Capstone Project

--Applied Data Science Capstone by IBM/Coursera

Introduction: Find a similar neighborhood

In this project we will try to find a similar **neighborhood** between Two cities, **Toronto** and **New York**.

Here is the scenario. James now is working in Toronto for a company and living in the Neighborhood **Cedarbrae** which in Borough **Scarborough**. He received another opportunity for a company locates on New York. He decided to accept this opportunity. Before he moves to New York, he wants to find a house with a neighborhood similar to which he is living in the Toronto.

We will compare the neighborhood with their **Venues** as the first criterion. Then rank the similar neighborhood by the distance from **his new company in Greenwich Village**.

We will use our data science powers to generate a most promising neighborhoods based on these criterias. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholder.

Data

Based on definition of our problem, factors that will influence our decission are:

- The similarity between different neighborhood according to its venues.
- Distance of neighborhood from his new company.

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

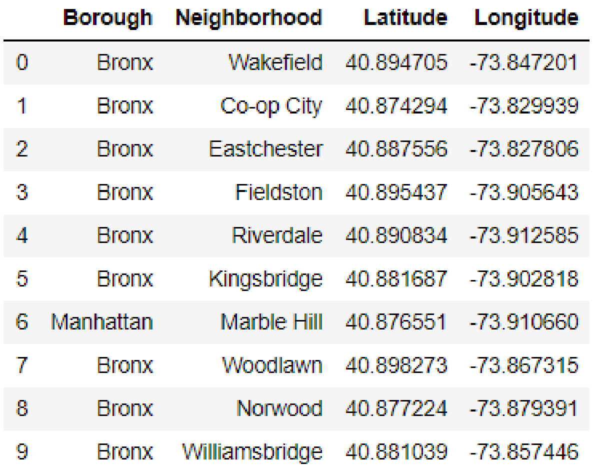
- Boroughs list of **Toronto** and **New York** from web(New York: https://cocl.us/new_york_dataset / Toronto:https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
)
- Same with above, we will obtain the neighborhood list from the web.
- Longitude and latitude will be obtained by using **Google Maps API reverse geocoding**
- Venues will be obtained using **Foursquare API**

Prepare the Neighborhood data for New York city

Let's get the data from below csv file:

https://cocl.us/new_york_dataset

Here is the what it looks like.



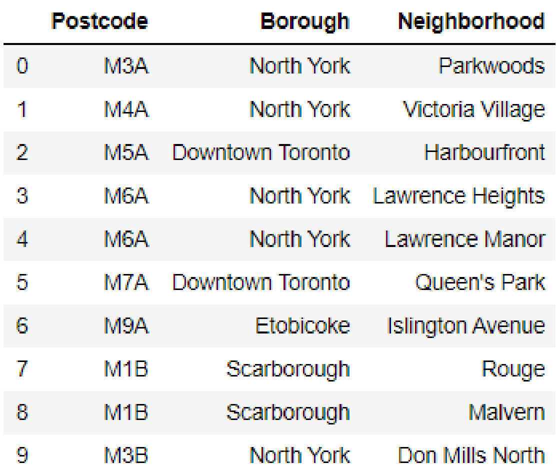
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446

Prepare Neighborhood the data for Toronto

We use the Beautiful Soup to grab the table from website(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

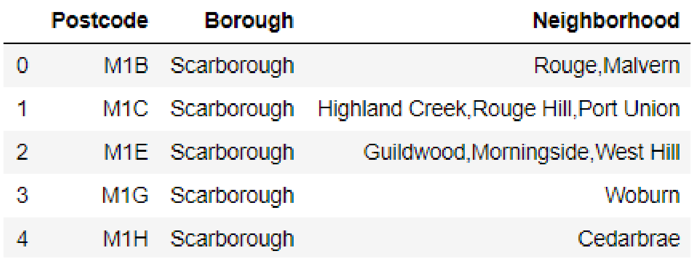
But there is a lot invalid Borough and "Not assigned". Let's get the valid Borough value list and remove the "Not assigned" in borough and use the borough name instead of the neighborhood name if it is "Not assigned"

Here is the data looks like.



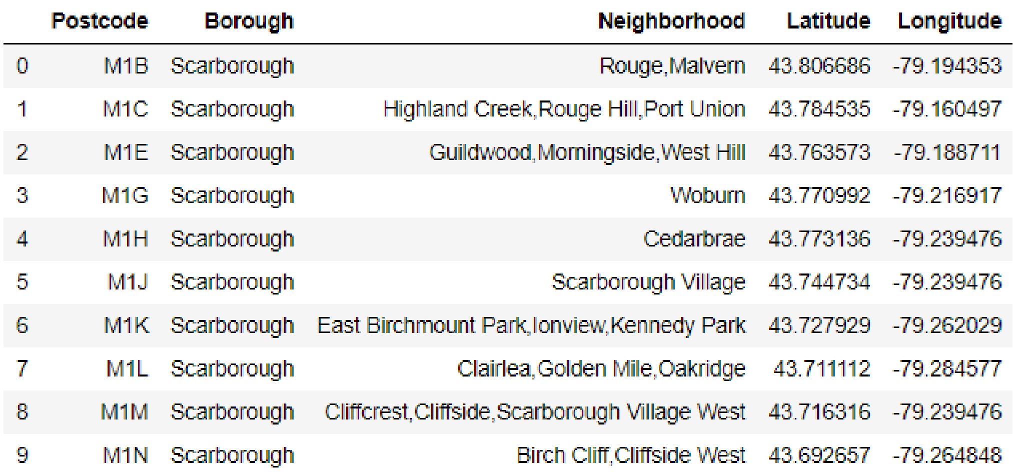
	Postcode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M6A	North York	Lawrence Heights
4	M6A	North York	Lawrence Manor
5	M7A	Downtown Toronto	Queen's Park
6	M9A	Etobicoke	Islington Avenue
7	M1B	Scarborough	Rouge
8	M1B	Scarborough	Malvern
9	M3B	North York	Don Mills North

Let’s combine the Neighborhood name if the Postcode and Borough is same and show the final table



	Postcode	Borough	Neighborhood
0	M1B	Scarborough	Rouge,Malvern
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union
2	M1E	Scarborough	Guildwood,Morningside,West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Finally let's combine the coordinates and Postcode from a csv file (http://cocl.us/Geospatial_data -O Torontocoordinates.csv).

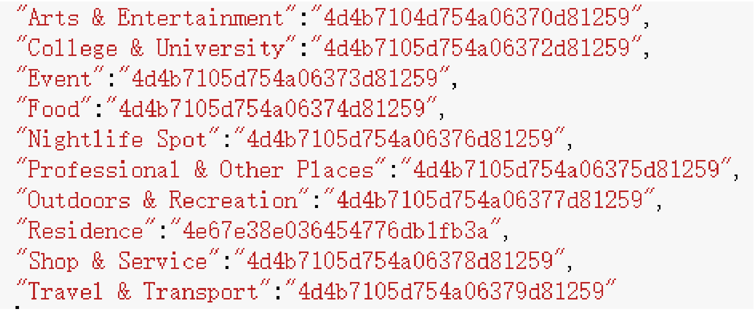


	Postcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park,Ionview,Kennedy Park	43.727929	-79.262029
7	M1L	Scarborough	Clairlea,Golden Mile,Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffcrest,Cliffside,Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff,Cliffside West	43.692657	-79.264848

Grab the venues from the Foursquare API

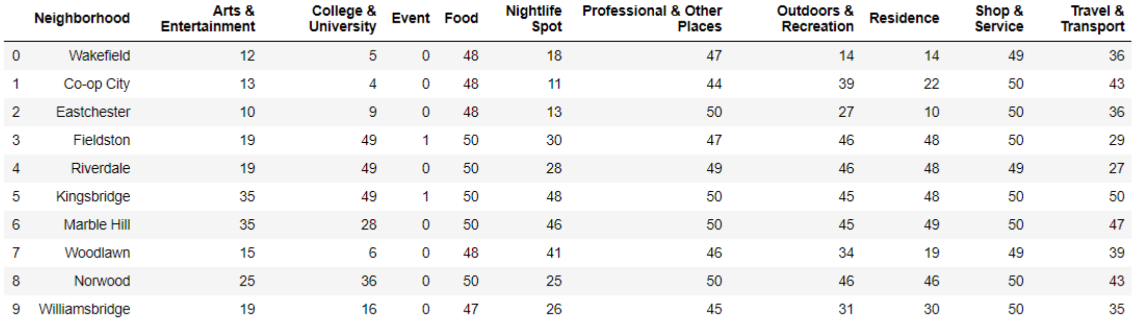
Now we already have the coordinates data for every single Neighborhood. Next, we can use the coordinates information to grab the number venues of very categories that most currently checked in. We consider that the venues of a neighborhood is the reason why that people will decide to live there. Then if two neighborhoods have similar venues, we can consider they are similar neighborhood.

Here are the top-level categories we will use and its id.



```
"Arts & Entertainment":"4d4b7104d754a06370d81259",
"College & University":"4d4b7105d754a06372d81259",
"Event":"4d4b7105d754a06373d81259",
"Food":"4d4b7105d754a06374d81259",
"Nightlife Spot":"4d4b7105d754a06376d81259",
"Professional & Other Places":"4d4b7105d754a06375d81259",
"Outdoors & Recreation":"4d4b7105d754a06377d81259",
"Residence":"4e67e38e036454776db1fb3a",
"Shop & Service":"4d4b7105d754a06378d81259",
"Travel & Transport":"4d4b7105d754a06379d81259"
```

We use the Place API from foursquare to grab the venues number of every categories of every neighborhood in New York. Here is what it looks like.



	Neighborhood	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Professional & Other Places	Outdoors & Recreation	Residence	Shop & Service	Travel & Transport
0	Wakefield	12	5	0	48	18	47	14	14	49	36
1	Co-op City	13	4	0	48	11	44	39	22	50	43
2	Eastchester	10	9	0	48	13	50	27	10	50	36
3	Fieldston	19	49	1	50	30	47	46	48	50	29
4	Riverdale	19	49	0	50	28	49	46	48	49	27
5	Kingsbridge	35	49	1	50	48	50	45	48	50	50
6	Marble Hill	35	28	0	50	46	50	45	49	50	47
7	Woodlawn	15	6	0	48	41	46	34	19	49	39
8	Norwood	25	36	0	50	25	50	46	46	50	43
9	Williamsbridge	19	16	0	47	26	45	31	30	50	35

Methodology

In this project we will focus on how to use the neighborhood data to find out the alternative neighborhoods that can meet our stakeholder's requests.

To find out the alternative neighborhods, the **first step** is collect the data: neighborhood list/venue category list of every neighborhood/coordinate of ervry neighborhood. In the Date section, we already pretreated the data that used in this project.

Second step is to build the algorithm that can meansure similarity bewteen different neighborhoods. In second step, there will be 2 parts. One, explore the data to see if any data we can discard. Two, use the number of every categories as the components of vector and calcuate the Euclidean distance.

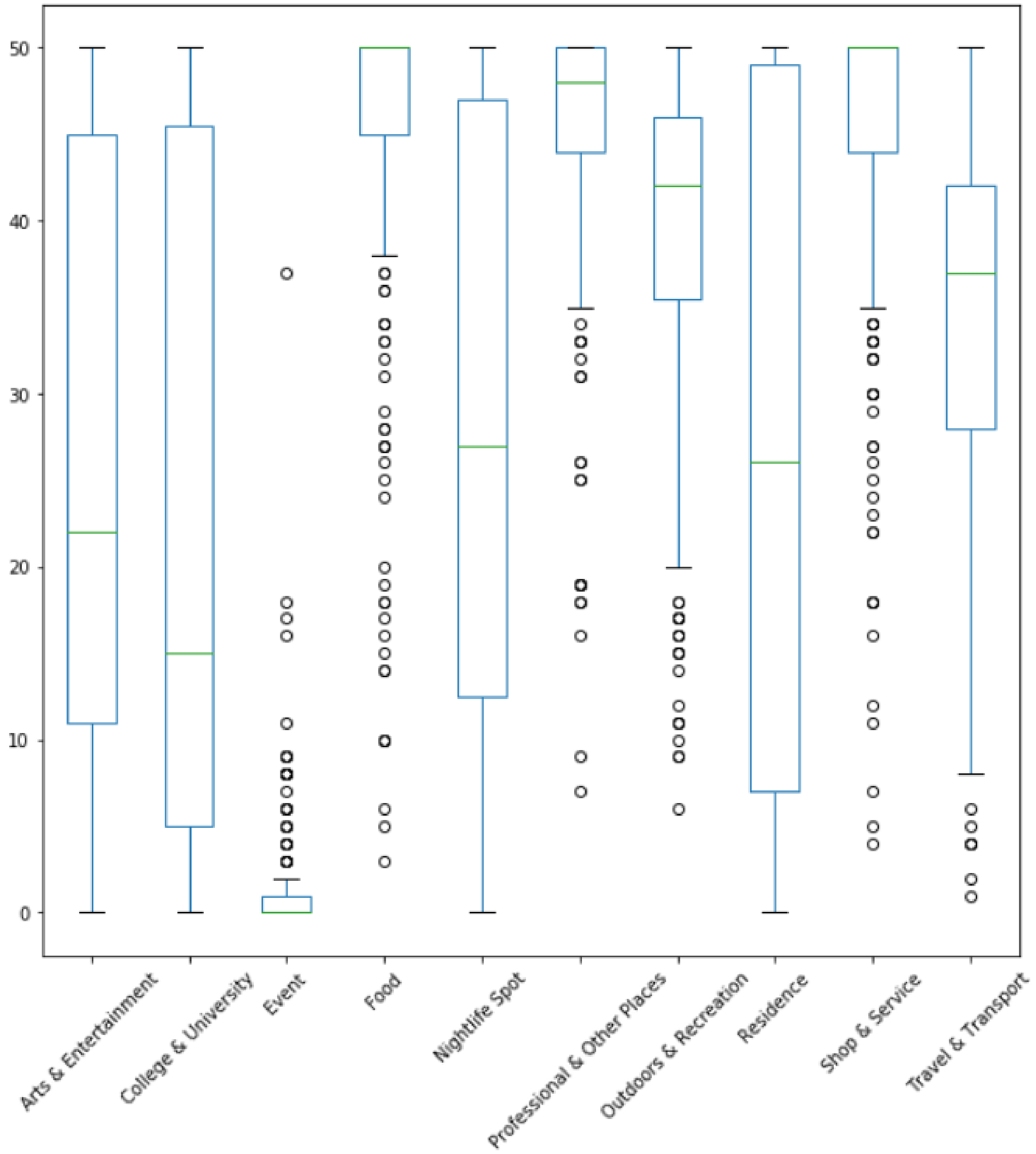
Third step will use the coordinate data to calcuate the distance between the company location and the Top 5 similar neighborhood and will rank them base on the distance.

Finally, **3 neighborhoods** can be recommended to our stakeholder which can meet the two criterions: with similar venues in the neigborhood and close to his company.

Analysis

Explore the Data

Let's use the boxplot to explore the data(showing the average count, spread and outliers).



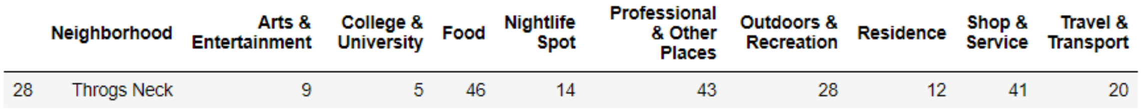
We limit the maximun venues number for every category to 50, and set the radius to 500 meters.

From the boxplot, we can see, the most frequent category **Food**, followed by **Professional & Other Places**. The 3rd place is **Outdoors & Recreation**. Event has very little data, so we'll discard it.

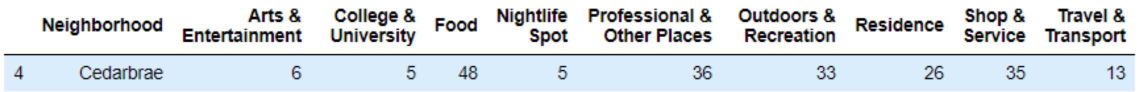
Calculate the Similarity

Now we have the venues number of every category, then we can use the venues numbers as the componentes of the vector. And we can calculate the Euclidean distance between every two vectors. If two vectors are close to ecah other, that means they are similar to each other. Then we can calculate the distance between every 2 vectors and sort the distance. After that, we can get the **Top 5** similar Neighborhoods.

From the above list, the most similar neighborhood with **Cedarbrae** is **Throgs Neck**. Let's output the venues nubmer of **Cedarbrae** and **Throgs Neck**



	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Professional & Other Places	Outdoors & Recreation	Residence	Shop & Service	Travel & Transport
28	Throgs Neck	9	5	46	14	43	28	12	41	20



	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Professional & Other Places	Outdoors & Recreation	Residence	Shop & Service	Travel & Transport
4	Cedarbrae	6	5	48	5	36	33	26	35	13

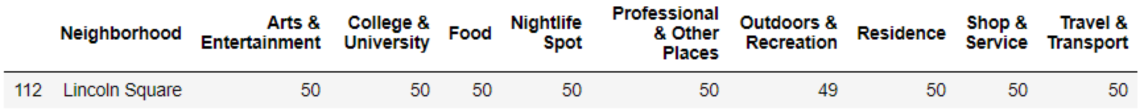
Now we can see the Top 3 venues categories of **Cedarbrae** are:
Food, **Professional & Other Places** and **Shop & Service**.

And the Top 3 venues categories of **Throgs Neck** are:

Food, **Professional & Other Places** and **Shop & Service**.

We can see the result matches our expectation and similar to Cedarbrae

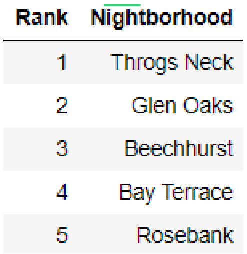
And we can also check the most dissimilar neighborhood **Lincoln Square**.



	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Professional & Other Places	Outdoors & Recreation	Residence	Shop & Service	Travel & Transport
112	Lincoln Square	50	50	50	50	50	49	50	50	50

we can see that **Lincoln Square** has different venues when compare to the **Cedarbrae**.

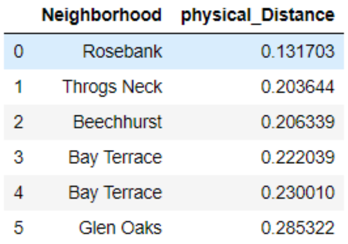
The final Top 5 neighborhoods which similar to **Cedarbrae**.



Rank	Nightborhood
1	Throgs Neck
2	Glen Oaks
3	Beechhurst
4	Bay Terrace
5	Rosebank

Calculate the Physical Distance

In this section, we can use the coordination of neighborhood to calculate the physical distance bewtween alternative neighborhoods and his new company in **Greenwich Village**. Let's rank the Top 5 by the distance

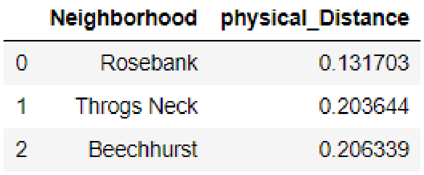


	Neighborhood	physical_Distance
0	Rosebank	0.131703
1	Throgs Neck	0.203644
2	Beechhurst	0.206339
3	Bay Terrace	0.222039
4	Bay Terrace	0.230010
5	Glen Oaks	0.285322

Result

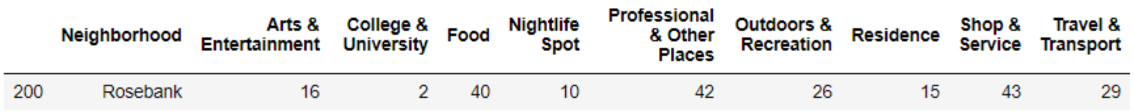
Here is the final Top 3 neighborhoods base on our criterions:

- The similarity between different neighborhood according to its venues.
- Distance of neighborhood from his new company.

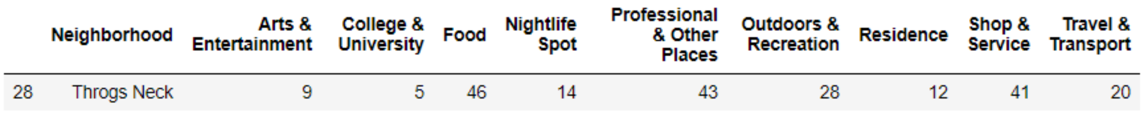


	Neighborhood	physical_Distance
0	Rosebank	0.131703
1	Throgs Neck	0.203644
2	Beechhurst	0.206339

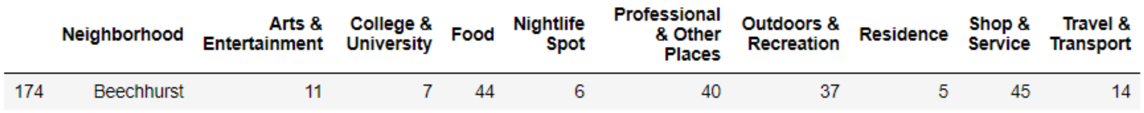
Let's see what Veneus in these 3 neighborhood: **Rosebank**, **Throgs Neck** and **Beechurst**.



	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Professional & Other Places	Outdoors & Recreation	Residence	Shop & Service	Travel & Transport
200	Rosebank	16	2	40	10	42	26	15	43	29



	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Professional & Other Places	Outdoors & Recreation	Residence	Shop & Service	Travel & Transport
28	Throgs Neck	9	5	46	14	43	28	12	41	20



	Neighborhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Professional & Other Places	Outdoors & Recreation	Residence	Shop & Service	Travel & Transport
174	Beechhurst	11	7	44	6	40	37	5	45	14

We can see all of them have similar categories when compare to our target neighborhood. And can be used this as the final result for our stakeholder.

Discussion

To be honest, the data we are using now is the Top-Level categories, which can't provide more details about the venues. Like the category **Food**, we still can be divided into different categories. If we can divide them basic on the region like **Chinese Restaurant**/**Japanese Restaurant**/**Italian Restaurant**, which can be more representative for our stakeholder's favor.

Conclusion

In this capstone project, we use the skills of Data science to analyse the data of neighborhoods.

- We grab the neighborhood data from web and get the venues data from Foursqaure.
- We clean the invalid data
- We use the knowledge of high dimensional data analysis to figure out the most similar neighborhoods.
- Finally, we get the final 3 alternative categories to our stakeholder.