

FINAL REPORT

Introduction : business problem

The decision about placement or destination is pivotal in many fields such as, among others, tourism, real estate sector and service business as well.

An interpretative model for consumers, citizens and business decision makers it's therefore the tool for supporting the intake decision process based on a rationale approach, avoiding personal or irrational decision.

Description

In this project we try to compare to different but at same time similar cities such as Toronto and New York focusing especially on the perspective of tourism. The comparison is made by the service resources the two towns have rooted in the territories (accommodation and hospitality, restaurants, entertainments, etc.)

Toronto and New York are the famous places in the world. They are diverse in many ways. Both are multicultural as well as the financial hubs of their respective countries. The project want to explore how much they are similar or dissimilar in aspects from a tourist point of view regarding food, accommodation, beautiful places, and many more.

Target audience and people who care about

Information in the Tourism sector is very important as can drive the selection of people among many destinations. The capabilities of central or local administration to give citizen or tourist a dashboard of synthetic but at the same time exhaustive information to select the correct destination based on their needs is a strong weapon to utilize in a website or portal for information or advertisement for tourist.

Therefore the target audience for this project could be i.e. a "tourist office" or a public administration who wants give specific information to their public. The project that has been created as a pilot model for comparison of two or more destination can be subsequently implemented in an automated application for website or mobile app.

This prototype can also be extended for other different targets. Let's imagine this model applied for the comparison of two cities (or neighborhoods in towns) in the sector of real estate. In this case the

information in the model would also consist of the houses (qualities and quantities) and the price per square meter. In this case the people who would care about can be the real estate companies interested to give more information to their clients and the final target of the “app” could be the people interested to get information about a new house.

Data & Sources

This project works with two sets of data. The first dataset consists of New York’s different neighborhoods and their respective geometric coordinates. The second dataset consists of Toronto’s different borough and their respective postcodes.

Per each of the two destinations the project requires the gathering of the information about the resources existing in the territory.

Foursquare API provides with an access to an enormous database consisting of venues from all around the world including rich variety of information such as addresses, tips, photos and comments. Having signed up for a Foursquare developer, using the Client ID and Client Secret, it is possible to make API requests in order in order to retrieve venue information.

Therefore per each destination and especially for two selected neighbourhoods chosen as case study(Toronto downtown and Manhattan, we have extracted by Foursquare API, information about different venues (Restaurants, Coffee shops, etc).

Cleaning

The raw data obtained by investigating the aforementioned sources must be cleaned and transformed to be analysed and submitted to a model of interpretation.

Data sources are in different format: CSV, xls, json and also data scraped from the web such as Wikipedia page that contains Postcode of the city of Toronto in a wikitable. The dataframe obtained consists of: PostalCode, Borough and Neighborhood.

After dropping lines where the column ‘Borough’ are not assigned the project has required the association of each cardinal info and therefore longitudinal and latitudinal data.

By cardinal info the project has moved to obtain the localization data about the territories. Obtained the different neighborhoods and their respective geometric coordinates for the city of New York and Toronto, we have come up with different venues that the different destinations have to offer.

After performing One-Hot-Encoding and grouping together the rows by neighborhoods, the NY dataset and Toronto dataset seemed to share 250 features. Both the dataframes were combined and compared to understand similarities and differences and therefore to gain knowledge about characteristics and specificity of the two destinations.

Here we show a case of what the data consists of and what features can be extracted from them.

The combination of data is synergic as, for instance, we compare data from a public website Wikipedia page that contains Postcode of Toronto like this

	Postcode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront, Regent Park
3	M6A	North York	Lawrence Heights, Lawrence Manor
4	M7A	Queen's Park	Queen's Park

With a source of rich qualitative and quantitative information like foursquare as done in this example.

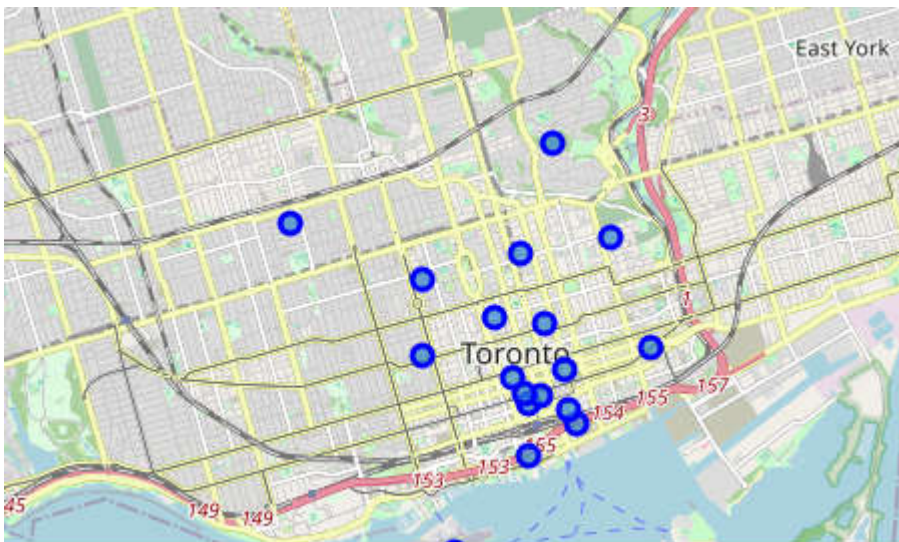
	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue
0	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636	0	Coffee Shop
1	Downtown Toronto	Ryerson, Garden District	43.657162	-79.378937	0	Café
2	Downtown Toronto	St. James Town	43.651494	-79.375418	4	Gastropub
3	Downtown Toronto	Berczy Park	43.644771	-79.373306	2	Seafood Restaurant
4	Downtown Toronto	Central Bay Street	43.657952	-79.387383	1	Coffee Shop

In this extract of the analysis is shown that per the Downtown Toronto (first column) and for each Neighborhood belonging to (second column) we can understand and appreciate the first and more important (in terms of frequency) venue existing in the territory (last column).

The combination of these data can be done by an intermediate step that is the link of the location info (long and latitude) per each destination in this case Neighborhood like this

	Borough	Neighborhood	Latitude	Longitude
0	Downtown Toronto	Harbourfront,Regent Park	43.654260	-79.360636
1	Downtown Toronto	Ryerson,Garden District	43.657162	-79.378937
2	Downtown Toronto	St. James Town	43.651494	-79.375418
3	Downtown Toronto	Berczy Park	43.644771	-79.373306
4	Downtown Toronto	Central Bay Street	43.657952	-79.387383

Finally we can also plot on a geographical map the issue we are studying like done in the example here reported.



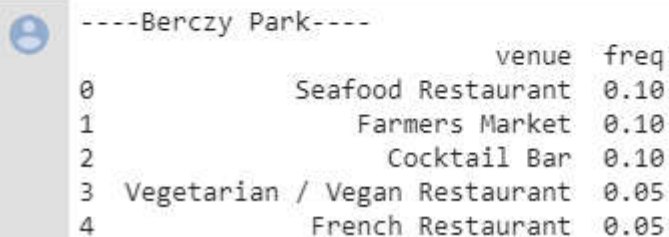
Methodology section

For this problem, we will get the services of Foursquare API to explore the data of two cities, in terms of their neighborhoods. The data also include the information about the places around each neighborhood like restaurants, hotels, coffee shops, parks, theaters, art galleries, museums and many more. We selected one Borough from each city to analyze their neighborhoods. Manhattan from New York and Downtown Toronto from Toronto. We will use machine learning technique, “Clustering” to segment the neighborhoods with similar objects on the basis of each neighborhood data. These objects will be given priority on the basis of foot traffic (activity) in their respective neighborhoods. This will help to locate the tourist’s areas and hubs, and then we can judge the similarity or dissimilarity between two cities on that basis.

We visualize the data by mapping in order to get a knowledge of the object investigated.

We analyse both boroughs neighborhoods through one hot encoding (giving ‘1’ if a venue category is there, and ‘0’ in case of venue category is not there). On the basis of one hot encoding, we calculate mean of the frequency of occurrence of each category and picked top ten venues on that basis for each neighborhood. It means the top venues are showing the foot traffic or the more visited places like in the following extract as example.

```
for hood in downtown_toronto_grouped['Neighborhood']:
    print("----"+hood+"----")
    temp = downtown_toronto_grouped[downtown_toronto_grou
    temp.columns = ['venue', 'freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset
    print('\n')
```



	venue	freq
0	Seafood Restaurant	0.10
1	Farmers Market	0.10
2	Cocktail Bar	0.10
3	Vegetarian / Vegan Restaurant	0.05
4	French Restaurant	0.05

Then we create a dataframe, as shown in the next picture, on which we run the unsupervised machine learning method to create clusters of Neighborhood based on the major frequency of venue.

	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	Adelaide,King,Richmond	Steakhouse	Asian Restaurant	Coffee Shop	Bar	Concert Hall	Café	Seafood Restaurant	Pizza Place	Speakeasy	Hotel
1	Berczy Park	Seafood Restaurant	Farmers Market	Cocktail Bar	Vegetarian / Vegan Restaurant	Italian Restaurant	Concert Hall	Liquor Store	Museum	Breakfast Spot	Belgian Restaurant
2	CN Tower,Bathurst Quay,Island airport,Harbourf...	Airport Service	Airport Lounge	Airport Terminal	Plane	Airport	Airport Food Court	Airport Gate	Boutique	Sculpture Garden	Boat or Ferry

Clustering neighborhood as said has been done by the frequencies of types of venue in the territories.

We decided to run the methodologies with the selection of 5 clusters, as done in the learning environment and equally both for Downtown Toronto and Mahanattan.

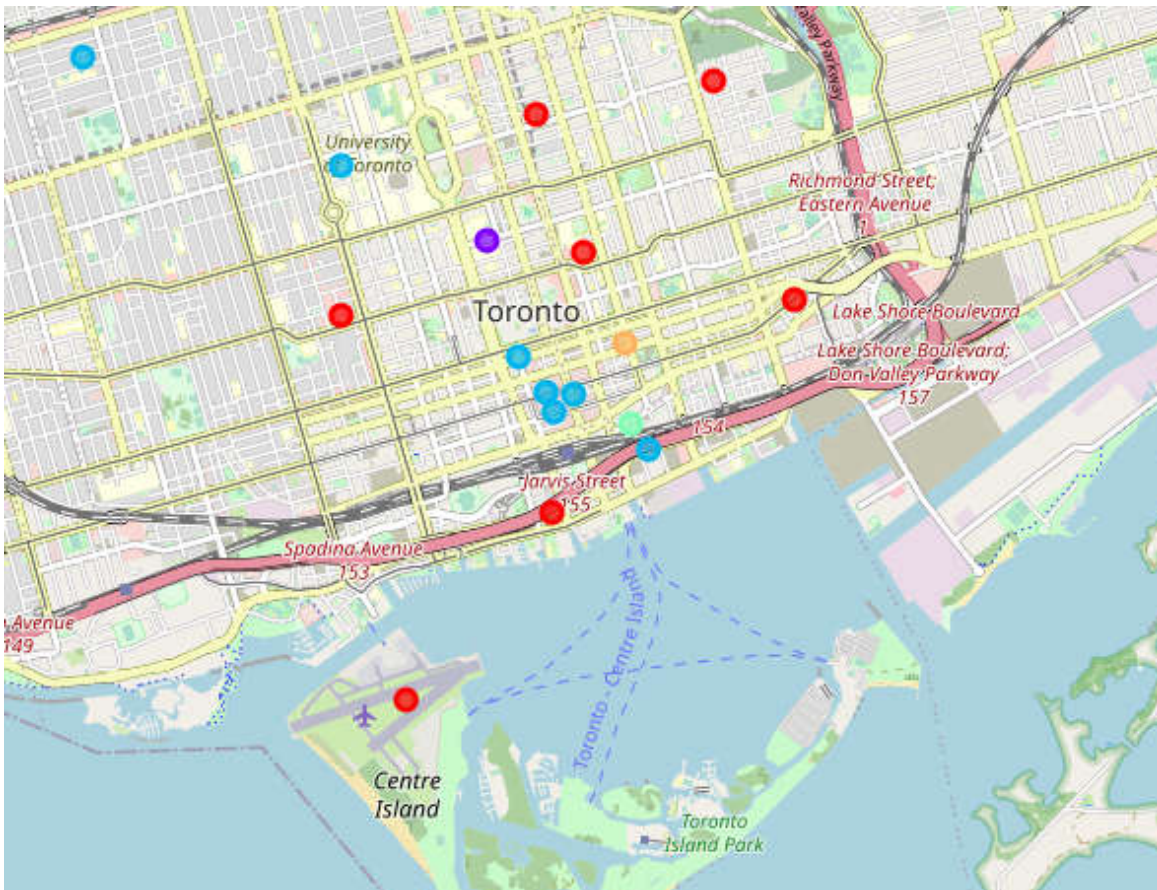
In the following it is shown the extract of the output of the function launched on the Toronto dataframe.

```
# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
downtown_toronto_merged = downtown_toronto_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

downtown_toronto_merged.head() # check the last columns!
```

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	Downtown Toronto	Harbourfront,Regent Park	43.654260	-79.360636	0	Coffee Shop	Park	Breakfast Spot	Spa	Performing Arts Venue	Chocolate Shop	Pub	Mexican Restaurant	Restaurant	E
1	Downtown Toronto	Ryerson,Garden District	43.657162	-79.378937	0	Café	Steakhouse	Burrito Place	Burger Joint	Movie Theater	Pizza Place	Beer Bar	Plaza	Clothing Store	R
2	Downtown Toronto	St. James Town	43.651494	-79.375418	4	Gastropub	Coffee Shop	Restaurant	BBQ Joint	Food Truck	Gym	Hotel	Italian Restaurant	Japanese Restaurant	Cn
3	Downtown Toronto	Berczy Park	43.644771	-79.373306	2	Seafood Restaurant	Farmers Market	Cocktail Bar	Vegetarian / Vegan Restaurant	Italian Restaurant	Concert Hall	Liquor Store	Museum	Breakfast Spot	B
4	Downtown Toronto	Central Bay Street	43.657952	-79.387383	1	Coffee Shop	Bubble Tea Shop	Art Museum	Chinese Restaurant	Modern European Restaurant	Gastropub	Pizza Place	Ramen Restaurant	Bar	Jap

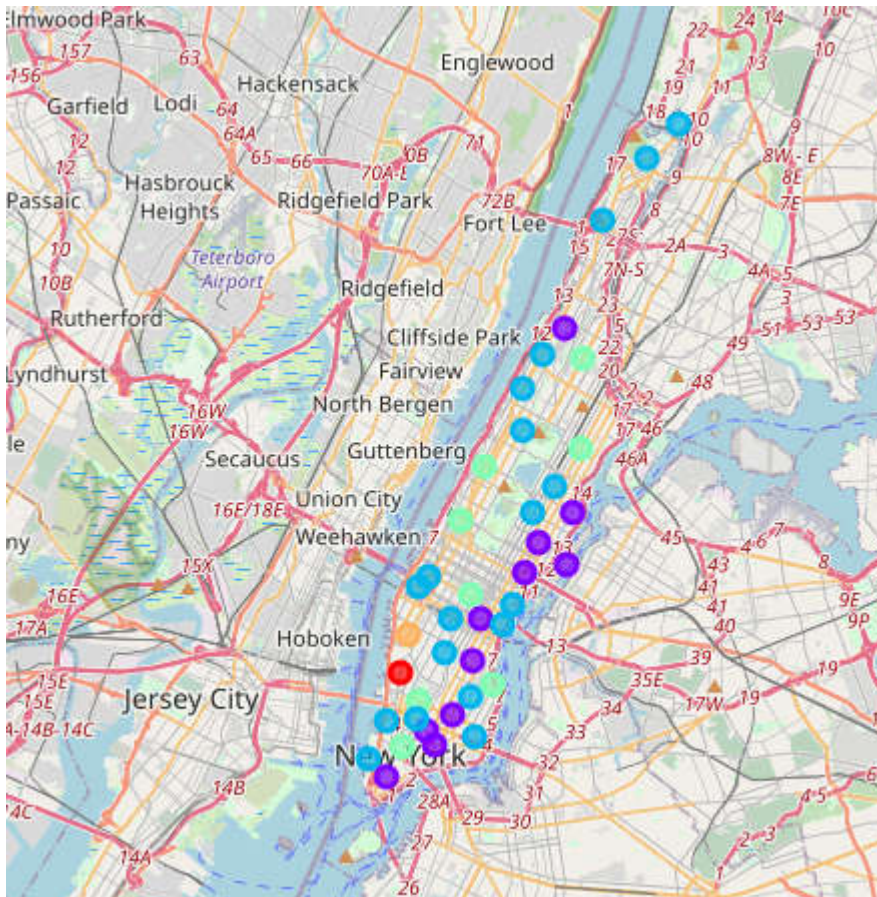
In the next picture the 18 Downtown Toronto Clusters are outlined by 5 colours as the number of clusters.



The interpretation of the cluster is given below.

- Cluster 1 (●) (Airport Lounge, Coffee Shop, Cafe, Restaurants & Grocery Store) 7 neighbourhoods
- Cluster 2 (●) A (Gastropubs) 2 neighbourhoods
- Cluster 3 (●) (Cafè) 7 neighbourhoods
- Cluster 4 (●) (Coffee Shop, Cafe, Park & Japanese Restaurant) 1 neighbourhoods
- Cluster 5 (●) (Seafood, steakhouse, Hotel & Cafe) 1 neighbourhoods

Before giving a qualitative interpretation of the clusters as basic activity to perform the comparison between the two towns we decide to depict the final map of clusters of Manhattan as final step of the descriptive analysis run completely in the Jupiter notebook.



The final map of clusters of Manhattan are here depicted in a qualitative approach.

- Cluster 1 (●) (Residential) 1 orange neighbourhoods
- Cluster 2 (●) (Commercial places) 11 blu neighbourhoods
- Cluster 3 (●) (Tourist Areas & Hubs) 19 neighbourhoods
- Cluster 4 (●) (Center Activity) 8 neighbourhoods
- Cluster 5 (●) (Cultural & Going Out Places) 1 neighbourhoods

Results section and discussion

In this project we aimed to find a model enabling the comparison of two destinations in a way to understand better decision about living, tourism etc. We have taken the city of Toronto and New York as cases to develop and test the idea behind the project. Both the towns are major

cities of the respective countries and are as well destination for people visiting for tourism and for living (work especially) coming from outside. To be more focused in first initial pilot phase we have selected two boroughs per each city : Downtown Toronto for Toronto and Manhattan for New York. This decision is based on the fact that these two boroughs are the “heart” of the two cities.

Therefore they have in common many things such as tourism, business and also center for art and entertainment.

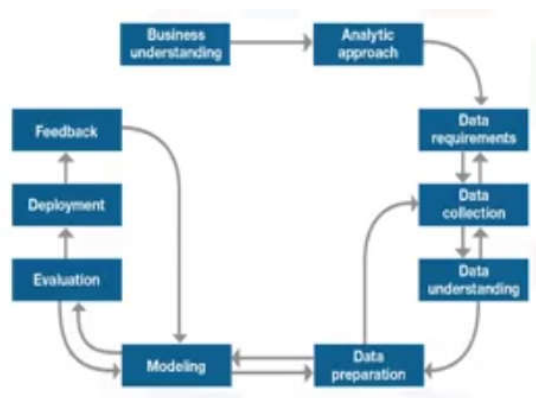
Our analysis has shown these commonalities as the most important clusters for Manhattan are the “tourist areas” (19 neighbourhoods) and the “commercial places” (11 neighbourhoods) in a similar way Downtown Toronto has many neighbourhoods devoted to tourism (7 from cluster 1 and other 7 in cluster 3).

On the other hand the analysis shows that Manhattan is more devoted to “center for activities” such as entertainment, gym, spa and recreational or also for accomodations (hotels and so on). While the prevalent venues in Downtown Toronto are related to eat&drink services such as caffè, restaurants, bakery etc.

Discussion section: observations and recommendations.

This pilot project has given the opportunity to complete a process of analysis based on data search, structured analysis (ML) and descriptive statistical methodologies.

The development of the work done has shown the importance to follow a structured approach during the steps of research analysis. Relying in the The CRISP-DM model here reproduced is pivotal in order



to have more confidence in the approach and the robustness of the results obtained.

In this respect an observation or recommendation emerging from the work done is related to the decision about cluster.

We have decided to define 5 clusters per each of the two boroughs to standardize the approach but probably it would be better to define the number of the clusters based on the situation investigated.

Conclusion section where you conclude the report.

Although the project has been structured in a pilot approach and therefore it requires a lot of improvement both in the precision of the analysis and the deployment of the applications, the outcome shows that the study to standardize and consequently automate the comparison of consistency of two destination can be helpful for decision maker and for final users such as consumers, tourists, citizens generally speaking.

The level of the automation of the rules included in the Jupiter notebook leave the possibility to think that this work can also be rooted in a structure of a website or a mobile app in order to give the chance to spread the diffusion and application on a large basis of the tool.