**Introduction: Business Problem**

Toronto is the most important and populous city in Canada. It is very diverse and serves as the financial capital of Canada. The city is a center for trade, finance, governmental services, finance, arts, real estate and media in the Canada. The owners of a very popular restaurant chain in the United States want to expand their chain to Canada and would like to open some delis in Toronto to begin their Canadian operations.

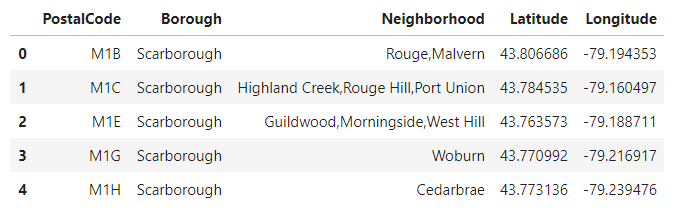
Toronto is the fourth largest in North America with more than 2.9 million inhabitants and a metropolitan area population of 5.93 million. It’s made up of more than a hundred neighborhoods with different characteristics. Marketing consultants has determined that we must pay special attention to entertainment services, because they are considered attractive to potential customers.

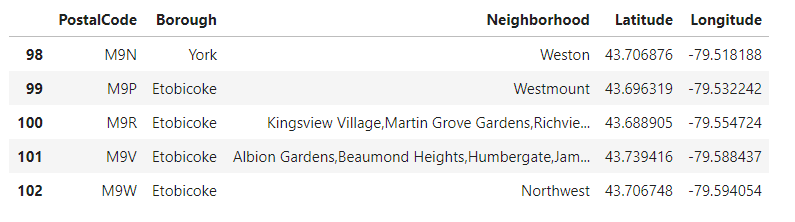
The project will consist of obtaining information of the neighborhoods and make recommendations about best places to start the deli/bodegas chain expansion.

**Data**

Required data will be sourced from:

* Toronto information, including boroughs, postal codes and neighborhoods, can be obtained from Toronto (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>)
* In order to make recommendations about where to situate the delis, Toronto segmentation will be made based on different neighborhoods.
* Full list of neighborhoods can be obtained from Wikipedia Toronto information, but only their names will be used. They must be geolocated in order to use Foursquare services for obtaining venues.
* For geolocation of neighborhoods Python geocoder will be used.
* Every neighborhood obtained will be geolocated using python geocoder package, using neighborhood name plus city and country.
* Geocoder returns longitude and latitude information for every neighborhood center, then it will be used as main Foursquare input.
* In order to obtain venues and their categories FOURSQUARE (<https://foursquare.com/>) will be used.
* Using services provided by Foursquare one can obtain venues for every neighborhood. Such services require as input geo-localization, it means the latitude and longitude obtained in previously described step.

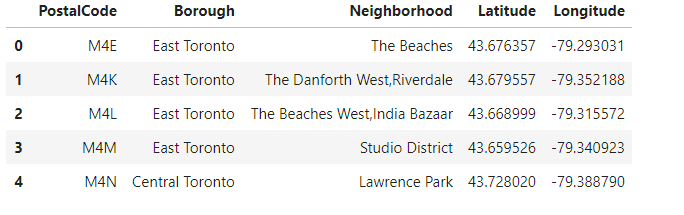


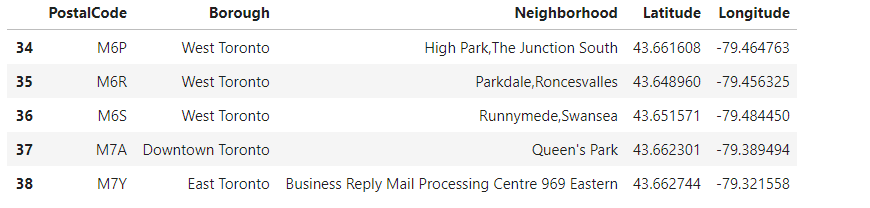


**Methodology**

A Jupyter Notebook will be developed in order to process data and segment the neighborhoods. Following steps will be implemented:

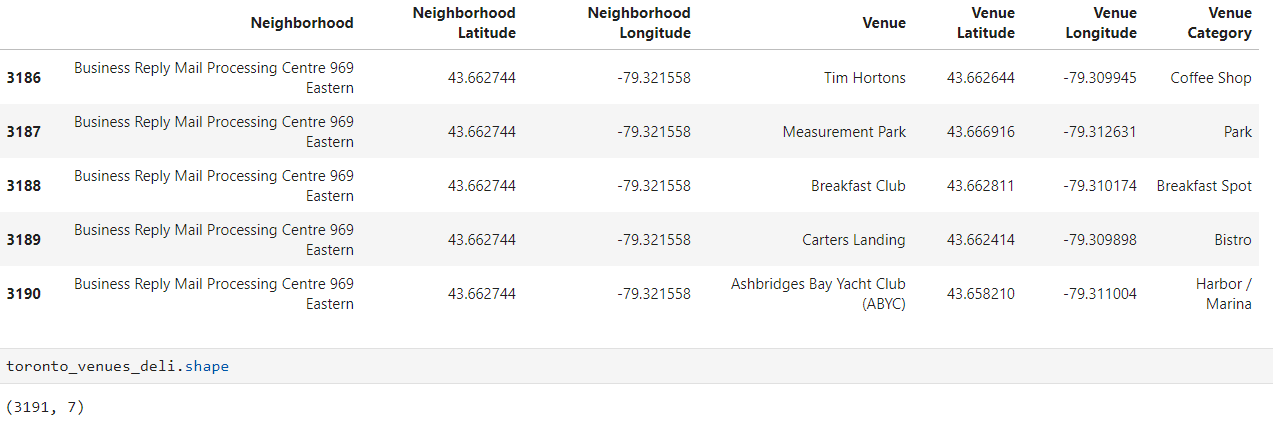
1. **Build neighborhoods list**  
   A list of districts is obtained from Toronto Wikipedia page. That list contains the names of the neighborhoods, boroughs and postal codes for the city. The data for them is then wrangled as in the Peer Reviewed assignment on Toronto clustering and segmentation.
2. **Neighborhoods geolocation**  
   Every element in the neighborhoods dataset is geolocated using *Python Geolocator* and two columns are updated containing latitude and longitude for each district or neighborhood by using the name of boroughs containing the word *Toronto*.  
   The geolocator service has some problems, many times gives time out error, for this reason in this step the information obtained is saved in a text file in CSV format.

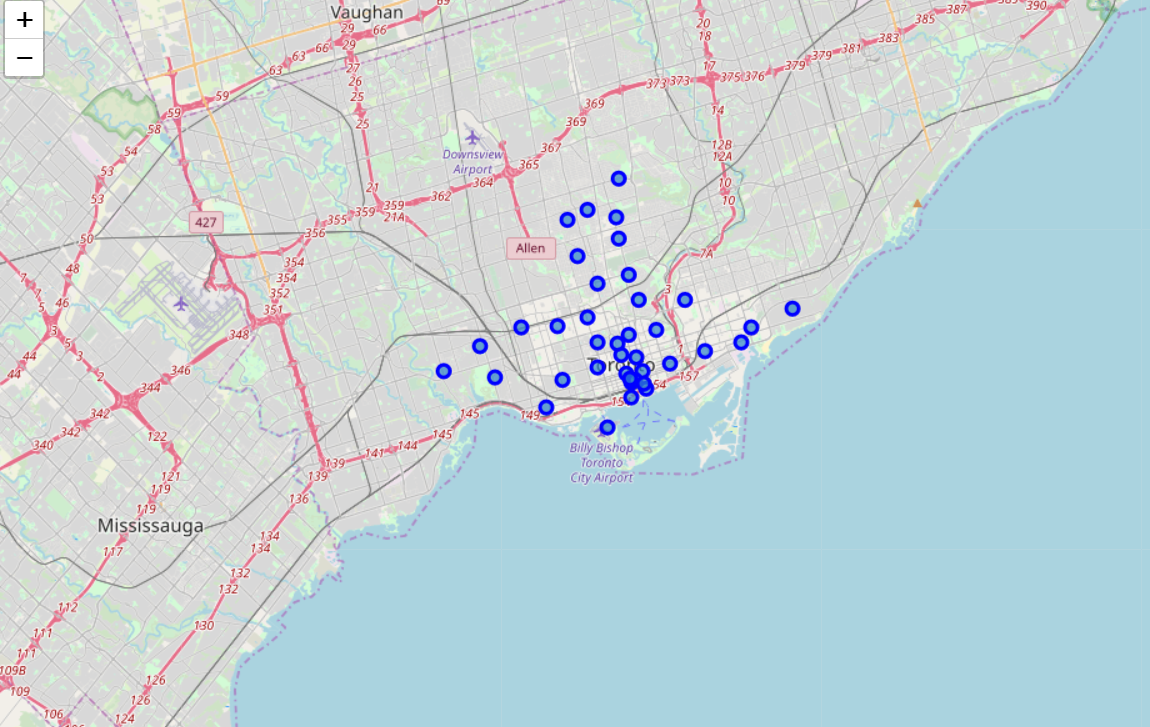




1. **Venues compilation**  
   As next step Foursquare services are used for obtaining venues for every neighborhood. The output is a new dataset with many records for every neighborhood containing the venues found for each of them.  
   A free Foursquare service with limited count of calls is used. In order to minimize the usage of Foursquare, the information is saved in a text (CSV) file. This was also done using Deli/Bodega category Id 4bf58dd8d48988d146941735 to retrieve data from Foursquare API. After this the explore function is used to get deli categories in the neighborhoods.

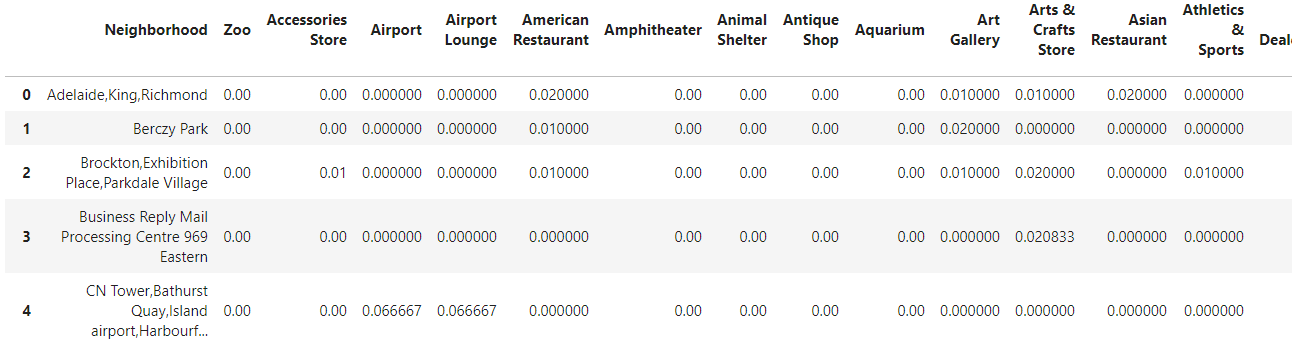


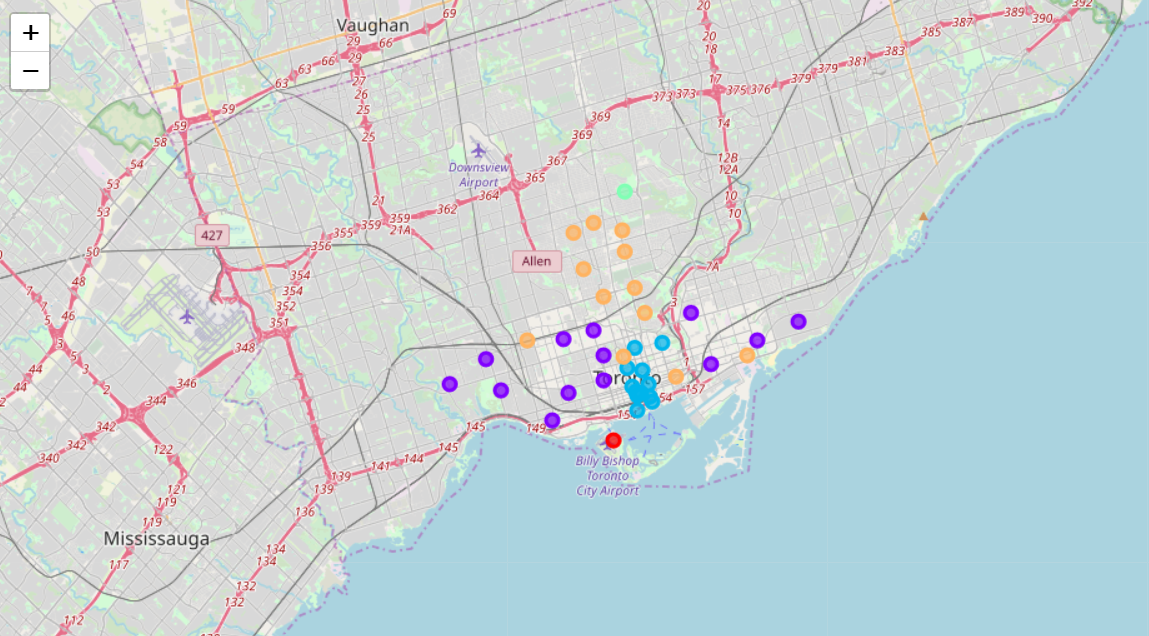




Deli restaurants in Toronto

1. **Neighborhoods segmentation**  
   The problem on hand is a case of unsupervised segmentation and, from the possible machine learning algorithms, K-means was chosen.
   * Taking in account that the venues information obtained from Foursquare is categorical, it must be previously processed in order to be handled by K-means algorithm. For this *"pandas.get\_dummies"* is used for dummies variables.
   * The list of dummy variables obtained are then grouped as features of every neighborhood.
   * K-means algorithm is executed with K = 5 chosen.
   * Next step is to build the segmentation dataframe, composed of the top venues for every neighborhood plus a segment label determined by K-means.





The feature above is used to group the neighborhoods into clusters using K-means clustering algorithm. The Folium library helps to visualize the neighborhoods in Toronto and its emerging clusters.

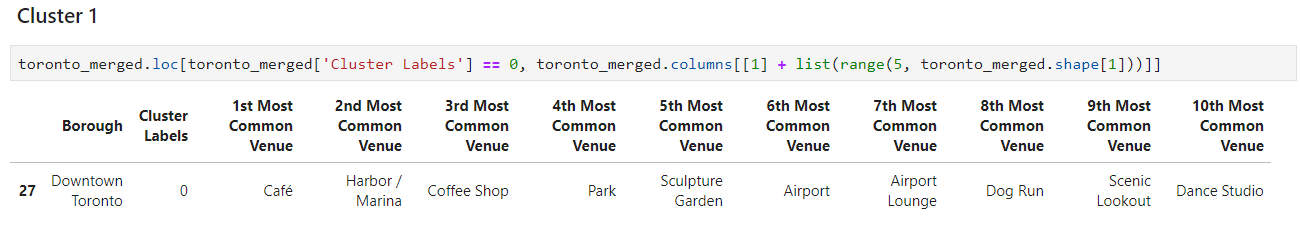


1. **Segments analysis**  
   Every segment is printed individually, were different characteristics can be observed for each group.  
   Next section describes the results.

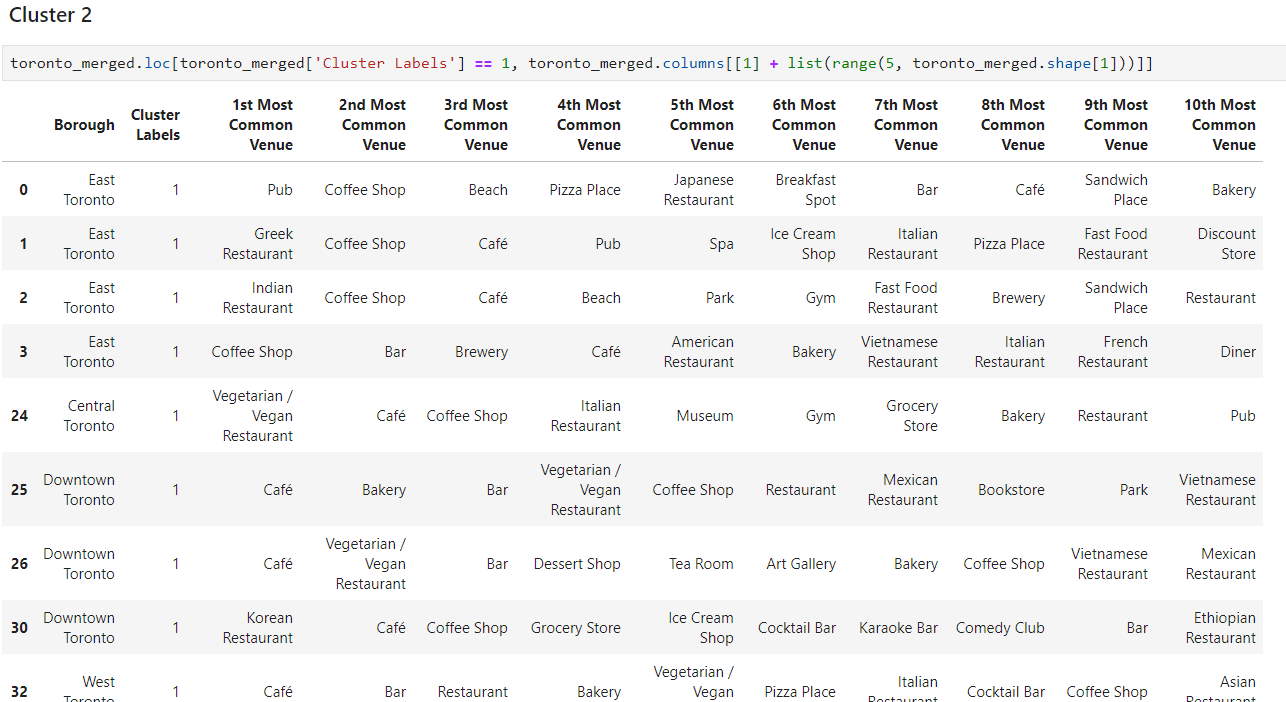
**Results**

As a result of segmenting Toronto, five clusters where defined.  
Following their characteristics are shown:

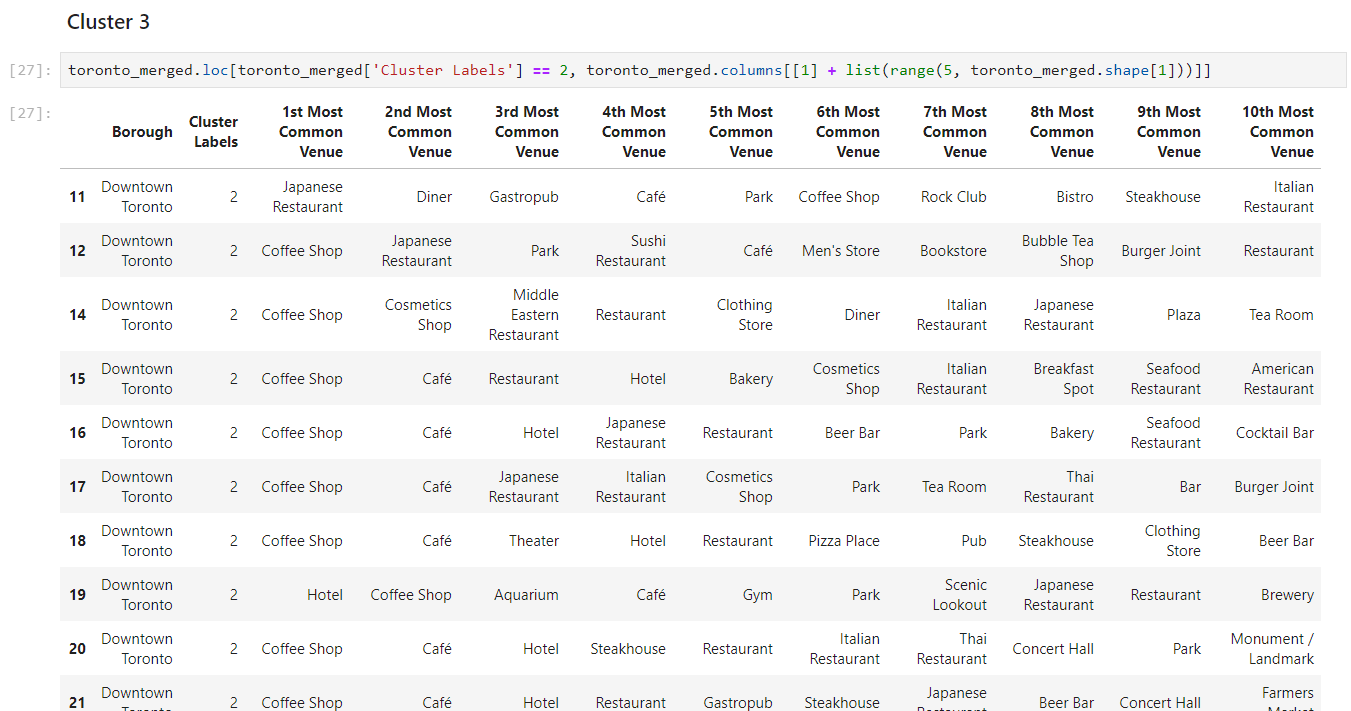
**Cluster 1**

****

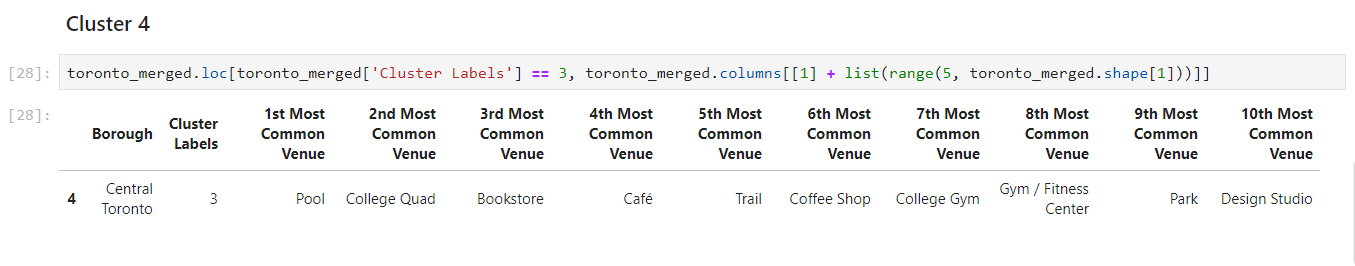
**Cluster 2**

****

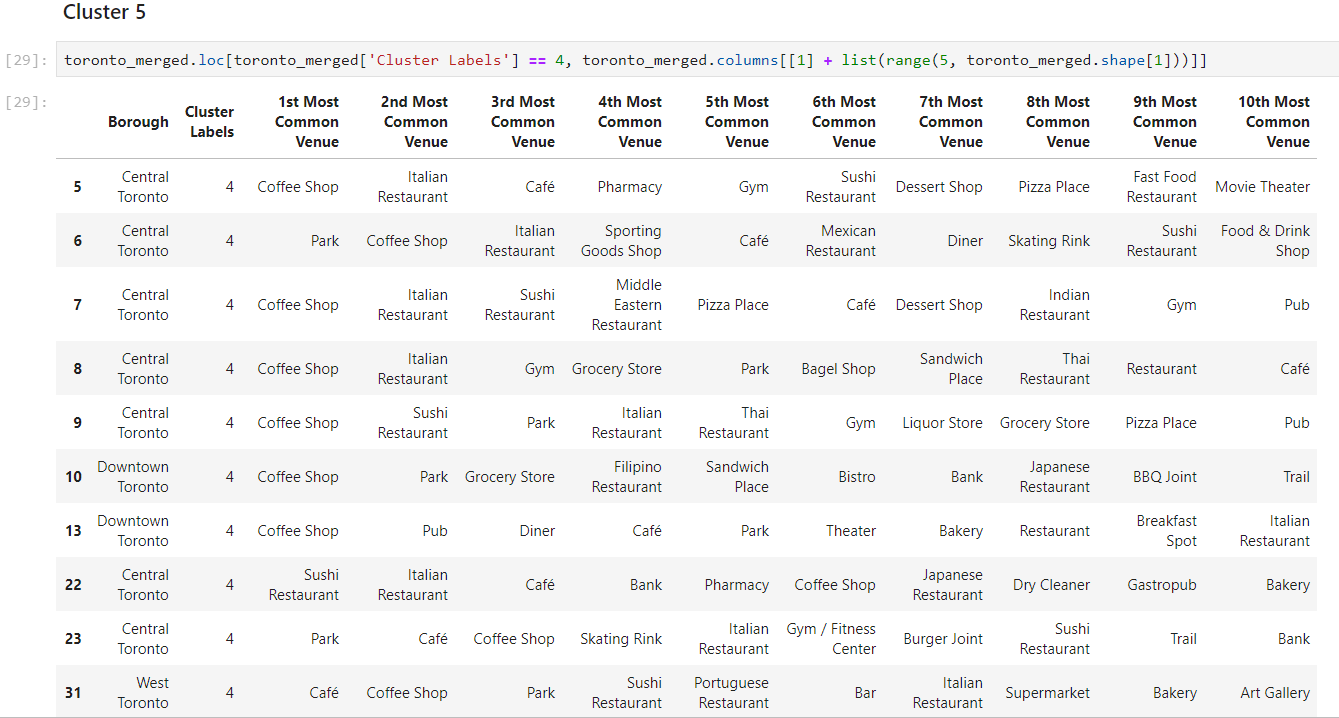
**Cluster 3**

****

**Cluster 4**

****

**Cluster 5**

****

**Discussion**

The aim of this project was to help a deli business trying to open new restaurants in Toronto find the best locations to set up in order to maximize their return on investments. It was observed that both **Clusters 1** (Downtown Toronto) and **4** (Central Toronto) did not have a lot of deli places and hence it would be worthwhile on the part of the company to set up their restaurants in those places. **Clusters 2**, **3** and **5** seem to be places of extremely high competition among deli businesses and hence would be risky to open the business from there.

**Conclusion**

Data was gathered from trusted sources and strong methodology was used for processing. The goals of the project were met and with some additional work can be developed to support the opening of a deli restaurant in an unknown location in a city. The future of this project includes taking other factors such as the cost of living in that cluster into consideration in order to recommend an area for opening the deli business.