INTRODUCTION

We live in a global world where opportunities are not limited to the country we reside in. We apply for jobs in new cities, start business endeavors across of the globe, and have a seemingly endless list of places we can travel to.

New York City and London are no exception. Both financial centers are filled with booming economies, lively cultural experiences, and a wealth of resources. As such, it is not uncommon for people to consider visiting or even moving there. Say an employee must move from New York City to London to take on a new role at her company. She wants to live in a neighborhood that offers many of the same amenities as her hometown. While the decision to move may be a difficult one, it may be harder to know what neighborhood to settle down in. This comes as no surprise as these metropolises are home to a population of more than 8 million and span roughly 300 square miles and 600 square miles, respectively.

With so many neighborhoods to choose from, it's tough for an individual to weigh all the options. That's where the role of data science comes in: we will use the computing power of machines and the insights from statistical models to help the user make a more informed decision. Through this project, a user will be able to see how the various neighborhoods in New York City and London are similar based on the venues they offer.

DATA SOURCE

- For the New York City portion of this project, I will be using the data set provided in the IBM lab that is sourced from NYU. The NYC data set is a json file that contains the names of the neighborhoods as well their boroughs and geographic coordinates.

- For the London portion of this project, I will be using this Wikipedia page on areas in London. The data set will be web scraped from the table under "Other use of place names" and for the sake of comparison I will assume the district location is synonymous to the neighborhood value in the NYC data set. This table also contains the names of the boroughs and grid ref geographic coordinates.

- To identify venues within the neighborhoods, I will be using the Foursquare API. I will then conduct an analysis using K-Means clustering across both cities to find similar neighborhoods on the basis of mean venue frequency.

METHODOLOGY

Before I began to analyze the data set, I had to extract the relevant information from the data set and transform it into a Pandas data frame. The resulting data frames for New York City and London, respectively, were as follows:

Borough	Neighborhood	Latitude	Longitude
Bronx	Wakefield	40.894705	-73.847201
Bronx	Co-op City	40.874294	-73.829939
Bronx	Eastchester	40.887556	-73.827806
Bronx	Fieldston	40.895437	-73.905643
Bronx	Riverdale	40.890834	-73.912585

Figure 1: New York City initial data frame

Borough	Neighborhood	PostTown	OSGridRef
Bexley, Greenwich	Abbey Wood	LONDON	TQ465785
Ealing, Hammersmith and Fulham	Acton	LONDON	TQ205805
Croydon	Addington	CROYDON	TQ375645
Croydon	Addiscombe	CROYDON	TQ345665
Bexley	Albany Park	BEXLEY, SIDCUP	TQ478728

Figure 2: London initial data frame

In the London data set, I converted the “OSGridRef” to the Latitude and Longitude coordinates using the *OSGridConverter* package and *grid2latlong* library. This was necessary to use the Foursquare API in later steps.

I filtered the London data set to only include the “PostTown” London to narrow down the results from 531 neighborhoods to 297. This was done to try and even the playing field, as New York City has fewer neighborhoods and is a smaller than London.

Upon inspecting the data frame, found that there were neighborhoods with the same name both within and across the two cities. This would become an issue in later steps when I would group the venues by neighborhood. Because the name of the neighborhood would not be considered a unique key, venues would be wrongly merged. To solve this, I concatenated the neighborhood and borough columns. I also added a column to specify the city as the boroughs might not be as well known to the user. The resulting data frame is shown below:

City	Neighborhood	Latitude	Longitude
London	Chelsea (Kensington and Chelsea)	51.482096	-0.164778
New York City	Chelsea (Manhattan)	40.744035	-74.003116
New York City	Chelsea (Staten Island)	40.594726	-74.189560
London	Childs Hill (Barnet)	51.563647	-0.204811
New York City	Chinatown (Manhattan)	40.715618	-73.994279
London	Chinatown (Westminster)	51.511252	-0.131875
London	Chinbrook (Lewisham)	51.431244	0.029005
London	Chingford (Waltham Forest)	51.632025	0.014819
London	Chiswick (Hounslow, Ealing, Hammersmith and Fu...	51.492616	-0.265268
London	Church End (Barnet)	51.599373	-0.188951
London	Church End (Brent)	51.492616	-0.265268

Figure 3: Combined data frame with updated "Neighborhood" column

I then mapped the cities using *Folium* and added labels to mark the neighborhoods in New York City and London, respectively:

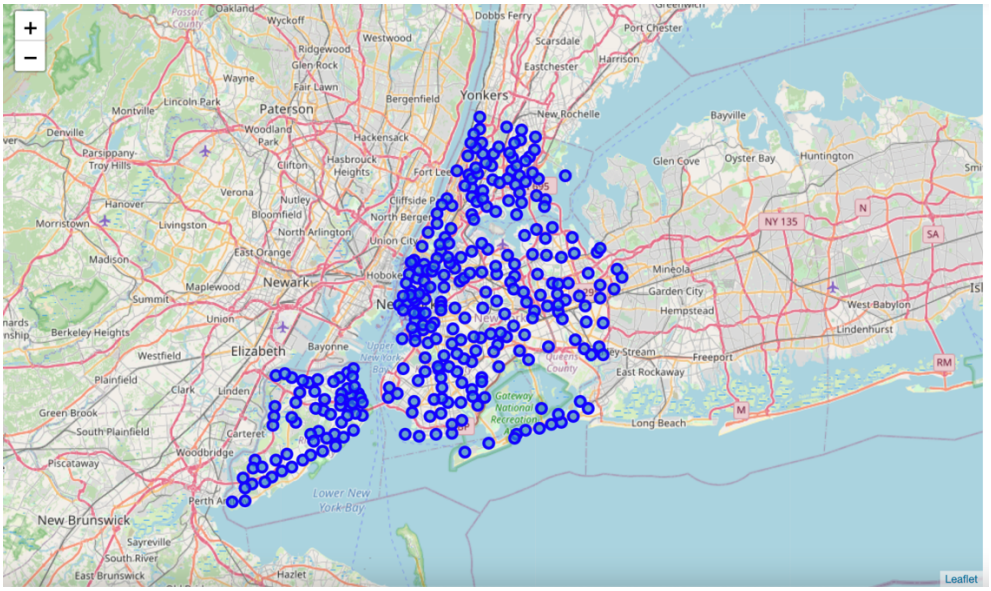

Figure 4: Map of New York City and its 306 neighborhoods

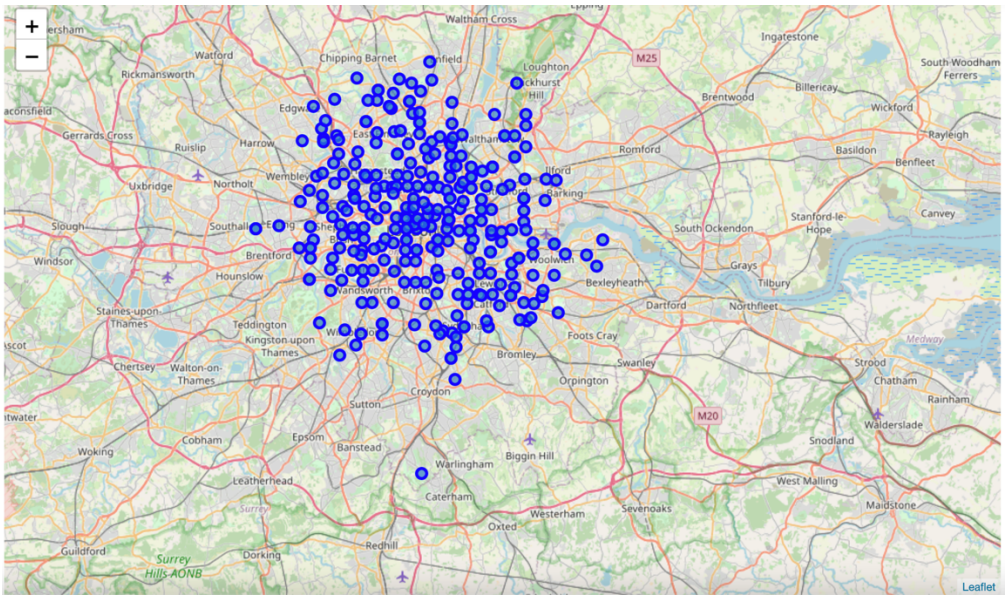

Figure 5: Map of London narrowed down to 297 neighborhoods

I then used the Foursquare API to get the top 100 venues within a 750-meter (approximately 0.5 mile) radius of the neighborhood. From those 29,507 venues, I extracted the venue category and created a dummy variable for each of the 525 unique categories. After that, I grouped the results by the now unique neighborhood column. I then found the mean frequency of occurrence of each category and focused on the top 10 categories. Lastly, I used K-Mean clustering to group the similar neighbors into 5 clusters.

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	9th Most Common Venue	10th Most Common Venue
Abbey Wood (Bexley, Greenwich)	Playground	Grocery Store	Campground	Indian Restaurant	Fabric Shop	Factory	Falafel Restaurant	Farm	Financial or Legal Service	Farmers Market
Acton (Ealing, Hammersmith and Fulham)	Gym / Fitness Center	Pub	Grocery Store	Indian Restaurant	Train Station	Park	Fast Food Restaurant	Supermarket	Chinese Restaurant	Bakery
Aldgate (City)	Coffee Shop	Hotel	Gym / Fitness Center	Restaurant	Middle Eastern Restaurant	Cocktail Bar	Café	Italian Restaurant	Food Truck	French Restaurant
Aldwych (Westminster)	Hotel	Theater	Coffee Shop	Restaurant	Café	Ice Cream Shop	Bakery	Steakhouse	Museum	History Museum
Allerton (Bronx)	Donut Shop	Pizza Place	Sandwich Place	Supermarket	Food	Fast Food Restaurant	Bus Station	Pharmacy	Discount Store	Gas Station

Figure 6: Top 10 venues per neighborhood using Foursquare API

RESULTS & DISCUSSION

These are the 5 different neighborhood clusters:

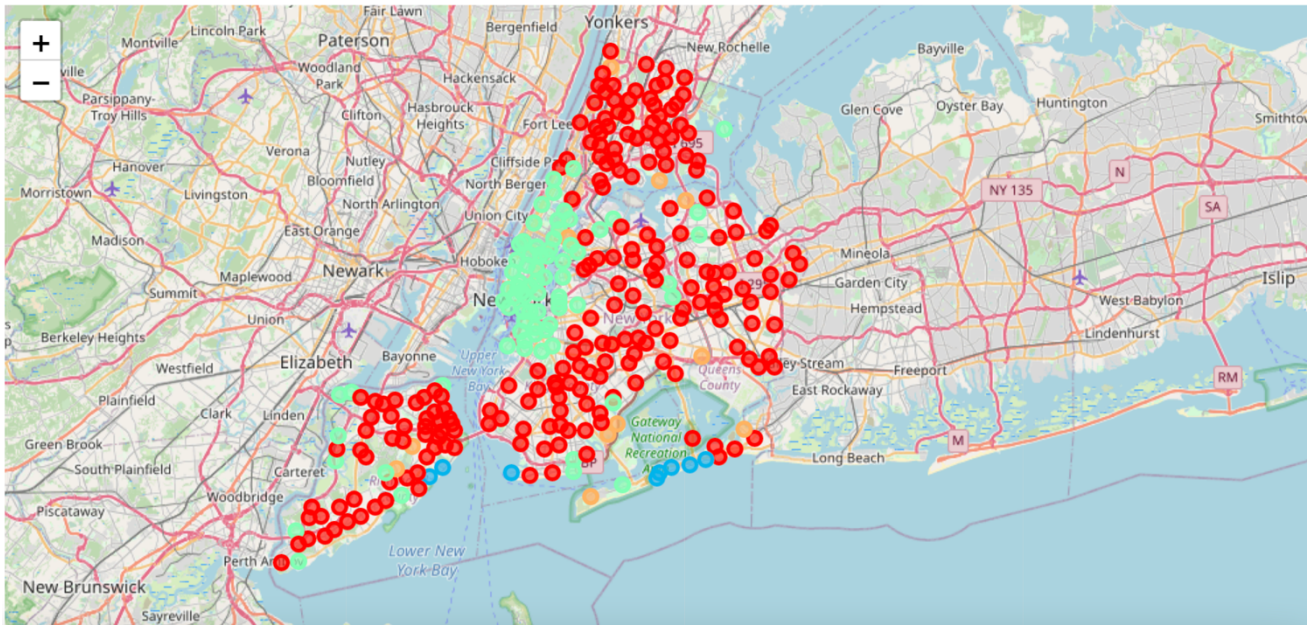

Figure 7: Cluster map of New York City showing only 4 of the 5 clusters

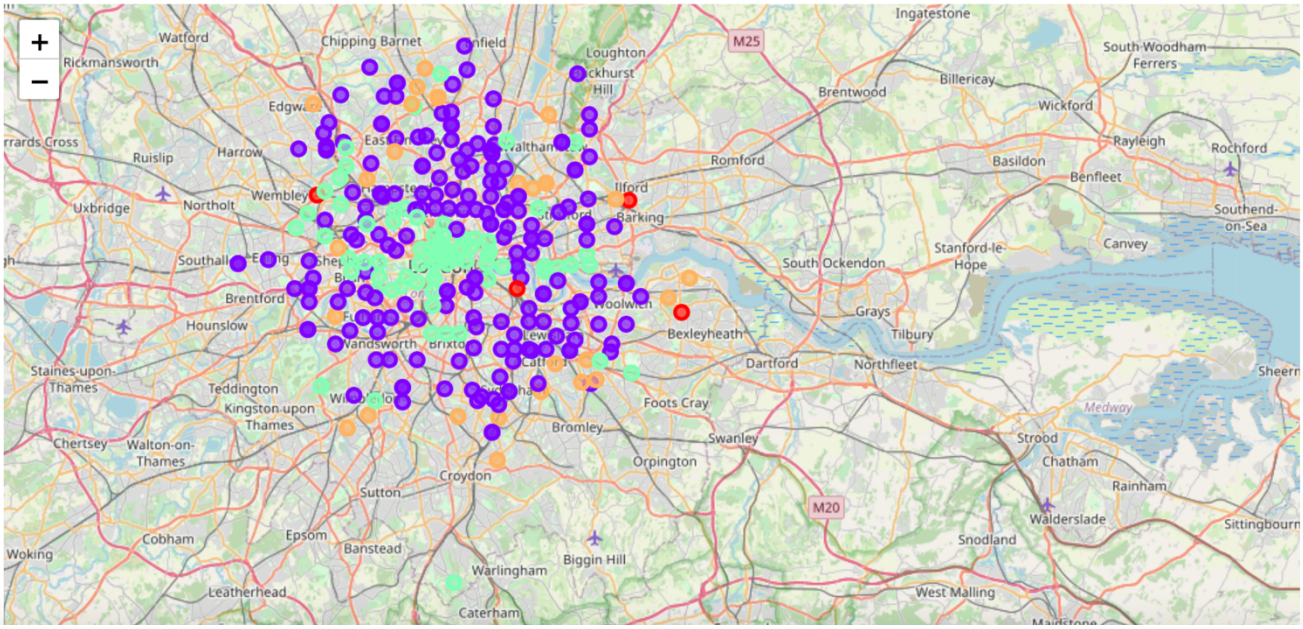

Figure 8: Cluster map of London showing only 4 of the 5 clusters

To better understand how to label the clusters beyond their colors, let's examine each cluster:

Red (Cluster 0):

```
merged.loc[merged['Cluster Labels'] == 0, merged.columns[[1] + list(range(5, merged.shape[1]))]]
```

	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	9th Most Common Venue	10th Most Common Venue
0	Wakefield (Bronx)	Pharmacy	Caribbean Restaurant	Supermarket	Fast Food Restaurant	Gas Station	Bagel Shop	Ice Cream Shop	Food	Bakery	Dessert Shop
1	Co-op City (Bronx)	Mattress Store	Accessories Store	Pizza Place	Fast Food Restaurant	Shopping Mall	Pharmacy	Bakery	Bank	Harbor / Marina	Seafood Restaurant
2	Eastchester (Bronx)	Caribbean Restaurant	Fast Food Restaurant	Diner	Burger Joint	Shopping Mall	Grocery Store	Cocktail Bar	Seafood Restaurant	Pizza Place	Sandwich Place
4	Riverdale (Bronx)	Bank	Sandwich Place	Bar	Medical Supply Store	Mexican Restaurant	Pharmacy	Pizza Place	Japanese Restaurant	Diner	Donut Shop
5	Kingsbridge (Bronx)	Pizza Place	Bar	Sandwich Place	Mexican Restaurant	Deli / Bodega	Burger Joint	Diner	Bakery	Donut Shop	Park
6	Marble Hill (Manhattan)	Spanish Restaurant	Bank	Sandwich Place	Pizza Place	Donut Shop	Supplement Shop	Pharmacy	Mexican Restaurant	Bakery	Supermarket
7	Woodlawn (Bronx)	Pizza Place	Pub	Food Truck	Deli / Bodega	Bar	Rental Car Location	Ice Cream Shop	Train Station	Bakery	Donut Shop

In the Red cluster ***Pizza Places, Delis/Bodegas & Pharmacies*** are the most prevalent. This cluster covers most of the New York City with the exception of the city center. There are very few of these in London. This is not surprising as New York City is known for its pizza and delis/bodegas.

Purple (Cluster 1):

```
merged.loc[merged['Cluster Labels'] == 1, merged.columns[[1] + list(range(5, merged.shape[1]))]]
```

	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	9th Most Common Venue	10th Most Common Venue
307	Acton (Ealing, Hammersmith and Fulham)	Gym / Fitness Center	Pub	Grocery Store	Indian Restaurant	Train Station	Park	Fast Food Restaurant	Supermarket	Chinese Restaurant	Bakery
312	Archway (Islington)	Pub	Coffee Shop	Italian Restaurant	Grocery Store	Café	Pizza Place	Japanese Restaurant	Seafood Restaurant	Asian Restaurant	Pool
314	Balham (Wandsworth)	Coffee Shop	Pub	Grocery Store	Pizza Place	Bakery	Italian Restaurant	Indian Restaurant	Bar	Fast Food Restaurant	Supermarket
315	Bankside (Southwark)	Pub	Coffee Shop	Italian Restaurant	Café	Wine Bar	Seafood Restaurant	Bakery	Street Food Gathering	Asian Restaurant	Gym / Fitness Center
318	Barnsbury (Islington)	Grocery Store	Pub	Park	Coffee Shop	Café	Theater	Brewery	Rental Car Location	Breakfast Spot	Bar
319	Battersea (Wandsworth)	Coffee Shop	Grocery Store	Italian Restaurant	Pub	Bar	Hotel	Park	Track	Cocktail Bar	Furniture / Home Store

In the Purple cluster ***Pubs, Cafés, Coffee Shops, Grocery Stores & Parks*** are the most prevalent. This cluster covers most London with the exception of the city center. This cluster does not exist in New York City. That's not to say these venues don't exist, but rather they are not the most popular types of venues within the neighborhoods in New York City.

Blue (Cluster 2):

```
merged.loc[merged['Cluster Labels'] == 2, merged.columns[[1] + list(range(5, merged.shape[1]))]]
```

	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	9th Most Common Venue	10th Most Common Venue
85	Sea Gate (Brooklyn)	Beach	Supermarket	Paper / Office Supplies Store	Park	Construction & Landscaping	Spa	Falafel Restaurant	Farm	Factory	Financial or Legal Service
178	Rockaway Beach (Queens)	Beach	Ice Cream Shop	Latin American Restaurant	Bar	Bagel Shop	Food Stand	BBQ Joint	Restaurant	Pharmacy	Hotel
179	Neponsit (Queens)	Beach	Park	Bus Stop	Fast Food Restaurant	Financial or Legal Service	Film Studio	Filipino Restaurant	Field	Zoo Exhibit	Fish Market
190	Belle Harbor (Queens)	Beach	Spa	Boutique	Italian Restaurant	Deli / Bodega	Pub	Event Space	Chinese Restaurant	Bakery	Bagel Shop
191	Rockaway Park (Queens)	Beach	Italian Restaurant	Spa	Bagel Shop	Pizza Place	Donut Shop	Deli / Bodega	Bar	Pub	Boutique
204	South Beach (Staten Island)	Beach	Pier	Athletics & Sports	Theme Park	Skate Park	Deli / Bodega	Soccer Field	American Restaurant	BBQ Joint	Food
232	Midland Beach (Staten Island)	Baseball Field	Beach	Other Great Outdoors	Basketball Court	Food	Bookstore	Chinese Restaurant	Bus Stop	Bagel Shop	Deli / Bodega
302	Hammels (Queens)	Beach	Taco Place	Supermarket	Wine Shop	Fried Chicken Joint	Gym / Fitness Center	Farmers Market	Fast Food Restaurant	Bakery	Pharmacy

In the Blue cluster ***Beaches*** are the most prevalent. This cluster only exists in New York City and is located on the edges of the city by the water. This makes a lot of sense as New York is located on a natural harbor and surrounded by water, unlike London.

Green (Cluster 3):

```
merged.loc[merged['Cluster Labels'] == 3, merged.columns[[1] + list(range(5, merged.shape[1]))]]
```

	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	9th Most Common Venue	10th Most Common Venue
12	City Island (Bronx)	Harbor / Marina	Boat or Ferry	Seafood Restaurant	Deli / Bodega	American Restaurant	Park	Italian Restaurant	Flower Shop	Fish Market	Spanish Restaurant
49	Greenpoint (Brooklyn)	Bar	Coffee Shop	Cocktail Bar	Yoga Studio	Pizza Place	Café	Record Shop	Mexican Restaurant	French Restaurant	Italian Restaurant
52	Sheepshead Bay (Brooklyn)	Sandwich Place	Turkish Restaurant	Seafood Restaurant	Italian Restaurant	Grocery Store	Harbor / Marina	Bar	Bank	Russian Restaurant	Pharmacy
59	Prospect Heights (Brooklyn)	Bar	Mexican Restaurant	Plaza	Wine Shop	Sushi Restaurant	Bakery	Cocktail Bar	Coffee Shop	New American Restaurant	Ice Cream Shop
61	Williamsburg (Brooklyn)	Pizza Place	Coffee Shop	Wine Bar	American Restaurant	Latin American Restaurant	Bar	Cocktail Bar	Nightclub	South American Restaurant	Bakery
64	Brooklyn Heights (Brooklyn)	Park	Wine Shop	Coffee Shop	Yoga Studio	Italian Restaurant	Pizza Place	Cocktail Bar	Middle Eastern Restaurant	Deli / Bodega	Gym / Fitness Center

In the Green cluster ***Coffee Shops, Hotels & a variety of other venues*** are prevalent. This cluster only exists in the city centers of New York City and London. This can be explained by the concentration of venues in the heart of the cities that cater to the needs of the residents, workers, and tourists.

Orange (Cluster 4):

```
merged.loc[merged['Cluster Labels'] == 4, merged.columns[[1] + list(range(5, merged.shape[1]))]]
```

	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	9th Most Common Venue	10th Most Common Venue
3	Fieldston (Bronx)	Bus Station	Park	Plaza	Coffee Shop	River	Playground	Café	Art Gallery	Pizza Place	Medical School
27	Clason Point (Bronx)	Park	Scenic Lookout	Recording Studio	Pool	Boat or Ferry	Playground	Grocery Store	Falafel Restaurant	Farm	Film Studio
76	Mill Island (Brooklyn)	Golf Course	Pool	Harbor / Marina	Gym Pool	BBQ Joint	Playground	Park	Seafood Restaurant	Tourist Information Center	Flower Shop
91	Bergen Beach (Brooklyn)	Harbor / Marina	Playground	Baseball Field	Park	Gym	Comfort Food Restaurant	Athletics & Sports	Falafel Restaurant	Farm	Film Studio
110	Roosevelt Island (Manhattan)	Park	Bus Line	Deli / Bodega	Athletics & Sports	Boat or Ferry	Pizza Place	Gym	Sandwich Place	Tram Station	Pharmacy
148	South Ozone Park (Queens)	Park	Bar	Hotel	Fried Chicken Joint	Donut Shop	Deli / Bodega	Intersection	Moving Target	Food Truck	Fast Food Restaurant

In the Orange cluster ***Parks & Grocery Stores*** are the most prevalent. This cluster exists in the outskirts of New York City and London.

As expected, the neighborhoods in London are more similar to each other than they are to those in New York City, but that's not to say that you can't find those same venues. If someone lived in New York City and frequently stopped by a pizza place when visiting her friends, she can still do that in London, but she might have to go down a few blocks. Likewise, the London city center will still have the hustle and bustle of a large city. That's where the two cities will feel quite similar in their choice of venues. In the future, it would be interesting to group these venues into different categories to understand

what sorts of activities are popular in the neighborhood. That way a user can better decide if they want a quiet neighborhood with natural landscapes (beaches, parks, and grocery stores), a neighborhood with a vibrant nightlife (pubs, bars, theaters, concerts), or one with a multitude of food options (restaurants, bakeries, grocery stores).

CONCLUSION

As the world continues to become more connected and filled with more and more data, humans will be faced with more decisions to make. We will have to consider more than just a change in venue options. But as these decisions continue to grow more complex, we will have advanced technological capabilities. This will aid in exploring data in a more efficient and effective manner to uncover insights that were nearly impossible or extremely time consuming for humans to do on their own.