Coursera Applied Data Science Capstone Project

Battle of the Boroughs

Report

By Malte Jörg

Table of Contents

1 Introduction

In this project I will analyze and compare the two major boroughs of the cities of Toronto and New York in terms of their venue diversity.

Toronto is Canada’s largest city and a world leader in such areas as business, finance, technology, entertainment and culture. Its large population of immigrants from all over the globe has also made Toronto one of the most multicultural cities in the world.

New York is also the largest city in its country, the U.S., and an architectural marvel with plenty of historic monuments, magnificent buildings and countless dazzling skyscrapers. The city is home to numerous museums, parks, trendy neighborhoods and shopping streets. As an addition to that New York is a financial powerhouse with the largest stock exchange in the world.

Both cities are located in North America as well as in the western parts of their country. Their description, location and history sound very similar, but the question is if that is true. Venues are the heart of a city and are can describe the livability when it comes to data. Therefore it can be used to describe differences in two cities, especially the city center, here called the major boroughs of a city. The goal of this project is to detect differences in the main centres of these two cities. The insights of this analysis can help to determine cultural differences between the population of the cities as well as give tourists or people, who want to move to these cities, further information.

2 Data Description

The data used in this project is gathered from multiple sources and is used differently to accomplish the above mentioned goal.

2.1. Location Data

The location data is the basis of the exploration of venues in the two cities. Each Borough in a city consists of multiple neighborhoods of which all should have a geographical location in form of a latitude and longitude value. There can be further information of the neighborhoods like the postal code which can also be set into relation to a geographical location in the city. For example the city of New York has the borough *Manhattan* which consist of multiple neighborhoods, postal codes, which can be set into relation to a geographical position. These positions are spread out all over the borough and are supposed to cover the whole area as good as possible.

2.2. Venue Data

Venue data is gathered via the Foursquare API. Each venue consists of a name, adress, location data, category and rating as well as possible photos. In terms of the location, the venue can be set into relation to a borough or neighborhood in its city. The relation to the data is created by exploring a specific location in the city with a defined radius where multiple venues can be assigned to. Through this relationship there can be made a direct link between the borough name and its consisting venue categories. For example in a neighborhood in 'Manhatten' with the location x/y, there is 'Bobby's Pizza' inside a 500 meter radius. 'Bobby's Pizza' is in 'Random Street 123' and categorized as a 'Italian Restraunt' with the location i/o.

2.3.Venue Category Data

Foursquare privides a list of venue categories which consist of main categories and sub-categories. This list can be split into a dataset which contains a specific relation between all sub-categories to a main category. For example an 'Italian Restraunt' and a 'Chinese Restraunt' both can be set into the main category 'Food'.

3 Methodology

3.1.Data Collection and Preprocessing

First we obtained the location data of the two cities of New York and Toronto. New Yorks location data is provided by the coursera course and therefore loaded from a .json file. The dataset consists of the features Borough, Neighborhood, Latitude and Longitude. The last two are the geographical locations of the neighborhood, which will be later used to gather surrounding venues. Toronto's Neighborhood and Borough data is scraped from a wikipedia page. The corresponding location data is also provided by coursera and loaded from a .csv file. The data of the two dataframes is matched and joined together so that the features of the dataset are similar to the ones of the New York dataset. The New York dataset consists of 306 locations, whereas the Toronto neighborhood consists of 103 locations to analyze. Locations in the Toronto dataset where dropped, because a Neighborhood name was 'Not assigned'.

The venue data was gathered via the Foursquare API. Here we explored venues around a given location with a specific radius. The radius is set to 500 meters and each location which is explored by the API is assigned to the Neighborhood and Borough of the location data in the location datasets gathered above. In a next step, multiple venues inside the same borough are dropped from the dataset to assure that no venue is considered multiple times during futhrer analysis.

By exploring the locations of the two cities, we have found 1700 venues in Toronto and 9458 venues in New York (5.56 more than Toronto). This is a huge difference in the amount of venues of the two cities, which is a first fact to consider in the analysis. Toronto has 2.6 million citizens, whereas New York has 8.4 million which is about 3 times more than Toronto (Source: wikipedia.com). Considering this comparison there are defenetly more venues listed in New York than in Toronto. Besides, this difference has no impact in our analysis, if the venue data is big enough to be representative for the city. We will have a look at this during data examination.

Venue categories in Foursquare are very specific and diverse, which makes them hard to compare and analyze. Therefore we will get a venue category list from Foursquare with all possible venue categories. The dataset is build like a tree, where the root is the main category consisting of many sub-categories. We put each of these sub-categories in a dataframe and assign their main category to it. In a next step we matched these sub-categories to each of the venue categories to add the actual main category of the venue to the venue dataset.

This conlcudes the data collection and preprocessing process of this project. We now have two venue dataset of each of the two cities and two location datasets. Next we will proceed to the data examination step.

3.2. Data Examination

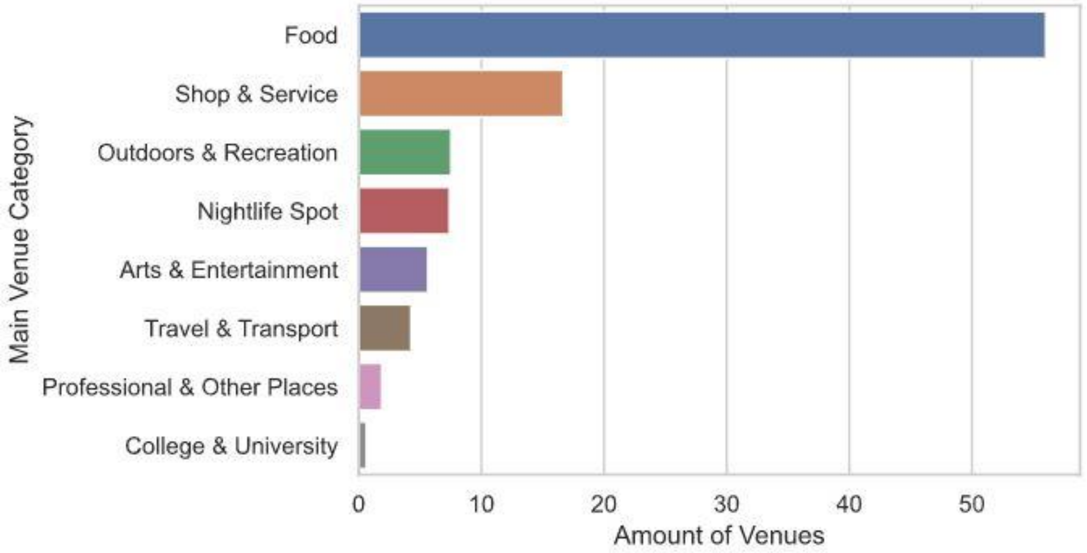
In this part of the project we examine the above gathered data to gain first insights. First of all we determine all boroughs of New York and Toronto. In a next step we check the spread of the main venue categories in the cities. Based on the amount of venues in each of the cities Boroughs we select the two top Boroughs. Both selected boroughs will be further examined to make a first analysis.

Toronto		**New York**	
Central Toronto	111	Bronx	1139
Downtown Toronto	**798**	Brooklyn	2507
East Toronto	128	**Manhattan**	**2924**
Etobicoke	66	Queens	2091
Mississauga	12	Staten Island	797
North York	19		
Scarborough	88		
West Toronto	158		
York	19		

Table 1: Boroughs of Toronto and New York and their amount of venues

As shown in table 1, there are 10 Boroughs in Toronto and 5 in New York. Each of these boroughs consists of venues. The two boroughs, on in each city, is selected by the highest amount of venues for further analysis. Manhattan and Downtown Toronto are selected, which can be seen in table 1 highlighted in bold. There is a still a significant difference in the amount of venues between these two cities. Therfore we examined the relative amount of main venue categories.

Downtown Toronto



Manhattan

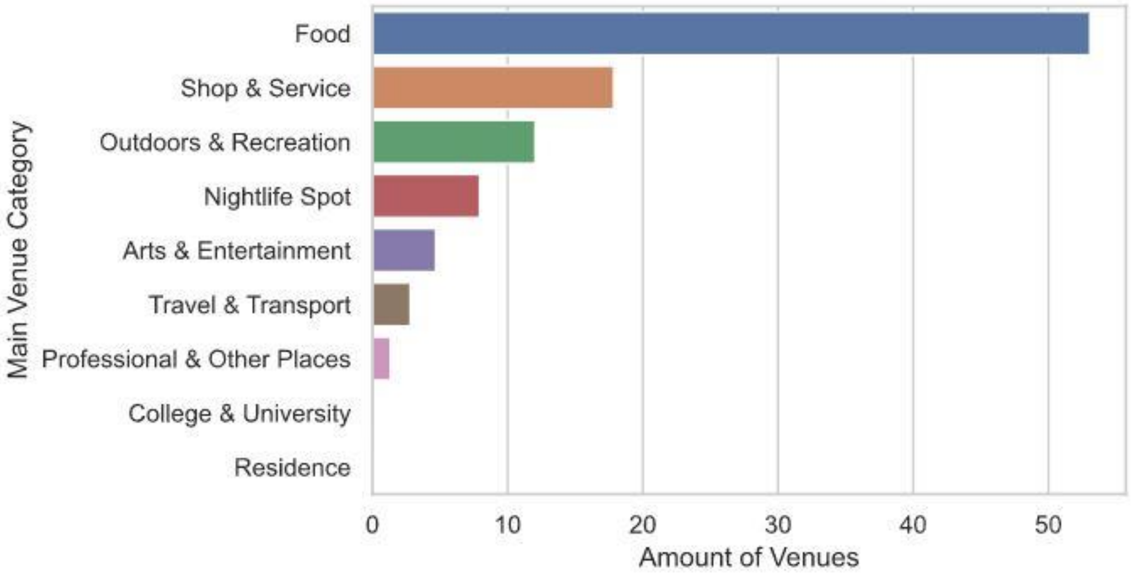


Illustration 1: Relative amount of venues in main boroughs

Illustration 1 shows the relative amount of main venue categories compared to each other. The statistics show that food is by far the major venue catorgy in both of the city centres followed by shops and recreation. It gets interesting at 3rd place which is Outdoors & Recreation. Manhattan New York has relatively more Outdoors & Recreation Venues as Downtown Toronto, which could be a sign for a much more livable and green city. Therefore Downtown Toronto has more Travel & Transport venues as Manhattan. This on the other hand could be an effect caused by the low amount of venues in the dataset. Let's assume that both cities have a decent transport system, so the density of transport venues in the area should be euqal. With less venues in Toronto sorrounding these transport venues the relative amount of transport venues in Toronto is much higher than in New York. This effect should therefore not play a big role during further analysis compared to the outdoors & recreation effect.

Venue Category	**Downtown Toronto**	**Manhattan**
1.	Coffee Shop	Coffee Shop
2.	Café	Italian Restraunt
3.	Restraunt	Café
4.	Japanese Restraunt	Pizza Place
5.	Italian Restraunt	American Restraunt
6.	Hotel	Bakery
7.	Park	Park
8.	Bakery	Bar
9.	Sushi Restraunt	Hotel
10.	Pizza Place	Gym / Fitness Center

Table 2: Top 10 venue categories

Table 2 shows the Top 10 venues in both boroughs. Coffee Shops are top in both lists followed by Cafés and different type of Restraunts. Manhatten has a lot american and italian Restraunts whereas Downtown Toronto has a more diverse spread with normal, japanese, italian and shishi restraunts. This is a hint for a more diverse culture with a large japanese, italian and canadian population. Compared to Manhatten with no asian food places in the top 10 list it only shows a large italian and american food culture.

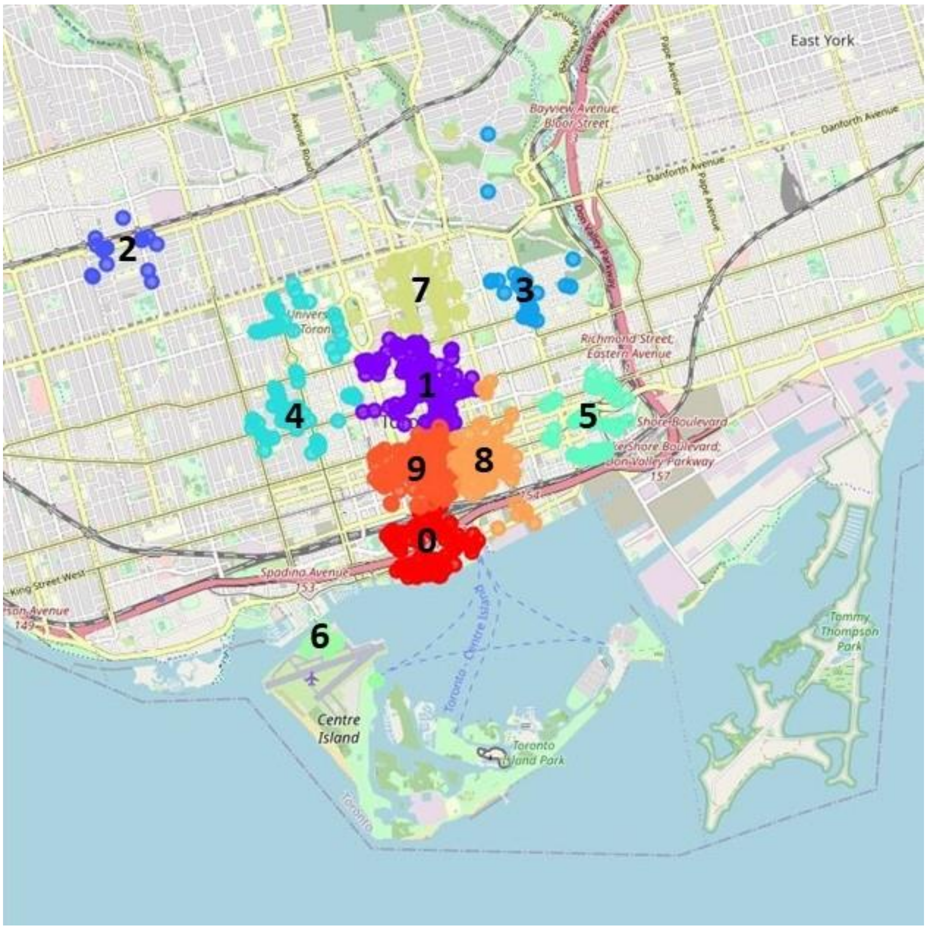
Next to cafés and restraunts, Manhattan shows a large amount of Bars and Gymns in its top 10 list. This supports the thesis of a more livable and active borough from above.

This concludes the examination of the data. Next up is the modeling process.

3.3.Modeling

In this project step we will dig deeper into New Yorks and Torontos city centres. Both boroughs consist of a too large amount of neighborhoods, Downtown Toronto 19 and Manhattan 40, to use this feature for clustering. Therefore we will cluster the borough venues into each 10 clusters by their geographical location. With this procedure we can divide the boroughs into 10 specific areas which we can analyze by themselves and compare to each other. Taking every aspect into consideration we will get a more clear understanding of the two city centres and the cities respectively.

Downtown Toronto



Manhattan

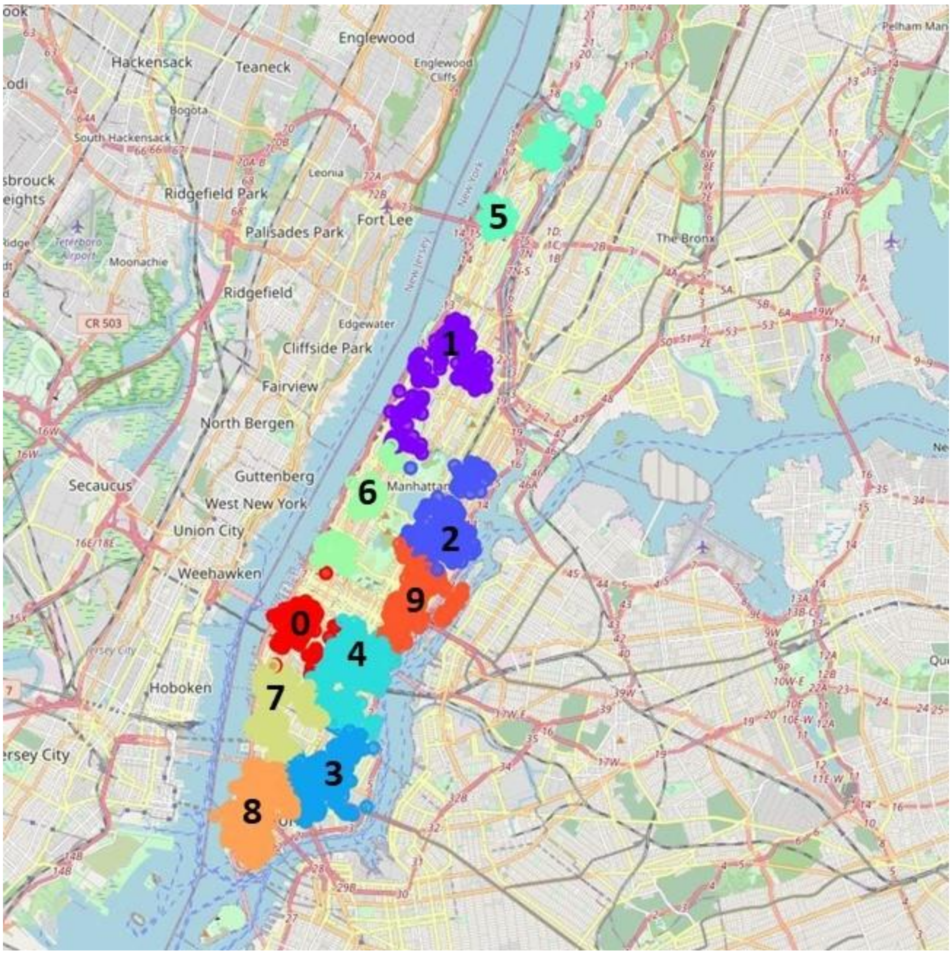


Illustration 2: Clustering of Downtown Toronto and Manhattan

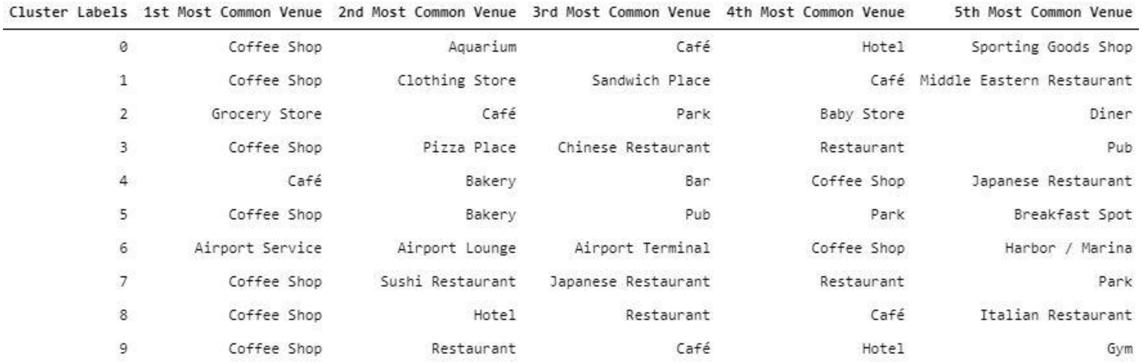
Illustration 2 shows the clustering of the two major boroughs. The clustering with the kMeans algorithm with the geographical features allowed a clear partitioning in 10 parts of each borough. These boroughs can now be analyzed by their most common venues and main venue categories. This will take place in the next part of this report, the results section.

4 Results

In the last section we clustered the venues in each of the two boroughs of the cities of Toronto and New York in 10 clusters. These represent, as seen in illustration 2, specific areas of the borough. In this section we analyze each of these clusters and compare them to each other as well to the clusters in the other city. Therefore we create a top 5 list of the venue categories of each cluster.

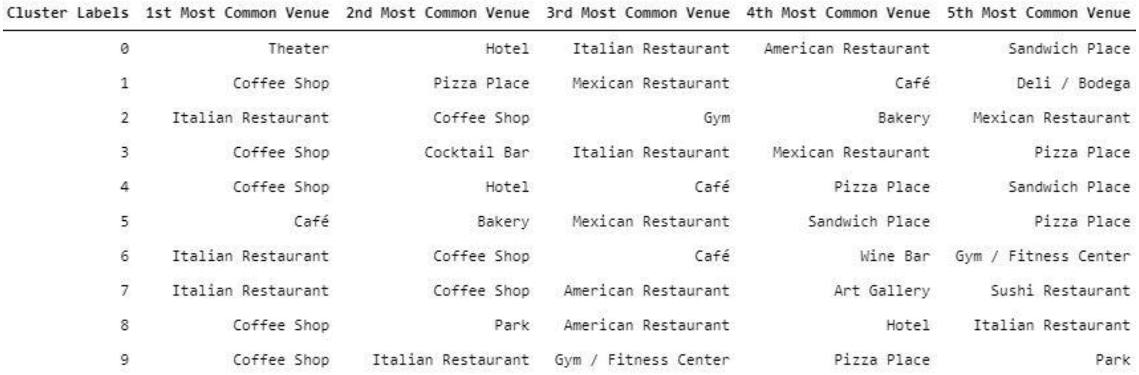
Table 3 shows the top 5 list of each cluster and a clear description of the areas in each of the two boroughs. At first sight this looks like a very common description because venues like Coffee Shops and Restraunts appear multiple times in each clusters. Here to mention though are the unique venue categories which show a clear difference in the clusters and should be a help to people who move to one of these cities.

Downtown Toronto



Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Coffee Shop	Aquarium	Café	Hotel	Sporting Goods Shop
1	Coffee Shop	Clothing Store	Sandwich Place	Café	Middle Eastern Restaurant
2	Grocery Store	Café	Park	Baby Store	Diner
3	Coffee Shop	Pizza Place	Chinese Restaurant	Restaurant	Pub
4	Café	Bakery	Bar	Coffee Shop	Japanese Restaurant
5	Coffee Shop	Bakery	Pub	Park	Breakfast Spot
6	Airport Service	Airport Lounge	Airport Terminal	Coffee Shop	Harbor / Marina
7	Coffee Shop	Sushi Restaurant	Japanese Restaurant	Restaurant	Park
8	Coffee Shop	Hotel	Restaurant	Café	Italian Restaurant
9	Coffee Shop	Restaurant	Café	Hotel	Gym

Manhattan



Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Theater	Hotel	Italian Restaurant	American Restaurant	Sandwich Place
1	Coffee Shop	Pizza Place	Mexican Restaurant	Café	Deli / Bodega
2	Italian Restaurant	Coffee Shop	Gym	Bakery	Mexican Restaurant
3	Coffee Shop	Cocktail Bar	Italian Restaurant	Mexican Restaurant	Pizza Place
4	Coffee Shop	Hotel	Café	Pizza Place	Sandwich Place
5	Café	Bakery	Mexican Restaurant	Sandwich Place	Pizza Place
6	Italian Restaurant	Coffee Shop	Café	Wine Bar	Gym / Fitness Center
7	Italian Restaurant	Coffee Shop	American Restaurant	Art Gallery	Sushi Restaurant
8	Coffee Shop	Park	American Restaurant	Hotel	Italian Restaurant
9	Coffee Shop	Italian Restaurant	Gym / Fitness Center	Pizza Place	Park

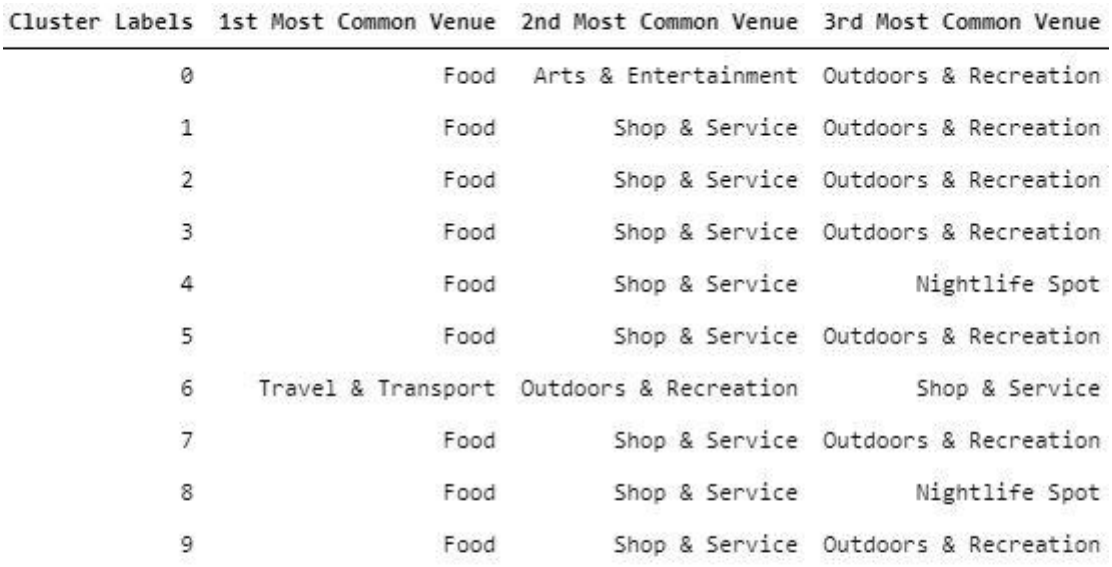
Table 3: Top 5 common venue in clusters

In Toronto for example there is an Airport cluster (Cluster 6) with outliers in the dataset, which can be clearly seen in the maps above. Cluster 2 on the other hand seems be very good for young families to go to because there are a lot of grovery and baby stores in the area. Cluster 7 on the other hand sounds like an asian neighborhood because of the specific restraunt types.

In Manhatten Cluster 0 is a cultural centre with a lot theatres and adjacent hotels. Cluster 2 and 9 are more for sports whereas cluser 9 is a more livable area due to its parks. Taking parks as the main objective, cluster 8 is the place to go to. If there is a need for diversity in food places, cluster 5 is a unique area compared to other clusters of Manhattan.

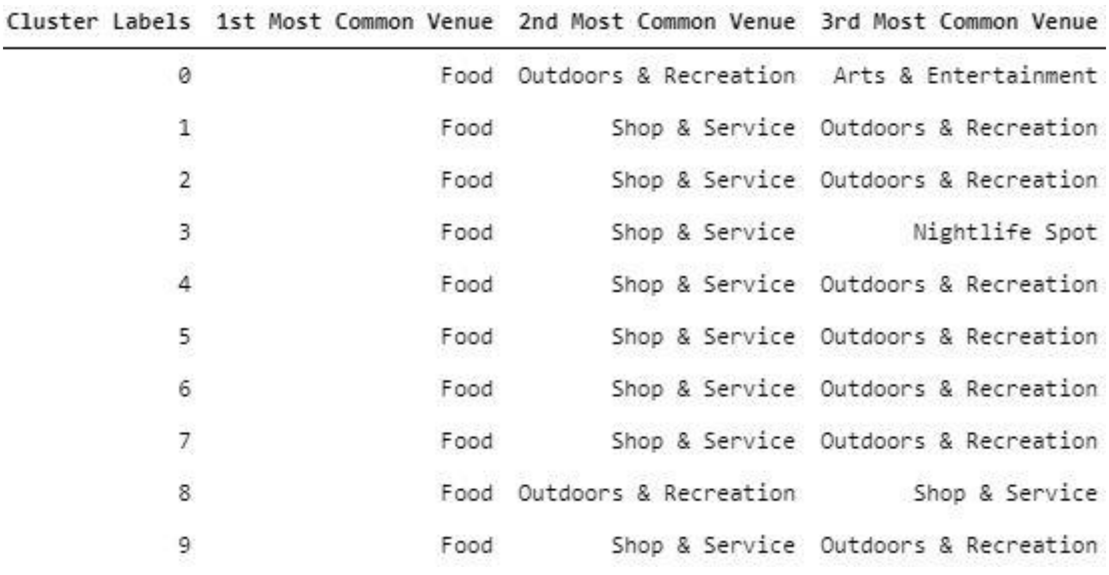
To get a more precise view of the areas, we can analyze each cluster by their main venue category in two different appraches next.

Downtown Toronto



Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Food	Arts & Entertainment	Outdoors & Recreation
1	Food	Shop & Service	Outdoors & Recreation
2	Food	Shop & Service	Outdoors & Recreation
3	Food	Shop & Service	Outdoors & Recreation
4	Food	Shop & Service	Nightlife Spot
5	Food	Shop & Service	Outdoors & Recreation
6	Travel & Transport	Outdoors & Recreation	Shop & Service
7	Food	Shop & Service	Outdoors & Recreation
8	Food	Shop & Service	Nightlife Spot
9	Food	Shop & Service	Outdoors & Recreation

Manhattan



Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Food	Outdoors & Recreation	Arts & Entertainment
1	Food	Shop & Service	Outdoors & Recreation
2	Food	Shop & Service	Outdoors & Recreation
3	Food	Shop & Service	Nightlife Spot
4	Food	Shop & Service	Outdoors & Recreation
5	Food	Shop & Service	Outdoors & Recreation
6	Food	Shop & Service	Outdoors & Recreation
7	Food	Shop & Service	Outdoors & Recreation
8	Food	Outdoors & Recreation	Shop & Service
9	Food	Shop & Service	Outdoors & Recreation

Table 4: Top 3 common main venue categories in cluster

The two top 3 lists of the main venue categories (Table 4) support the insights gathered above. New for both cities is the category nightlife spot. Downtown Toronto shows here two potential areas with cluster 4 and 8 and Manhattan New York one with cluster 3. To get a different view of the clusters we will approach the analysis from the category perspective and determine which cluster is the most common for each category.

Main Venue Category	Downtown Toronto	Manhattan
Arts & Entertainment	0	0
College & University	4	6
Food	9	1
Nightlife Spot	7	0
Outdoors & Recreation	2	8
Professional & Other Places	5	8
Shop & Services	2	5
Travel & Transport	6	0
Residence	-	9

Table 5: Top clusters for main venue categories

This list, as shown in table 5, is most usable for Tourists and people who want to move to one of these cities. It clearly shows the hotspots for each category. It gets the most interesing when you add up clusters which appear multiple times in the list.

In Manhattan Cluster 0 is a hotspot for Arts & Entertainment, Nighlife and Travel & Transport, which makes it a very touristy area. Cluster 8 on the other hand is a hotspot for Outdoors & Recreation as well as Professional & Other Places. This could be a sign for an area where you will meet locals more likely, because the area appears to be very livable and diverse.

In Downtown Toronto only cluster 2 stands out with Outdoors & Recreation and Shop & Service as its main category. This could also be a sign for an area with locals as cluster 8 in Manhattan. A sign for the more diverse main cluster categorization could be the much smaller area of the borough we area observing.

This concludes the presentation of the results of this project. Next we will do a full analysis of the gathered insights with a discussion followed by a conclusion.

5 Discussion

The Foursquare data of New york and Toronto and especially the two city centres which were selected for this analysis gives a good first look about the cities culture. Food is the main thing which you can find tons in both of the areas. The specific type of foods shows a more diverse culture in Toronto than in New York which is surprising. New Yorks famous chinatown cannot be seen in the data in contrast to the japanese and sushi area ín Downtown Toronto. Further more there is a vast amount of Italian restraunts and Pizza places in New York. Coffee Shops, a major category in both cities, show that these centres are a working place powerhouse. This was already shown in the buisness understanding section, where both cities are described as their countries financial centres. Sadly there is no other sign than Food and Coffee for this description. This is because Foursquare provides no or less data of regular buisnesses.

Manhatten appears to be a more livable city. Parks, gyms and other outdoor activities appear more likely in different areas of the borough. Downtown Toronto is not a place for living rather than to work, nightlife and tourist activities. Still, Manhattan the larger area if you compare the size of the two boroughs. This is a fact which should not be taken out of consideration. Manhatten provides a much larger space for people to live and therefore has a higher demand of venues for local people. Downtown Toronto could on the other hand attract more buisnesses and people who travel into the city. This shows the relative high amount of Travel & Transport in Downtown Toronto compared to Manhatten.

6 Conclusion

Taking everything into consideration Manhatten New York and Downtown Toronto seem to be similar cities when it comes to their venunes. Both boroughs are flooded with food and coffee places as well as venues for nightlife, culture, sport and parks. The analysis showed partly differences in certain areas of the boroughs and the city itself. It should give Tourists or people who want to move to one of those cities a first insight on what to expect in certain areas of these cities. The clustering of the city venues showed a more precise partitioning of the area. The analysis of the clusters in terms of venue categories and main categories showed differences in different part of the cities. Because of the vast amount of food and coffee places this analysis wasn't always easy. The interpretation on the data relied mostly on unique categories in on of the clusters and led to the conclusions made in the analysis above. Because of that, statements about cultural differences between those cities were kept low. The canadian and american culture, especially in english speaking parts of Canada are very similar. Differences could be seen in migrating cultures like italian and asian cultures. The statements are based on the amount of these specific restraunts. Surprisingly an asian area could not been found in New York, which was highly expected due to well known neighborhoods. This fact leads to a point which could be done in an analysis following this project. The question could be how the data shows specific cultural venues and if they could be clustered in specific parts of the city. Maybe there are very diverse areas of the city with a large diversity of venues related to countries or cultures. This will defenetly give more insights about the culture in these two cities as this analysis could provide.