OPENING A NEW INDONESIAN RESTAURANT IN TORONTO, CANADA

APRIL 26

COURSERA IBM DATA SCIENCE
Authored by: Ozi Priawadi

1. Introduction

1.1 Background

According to www.yelp.ca, there are more than 15,000 restaurants in Toronto and about 3 million people (2017). That’s why opening a new restaurant there can be an extremely challenging task. According to several surveys, up to 40% of such start-ups fail in the very first year.

Let's suppose, an investor has enough time and money, as well as a passion to open the best eating spot in Toronto. What would be the best place for it? Is there a better way to answer these questions rather than guessing?

What if there is a way to cluster city neighborhoods, based on their near-by restaurant similarity? What if we can visualize these clusters on a map? What if we might find where Asian Restaurant is the most and least popular? Equipped with that knowledge, we might be able to make a smart choice from that data.

Let us allow machine learning to get the job done. Using reliable venue data, it can investigate the city neighborhoods, and show us unseen dependencies. Dependencies that we are not aware of.

> ***“According to www.yelp.ca, there are more than 15,000 restaurants in Toronto and about 3 Million.***

1.2 Business Problem

The objective of this capstone project is to find the most suitable Location for Entrepreneur to open a new Indonesian Restaurant in Toronto, Canada. By using Data Science and Machine Learning methods such as Clustering. This project aims to provide solutions to answer the business question: In Toronto, if an investor, entrepreneur, or chefs wants to open an Indonesian Restaurant, where should they consider opening it?

1.3 Target Audience

Investors, Entrepreneurs, or Chefs who interested to open a new restaurant and may need a piece of objective advice regarding the right location would be most successful to Open Indonesian Restaurant in Toronto, Canada.

2. Data

2.1 Data Source

To solve the problem, we will need data below:

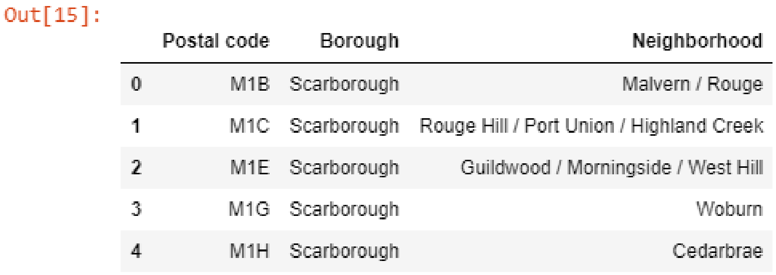
- List of Neighborhoods in Toronto, Canada.
- Latitude and Longitude of these Neighborhoods.
- Venue data related to Asian restaurants. This will help us to find the Neighborhoods that are most suitable to open an Indonesian Restaurant.

2.2 Extracting the Data

- Scrapping of Toronto neighborhoods via Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package (http://cocl.us/Geospatial_data)
- Using Foursquare API to get venue data related to these neighborhoods (https://developer.foursquare.com/docs)

3. Methodology

First, collecting neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia page, then cleansing the data. (*https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M*)



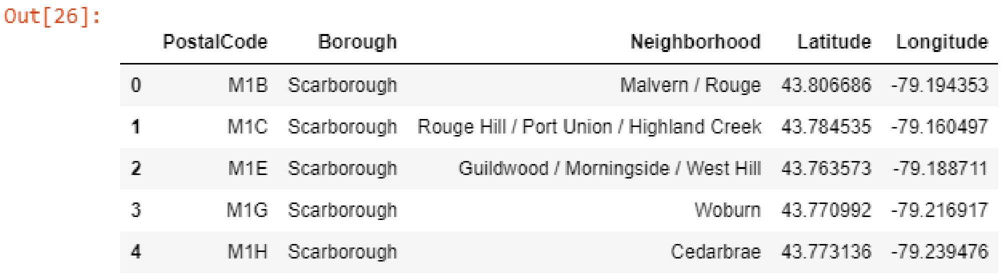
Out[15]:

	Postal code	Borough	Neighborhood
0	M1B	Scarborough	Malvern / Rouge
1	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek
2	M1E	Scarborough	Guildwood / Morningside / West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Figure 1 Data after Cleansing

We are using Pandas to collecting data from URL. Web scraping by utilizing pandas html table scraping method it's easier and more convenient to pull tabular data directly from a web page into dataframe.

However, we only get list of neighborhood names and postal codes from Wikipedia. So, we will need to get and adding coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, we are using Geocoder package to get the coordinates of Toronto neighborhoods by matching their Postal Code.



Out[26]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern / Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill / Port Union / Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood / Morningside / West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 2 Data After Adding Coordinates

After adding all these coordinates, we visualized the map of Toronto using Folium package to verify whether these are correct coordinates.

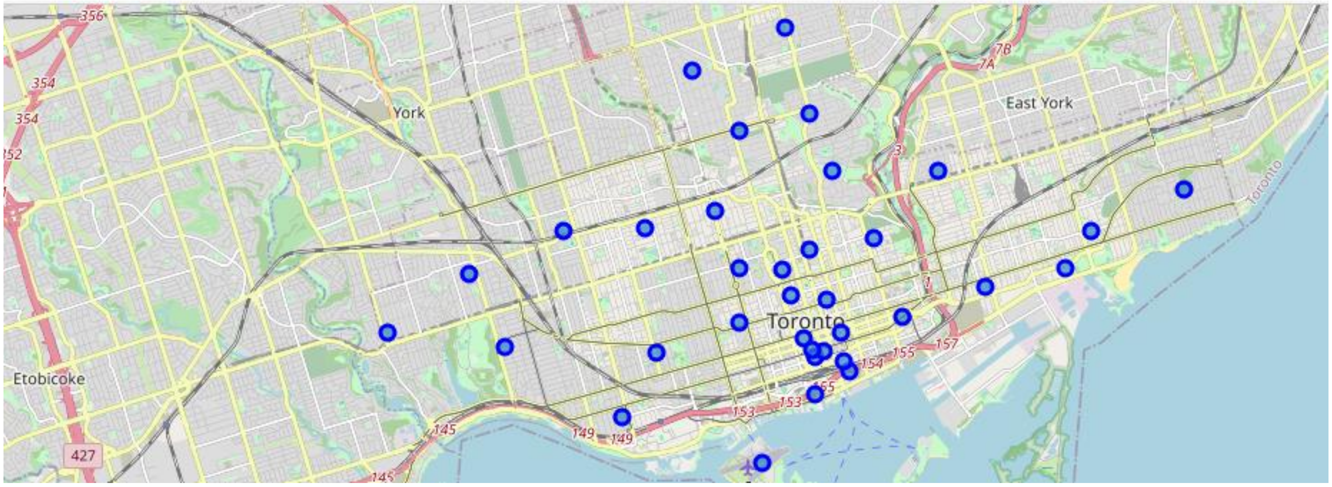
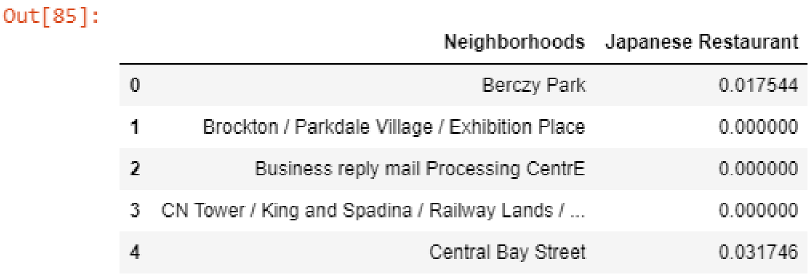


Figure 3 Visualized Maps of Toronto

Next, we are using Foursquare API to pull the list of top 100 venues within 500 meters radius. From Foursquare API, we are able to pull the names, categories, latitude and longitude of the venues. With this data, we can check how many unique categories from these venues. Then, we analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

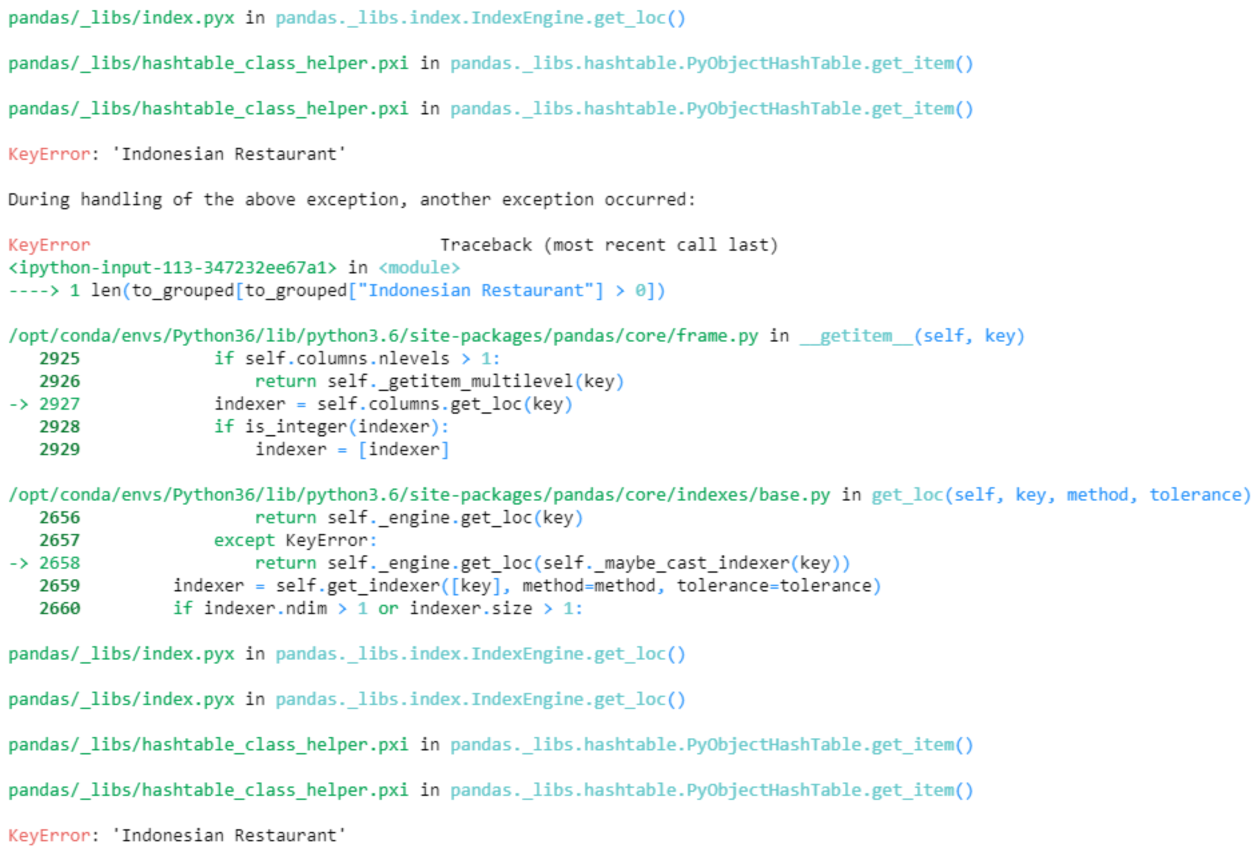


Out[85]:

	Neighborhoods	Japanese Restaurant
0	Berczy Park	0.017544
1	Brockton / Parkdale Village / Exhibition Place	0.000000
2	Business reply mail Processing CentrE	0.000000
3	CN Tower / King and Spadina / Railway Lands / ...	0.000000
4	Central Bay Street	0.031746

Figure 4 Grouping Japanese Restaurants by Neighborhoods

From this step, we made a justification to specifically look for “Japanese Restaurants”. Previously, when we ran the model, we are looking for “Indonesian Restaurants” but because getting an Error. So, we changed it and looked for the restaurants closest to Asian Restaurant and we choose “Japanese Restaurants” because taste of Japanese and Indonesian is almost same.



```
pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'Indonesian Restaurant'

During handling of the above exception, another exception occurred:

KeyError                                  Traceback (most recent call last)
<ipython-input-113-347232ee67a1> in <module>
----> 1 len(to_grouped[to_grouped["Indonesian Restaurant"] > 0])

/opt/conda/envs/Python36/lib/python3.6/site-packages/pandas/core/frame.py in __getitem__(self, key)
   2925             if self.columns.nlevels > 1:
   2926                 return self._getitem_multilevel(key)
-> 2927             indexer = self.columns.get_loc(key)
   2928             if is_integer(indexer):
   2929                 indexer = [indexer]

/opt/conda/envs/Python36/lib/python3.6/site-packages/pandas/core/indexes/base.py in get_loc(self, key, method, tolerance)
   2656                 return self._engine.get_loc(key)
   2657             except KeyError:
-> 2658                 return self._engine.get_loc(self._maybe_cast_indexer(key))
   2659         indexer = self.get_indexer([key], method=method, tolerance=tolerance)
   2660         if indexer.ndim > 1 or indexer.size > 1:

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas/_libs/hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 'Indonesian Restaurant'
```

Figure 5 Getting Error Grouping Indonesian Restaurant

Lastly, we performed the clustering method using K-Means Clustering Algorithms. K-Means algorithm is one of the most common cluster methods of unsupervised learning and it is highly suited for this project as well. K-Means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

We clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Japanese Food”. Based on the results (the concentration of clusters), we will be able to recommend the ideal location to open the new Indonesian Restaurant.

4. Results

4.1 Visualizing Cluster

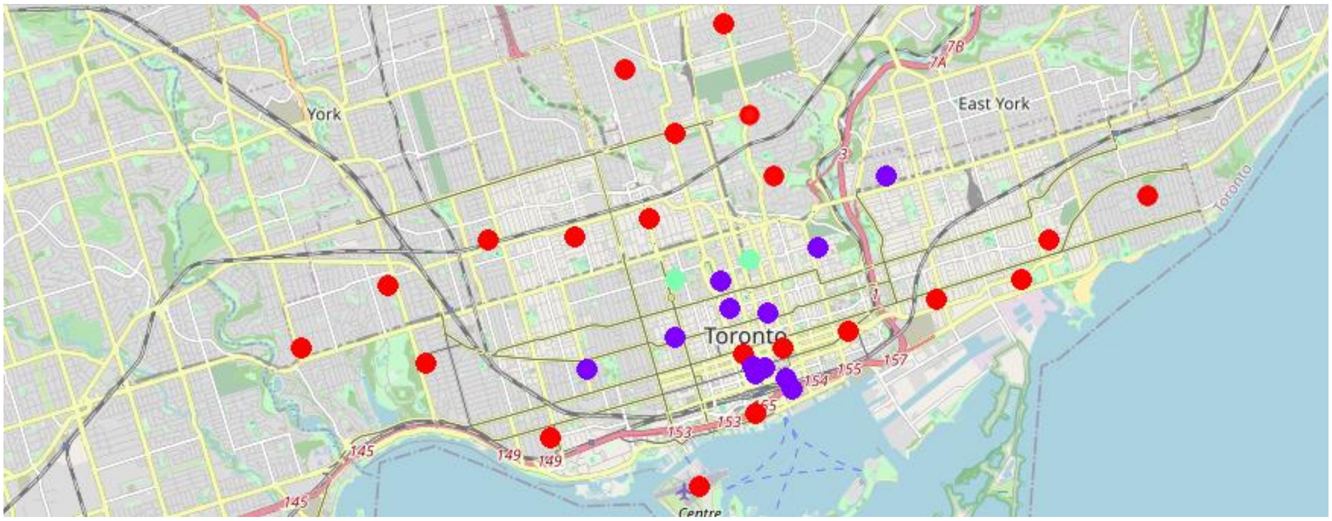


Figure 6 K-Means Cluster

MAP LEGEND
Cluster 0 - Red dots
Cluster 1 - Purple dots
Cluster 2 - Light Green dots

The results from K-Means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Japanese Restaurants are in each neighborhood:

- Cluster 0: Neighborhoods with lowest number to no existence of Japanese Restaurant
- Cluster 1: Neighborhoods with high number of Japanese Restaurants
- Cluster 2: Neighborhoods with high number of Japanese restaurants

5. Discussion and Recommendation

5.1 Discussion

- Most of Japanese Restaurant are concentrated in the Garden District, Ryerson
- Highest number of Japanese Restaurant can be found in Cluster 1 and Cluster 2
- Cluster 0 has very low number to no existence of Japanese Restaurant

- Cluster 0 mostly comes from Harbourfront East / Union Station and The Annex / North Midtown / Yorkville

5.2 Recommendation

- Open New Indonesian Restaurant in Cluster 0 with lowest number to no existence competition
- Avoid Neighborhood in Cluster 1 and 2, already high concentration of Japanese Restaurant and Intense Competition
- Nonetheless, if the food is authentic, affordable and good taste, I am confident that it will have great following everywhere

6. Conclusion

- Answer the business question: The neighborhoods in Cluster 0 are the most preferred locations to open New Indonesian Restaurant
- Findings of this project will help the relevant stakeholders (example: Investors, Entrepreneurs, or Chefs) to capitalize on the opportunities on High Potential Locations while avoiding overcrowded areas in their decisions to Open New Indonesian Restaurant
- In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.

7. References

- List of neighborhoods in Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Foursquare Developer Documentation: https://developer.foursquare.com/docs