Investigation of opening new restaurant in Manhattan and Berczy Park by K-means clustering

# 9 Nov 2020

## Introduction

* 1. Background

In this project, I want to capture opportunities in two of my favourite neighbourhoods- Manhattan in New York and Berczy Park in Toronto to open my own restaurant. However, I want to investigate whether I should open the restaurants in both places, or just open one restaurant in these two opinions. According to my experience, keeping a simple menu and selling food to specify targeted customers are the key to success.

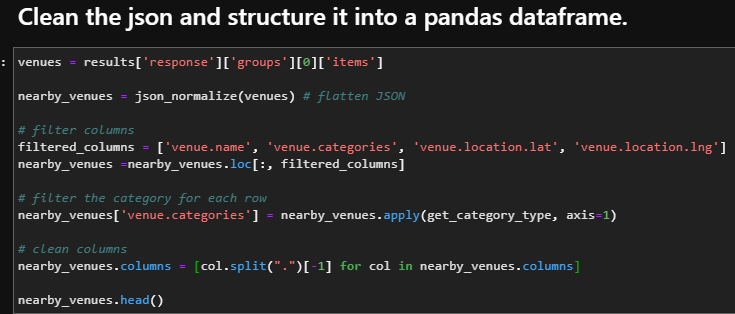
* 1. Business Problem

People living in the similar Neighbourhood would share similar interest and taste on food due to cultural and background similarities. Therefore, via clustering, I can determine whether neighbourhoods near my two opinions are highly concentrated in one segment.

# Data source and cleaning

2.1 Foursquare API

Huge amount of geographical data have to be obtained to support my analysis. Furthermore, the data have to be reliable and accurate. Calling Foursquare API is excellent options for me. I can obtain many details of venues near my targeted neighbourhoods. Those data are return by json format which is very convenience for data cleaning and transformation.



2.2 Scraping data from various website

Foursquare API is powerful but it is not totally perfect to fulfil my need. Therefore, I need to scrapy data from webpage. The table below on wiki page - <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> is very useful.



Luckily, by using powerful python library – Beautiful Soup, it is very easy to retrieve data on the table and transform it to pandas data frame.



2.3 Data cleaning

For Data cleaning, it is the most time consuming and critical process. To avoid the problem of “garbage in garbage out”, I did it very carefully and select many powerful python libraries below.

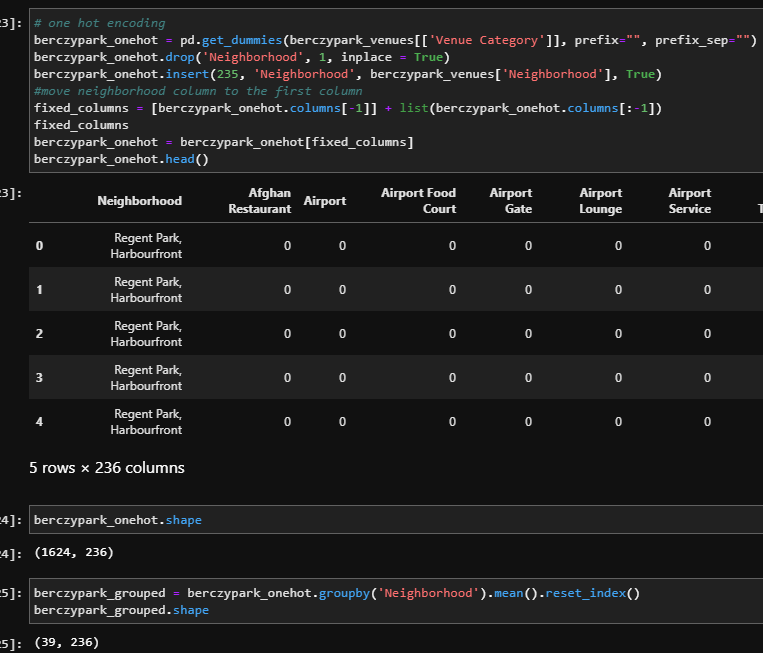
* Pandas
* Numpy
* Json

Json library is mainly used for transforming data returned by Foursquare API. After that, I would use Numpy to do message and turn it to any useful array. The efficiency of Numpy can shapely reduce processing time of handling huge amount of data. Finally, Pandas can help to transform data to various well-structured data frame. By this library, I can also clean some abnormal data, understanding the shape of data and easily transform data to any structure that can be processed for data visualization and date analysis model. For details steps, please refer to my Jupiter Book.

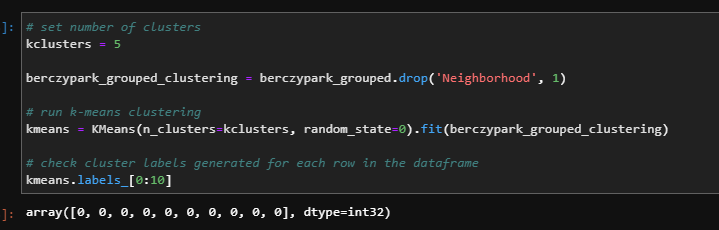
# Methodology

I choose to use K-mean model as it is very sample and reliable algorithms. Especially, the data types obtained so far are not very comprehensive. Using simple machine learning algorithms can minimize the risk to input too little and simple data to a complex algorithm which criteria can’t me fullfited.

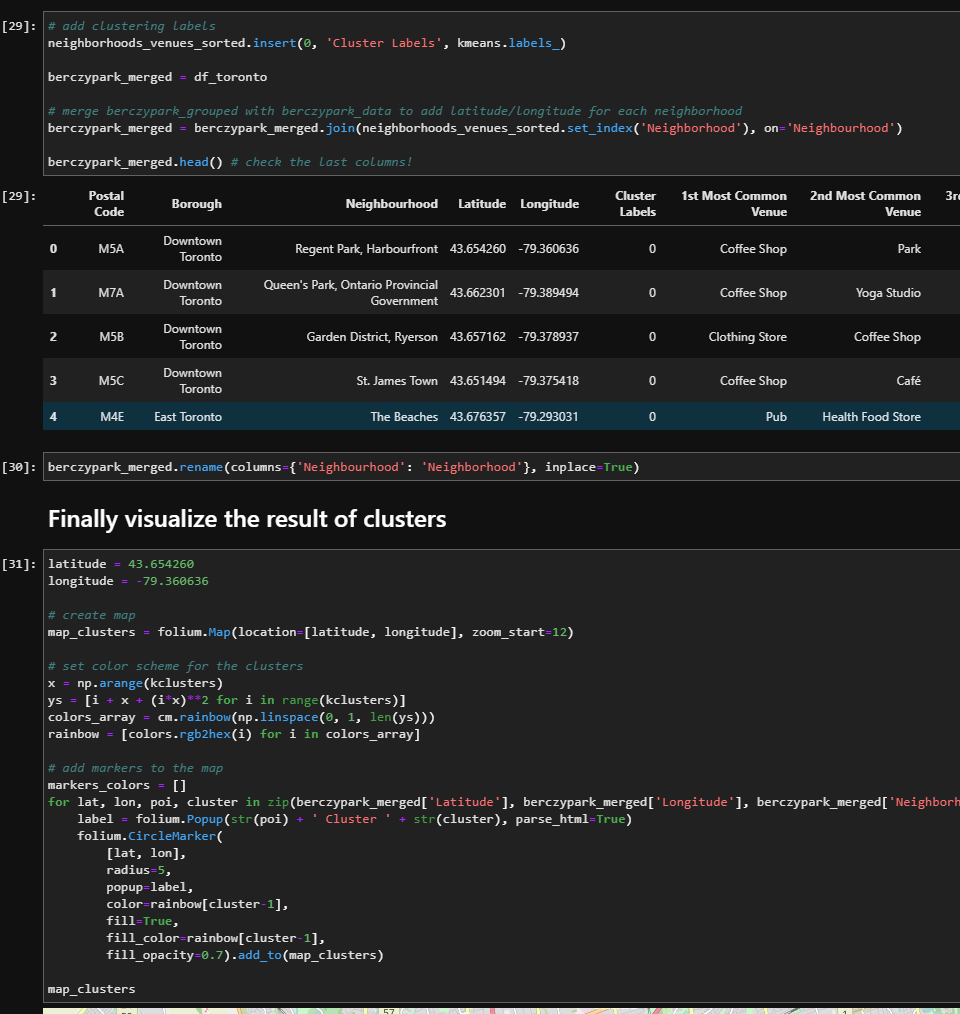
After Data cleaning, I choose to use K-mean model to perform clustering model. First of all, I transformed data for both neighbourhoods in below shape b Pandas. The reason is K-mean model is the unsupervised machine learning algorithms.



After that, input the data frame to K-mean clustering model



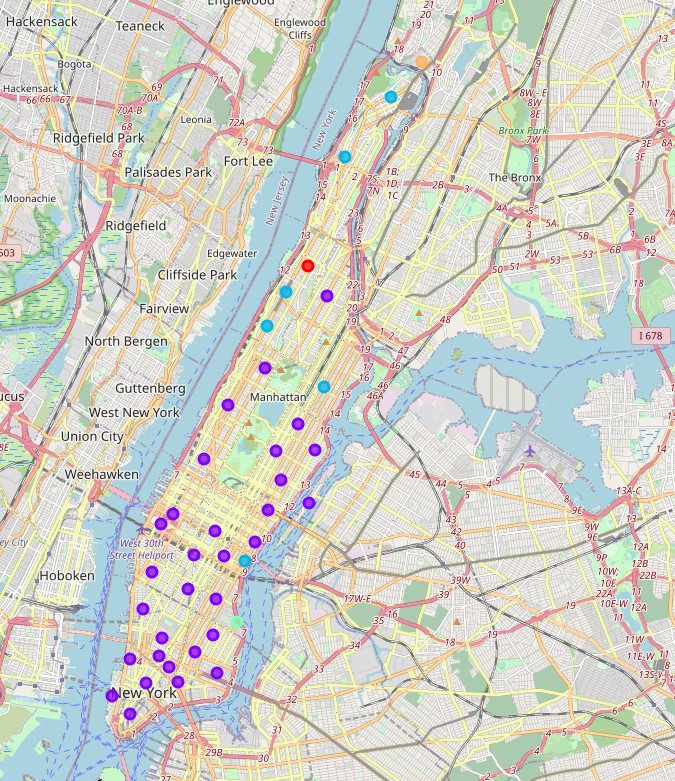
Combine clustering result to existing data frame for data visualization and summary.



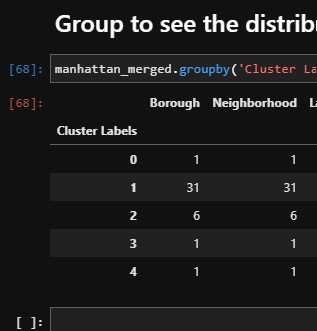
# Result

By the result, I generate summary tables and distribution of those clusters on map. From these two outputs, it is very easy to see whether neighbourhoods near my two opinions are highly concentrated in one segment.

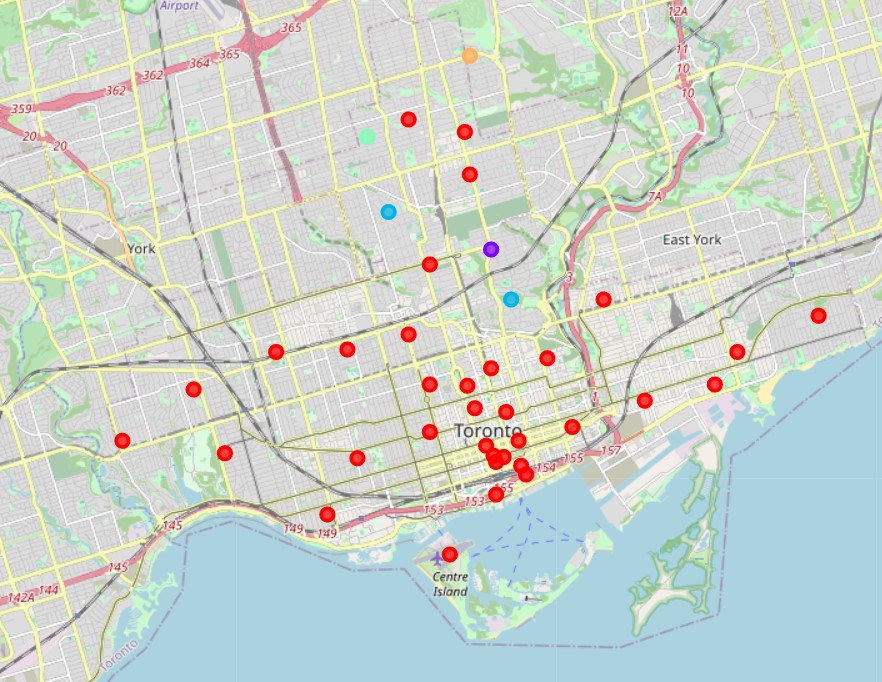
*Distribution of cluster – Manhattan (The color represent the cluster group)*



*Summary Table – Manhattan*



*Distribution of cluster –* *Berczy Park (The color represent the cluster group)*



*Summary Table – Berczy Park*



# Discussion

The results of Manhattan and Berczy Park are seem to be very well for me.

For Manhattan, 31 out of 40 (77.5%) neighbourhoods are classified as cluster 1. It means that, the similarity is very strong. There is stronger similarity shown in Berczy Park. 34 out of 39 (87.18%) are classified as cluster 0.

This preliminary finding show both targeted neighbourhoods can fulfil our requirement.

However, it is not good enough to conclude our decision.

To ensure customers can reach our restaurant easily, the location of each neighbourhoods have to be highly concentrated with each other. From two maps above, Manhattan may be better than Berczy Park.

# Conclusion

According to our result, I would suggest opening the restaurant in Manhattan first as it should be a better place to fulfil our needs and easier for customer to reach our restaurant.

It does not mean that we need to give up our plan in Berczy Park. As our restaurant in Manhattan become famous with good brand name. One more restaurant can be opened in Berczy Park to capture this opportunity too.