Ice Cream Shop	Pizza Place	
21	Tribeca	Italian Restaurant	Spa	Park	Café	American Restaurant	Boutique	
22	Little Italy	Bakery	Café	Bubble Tea Shop	Clothing Store	Sandwich Place	Salon / Barbershop	Mec
23	Soho	Clothing Store	Boutique	Art Gallery	Shoe Store	Women's Store	Italian Restaurant	G
24	West Village	Italian Restaurant	New American	American Restaurant	Cosmetics Shop	Wine Bar	Jazz Club	

					Restaurant		
27	Gramercy	Bar	Italian Restaurant	Cocktail Bar	American Restaurant	Pizza Place	Bagel Shop
28	Battery Park City	Park	Coffee Shop	Hotel	Gym	Memorial Site	Boat or Ferry
29	Financial District	Coffee Shop	Hotel	Wine Shop	Steakhouse	Gym	Cocktail Bar
30	Carnegie Hill	Coffee Shop	Pizza Place	Bar	Yoga Studio	Spa	Bakery
31	Noho	Italian Restaurant	French Restaurant	Sushi Restaurant	Cocktail Bar	Bookstore	Grocery Store
32	Civic Center	Gym / Fitness Center	Italian Restaurant	Coffee Shop	French Restaurant	Hotel	Sandwich Place
33	Midtown South	Korean Restaurant	Hotel	Dessert Shop	American Restaurant	Japanese Restaurant	Hotel Bar
34	Sutton Place	Gym / Fitness Center	Italian Restaurant	Furniture / Home Store	Indian Restaurant	Juice Bar	Gym
35	Turtle Bay	Italian Restaurant	Steakhouse	Sushi Restaurant	Coffee Shop	Wine Bar	Ramen Restaurant
38	Flatiron	Gym	Yoga Studio	American Restaurant	Japanese Restaurant	Clothing Store	Gym / Fitness Center
39	Hudson Yards	American Restaurant	Italian Restaurant	Gym / Fitness Center	Café	Spanish Restaurant	Restaurant

Cluster 3

```
In [65]: manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 2, manhattan_
```

Out[65]:

	Neighborhood	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 Mos Commo Venu
2	Washington Heights	Café	Bakery	Mobile Phone Shop	Grocery Store	Spanish Restaurant	Coffee Shop	Tapa Restaurar
3	Inwood	Mexican Restaurant	Lounge	Café	Pizza Place	Deli / Bodega	Wine Bar	America Restaurar
5	Manhattanville	Italian Restaurant	Mexican Restaurant	Seafood Restaurant	Park	Coffee Shop	Deli / Bodega	Supermark
20	Lower East Side	Coffee Shop	Ramen Restaurant	Pizza Place	Café	Japanese Restaurant	Art Gallery	Bake
26	Morningside Heights	Coffee Shop	American Restaurant	Park	Bookstore	Burger Joint	Food Truck	Del Bodeg
36	Tudor City	Park	Mexican Restaurant	Café	Greek Restaurant	Hotel	Dog Run	Sus Restaurar

Cluster 4

```
In [66]: manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 3, manhattan_
```

Out[66]:

	Neighborhood	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	8 C
37	Stuyvesant Town	Bar	Park	Playground	German Restaurant	Basketball Court	Baseball Field	Harbor / Marina	

Cluster 5

```
In [67]: manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 4, manhattan_
```

Out[67]:

	Neighborhood	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	
0	Marble Hill	Sandwich Place	Discount Store	Coffee Shop	Yoga Studio	Steakhouse	Supplement Shop	Shopping Mall	R

Analyze Central Toronto Neighborhood and Results

Run *k*-means to cluster the Central Toronto neighborhood into 5 clusters

Examine Central Toronto venue clusters

Examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, you can then assign a name to each cluster. I will leave this exercise to you.

Cluster 1

In [77]:
```
Toronto_merged.loc[Toronto_merged['Cluster Labels'] == 0, Toronto_merged
```

Out[77]:

	Neighborhood	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 Venu
1	Davisville North	Sandwich Place	Hotel	Breakfast Spot	Gym	Park	Clothing Store	Food Dr Sh
2	North Toronto West	Clothing Store	Coffee Shop	Yoga Studio	Salon / Barbershop	Bagel Shop	Chinese Restaurant	Dess Sh
3	Davisville	Coffee Shop	Sandwich Place	Dessert Shop	Italian Restaurant	Pizza Place	Café	Su Restaur
5	Deer Park,Forest Hill SE,Rathnelly,South Hill,...	Coffee Shop	Pub	American Restaurant	Sports Bar	Vietnamese Restaurant	Fried Chicken Joint	Light F Stat
8	The Annex,North Midtown,Yorkville	Sandwich Place	Coffee Shop	Café	Pizza Place	American Restaurant	Park	F

Cluster 2

In [78]: `Toronto_merged.loc[Toronto_merged['Cluster Labels'] == 1, Toronto_merged`

Out[78]:

	Neighborhood	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
6	Roselawn	Garden	Yoga Studio	Dessert Shop	Ice Cream Shop	Hotel	History Museum	Gym	G Restau

Cluster 3

In [79]: `Toronto_merged.loc[Toronto_merged['Cluster Labels'] == 2, Toronto_merged`

Out[79]:

	Neighborhood	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 Com
4	Moore Park,Summerhill East	Tennis Court	Park	Summer Camp	Playground	Yoga Studio	Dessert Shop	History Museum	

Cluster 4

In [80]: `Toronto_merged.loc[Toronto_merged['Cluster Labels'] == 3, Toronto_merged`

Out[80]:

	Neighborhood	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 M Comm Ve
0	Lawrence Park	Swim School	Bus Line	Park	Yoga Studio	Dessert Shop	Hotel	History Museum	C

Cluster 5

In [81]: `Toronto_merged.loc[Toronto_merged['Cluster Labels'] == 4, Toronto_merged`

Out[81]:

	Neighborhood	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 Con V
7	Forest Hill North,Forest Hill West	Trail	Jewelry Store	Sushi Restaurant	Mexican Restaurant	Yoga Studio	Dessert Shop	History Museum	

Result Discussion and Recommendation

Result Discussion

Based on the outlined analysis methodology, we will discuss the results as follows:

1. Total number of venue comparison between Manhattan and Central Toronto:

```
(1) Both Manhattan and Central Toronto show similarities in larg
e numbers of diverse venue in their neighborhoods, reflecting th
eir leading business and cultural center status in their respect
ive countries.
(2) However, based on individual neighborhood venue data compari
son, Manhattan has order of magnitude more of total number of ve
nue, reflecting that Manhattan has much more number of neighborh
ood (40 for Manhattan and 9 for Central Toronto) and much higher
working and residential population than Central Toronto.
(3) Based on the above comparison, there should be good location
options for new business office and family home in Central Toron
to. However, the selection options in Central Toronto could stil
l be much less than Manhattan since the total number of venue op
tions in Central Toronto is much less than the total number of v
enue in Manhattan.
```

2. Venue ranking information and their implications:

(1) For Financial district in Manhattan, the top ranking venues include coffee shop, hotel, and various restaurants, reflecting those venues serving working and traveling populations. Will sel ect new busienss office location options in Central Toronto with the similar venue ranking.

(2) For high-end residential neighborhood in Manhattan, such as Upper East Side next to famous New York Central Park, the top v enues include restaurant, exhibit and art gallery, and fitness f acilities. These venues are also preferred by the busienss manag er for their new family home location optons in Toronto.

(3) To match the venue preference, the new business office locat ion will clearly be at relatively small area of financial distri ct in Central Toronto where there are high concentration of offi ce buildings, hotels, restaurants and coffee shop.

(4) For family home location options in Central Toronto, the cho ices could be relatively hard to match Upper East Side neighborh ood in Manhattan that are close to Park, musium and restaurants. Although downtown area in Central Toronto could offer good resta urant and musium venue, but downtown could be far away fro large park venue comparable to Manhattan Central Park. On the other ha nd, the large park venue in Toronto could be more available in a far distance areas from Toronto downtown and from restaurant and musium venues.

3. K-means venue clustering results and their implications:

(1) K-means venue clustering analysis provides consistent result s with those from venue ranking analysis. It provides convincing confirmation regarding observations from the venue ranking analy sis.

(2) In addition, the clustering results reveal that the neighbor hood of Davisville North offers good venues of Park, grovery sho pping and dancing and fitness facilities. Such venue availabilit y could be suitable for new family home location in Central Toro nto.

Recommendations:

Based on the above result discussions, we will propose the following recommendations regarding potential location options in Central Toronto for US business manager to consider:
(1) For new business office in Central Toronto, the clear option is at the financial district of Central Toronto, with plenty venues of office building, hotel, and restaurants.
(2) For new family home, the potential location options could be in the neighborhood of Davisville North or Davisville, where there are revnues of park, grocery shopping, restaurants and fitness facilities.

Suggestions for further study:

```
(1) After presenting the project, need to solocit feedbacks from
the US business manager and his company's decision makers.
(2) To verify and to evaluate location options, suggest to arran
ge visits to Central Toronto by the US business manager and his
family.
(3) May need to expand this project to inlcude other important f
actors in location selection, such as cost and transportation co
nvinince.
```

Conclusion

Based on data collection and analysis, this project provides detail evaluation of relocation options for US business manager for office and family home location selection.

Although Toronto also offers large number of venue for business and family life, comparison to Manhattan in New York, there are still some limitations, especially for family home location considerations. There coudl be tradeoff considerations to balance the preference for various venues.

While the report presents recommendations, it also suggests to further improve this study by getting feedbacks, arranging visits to Toronto, and to consider other important factors.

With further iterative data analysis and evaluation, we believe that Toronto could offer very satisfactory options for both new business office and family home selection.