**The Battle of Neighbourhoods in Toronto City**

**(Restaurant Choice)**

## **Introduction/Business problem**

Toronto City attracts a large number of tourists as one of the biggest cities in Canada. For tourists, finding the right place to eat can be a challenge, though Asian food, Middle Eastern, Mexican and many others. This is just one motive for giving tourists a good overview about what to eat where.

Thus, the goal I want to reach with this exercise is to give a simple recommendation to tourists in Toronto: in which district of the city will you find a large number or even concentration of which types of restaurants? Where to eat Mediterranean food, where to find American food, where to get fast food? The target audience are foreign tourists.

## **Description of the data**

I will, as requested by the assignment task, use foursquare data about venues in Toronto and extract the restaurant information from it. Foursquare is a US tech company from New York focusing on location data. Their technology and data powers apps such as Apple's Maps, Uber, Twitter and many other household names.

I will use foursquare data such as the restaurant name, ID, location and category of food (Asian food, vegetarian, Italian etc.).

For the Toronto neighbourhood data, a Wikipedia page "[https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M"](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M%22) exists that has all the information we need to explore and cluster the neighbourhoods in Toronto. i will scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas dataframe so that it is in a structured format.

## **Methodology**

In this section, I will describe the data analysis and how I used the data to yield the results.

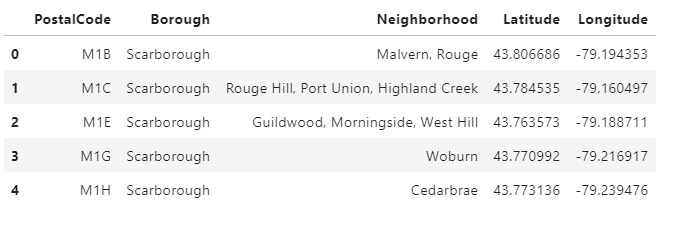
Starting out, I scraped data create a dataframe with the city neighbourhood of Toronto: “https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M For this, I used the BeautifulSoup read function to read the table in the page. I had to clean the resulting data frame in terms of unnecessary information or data that could not be handled in a data frame. The result is a nice data frame.

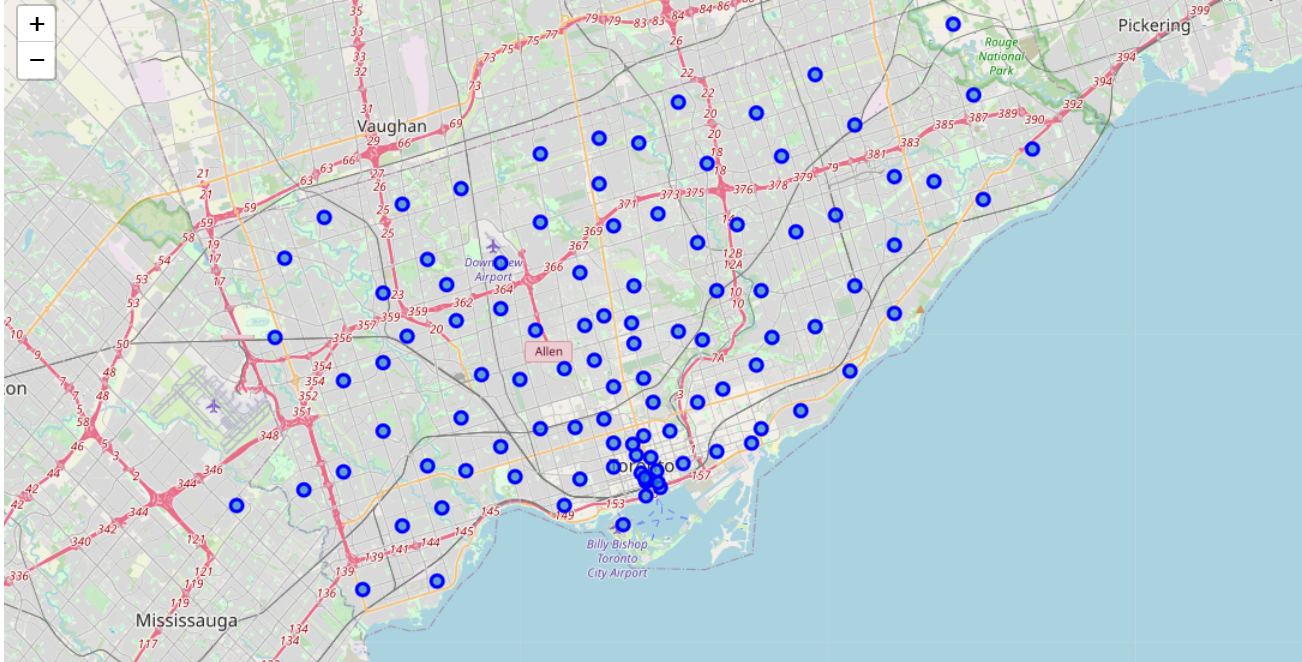
Then i had to read the exact locations (Latitude & Longitude) of all Toronto Postal Code from the following link

“https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs\_v1/Geospatial\_Coordinates.csv

I have used panada readcsv functions to convert data to a nice a data frame and then merge it with my original neighbour data frame.

Here is the resulted dataframe results with all neighbourhoods exact locations, then i have plot Toronto map with the neighbourhoods represented over it:

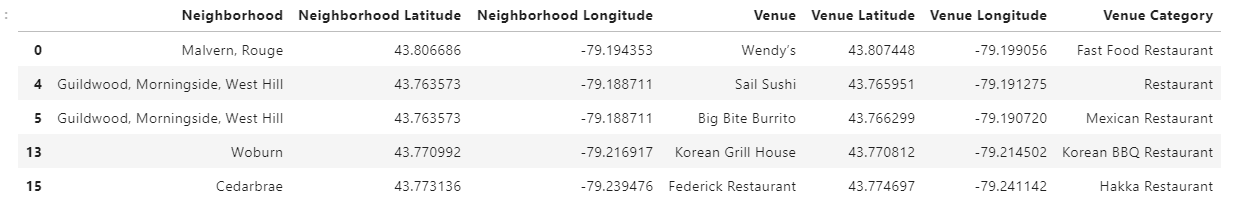




Next Step, I have retrieved the foursquare data for all venues on foursquare with a distance of less than 500 meters from each centre of each neighbourhood, as indicated as blue dots in the map above.

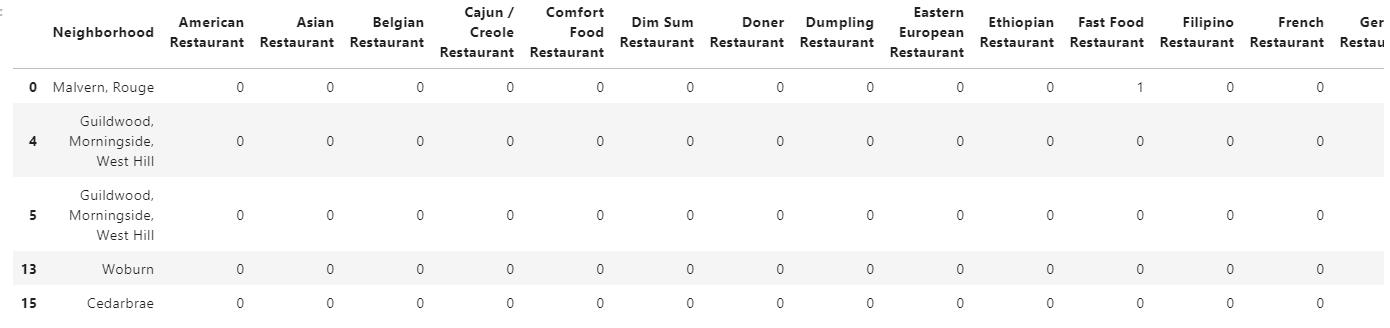
The result was a list of 2106 venues all over the city. Out of these 2106 venues (272 Unique Venue Category), 479 where restaurants. These 479 restaurants come from 49 unique restaurant categories, such as Italian, Vietnamese or French.

Then i grouped some similar restaurants type under one common category for better categorization, this results in 39 unique restaurant categories.

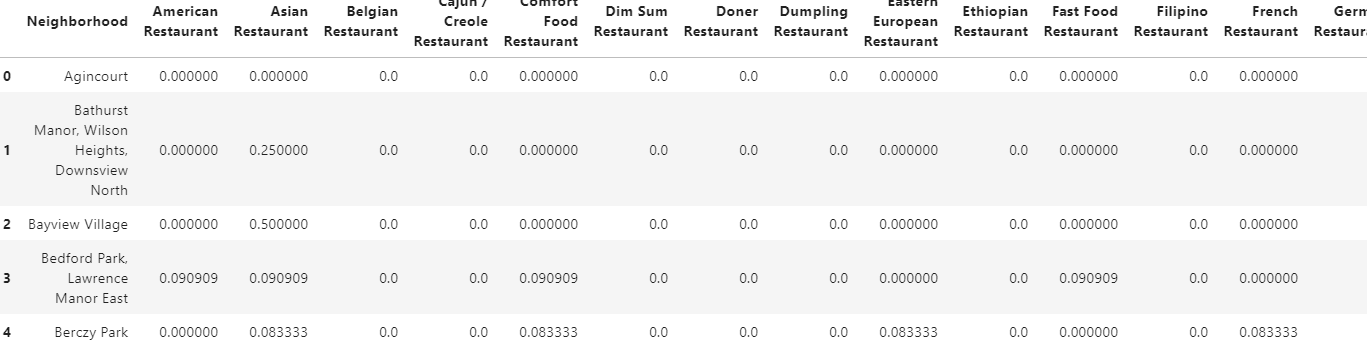




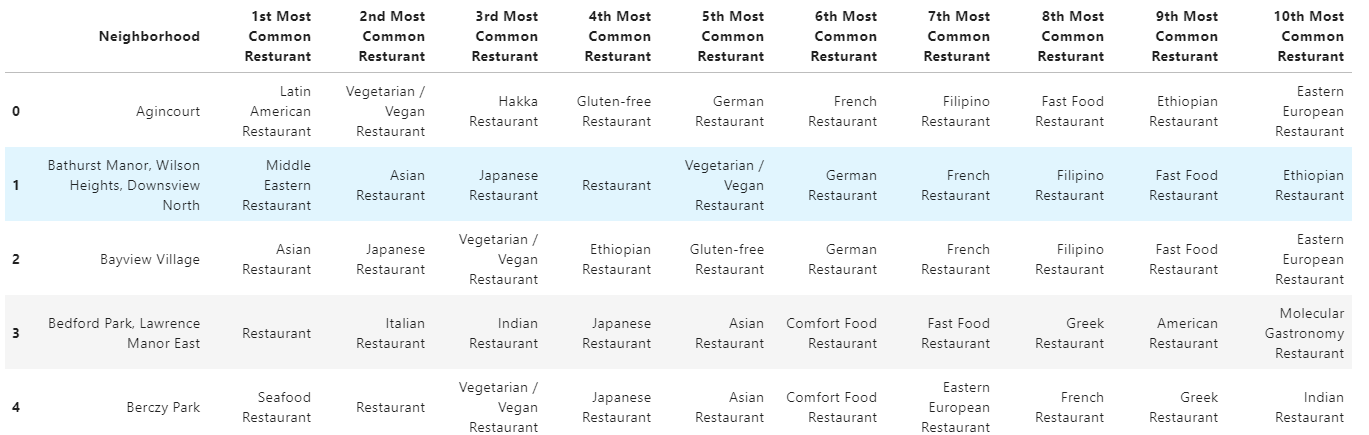
To find clusters of restaurant types in the different city districts, I first transformed the data frame with the restaurant venues, associated to city districts, by one-hot encoding (0/1), as seen in the picture below.



Next, I used grouping to show the frequency of each category of restaurants in each city district.



Then, I used this information to create a data frame in which you can see the most common restaurant venue types for each city district.

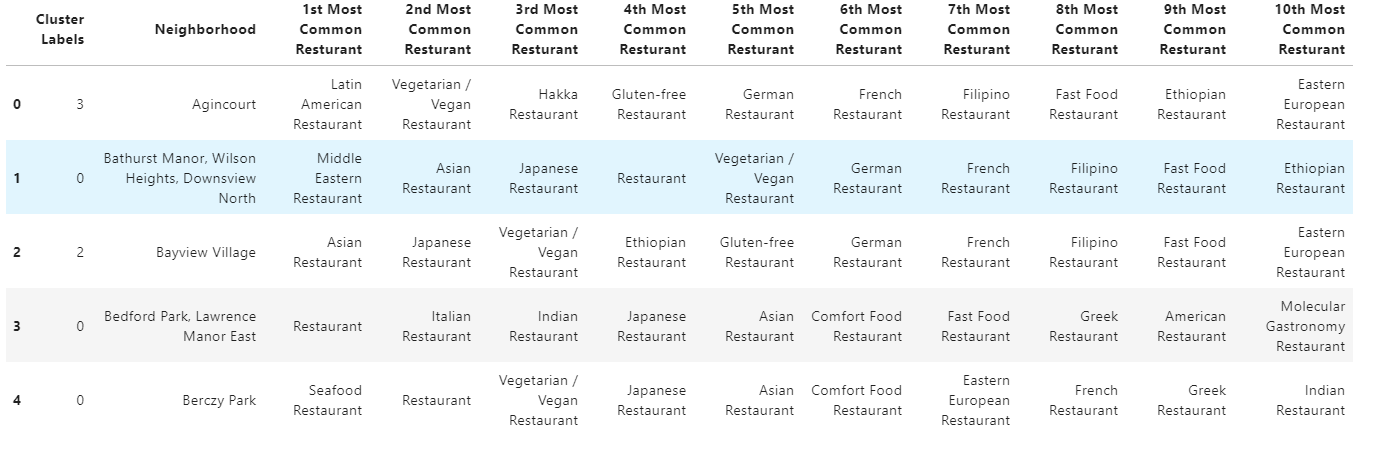


Now, with all this data, I could finally run an unsupervised machine learning algorithm, more specifically, a k-means clustering algorithm from the scikit-learn package. One could use the ellbow method to systematically define the k value, but I simply chose k to be 5, having been inspired by one of the coursera courses to do so.

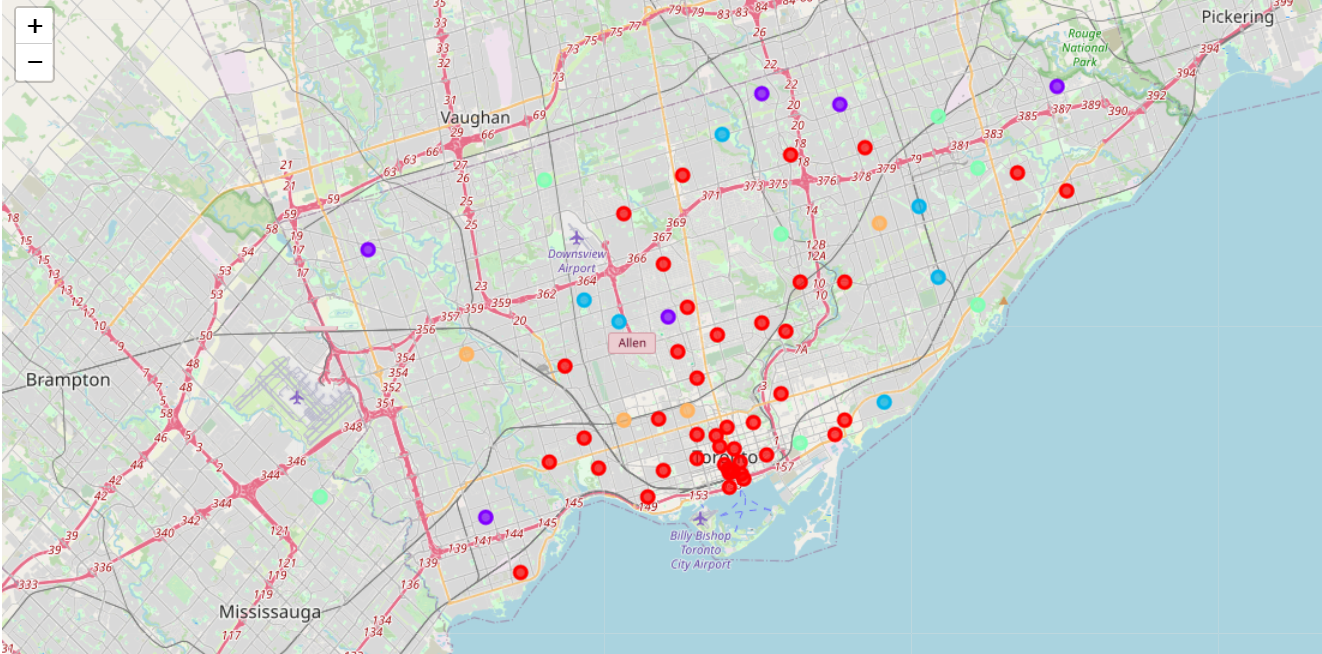
**Results**

And here already comes the result:

What we see in the table are the city districts and their most common restaurant venues, and they now have been assigned five different cluster labels from 0 to 4.

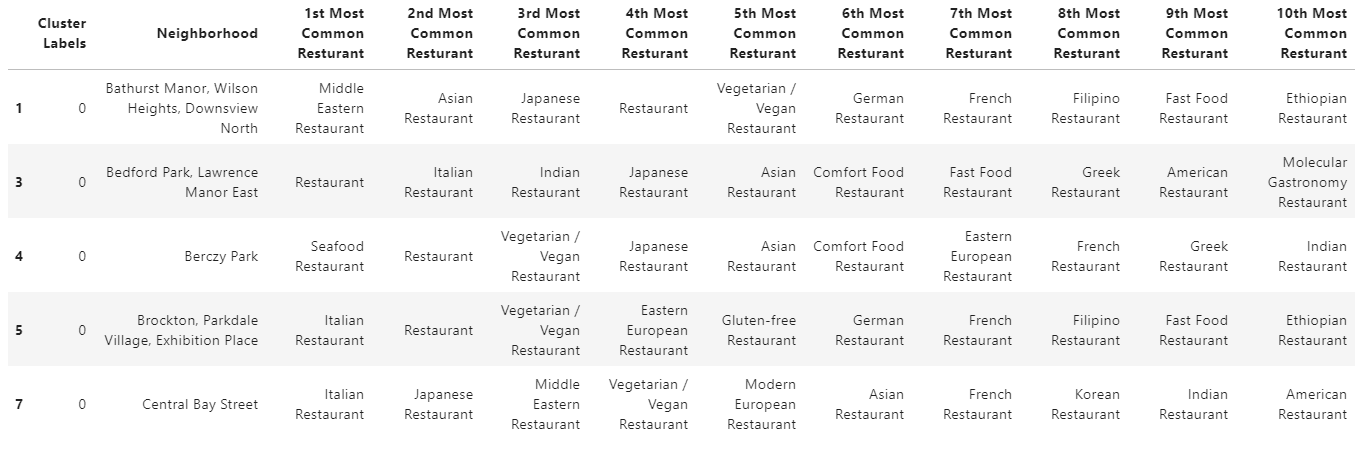


We can now use the cluster labels to show the city districts marked with a cluster-specific colour on a map.



Now, what is the final result of this exercise? We now can show five clusters of restaurant type concentrations for the city of Toronto, which I named according to the restaurant concentration the data shows.

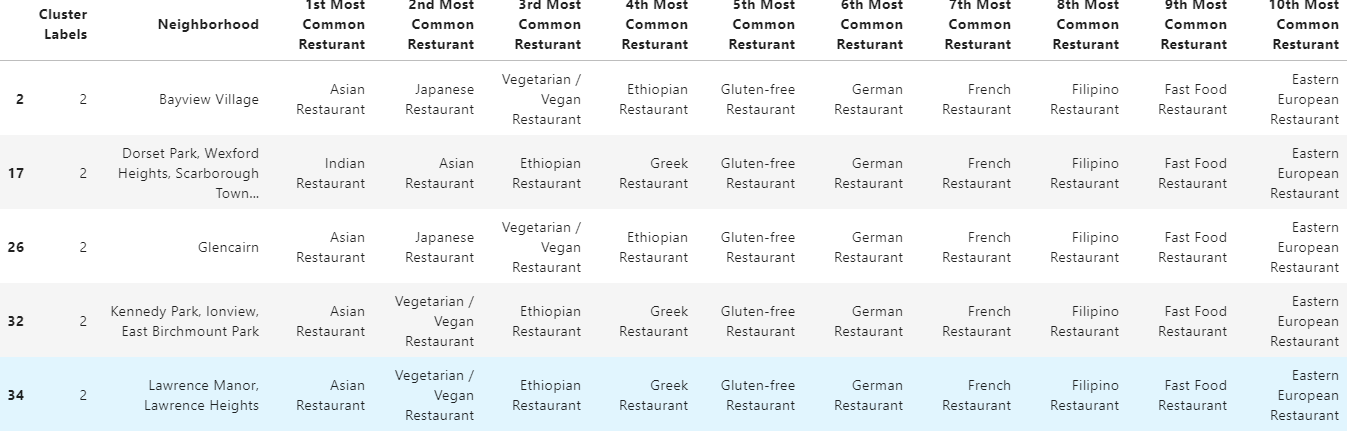
Cluster-1"Red" (Italian-European & Seafood lovers)



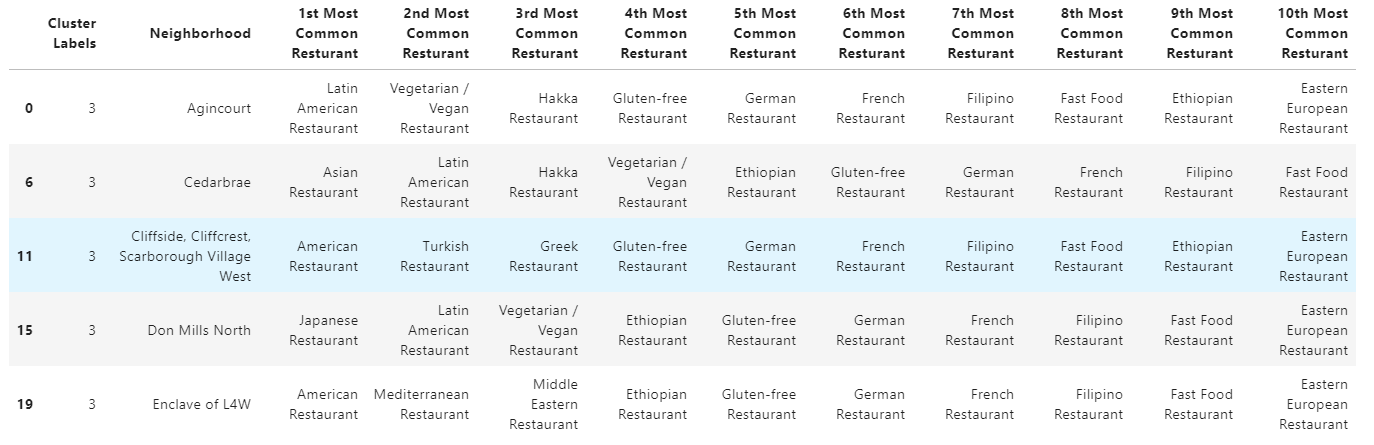
Cluster-2"Purple" (Fast food & Vegetarian seekers)



Cluster-3"Blue" (Asian flavour)



Cluster-4"Green" (American and Latin American Time)



Cluster-5"Yellow" (A taste from Middle East)



## **Discussion**

If I reflect the work necessary to create these results, what comes to my mind is that for typical ways of scraping, cleaning, handling, transforming and visualizing data, all the tools are simply there. We just have to get to know the available open source packages and learn how to use them. What I find fantastic is that nearly all of them are free of charge. Also, a simple notebook computer is enough.

## **Conclusion**

## We have achieved the goal presented at the outset of this project: tourists can see in the results Toronto city districts best match their food desires. This is just one example of fantastic data science uses cases one can realize applying technology which is available for free today!