CAPSTONE PROJECT REPORT

The Choice – NYU v/s UofT

Prepared by: Emma K

Date: May 2019

INTRODUCTION / BACKGROUND

Background:

DrivenByData is a new startup company that provides its clients detailed analysis and observations backed by relevant data, to help its subscribed users in making informed decisions.

Business Problem:

There is a recent request from Charlotte who resides in Paris, France. She is an international student wanting to pursue her post graduate degree in Fine Arts abroad. She received acceptance letters from both New York University | NYU (U.S) and University of Toronto | UofT (Canada). Charlotte who's never been to either cities before, would like to make an informed decision on which school to attend based on the proximity and availability of art related venues (such as performing centres, art galleries, art stores, etc.) to the university campus where she will attend classes. Charlotte would also like to have a choice of neighbourhoods based on her preference to be considered for residence.

DATA

The data scientist at DrivenByData has started assessing this business problem. He has concluded that the following data points will be required to provide a thorough analysis of both regions and recommendation to the client:

- Geo coordinates of both NYU and UofT
- Number of Art related venues around each of the NYU and UofT
- Distance of these venues from each of the universities
- Natural grouping of venue data on the basis of similarity

Data Description:

- FourSquare explorer API to search for Art related venues around NYU and UofT in their respective cities of New York City and Toronto.
- Publicly available dataset that contains neighbourhoods data of Manhattan (and New York City) at https://cocl.us/new_york_dataset

Approach to solve the problem:

- Convert addresses into their equivalent latitude and longitude values.
- Use the Foursquare API to explore neighbourhoods in New York City and Toronto
- Use the Foursquare API explore function to get the most common venue categories in each city
- Group the neighbourhoods into clusters using k-means clustering algorithm
- Use Follium library to visualize neighbourhoods and their emerging clusters.

METHODOLOGY

Data preparation

- After importing the necessary libraries and setting up the environment, use the Nominatim module to convert an address into latitude and longitude values.
- Given the addresses to both Universities as input:
  - NYU at 70 Washington Square S, New York, NY
  - UofT at 27 Kings College Circle, Toronto, ON

the following were returned as their respective geographical coordinates:

	City/Country	University	Latitude	Longitude
0	New York City / USA	NYU	40.729429	-73.997218
1	Toronto / Canada	UofT	43.660722	-79.395920

- Foursquare API requests are built to obtain results with art related venues in the vicinity of each university (radius of 1000m). The search query used was the word ‘art’ along with the geo coordinates.
- The resulting json response was analyzed based on the ‘venues’ component, then normalized into a DataFrame. This was done for both universities’ respective locations.
- The resulting DataFrame had many features: decided to keep only relevant one for this analysis, anything that include venue categories and location.
- In addition, the DataFrame was filtered on venue category and selected the following as they are of interest to our subject audience:

- Performing Arts Venue
- Art Gallery
- Arts & Crafts Store

- The filtered DataFrame then was sorted by distance: closest art venue first.
- They were 18 venues around NYU and 16 around UofT.

Top 10 art venues around NYU

```
0                           Blick Art Materials
1                         NYU Grey Art Gallery
2    Leslie+Lohman Museum of Gay & Lesbian Art
3                         BLICK Art Materials
4        The Brant Foundation Art Study Center
5          Storefront for Art and Architecture
6              New York University Art History
7                   La Sirena Mexican Folk Art
8                                         Artsy
9              Hulonthalo Gallery of Art in NY
Name: name, dtype: object
```

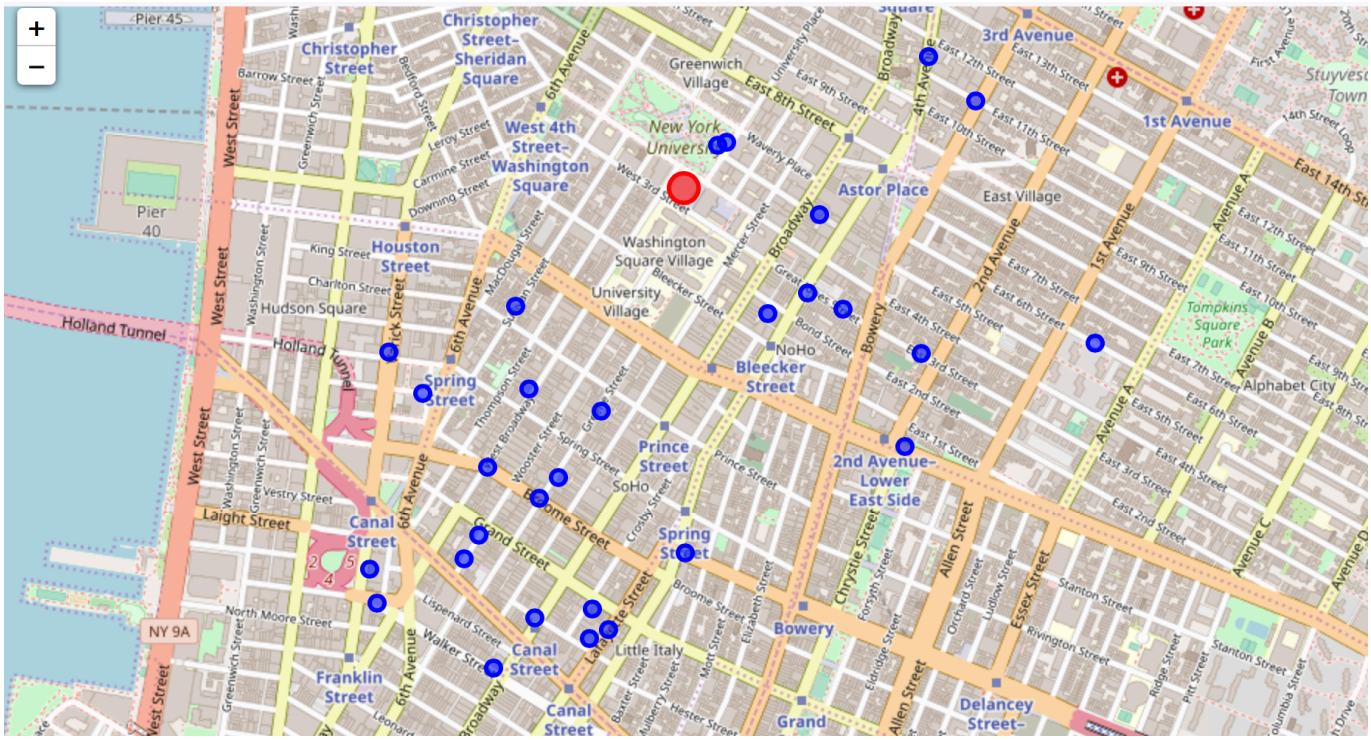
Top 10 art venues around UofT

```
0                              Art Gallery of Ontario
1    Ontario College of Art and Design University (...
2                           Aboveground Art Supplies
3                                  Department of Art
4                      University College Art Centre
5                             Curry's Art Store Ltd.
6                          Art Square Gallery & Cafe
7                                Toose Art Supplies
8                  University of Toronto Arts Centre
9                                      UTAC Art Lounge
Name: name, dtype: object
```

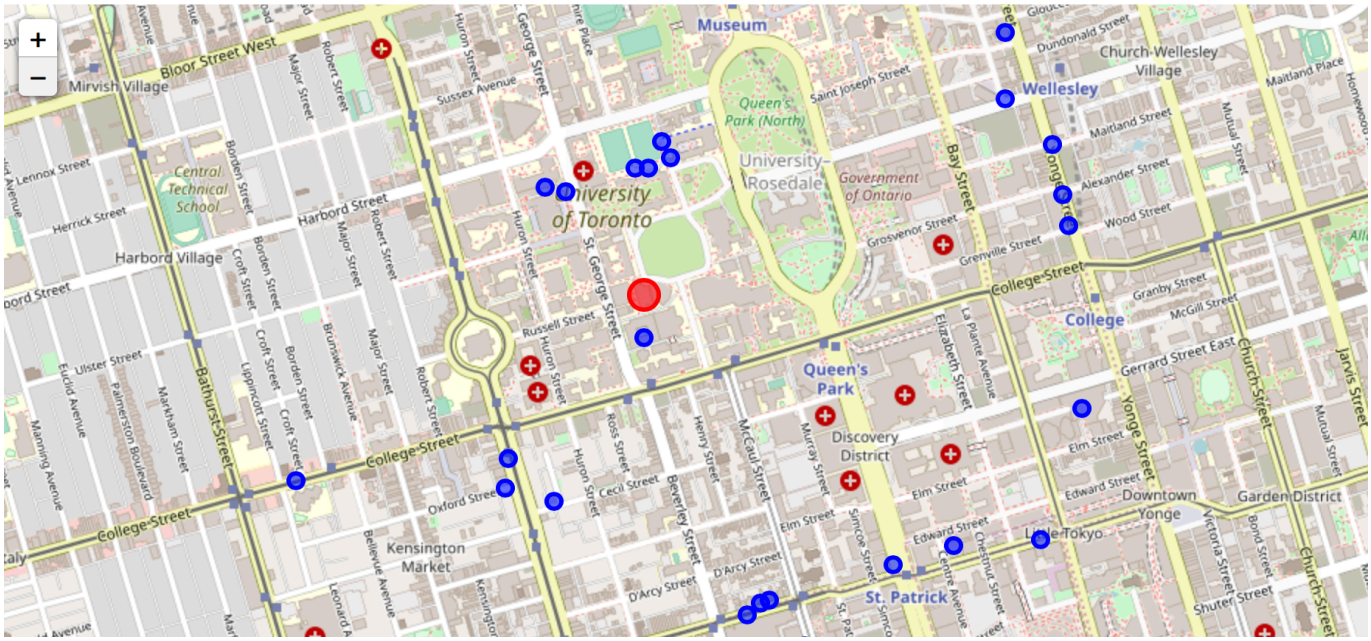
Exploratory Analysis

- Using Folium map plotting libraries, I created visualizations displaying the art venues around each of NYU and UofT. The red circle marker represents the university and the blue circles represent the art venues.

New York University



University of Toronto



By simply exploring both visualizations, we can easily conclude that while both cities data from Foursquare have resulted in the very similar number of venues (18 v/s 16), NYU has a more even distribution than UofT. This is useful insight when considering for where our client would like to live: NYU seems to have higher concentration of venues in nearby neighbourhoods, hence more options for our clients. UofT on the other hand seems to be more limiting as there seems to be no obvious concentration of art venues, they are dispersed and quite a distance from the main campus.

One key observation to make where the top 6 venues within 500 m of UofT are returned, by checking their names/address, one can easily tell that 4 out of 6 venues are on campus and tied to UofT. For the purpose of this analysis, the client's request was for options outside of her school environment.

	name	categories	address	cc	city	country	crossStreet	distance
7	Toose Art Supplies	Arts & Crafts Store	NaN	CA	Toronto	Canada	NaN	98
9	UTAC Art Centre	Art Gallery	15 Kings College Circle	CA	Toronto	Canada	NaN	297
10	UTAC Art Lounge	Art Gallery	15 King's College Circle	CA	Toronto	Canada	NaN	299
8	University of Toronto Arts Centre	Art Gallery	15 Kings College Circle	CA	Toronto	Canada	NaN	326
4	University College Art Centre	Art Gallery	NaN	CA	Toronto	Canada	King College Circle and Tower Rd	363
15	Gwartzman's Art Supplies	Arts & Crafts Store	448 Spadina Ave	CA	Toronto	Canada	College Street	495

On the other hand, from the top 5 list of venues within 500 m of NYU, only 1 out of 5 venues are on campus as seen below:

	name	categories	address	cc	city	country	crossStreet	distance
1	NYU Grey Art Gallery	Art Gallery	100 Washington Sq E	US	New York	United States	Washington Pl	154
15	Moniker Art Fair	Art Gallery	718 Broadway	US	New York	United States	NaN	340
0	Blick Art Materials	Arts & Crafts Store	1-5 Bond St	US	New York	United States	btwn Lafayette & Broadway	370
9	Hulonthalo Gallery of Art in NY	Art Gallery	670 Broadway	US	New York	United States	NaN	397
11	SoHo Gallery for Digital Art	Art Gallery	138 Sullivan St	US	New York	United States	at Prince St.	501

Based on the exploratory analysis, it seems NYU a better choice so far. Next we will further analyze New York City neighbourhoods for identifying neighbourhoods where our client could consider for her residence.

Clustering New York City

Next we will explore further the neighbourhoods of New York City mainly Manhattan for the purpose of our analysis.

First I downloaded the publicly available dataset that contains location data about New York neighbourhoods found at https://cocl.us/new_york_dataset. Then I created a DataFrame with all the neighbourhoods of Manhattan. Following that step, I used Foursquare API to explore the Manhattan neighbourhoods and segment them. I created a function to use Foursquare API for the identified neighbourhoods. There were a total of 1181 venues in Manhattan neighbourhoods. I had to limit the venues to only show the same art venue categories as we had done earlier.

The final grouping by neighbourhood resulted in a total of 14 art related venues grouped as follows:

Neighborhood	Venue
Battery Park City	1
Central Harlem	1
Hudson Yards	1
Lenox Hill	1
Lincoln Square	2
Lower East Side	3
Manhattan Valley	1
Manhattanville	1
Soho	2
Upper East Side	1

The neighbourhoods were analyzed by ‘one hot encoding’ and the data normalized by grouping rows of neighbourhood and the mean of the frequency of occurrence of each category:

	Neighborhood	Art Gallery	Arts & Crafts Store	Performing Arts Venue
168	Manhattanville	1	0	0
197	Central Harlem	1	0	0
251	Upper East Side	1	0	0
317	Lenox Hill	1	0	0
393	Lincoln Square	0	0	1

	Neighborhood	Art Gallery	Arts & Crafts Store	Performing Arts Venue
0	Battery Park City	0.000000	0.0	1.000000
1	Central Harlem	1.000000	0.0	0.000000
2	Hudson Yards	1.000000	0.0	0.000000
3	Lenox Hill	1.000000	0.0	0.000000
4	Lincoln Square	0.000000	0.0	1.000000
5	Lower East Side	0.666667	0.0	0.333333
6	Manhattan Valley	0.000000	1.0	0.000000
7	Manhattanville	1.000000	0.0	0.000000
8	Soho	0.500000	0.5	0.000000
9	Upper East Side	1.000000	0.0	0.000000

I then created a new DataFrame to display the top 3 venues (most common venues) for each neighbourhood. The resulting DataFrame looks as shown below:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Battery Park City	Performing Arts Venue	Arts & Crafts Store	Art Gallery
1	Central Harlem	Art Gallery	Performing Arts Venue	Arts & Crafts Store
2	Hudson Yards	Art Gallery	Performing Arts Venue	Arts & Crafts Store
3	Lenox Hill	Art Gallery	Performing Arts Venue	Arts & Crafts Store
4	Lincoln Square	Performing Arts Venue	Arts & Crafts Store	Art Gallery
5	Lower East Side	Art Gallery	Performing Arts Venue	Arts & Crafts Store
6	Manhattan Valley	Arts & Crafts Store	Performing Arts Venue	Art Gallery
7	Manhattanville	Art Gallery	Performing Arts Venue	Arts & Crafts Store
8	Soho	Arts & Crafts Store	Art Gallery	Performing Arts Venue
9	Upper East Side	Art Gallery	Performing Arts Venue	Arts & Crafts Store

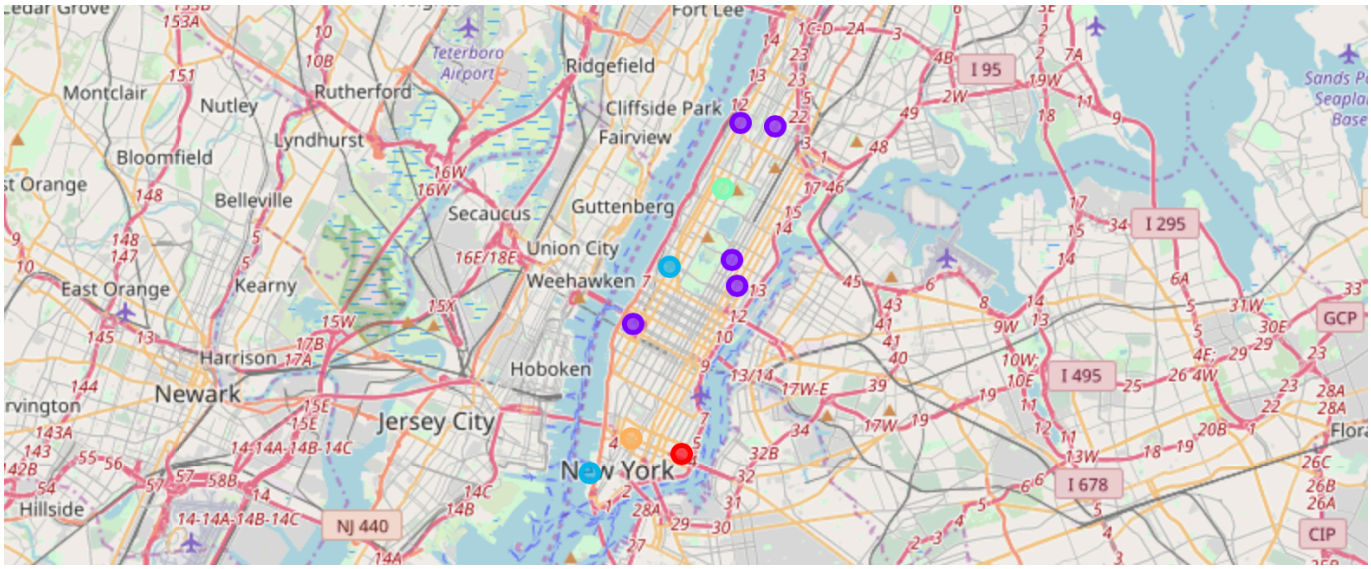
Machine learning

Clustering is a technique for finding similar grouping in a data, also referred to as clusters. It attempts to group by similarity, but not driven by a known goal/purpose. Clustering is an unsupervised learning, as you don't have prescribed labels in the data.

To make use of this algorithm, I built a dataset that merged both neighbourhoods coordinates data and the most common venues data as shown below:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
5	Manhattan	Manhattanville	40.816934	-73.957385	1	Art Gallery	Performing Arts Venue	Arts & Crafts Store
6	Manhattan	Central Harlem	40.815976	-73.943211	1	Art Gallery	Performing Arts Venue	Arts & Crafts Store
8	Manhattan	Upper East Side	40.775639	-73.960508	1	Art Gallery	Performing Arts Venue	Arts & Crafts Store
10	Manhattan	Lenox Hill	40.768113	-73.958860	1	Art Gallery	Performing Arts Venue	Arts & Crafts Store
13	Manhattan	Lincoln Square	40.773529	-73.985338	2	Performing Arts Venue	Arts & Crafts Store	Art Gallery
20	Manhattan	Lower East Side	40.717807	-73.980890	0	Art Gallery	Performing Arts Venue	Arts & Crafts Store
23	Manhattan	Soho	40.722184	-74.000657	4	Arts & Crafts Store	Art Gallery	Performing Arts Venue
25	Manhattan	Manhattan Valley	40.797307	-73.964286	3	Arts & Crafts Store	Performing Arts Venue	Art Gallery
28	Manhattan	Battery Park City	40.711932	-74.016869	2	Performing Arts Venue	Arts & Crafts Store	Art Gallery
39	Manhattan	Hudson Yards	40.756658	-74.000111	1	Art Gallery	Performing Arts Venue	Arts & Crafts Store

The algorithm was run for 5 clusters, here's a visualization of all on the map where each colour represents one cluster.



I examined the clusters to determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, here are the clusters:

Cluster #1

	Neighborhood	3rd Most Common Venue
20	Lower East Side	Arts & Crafts Store

Cluster #2

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
5	Manhattanville	Art Gallery	Performing Arts Venue	Arts & Crafts Store
6	Central Harlem	Art Gallery	Performing Arts Venue	Arts & Crafts Store
8	Upper East Side	Art Gallery	Performing Arts Venue	Arts & Crafts Store
10	Lenox Hill	Art Gallery	Performing Arts Venue	Arts & Crafts Store
39	Hudson Yards	Art Gallery	Performing Arts Venue	Arts & Crafts Store

Cluster #3

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
13	Lincoln Square	Performing Arts Venue	Arts & Crafts Store	Art Gallery
28	Battery Park City	Performing Arts Venue	Arts & Crafts Store	Art Gallery

Cluster #4

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
25	Manhattan Valley	Arts & Crafts Store	Performing Arts Venue	Art Gallery

RESULTS

- Comparing the surrounding neighbourhoods of both NYU and UofT, based on the exploratory analysis I provided, one can easily tell that there is no shortage of art related venues, galleries and art supply stores around both campuses. NYU has a total of 18 and UofT has 16, very comparable. However, plotting the results on the map, it is obvious to see that NYU has all 18 venues with uniform distribution compared to the visualization shown on the map surrounding the UofT campus. That said, UofT campus has 6 art venues within 500 m radius compared to NYU with only 4. Further investigation of both data sets, I noticed that 4 out of 6 venues are within the UofT campus and only 1 out of 5 was in NYU campus. This insight indicates that our choice towards NYU is the better one for the client.
- To address our client's request, we also had to provide an analysis of neighbourhoods with greater options of having nearby art venues, galleries and art supply stores, for her to reside in. This is where clustering of the New York city was very helpful. The only cluster that had all three types of art venues: Performing Art Venue, Art Gallery and Arts supply store was Cluster #2 with 5 different neighbourhood choices for our client.

DISCUSSION

- The analysis path that was taken resulted in clear path in favour of NYU compared to UofT. This will help our client in making her choice/decision.
- NYU has more art related venues within walking distance from the campus, which is an important decision factor for our client.
- NYU also has 5 neighbourhoods nearby that satisfy our client's request for selecting a residence place.

CONCLUSION

- This analysis depended on the accuracy of the data returned by Foursquare API.
- The client's request for having art related venues nearby was assumed to be venues of categories belonging to either a Performing Arts, Art Gallery and Arts & Crafts stores. The results could have been different if other venue categories were considered.
- A similar analysis on the Toronto neighbourhoods could also be done to improve the level of confidence in the recommendation.
- The analysis could be improved by adding data on demographics and rent prices for the resulting neighbourhoods.