Finding Similar Cities

Shubham Previndra Sharma

1. Introduction

1.1 Background

As the number of people travelling all around the global is increasing, it would be handy to know which places, cities, neighborhoods or location are best to enjoy say arts or what is food center or where one might enjoy best hospitality. There are times when you would look for someplace similar where you enjoyed the most or perhaps something very different. This information of similarity or dissimilarity comes in handy if you are either looking for same feeling like home or a place for an adventure.

1.2 Problem

The project aims to find which of the three cities Paris, Toronto and NYC are similar to each other or rather different from others and what makes them similar or dissimilar to each other.

1.3 Interest

This would be of interest to people who are travelling for variety of reasons, someone might look for place like home where one can look for similar place, one might look for a very different place where one could select dissimilar place. It is possible to look for places that offer specific kind of venues.

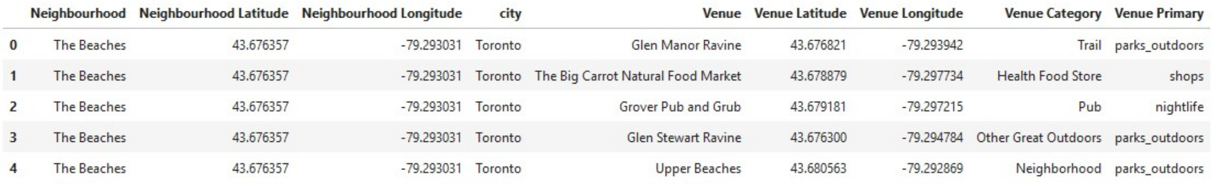
2. Data Acquisition and Preparation

2.1 Data Sources

We had data for the three cities from diverse sources. For NYC we used data from https://cocl.us/new_york_dataset. For Toronto we had postal codes data from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and coordinate data from https://cocl.us/Geospatial_data. For Paris we got neighborhood and post code data from https://en.wikipedia.org/wiki/Arrondissements_of_Paris and we got the coordinates using geocoders. For venue related data we used foursquare api.

2.2 Data Cleaning and Preparation

For having similar amounts of neighborhoods in all the cities we used "Manhattan data" for NYC and borough which had Toronto in their name for Toronto. We all took postal codes where no borough was allotted in Toronto neighborhood and used borough for neighborhood name if borough did not have a neighborhood. Then we merged all the data tables from each of the cities into a single table of 98 neighborhoods from the cities. We then discarded data related to postcode and boroughs.



	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	city	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Primary
0	The Beaches	43.676357	-79.293031	Toronto	Glen Manor Ravine	43.676821	-79.293942	Trail	parks_outdoors
1	The Beaches	43.676357	-79.293031	Toronto	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store	shops
2	The Beaches	43.676357	-79.293031	Toronto	Grover Pub and Grub	43.679181	-79.297215	Pub	nightlife
3	The Beaches	43.676357	-79.293031	Toronto	Glen Stewart Ravine	43.676300	-79.294784	Other Great Outdoors	parks_outdoors
4	The Beaches	43.676357	-79.293031	Toronto	Upper Beaches	43.680563	-79.292869	Neighborhood	parks_outdoors

Figure 1: Dataset

2.3 Feature selection

After cleaning data we scraped venue data from foursquare api and found that venue categories numbered around 280 which we reduced to 7 primary categories having similar kinds of venues under same category.

Primary Categories:

1. Arts and entertainment

2. Building

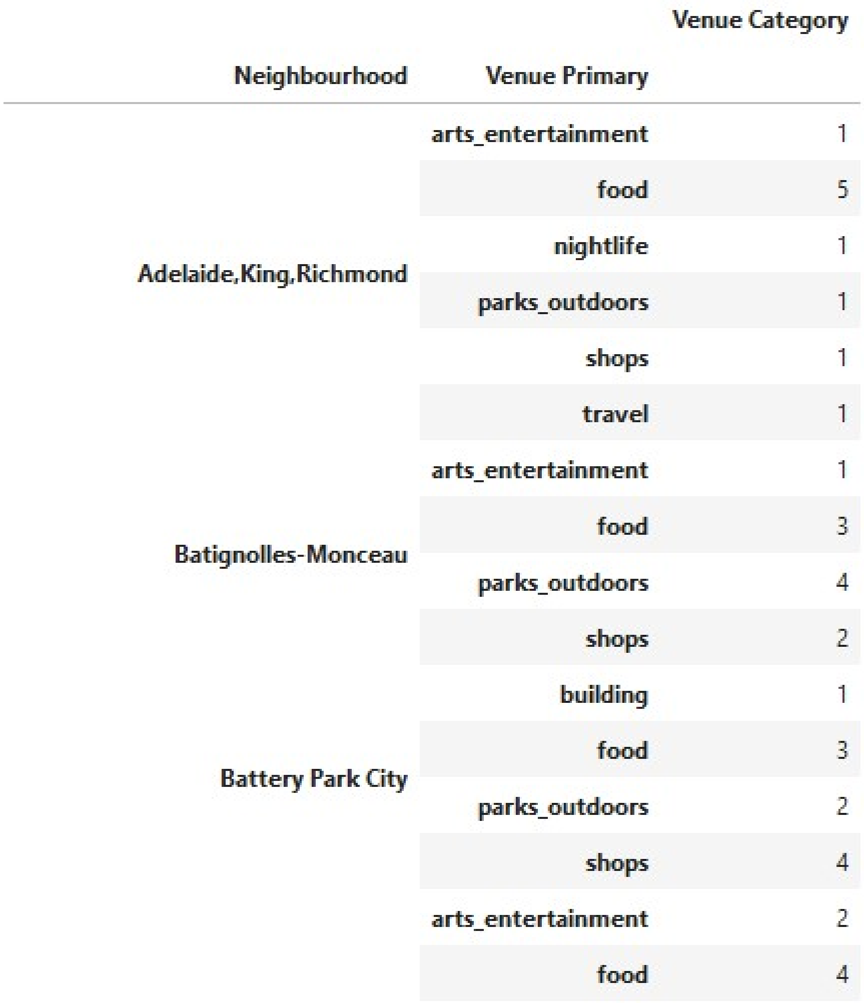
3. Food

4. Nightlife

5. Parks and outdoors

6. Shops

7. Travel



		Venue Category
Neighbourhood	Venue Primary	
Adelaide,King,Richmond	arts_entertainment	1
	food	5
	nightlife	1
	parks_outdoors	1
	shops	1
	travel	1
Batignolles-Monceau	arts_entertainment	1
	food	3
	parks_outdoors	4
	shops	2
Battery Park City	building	1
	food	3
	parks_outdoors	2
	shops	4
	arts_entertainment	2
	food	4

Figure 2: Primary Category

3. Predictive Modelling

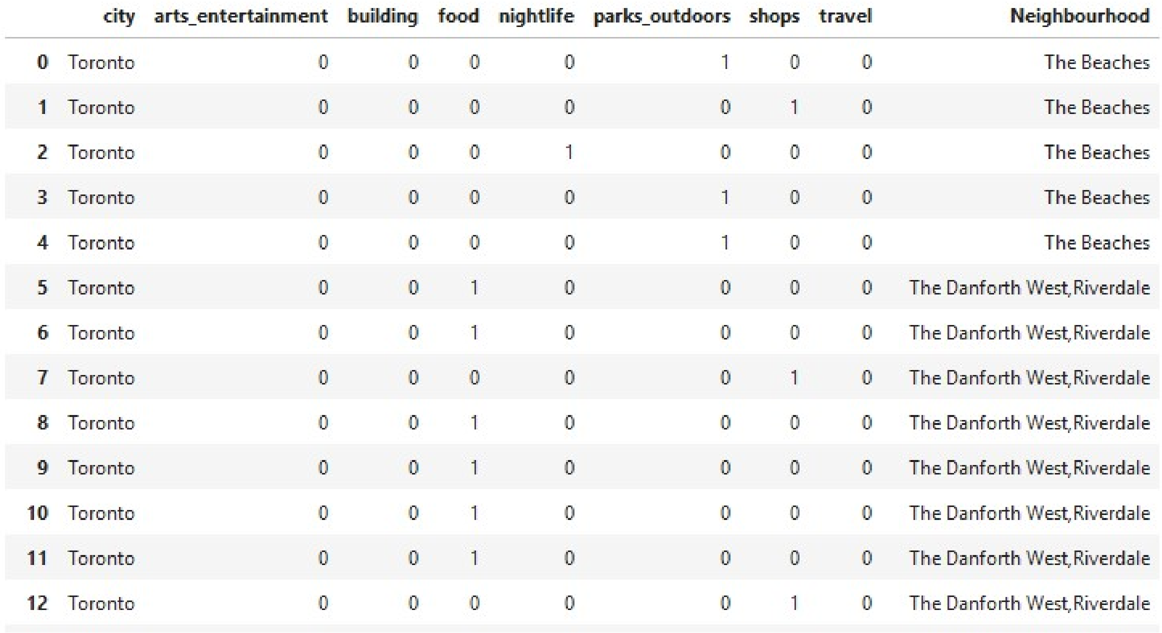
In the project clustering techniques (K-means) is used to cluster the neighborhoods into groups/clusters. The clusters hold different number of neighborhoods from the cities. Similar cities will have neighborhoods in the same cluster while dissimilar cities will have neighborhood in different clusters.

3.1 K-means

It is an unsupervised clustering technique which clusters together similar groups based on features without having prior knowledge of any labels which was suitable in this case as we had no prior information about similarity of cities. For features we classified the venues taken from foursquare api and then sorted them into the primary categories. We then one hot encoded the categories and used that to train the clustering model.

In the project the aim was to find which cities are more similar based on venues present in the neighborhood therefore we decide to go with two clusters. In one of case of two cluster it is very possible that majority of the neighborhoods of two cities will be in one cluster and neighborhood of one city in another which would be dissimilar city. In the other case there might be a condition where all neighborhoods are evenly spread across the cluster which would indicate either cities are very similar or dissimilar.

For the last case it would be prudent to check the cluster and its properties. In case if features in the clusters are not very distinct there are chances that cities are vey similar. However, if there is a clear pattern of count of features in the two clusters than that would indicate the reason why the cities are dissimilar.



	city	arts_entertainment	building	food	nightlife	parks_outdoors	shops	travel	Neighbourhood
0	Toronto	0	0	0	0	1	0	0	The Beaches
1	Toronto	0	0	0	0	0	1	0	The Beaches
2	Toronto	0	0	0	1	0	0	0	The Beaches
3	Toronto	0	0	0	0	1	0	0	The Beaches
4	Toronto	0	0	0	0	1	0	0	The Beaches
5	Toronto	0	0	1	0	0	0	0	The Danforth West,Riverdale
6	Toronto	0	0	1	0	0	0	0	The Danforth West,Riverdale
7	Toronto	0	0	0	0	0	1	0	The Danforth West,Riverdale
8	Toronto	0	0	1	0	0	0	0	The Danforth West,Riverdale
9	Toronto	0	0	1	0	0	0	0	The Danforth West,Riverdale
10	Toronto	0	0	1	0	0	0	0	The Danforth West,Riverdale
11	Toronto	0	0	1	0	0	0	0	The Danforth West,Riverdale
12	Toronto	0	0	0	0	0	1	0	The Danforth West,Riverdale

Figure 3: One hot encoded feature

4. Result

We found out that neighborhoods in cities of Toronto and NYC are mostly present in one cluster while neighborhoods of Paris are almost all in second cluster which indicates that Paris is different from the other two cities.

City	Neighborhood Count
NYC	31
Toronto	29

Table 1: Cluster 1

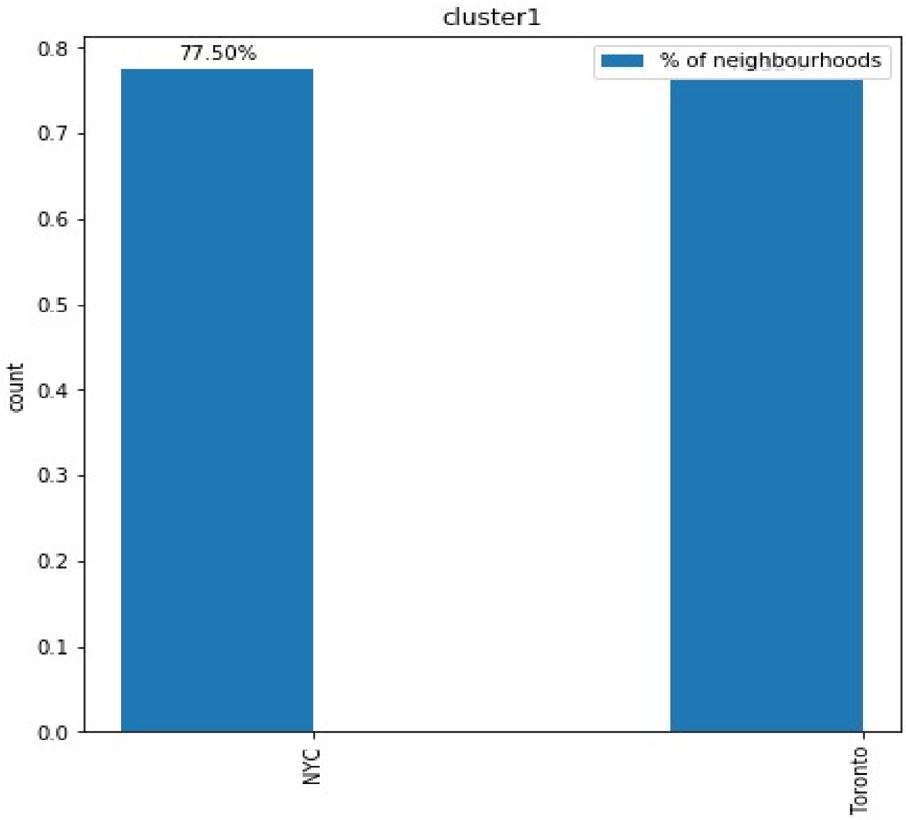


Figure 4: Cluster 1

City	Neighborhood Count
NYC	9
Paris	20
Toronto	9

Table 2: Cluster 2

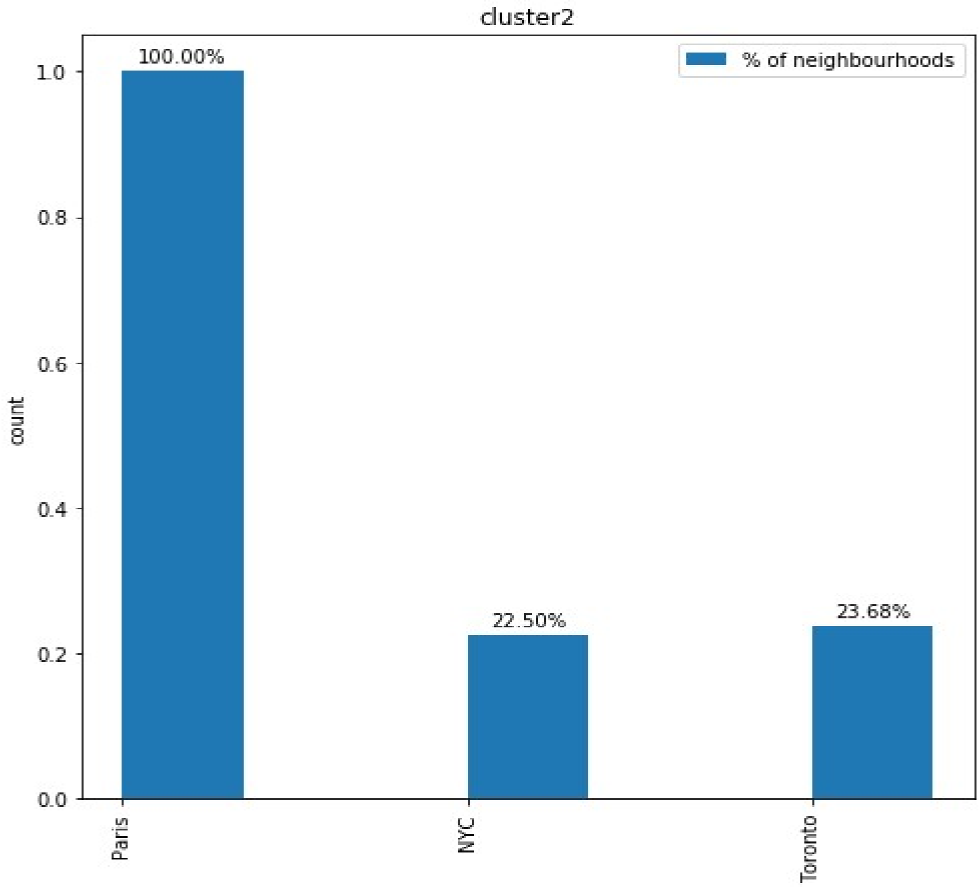
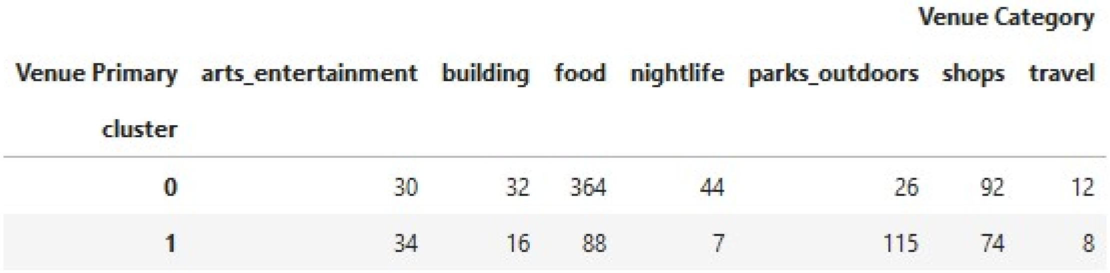


Figure 5: Cluster 2

Now that we know that Paris is clustered differently we use cluster analysis to find the reason.

One way is to see the common properties of neighborhood in the same clusters. Here we find venues that are more prelevant in the cluster.



	Venue Category						
Venue Primary	**arts_entertainment**	**building**	**food**	**nightlife**	**parks_outdoors**	**shops**	**travel**
cluster							
0	30	32	364	44	26	92	12
1	34	16	88	7	115	74	8

Figure 6: Cluster analysis

Here we see that cluster 1 which have neighborhoods from NYC and Toronto we have most of the venues related to food while cluster2 which has mostly neighborhood from Paris has venues that related to outdoors and parks followed by venues related to food.

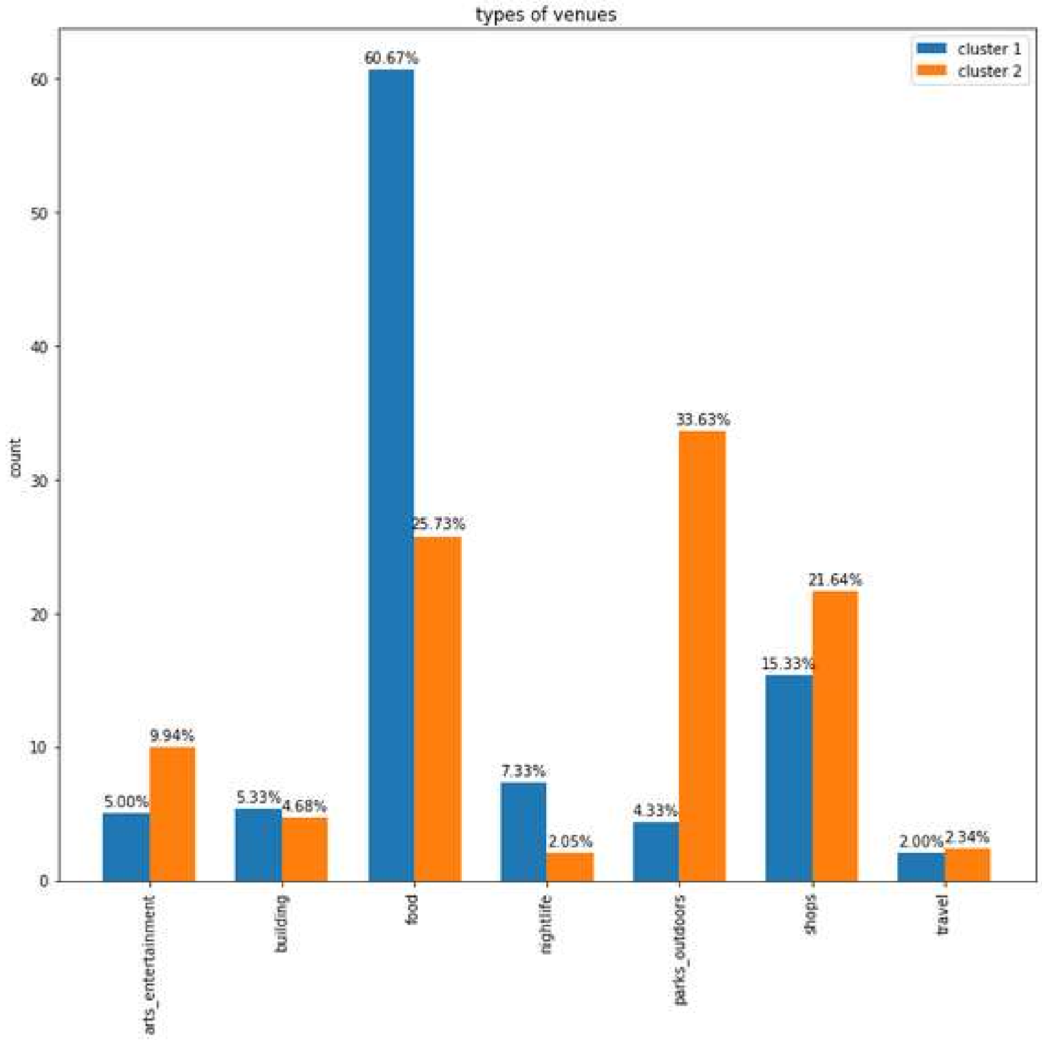


Figure 7: Types of venues in each cluster