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.

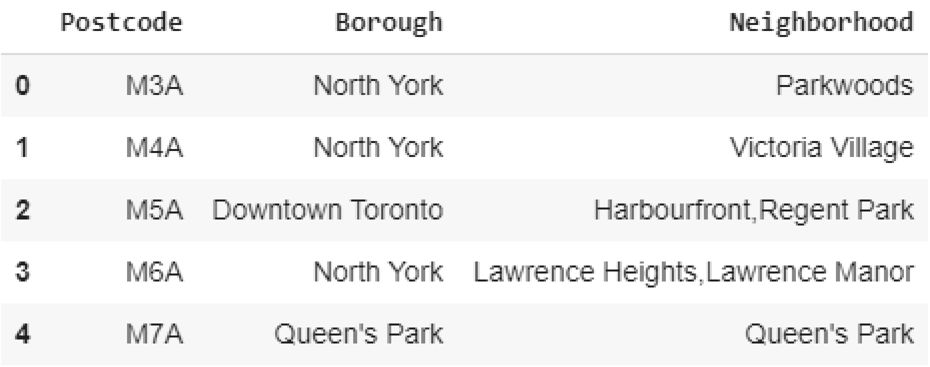
After dropping lines where the column ‘Borough’ are not assigned the project has required the association of each cardinal info and therefore longitudinal and latitudinal data.

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.

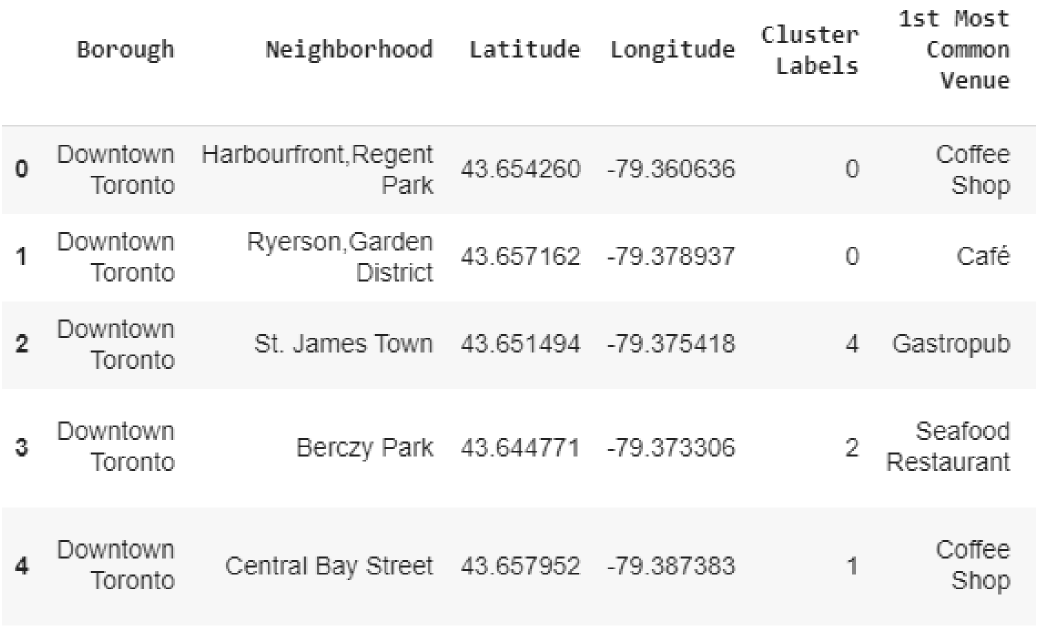
Here we show a case of what the data consists of and what features can be extracted from them.

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

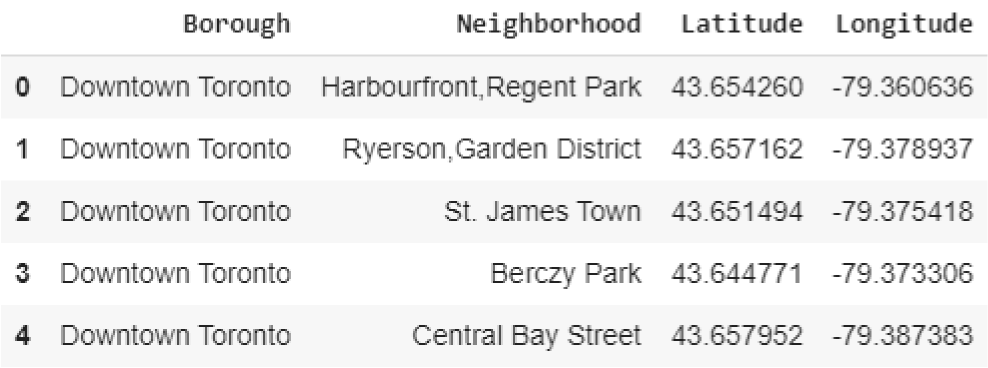
With a source of rich qualitative and quantitative information like foursquare as done in this example.



	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

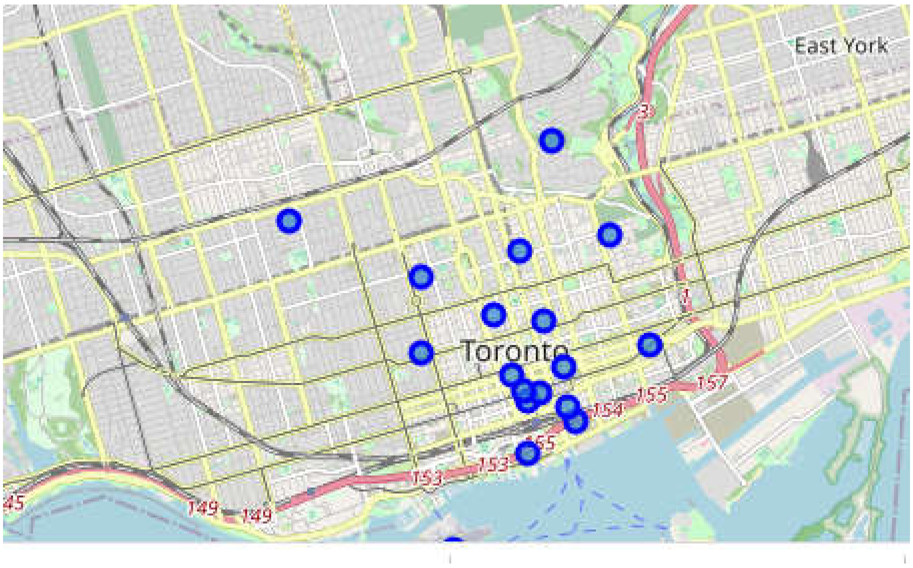
In this extract of the analysis is shown that per the Downtown Toronto (first column) and for each Neighborhood belonging to (second column) we can understand and appreciate the first and more important (in terms of frequency) venue existing in the territory (last column).

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.


East York
Toronto
145
149
149
153
153
155
154
155
157