Coursera Capstone Project – Battle of the Neighborhoods

To explore and analyze the venues in the neighborhoods of Philadelphia and Toronto using FourSquare venue data

1. Introduction

1.1 Background

When visiting or moving to a new city there can be difficulties when trying to choose or discover restaurants, stores and other local venues to explore in the neighborhoods. This can be daunting for both a user and business perspective when venturing forth into new territory.

When exploring new neighborhoods, there is a difficult task of analyzing various forms of criteria that can drastic impact your decision making. Such factors include, distance, price tier, ratings, photos and tips which is a lot of criteria too look through.

The Philadelphia area has many universities and colleges make it a top study destination, as the city has evolved into an educational and economic hub, with an estimated gross metropolitan product of $490 billion in 2019. This makes the twelve neighborhoods of Philadelphia a hot spot for new residents, travelers and businesses. The twelve neighborhoods of Philadelphia are Center City, South Philadelphia, Southwest Philadelphia, West Philadelphia, Lower North Philadelphia, Upper North Philadelphia, Bridesburg-Kensington-Port Richmond, Roxborough-Manayunk, Germantown-Chestnut Hill, Olney-Oak Lane, Near Northeast Philadelphia and Far Northeast Philadelphia.

Toronto is an international center for business, finance, arts and culture in the provincial capital of Ontario. Toronto is also an education and economic hub with diversified strengths in technology, design, food services, education and many more. This makes the ten boroughs a hot spot for new residents, travelers and businesses much like Philadelphia. The ten boroughs that make up Toronto are North York, Downtown Toronto, Etobicoke, Scarborough, East York, York, East Toronto, West Toronto, Central Toronto, Mississauga.

1.2 Problem

With the utilization of Foursquare venue data we can utilize this to explore and compare the city of Philadelphia and Toronto. This project will focus on how this data can be used to find new insights required for business opportunities in certain neighborhoods, cluster and segment venues to provide for a better user experience. We can then also compare and distinguish which city venues are better.

By segmenting and clustering venues based on our Foursquare venue data, we can compare and distinguish venue similarities to observe which neighborhoods would be good business ventures. Also, we will utilize factors such as user likes, ratings, tips, photos and distance to venues to seek if there are correlations that exist.

1.3 Interest

Interested parties might include business start-ups, stakeholders, and existing owners. Business start-ups might be looking for areas that lack certain categorical venues to further enhance a neighborhoods overall experience. Stakeholders would also be interested in this and also as a means to see how their invested interests are performing as a whole comparatively within city neighborhoods and to another city. Existing business will also benefit to see how the venues likes, tips and photos will bring effectiveness on their price tie and ratings.

2. Data Acquisition and Cleaning

2.1 Data Sources

Our two primary sources of data will be location information provided from data scrapped from Wikipedia on the neighborhood information for Philadelphia and Toronto plugged into python's geocoder library. Foursquare will provide our venue details to help explore, segment and analyze the neighborhoods of Philadelphia and Toronto. Below are the links of the Wikipedia sources for location data that was utilized in conjunction with Python's geocoder:

Philadelphia: https://en.wikipedia.org/wiki/List_of_Philadelphia_neighborhoods

Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2.2 Data Cleaning

After scraping the data from Wikipedia there was a lot of effort put forth in the cleaning and organizing of the data. Each set of data was cleaned to provide our location information. Before obtaining location data, the number of neighborhoods per borough were extracted for Frequency Distribution analysis.(Figures 1.1 & 1.2) The unique Boroughs' location information was obtained using the geocoder library to figure out the latitude and longitude for each of the boroughs. (Figure 2.1 & 2.2).

Figure 1.1 – Philadelphia Neighborhood Frequency Distribution by Borough

	Neighborhood	%-overall
South Philadelphia	29	16.02%
West Philadelphia	29	16.02%
Center City	21	11.6%
Near Northeast Philadelphia	19	10.5%
Far Northeast Philadelphia	16	8.84%
Southwest Philadelphia	15	8.29%
Lower North Philadelphia	12	6.63%
Germantown-Chestnut Hill	11	6.08%
Olney-Oak Lane	11	6.08%
Upper North Philadelphia	6	3.31%
Bridesburg-Kensington-Port Richmond	6	3.31%
Roxborough-Manayunk	6	3.31%

Figure 1.2 – Toronto Neighborhood Frequency Distribution by Borough

	Neighborhood	%-overall
North York	24	23.3%
Downtown Toronto	19	18.45%
Scarborough	17	16.5%
Etobicoke	12	11.65%
Central Toronto	9	8.74%
West Toronto	6	5.83%
East York	5	4.85%
East Toronto	5	4.85%
York	5	4.85%
Mississauga	1	0.97%

Figure 2.1 – Philadelphia Borough Localization Data

	Borough	Latitude	Longitude
0	Center City	39.952544	-75.165219
1	South Philadelphia	39.964110	-75.161050
2	Southwest Philadelphia	39.910040	-75.186370
3	West Philadelphia	40.053132	-75.028511
4	Lower North Philadelphia	39.964158	-75.198802
5	Upper North Philadelphia	40.059110	-75.052180
6	Bridesburg-Kensington-Port Richmond	39.980900	-75.099600
7	Roxborough-Manayunk	40.037990	-75.223080
8	Germantown-Chestnut Hill	40.078489	-75.211934
9	Olney-Oak Lane	40.041130	-75.124050
10	Near Northeast Philadelphia	40.092800	-74.987030
11	Far Northeast Philadelphia	40.092800	-74.987030

Figure 2.2 – Toronto Borough Localization Data

	Borough	Latitude	Longitude
0	North York	43.768260	-79.412630
1	Downtown Toronto	43.658200	-79.368320
2	Etobicoke	43.644360	-79.567130
3	Scarborough	43.772200	-79.256660
4	East York	43.691800	-79.327030
5	York	43.692080	-79.478630
6	East Toronto	43.659030	-79.349010
7	West Toronto	43.664712	-79.346346
8	Central Toronto	43.609727	-79.492844
9	Mississauga	43.587260	-79.644940

Next, we are able to append our venue information to our dataframes. We query foursquare to find our top 25 venues in each borough, making sure to return back information on each. This query focuses on obtaining the location information for each of the venues in the borough and their associated foursquare ID.(Figure 3.1 & 3.2)

Figure 3.1 – Philadelphia Borough Venue Localization & Categorization

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue ID	Distance	Venue Latitude	Venue Longitude	Venue Category
0	North York	43.76826	-79.41263	The Keg	5a35b4443abcaf37eb1a0d88	191	43.766579	-79.412131	Steakhouse
1	North York	43.76826	-79.41263	Konjiki Ramen	5a02789d0a464d3112a58785	144	43.766998	-79.412222	Ramen Restaurant
2	North York	43.76826	-79.41263	Toronto Centre for the Arts	4ad4c062f964a520c3f720e3	255	43.766228	-79.414115	Theater
3	North York	43.76826	-79.41263	Loblaws	4ae257cff964a520758d21e3	66	43.768722	-79.412101	Grocery Store
4	North York	43.76826	-79.41263	Satay Sate	57f92db0498ee70159702002	179	43.766690	-79.412100	Indonesian Restaurant

Figure 3.2 – Toronto Borough Venue Localization & Categorization

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue ID	Distance	Venue Latitude	Venue Longitude	Venue Category
0	North York	43.76826	-79.41263	The Keg	5a35b4443abcaf37eb1a0d88	191	43.766579	-79.412131	Steakhouse
1	North York	43.76826	-79.41263	Konjiki Ramen	5a02789d0a464d3112a58785	144	43.766998	-79.412222	Ramen Restaurant
2	North York	43.76826	-79.41263	Toronto Centre for the Arts	4ad4c062f964a520c3f720e3	255	43.766228	-79.414115	Theater
3	North York	43.76826	-79.41263	Loblaws	4ae257cff964a520758d21e3	66	43.768722	-79.412101	Grocery Store
4	North York	43.76826	-79.41263	Satay Sate	57f92db0498ee70159702002	179	43.766690	-79.412100	Indonesian Restaurant

Finally, we utilize the foursquare one last time to obtain each venues respective details for user likes, ratings, price tier, photo count and tips. This information is compiled in to one final dataframe that will be utilized for the remainder of the project to address the questions and problems we have. (Figure 4.1 & 4.2)

Figure 4.1 – Finalized Philadelphia Venue Localization & Information Dataset

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue ID	Distance	Venue Latitude	Venue Longitude	Venue Category	like_counts	rating	photo_count	reasons_count	Tips count
Center City	39.9525435	-75.165219	Dilworth Park	4bde0d566198c9b6c5cc12ff	49	39.9527718	-75.164723	Park	311	9.1	668	1	23
Center City	39.9525435	-75.165219	Philadelphia Film Center	47bdd66df964a520da4d1fe3	195	39.950835	-75.164683	Movie Theater	88	8.9	177	1	20
Center City	39.9525435	-75.165219	City Hall Courtyard	4f29e8e3e4b02f0aff55b2a7	138	39.9524842	-75.163592	Plaza	58	8.9	214	1	3
Center City	39.9525435	-75.165219	Del Frisco's Double Eagle Steak House	4ab2ac0bf964a520d66b20e3	177	39.9509564	-75.165459	Steakhouse	275	8.8	309	1	85
Center City	39.9525435	-75.165219	The Ritz-Carlton, Philadelphia	4a68db6ff964a52023cb1fe3	152	39.9514456	-75.164149	Hotel	199	8.8	599	1	63

Figure 4.2 – Finalized Toronto Venue Localization & Information Dataset

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue ID	Distance	Venue Latitude	Venue Longitude	Venue Category	like_counts	rating	photo_count	reasons_count	Tips count
North York	43.76826	-79.41263	The Keg	5a35b4443abcaf37eb1a0d88	191	43.7665789	-79.412131	Steakhouse	25	8.5	7	0	3
North York	43.76826	-79.41263	Konjiki Ramen	5a02789d0a464d3112a58785	144	43.7669977	-79.412222	Ramen Restaurant	39	8.3	68	1	8
North York	43.76826	-79.41263	Toronto Centre for the Arts	4ad4c062f964a520c3f720e3	255	43.7662283	-79.414115	Theater	46	8.1	145	1	13
North York	43.76826	-79.41263	Loblaws	4ae257cff964a520758d21e3	66	43.768722	-79.412101	Grocery Store	90	7.8	94	1	13
North York	43.76826	-79.41263	Satay Sate	57f92db0498ee70159702002	179	43.76669	-79.4121	Indonesian Restaurant	8	7.8	4	0	4

2.3 Data Limitations

Currently, the two limiting factors in a more robust data set is the fact that we have to limit the API calls for premium data, as a foursquare personal account restricts the premium call daily limit to 500. Thus, we look at the boroughs overall and only the top 25 venues in each borough. We are also limited by the amount of available data provided by foursquare at each of these venues, as they can be missing important information pertaining to our regression analyses. As such, this is more a modeling of what is possible with our the queriable data and could be expanded upon when limitations are lifted.

2.4 How the Data Will Be Utilized

Initially, an overall frequency table will be generated with percentages to distinguish the two important characteristics. First how many neighborhoods make up each borough and the overall percentage that borough contributes to the cities total neighborhood count. We can utilize this to understand if there are any significant differences between our cities.

The information extraction from Foursquare’s premium calls will allow us to utilize the information on user likes, tips, photo counts, rating and distance to see if there are any correlations that exist between these variables.

We will use the data to obtain information regarding the top 25 venue categories in each borough to observe the frequency distribution of what venue categories the borough is lacking. This will provide insight to potential new areas to recommend business start-ups looking to create new venues or add to a low category. We will utilize those that are one or less for our dataset.

Finally, we can utilize a k-means algorithm to cluster and segment our venues based on tips, photo counts, ratings, user likes and even distance. This can greatly enhance the user experience when selecting venues.