Coursera

IBM Applied Data Science Capstone Project

Opening a New Milk Tea & Street Food Store in Toronto, Canada

Yawen Liu

July 2021

1

2

[1] https://d1ralsognjng37.cloudfront.net/88d3e1e7-8c07-4058-8a24-a052ec9d1c4b.webp

[2] https://www.torani.com/sites/default/files/recipes/illustration/GettyImages-1265715517-min.jpg

1. Introduction

1.1 Background

Nowadays, milk tea (or bubble tea) has become an essential drink, especially in younger age people's daily life. Due to its versatility (i.e. flavors, toppings, kinds of milk), convenience, and successful social media marketing, milk tea has become a very popular affordable drink around the world, specifically in a city with a large young population like Toronto, Canada. Many people like to grab a milk tea and enjoy it while shopping, work, and hangout.

However, with the development of the milk tea shop market, stores selling milk tea only are not competitive anymore. Milk tea stores selling street food (i.e. popcorn chicken, lunch bento box) are more favored by consumers as consumers can get both food and drinks conveniently without going to another food store. It is preferred by people who enjoy Asian flavors and want to have a faster order process. Opening a milk tea & street food store would be very profitable and promising from a business standpoint with the appropriate selection of store locations in megacities.

[3]

[3] https://bonteacafe.com/media/tiger_milk_tea.jpg

1.2 Interests & Business Problem

One of my friends just moved to Toronto and wants to start her own business which is opening a milk tea & street food store to serve people who enjoy milk tea and Asian street food. As Toronto is the most populous city in Canada with a highly diversified and strong economy, on one hand, there will be no problems in finding consumers. On the other hand, the competition in the milk tea industry is fierce as well. To succeed, the location of the store is of vital importance.

As a result, in this project, we utilize the previously learned Data Science & Analysis tools to explore which area(s) is(are) a better location for my friend to open her milk tea & street food store.

After a thorough discussion with my friend, we decided to **focus on two areas when selecting the location of the store to attract consumers as much as possible**:

1. The store is expected to be around/near shopping malls
2. The store is expected to be far from similar milk tea stores

[4]

[4] https://www.frommanilawithloveblog.com/2019/08/this-chem-lab-inspired-milk-tea-shop.html

2. Data Acquisition and Cleaning

2.1 Required Data

To solve this problem, we need the following data to obtain insights:

- **List of Neighborhoods in Toronto, Canada**. For this project, the scope is confined to the city of Toronto, Canada which is the largest city in Canada and a world leader in business, finance, technology, entertainment, and culture with a large population. This data is obtained by web scraping of a relevant Wikipedia website.
- **Geospatial Data of Toronto's Neighborhoods**. The geospatial data include the Latitude and Longitude coordinates of each neighborhood in Toronto. This data is required to plot the map for visualization and to obtain the nearby venue data. This data is obtained from the Python Geocoder package.
- **Nearby Venue data of Each Neighborhood**. This data is utilized to perform clustering particularly related to shopping malls and similar milk tea (bubble tea) places. We make decisions based on the result of clustering. This data is extracted from the Foursquare API.

2.2 Data Cleaning

After the data is scraped from the website, data cleaning is executed for the preparation of geospatial data acquisition and data clustering.

a) Data downloaded and scrapped were first extracted and assigned to new lists with columns of Postal code, Borough, and Neighborhood respectively.
b) Then the newly created three data lists were merged into a new data frame which is the foundation for our project.
c) The Borough with "Not assigned" value was dropped and ignored in this project.

d) If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.
e) Reset index for the updated data frame then prints it out to check.

2.3 Feature selection

In this project, we clustered the neighborhoods based on the frequency of occurrence of "Shopping Mall" and "Bubble Tea Shop".

- The new milk tea store is expected to be close to the shopping mall as much as possible for high customer flow. As a result, we want to select the cluster locations with a high frequency of occurrence for the shopping mall.
- The new milk tea store is expected to be far from similar milk tea shop/bubble tea shop as much as possible to reduce fierce competition. In this case, we want to select the cluster locations with a low frequency of occurrence for Bubble Tea shop.

As a result, we are going to select the cluster locations with a combination of "**high frequency of occurrence for shopping mall + low frequency of occurrence for Bubble Tea shop**".

3. Methodology

We went through the following steps to finish this project:

1. **The List of Neighborhoods in Toronto, Canada** data was obtained from the following Wikipedia page: https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050.

We used the web scraping technique to extract data from the above Wikipedia page. The Beautifulsoup package was used for web scraping with the help of Python requests.

2. After web scrapping, we used Python libraries like Pandas and Numpy to go through the data cleaning process.

3. When the data cleaning was finished, we called the Python Geocoder package to obtain the geospatial information (i.e. Latitude, Longitude coordinate) of the neighborhoods on the list.

4. Once we have the geospatial information of neighborhoods, the Foursquare API was connected to obtain the venue data for neighborhoods on the list. This step has given us detailed venue information (i.e. name, category) as we specified. For each neighborhood in this project, we set to get the top 100 venues that are within a radius of 2000 meters. We filtered "Shopping Mall" and "Bubble tea shop" as target venue categories for neighborhoods.

5. With the help of further data wrangling from obtained JSON data, we went through the data clustering process and visualized the map data using the Folium package. We used K-means clustering for data clustering. In this project, we clustered the neighborhoods into 4 clusters based on their occurrence frequency for "Shopping mall" and "Bubble tea shop".

6. The data clustering results gave us insights on which neighborhoods have more "Shopping malls" and which neighborhoods have fewer "Bubble tea" places as competitors. By analyzing the result, we have narrowed down the location selection of the new milk tea and street food shop to one cluster.

4. Results

From the results of K-Means clustering, we have categorized the neighborhoods into four clusters based on the frequency of occurrence for "Shopping Mall" and "Bubble Tea Shop". A brief observation of four clusters is showing below:

- **Cluster 0**: Lowest frequency of occurrence for Shopping Mall, very low frequency of occurrence for Bubble Tea Shop.
- **Cluster 1**: Higher frequency of occurrence for Shopping Mall, the highest frequency of occurrence for Bubble Tea Shop.
- **Cluster 2**: Higher frequency of occurrence for Shopping Mall, low frequency of occurrence for Bubble Tea Shop.
- **Cluster 3**: Highest frequency of occurrence for Shopping Mall, very low frequency of occurrence for Bubble Tea Shop.

From the observation results of four clusters, Cluster 3 is the best choice of new milk tea location as it has more shopping mall concentration and fewer bubble tea shops as competitors.

The results of clustering are visualized in the map using the Folium package. A screenshot of the created map is attached on the next page:

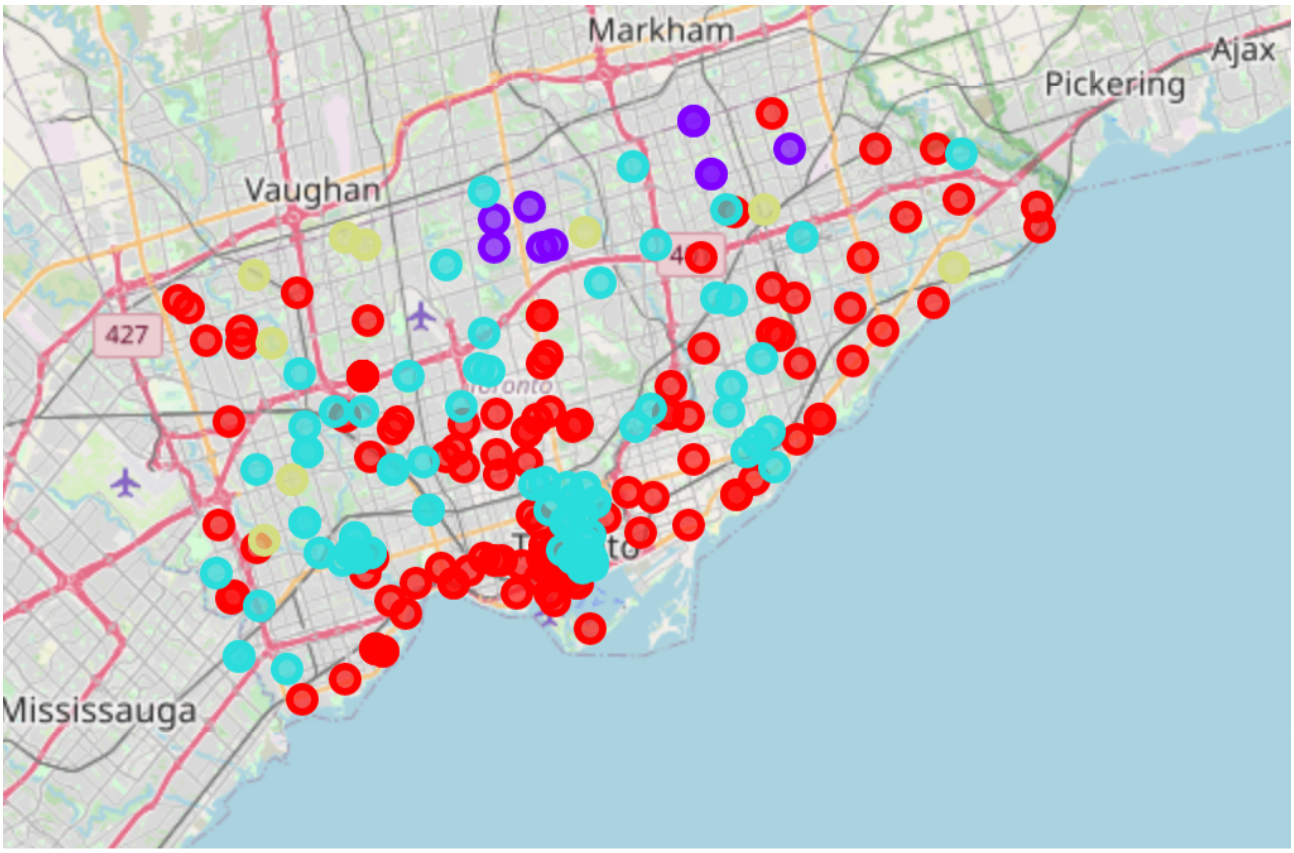

Figure 1. Screenshot of Superimposed Cluster Mark

Marker Color on the map:

- Cluster 0: Red
- Cluster 1: Purple
- Cluster 2: Blue
- Cluster 3: Yellow

5. Discussion

From the clustering results, the shopping mall is generally uniformly spread in the neighborhoods of Toronto, this is easy to understand given the situation that Toronto is the most populous city in Canada and the largest urban and metro area with a population density of 4,149.5 people per square kilometer (10,750/sq mi)[5]. However, the occurrence frequency of shopping malls is lower along the north-eastern side of the bay (circled in green in **Figure 2**), this might be due to the lower population density along the bay area[6] as the high population density might be more preferred by the developers. A screenshot of the Toronto population density map is shown as follows:

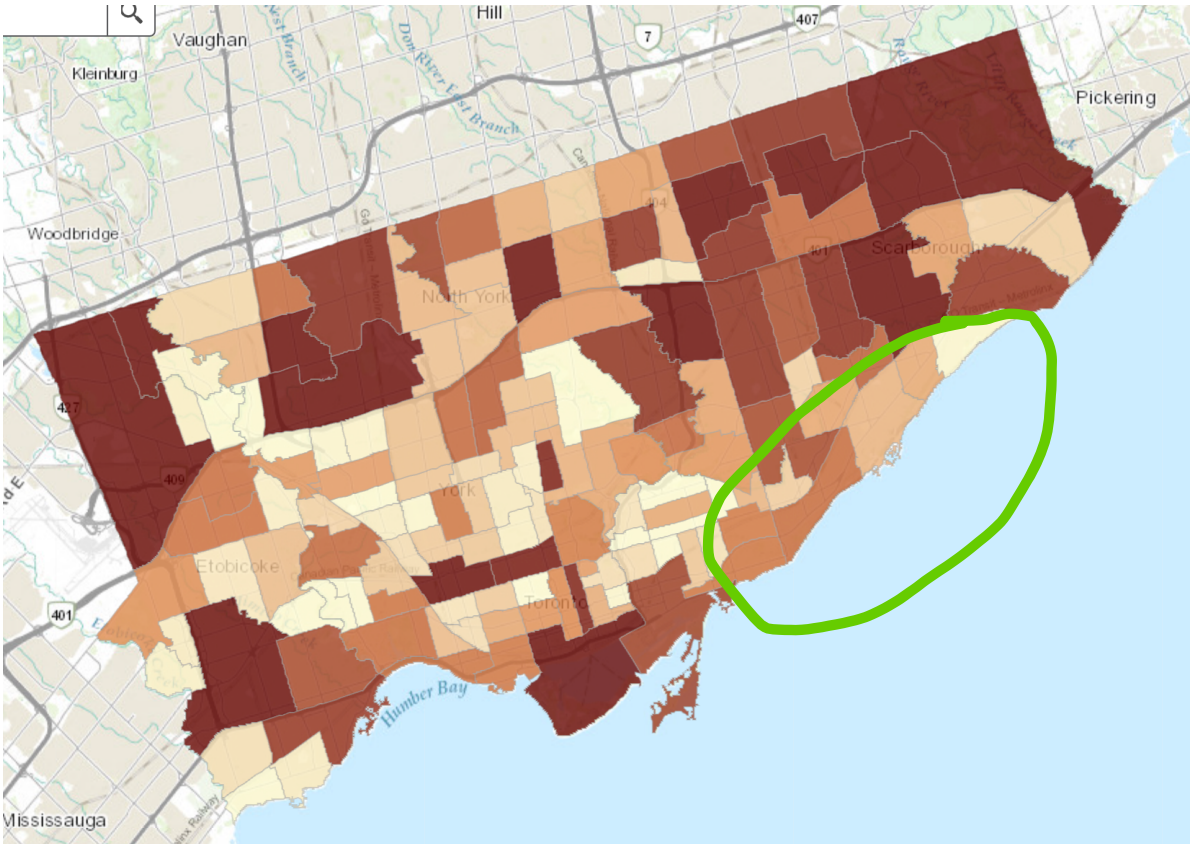

Figure 2. Screenshot of Toronto Population Density Map

[5] https://worldpopulationreview.com/canadian-cities/toronto-population
[6] https://www.arcgis.com/apps/webappviewer/index.html?id=1535b9fca54f46b3954bca6aaf3ab3f5

As shopping mall attracts thousands of consumers and provides enough time for consumers to make choices as well as recreational means of shopping, it would be very beneficial for us to build our new milk tea & street food store around/near or in the shopping mall to take advantages of the high customer flow. Given the above consideration, we should avoid the north-eastern side of the bay area as it has a lower occurrence frequency of shopping malls.

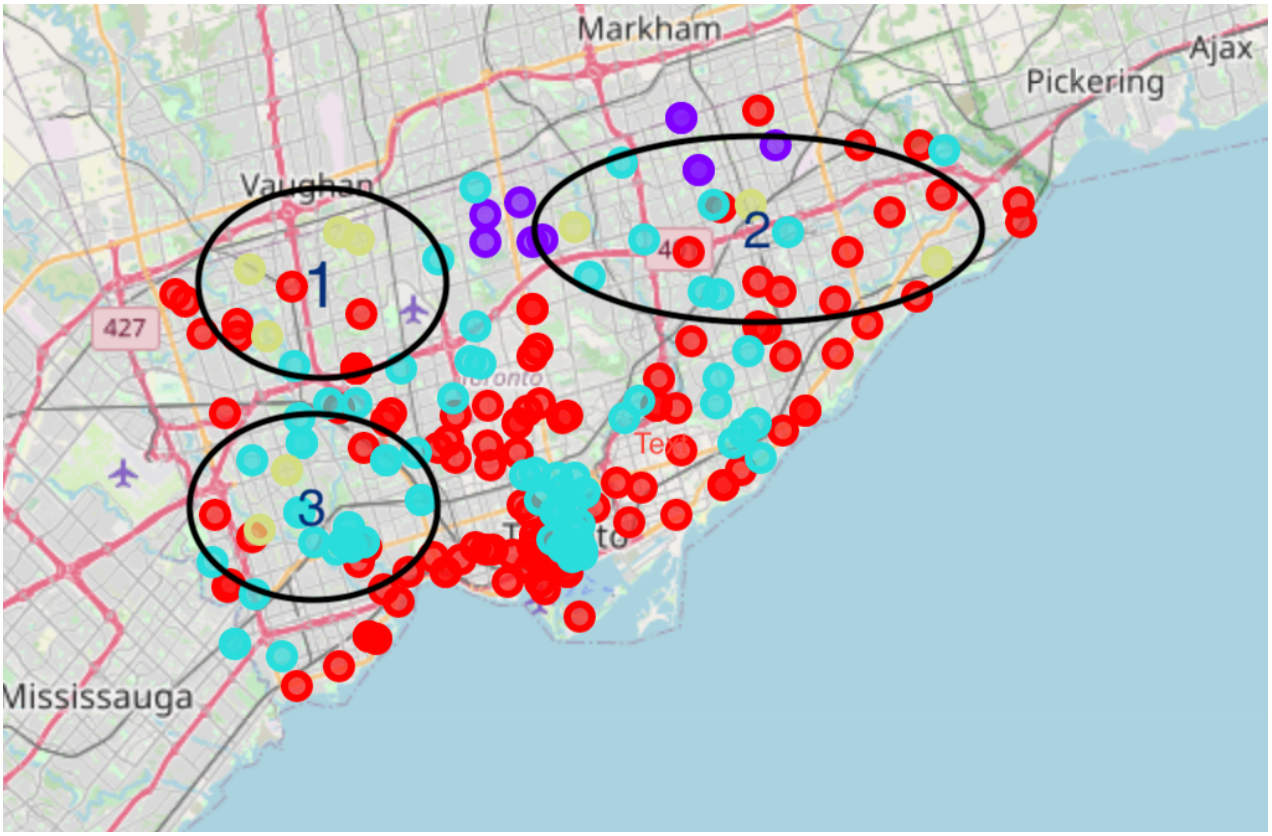

Figure 3. Screenshot of Clustering Results (Marked)

From the brief observation of clustering results, Cluster 3 (Yellow mark on **Figure 1**) is the best choice of location for the new milk tea & street food store because it has the highest frequency of occurrence for shopping malls and very low frequency of occurrence for a bubble tea shop. By targeting the new milk tea & street food store near Cluster 3, my friend (the new store owner) can expect a large customer flow and less competition as the shopping mall attracts plenty of consumers. With less

concentration of similar bubble tea shops and high quality of product, there is a chance that my friend's new milk tea store will be popular.

In **Figure 3**, the neighborhoods in Cluster 3 (Yellow mark on **Figure 3**) are divided into 3 groups further (circled in black and marked with a number on the map):

- Group 1
- Group 2
- Group 3

By taking the Toronto Population Density map (**Figure 2**) into consideration, it would be preferable to build the new milk tea & street food store in Group 2 area as it covers a higher population density area compared to the other two groups. Group 2 roughly contains three neighborhoods (three yellow dots in **Figure 3**). We can select further among these three neighborhoods area based on more factors like rent and traffic.

6. Conclusion

In this project, we went through the exploratory data analysis process to select the location for my friend's new milk tea & street food store in Toronto, Canada. We describe the business interests and problem, specify the data required, extract and prepare the data, clean the data, conduct machine learning by doing K-Mean data clustering based on the similarity of data, in the end with the help of the Toronto Population Density map, we are able to provide recommendations to my friend for the location of new milk tea & street food store. The recommendation to my friend is to choose the new milk tea & street food store from the Group 2 neighborhoods in Cluster 3.

This should act as a wonderful reference when my friend takes more factors into consideration in the future for the new store (i.e. rent, traffic, safety, and household income). By collecting more data in future research, we will be able to use more machine learning techniques (i.e. regression, classification) and methodology to locate the new milk tea & street food store.