IBM Data Science Final Capstone Project

Restaurant Recommendations for University of Toronto Students

Christopher Cheng

July 9th, 2020

Table of Contents

Introduction 3
Business Problem 3
Target Audience 3
Data 4
Data Acquisition 4
Methodology 4
Major Assumption in the Methodology 5
Limitations of this Methodology 9
Results 10
Discussion and Recommendations 12
Conclusion 13
References 14

Introduction

The University of Toronto (UofT) is regarded as one of the most highly ranked school in all of Canada, placing first in all of Canada and top 20 in the world according to the prestigious *Times Higher Education* [1]. Being such a highly regarded institution, it is clear why many students in Toronto and across the world take the leap to study here and proudly wear their UofT sweaters and hoodies! The University of Toronto has 3 campuses across Toronto (St. George, Scarborough, Mississauga) where the most popular campus is St. George, located in the heart of Downtown Toronto.

During my studies at the St. George campus, I found myself often asking what I want to have for lunch and what to do after I complete my exams. Even as a Toronto native myself, I found myself being confused with the sheer number of restaurants and venues located so close by! I find myself asking this question, as well as many of my peers when I ask what they want to eat for lunch. It is nice to have variety and have a nice pace of change every so often, but many people like to stick to their one favourite spot because of the fear of spending money on bad food (the worst feeling!). How can students reliably choose restaurants worthwhile while studying in Downtown Toronto? For those studying abroad and new to the area, what are some highly recommended places for them to eat from?

Business Problem

For this project, I would like to create a guide that can help freshmen's make healthy and delicious changes to their diets while studying at the University of Toronto. After all, a statistic of 8480 undergraduates have joined the University of Toronto St. George campus during the Fall 2019-20 semester alone. Accounting with the fact that over 23000 international students [2], it can be very hard for freshman and international students to navigate through Downtown Toronto and happily experiment with new foods. **The goal of the guide is to be able to tell students at the University of Toronto (St. George campus) which restaurants are best to eat at in the Downtown area, categorized by their food type.**

Target Audience

This project is hypothetically intended to be a guide perhaps in a Frosh handbook to help freshmen students, those unfamiliar with downtown, or even people wanting to just try something new!

Data

To perform the analysis to make the recommendation for students, I will need the following information:

1. Geo-coordinates of Downtown Toronto. Will need coordinates to pass to the Foursquare API so it knows where to search.
2. Top venues in Downtown Toronto. Return the venue's name, location, type of cuisine, and number of likes

Data Acquisition

To acquire the data necessary, I will use the following methods:

1. Geo-coordinates of UofT will be obtained using the Geocoder API via Google service We can confirm and be precise by searching the coordinates online.
2. The top venues in Downtown Toronto will be obtained using Foursquare through an API. Through the API, we can extract DT venue's names, ID, location, category, and number of likes

Methodology

To accomplish what I needed to solve, I first have to obtain the coordinates of the University of Toronto using Google's API Geolocator. To ensure the API returned the correct location, I call Folium to return a map of those coordinates.

Geocoder returned: (43.6607225, -79.39591980951508)

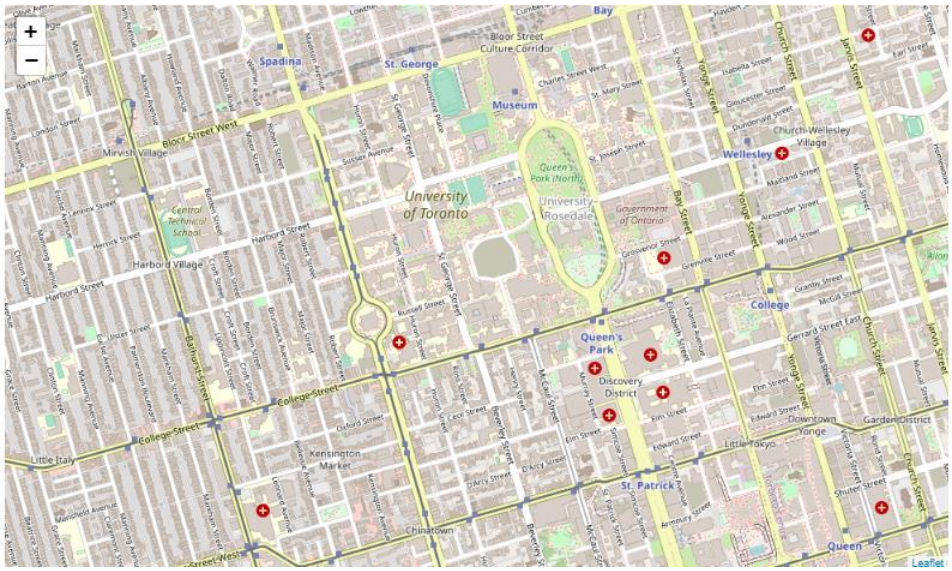

Figure 1 Map of the University of Toronto Campus (Zoom Level 15)

Having this information, we can now collect venue information close in the vicinity by calling Foursquare and accessing its information it has instore. When calling our API, we set a radius of

1km to allow students across all ends of the campus to easily choose a restaurant close to them.

Most importantly we include an explicit **search query "food"** when calling our API to only retrieve restaurants, and not other recreational venues such as gyms or parks.


```
{'meta': {'code': 200, 'requestId': '5f076c6c64a5d52760bdcc2b'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'University of Toronto',
  'headerFullLocation': 'University of Toronto, Toronto',
  'headerLocationGranularity': 'neighborhood',
  'query': 'food',
  'totalResults': 186,
  'suggestedBounds': {'ne': {'lat': 43.66972250900001,
    'lng': -79.38350247343162},
   'sw': {'lat': 43.65172249099999, 'lng': -79.40833714559854}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4aeb711ef964a52017c221e3',
       'name': 'Vegetarian Haven',
       'location': {'address': '17 Baldwin St',
```


Figure 2 Snippet code of the returned JSON file from Foursquare

With the returned JSON file, we clean, extract certain columns we want, and structure it into a panda's data frame.

	name	id	categories	lat	lng
0	Vegetarian Haven	4aeb711ef964a52017c221e3	Vegetarian / Vegan Restaurant	43.656016	-79.392758
1	Yasu	5362c366498e602fbe1db395	Japanese Restaurant	43.662837	-79.403217
2	Rasa	527d450111d25050de4ea0d8	Restaurant	43.662757	-79.403988
3	Blackbird Baking Co	535163cf498ea10a3b9582b5	Bakery	43.654764	-79.400566
4	Seven Lives - Tacos y Mariscos	50427a03e4b08d9f5931f593	Mexican Restaurant	43.654418	-79.400545

Figure 3 Snippet of Pandas Data frame of the Restaurants near UofT

Major Assumption in the Methodology

The goal of this project is to recommend students a good place to eat on campus such that they do not have to worry or be skeptical if the food will be good. My assumption for this model is that the number of likes, also known as the number of Foursquare users who actively recommended the venue is the sole criteria for the restaurant's quality. The assumption I am making is that the more likes a restaurant receives, the greater the quality experience students will have during their meal!

With this assumption in mind, we will need to retrieve the number of likes each restaurant has by calling Foursquare for each venue we have in our data frame. We would call Foursquare with each venue's ID to retrieve its number of likes, and we store it in our data frame.

	name	id	categories	lat	lng	likes
0	Vegetarian Haven	4aeb711ef964a52017c221e3	Vegetarian / Vegan Restaurant	43.656016	-79.392758	58
1	Yasu	5362c366498e602fbe1db395	Japanese Restaurant	43.662837	-79.403217	46
2	Rasa	527d450111d25050de4ea0d8	Restaurant	43.662757	-79.403988	79
3	Blackbird Baking Co	535163cf498ea10a3b9582b5	Bakery	43.654764	-79.400566	62
4	Seven Lives - Tacos y Mariscos	50427a03e4b08d9f5931f593	Mexican Restaurant	43.654418	-79.400545	299

Figure 4 Venue information with Number of Likes appended

With our number of likes, we can analyze its distribution through a histogram. From figure 5, we can see that many restaurants have received less than 33 likes on Foursquare. To contrast, there's a small number of restaurants who have over 269 total likes on Foursquare. We can review these statistics more carefully so we can make some claims about a good restaurant.

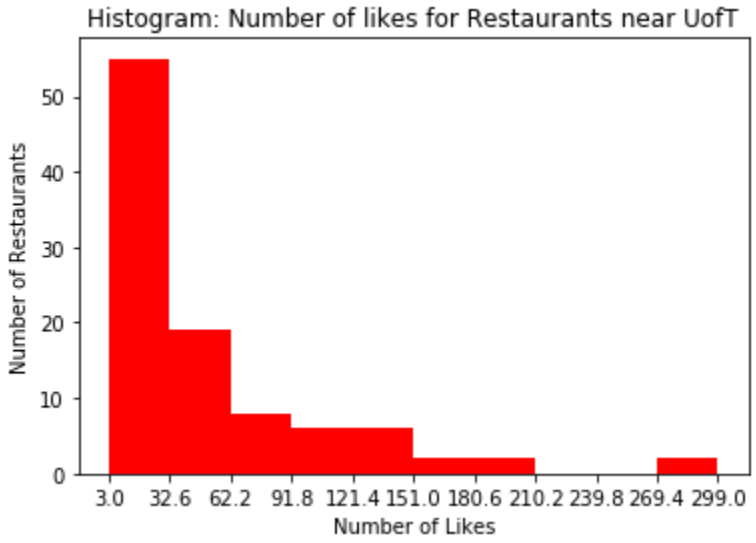

Figure 5 Histogram for Restaurants: Number of Likes

```
25%: 14.75 | 50%: 30.5 | 75%: 64.25
```

Figure 6 Detailed statistics for the Number of Likes Distribution

Using the percentile as the basis for categorizing and our metric for quality, we will set labels for each restaurant indicating how it rates according to its number of likes. The assumption in this methodology I will be using is declaring the bottom 25% of restaurants as "Not Enough Recommendations", 25-50% "Below Average Recommendations", 50-75% "Above Average Recommendations", and 75%+ "Very Recommended".

	name	id	categories	lat	lng	likes	Popularity
0	Vegetarian Haven	4aeb711ef964a52017c221e3	Vegetarian / Vegan Restaurant	43.656016	-79.392758	58	Above Average Recommendations
1	Yasu	5362c366498e602fbe1db395	Japanese Restaurant	43.662837	-79.403217	46	Above Average Recommendations
2	Rasa	527d450111d25050de4ea0d8	Restaurant	43.662757	-79.403988	79	Very Recommended
3	Blackbird Baking Co	535163cf498ea10a3b9582b5	Bakery	43.654764	-79.400566	62	Above Average Recommendations
4	Seven Lives - Tacos y Mariscos	50427a03e4b08d9f5931f593	Mexican Restaurant	43.654418	-79.400545	299	Very Recommended

Figure 7 Categorizing Restaurants based on Popularity

After popularity, I chose to categorize Foursquare's predefined categories with more general categories. For example, Japanese restaurants and sushi places are categorized under a new category, Asian.

	name	id	categories	lat	lng	likes	Popularity	Combined Categories
0	Vegetarian Haven	4aeb711ef964a52017c221e3	Vegetarian / Vegan Restaurant	43.656016	-79.392758	58	Above Average Recommendations	Other_Food
1	Yasu	5362c366498e602fbe1db395	Japanese Restaurant	43.662837	-79.403217	46	Above Average Recommendations	Asian_Food
2	Rasa	527d450111d25050de4ea0d8	Restaurant	43.662757	-79.403988	79	Very Recommended	Other_Food
3	Blackbird Baking Co	535163cf498ea10a3b9582b5	Bakery	43.654764	-79.400566	62	Above Average Recommendations	Snacks_Food
4	Seven Lives - Tacos y Mariscos	50427a03e4b08d9f5931f593	Mexican Restaurant	43.654418	-79.400545	299	Very Recommended	Latin_Food

Figure 8 Column to group each restaurant to a general category

Using the general categories as well as the popularity, we convert the columns and its values into dummy variables via the one hot encoding method. Thus, in our "onehot" data frame, we will have the combined categories and the popularity of the restaurant in the form of dummy variables.

Before we can move with our clustering model with the k-means algorithm, we have to determine the optimal level of clusters we should decide for our model. Although we're not entirely predicting anything, we can still determine its performance using a metric called the Silhouette Score, which gives a score depending on how far the inter-clusters are from each other, and how far the intra-clusters are from each other.

Plotting the silhouette score with our current model:

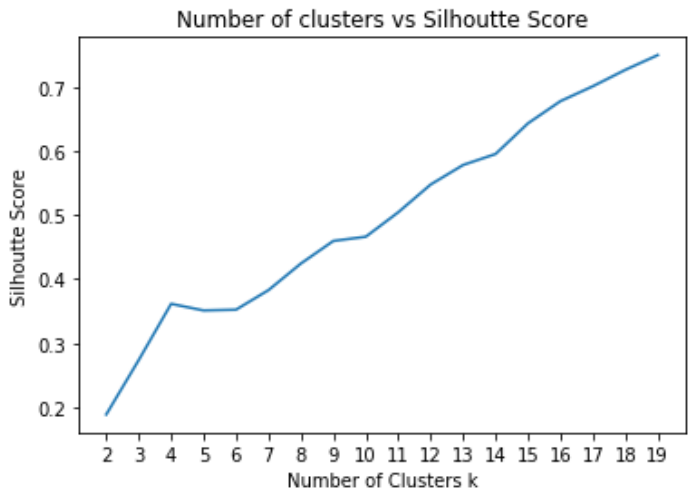

Figure 9 Silhouette Score vs. Number of Clusters: Combined Categories

From figure 9, we can see that there's an issue in our model because the silhouette score keeps increasing as k-clusters increases. There should be an elbow point to indicate the optimal number of clusters, but even at k = 20, the trend tends to keep increasing!

Thus, my assumption to categorize categories is wrong, and I will need to revert these changes. This is most likely because with so few columns and generalizing categories on my empirical basis, it is hard for the algorithm to accurately cluster these restaurants.

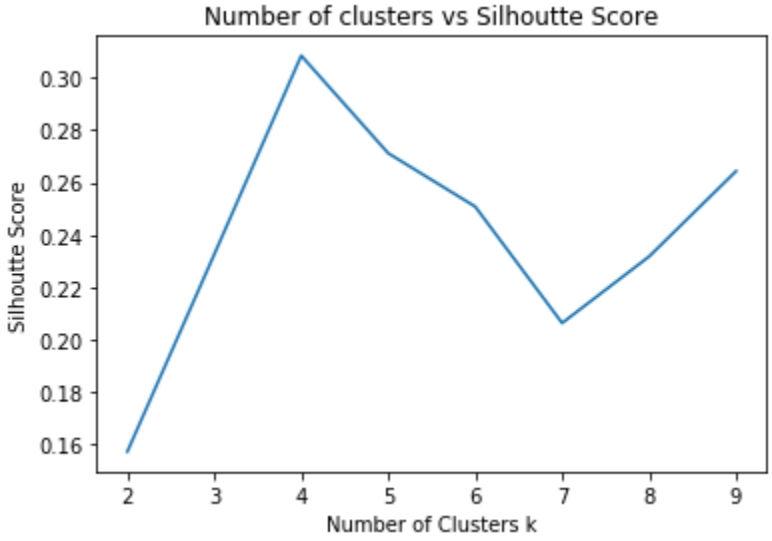

Figure 10 Silhouette Score vs Number of Clusters - All Categories

From figure 10, we can see a more reasonable Silhouette trend where an elbow point does occur. Taking k-cluster with the highest silhouette score, we will use k = 4 to fit our k-means model.

After applying k = 4 in our K-means algorithm, we can extract the labels the algorithm derived and tag it back to our data frame.

	name	categories	lat	lng	likes	Popularity	Cluster Label
0	Vegetarian Haven	Vegetarian / Vegan Restaurant	43.656016	-79.392758	58	Above Average Recommendations	2
1	Yasu	Japanese Restaurant	43.662837	-79.403217	46	Above Average Recommendations	2
2	Rasa	Restaurant	43.662757	-79.403988	79	Very Recommended	1
3	Blackbird Baking Co	Bakery	43.654764	-79.400566	62	Above Average Recommendations	2
4	Seven Lives - Tacos y Mariscos	Mexican Restaurant	43.654418	-79.400545	299	Very Recommended	1

Figure 11 Cluster labels assigned to each restaurant

With all of our restaurants assigned a cluster label, we can plot it using Folium.

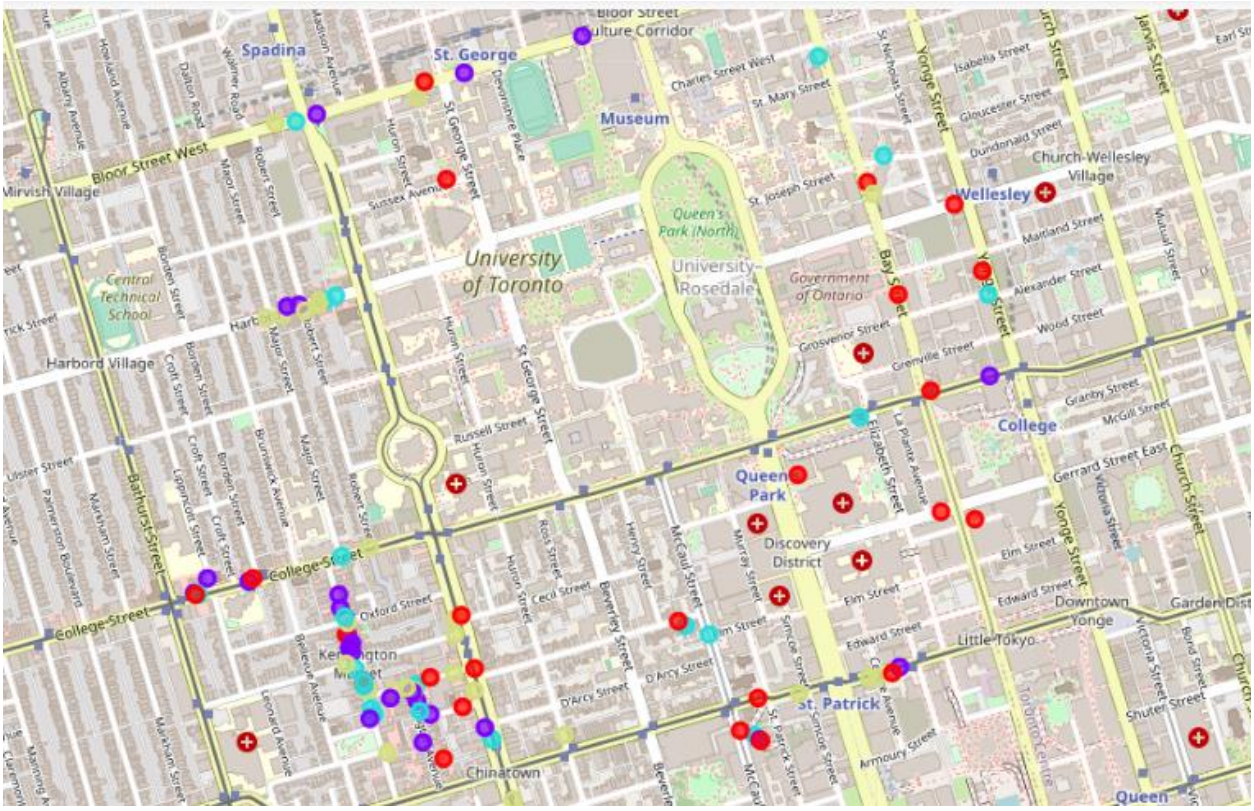

Figure 12 Folium map with 4 clusters

Finally, we can examine our clusters and understand trends the algorithm has deducted in the Results Section.

Limitations of this Methodology

Foursquare does indeed have ratings which are better suited to determine the quality of a restaurant since number of likes is used more of a popularity/trendy metric rather than qualitative. The problem is that accessing a venue's ratings and other detailed information are classified as premium calls, which I am unable to obtain.

Results

Understanding the clutters we obtained from the K-means algorithm, we can deduct a lot of information about how it classified data.

Reviewing Cluster 1 and 4

With cluster 1 and 4, we can see that the histogram shows most restaurants within this cluster are not popular or recommended enough on Foursquare.

	name	categories	lat	lng	likes	Popularity	Cluster Label
15	Sambuca Grill	Italian Restaurant	43.656110	-79.392946	10	Not Enough Recommendations	0
20	Anh Dao	Vietnamese Restaurant	43.656217	-79.399265	13	Not Enough Recommendations	0
24	Saigon Lotus Restaurant	Vietnamese Restaurant	43.654311	-79.399225	9	Not Enough Recommendations	0
27	Livelihood Cafe	Café	43.655821	-79.402629	11	Not Enough Recommendations	0
38	Xam Yu	Chinese Restaurant	43.655108	-79.398882	8	Not Enough Recommendations	0
52	Innis Cafe	Café	43.665401	-79.399715	6	Not Enough Recommendations	0
54	Somethin' 2 Talk About	Middle Eastern Restaurant	43.658395	-79.385338	6	Not Enough Recommendations	0
55	Dipped	Donut Shop	43.654920	-79.400154	9	Not Enough Recommendations	0

Figure 13 Data frame with Cluster Label = 0

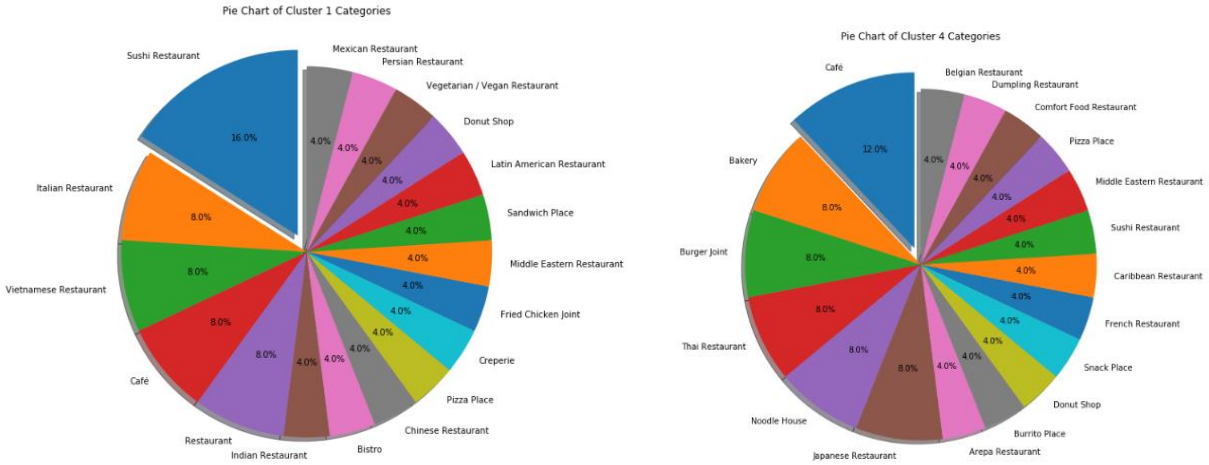

Figure 14 Pie Chart of Categories in Cluster 1 and Cluster 4

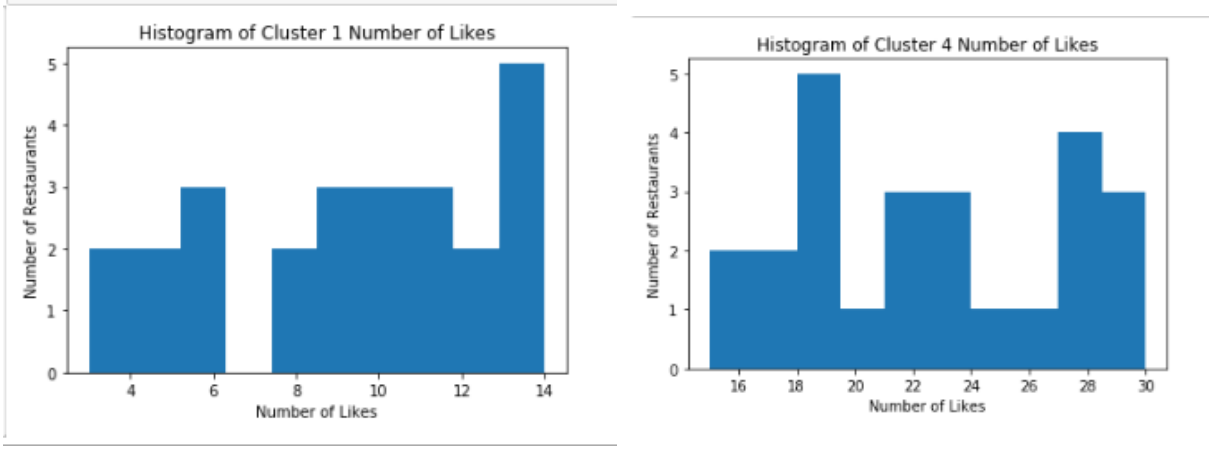

Figure 15 Histogram of Number of Likes Cluster 1 and Cluster 4

Reviewing these two clusters, we can ignore the restaurants within these clusters.

Cluster 2 and 3

	name	categories	lat	lng	likes	Popularity	Cluster Label
2	Rasa	Restaurant	43.662757	-79.403988	79	Very Recommended	1
4	Seven Lives - Tacos y Mariscos	Mexican Restaurant	43.654418	-79.400545	299	Very Recommended	1
5	Otto's Berlin Döner	Doner Restaurant	43.656387	-79.402788	121	Very Recommended	1
10	The Moonbean Cafe	Café	43.654147	-79.400182	150	Very Recommended	1
11	El Trompo	Mexican Restaurant	43.655832	-79.402561	99	Very Recommended	1
13	Hibiscus	Vegetarian / Vegan Restaurant	43.655454	-79.402439	89	Very Recommended	1
14	Jimmy's Coffee	Café	43.654493	-79.401311	200	Very Recommended	1

Figure 16 Cluster 2 data frame

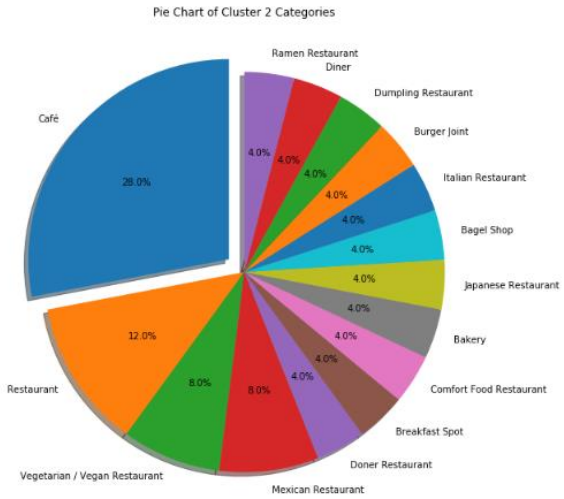

Figure 17 Pie chart of Cluster 2

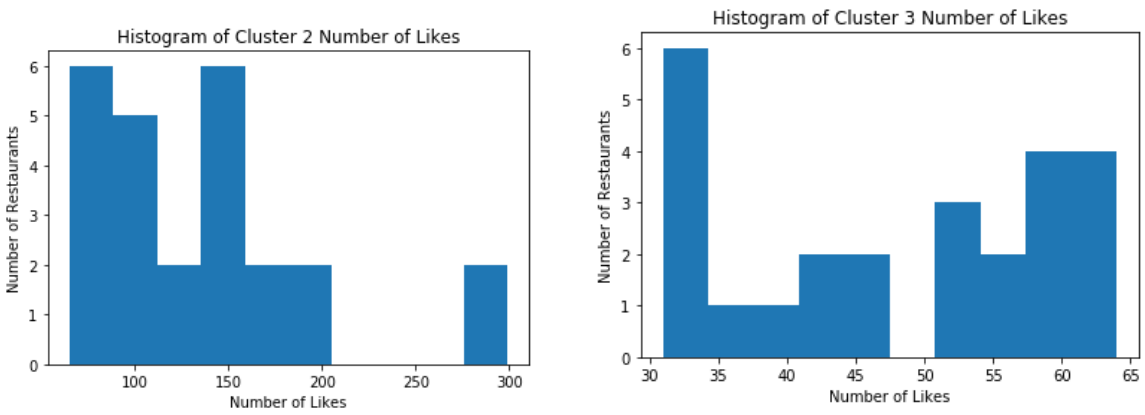

Figure 18 Histogram of Cluster 2 and 3

From the histogram, we can see that cluster 2 and 3 have fair amount of likes and recommendations from Foursquare users! In particular in cluster 2, we can see that the cafes near the University of Toronto are well received and liked by people!

Discussion and Recommendations

Based on what we have learnt from the clusters and according to our metric for evaluation, we can reliably recommend any restaurant from clusters 2 and 3 if a student is looking for somewhere to eat. I grouped restaurants by categories to recommend the highest liked restaurant, depending on what they're feeling that day (see Figure 19)!

From the cluster's histogram, it's clear to see that the K-means algorithm categorized the restaurants dependent on the empirical labels we set based on their number of likes. If I were to do this experiment again, I would like to see how the algorithm creates clusters if I did not create those attributes, or if the number of clusters have increased or decreased. But with k = 4, the clusters are extremely intuitive!

Another thing I would like to know if I had the resource is if I sorted the restaurants by ratings rather than their number of likes. Ratings are more truthful considering they are average and weighted (but they do have their downfalls). On the other hand, rating their quality by their number of likes is kind of like a popularity contest, where only the positive responses are shown, and the negative critical comments are not revealed.

Categories	Name	Latitude	Longitude	Likes	Popularity	Cluster label
Bagel Shop	Nu Bügel	43.655547	-79.402528	79	Very Recommended	1
Bakery	Wanda's Pie in the Sky	43.656163	-79.400566	136	Very Recommended	2
Breakfast Spot	Karine's	43.657002	-79.390743	162	Very Recommended	2
Burger Joint	The Burgernator	43.655642	-79.402440	118	Very Recommended	1
Burrito Place	Burrito Bandidos	43.662962	-79.383956	57	Above Average Recommendations	2
Café	Voodoo Child	43.667963	-79.388873	200	Very Recommended	2
Caribbean Restaurant	Rasta Pasta	43.654207	-79.400469	58	Above Average Recommendations	2
Comfort Food Restaurant	The Dirty Bird Chicken + Waffles	43.655109	-79.400674	70	Very Recommended	2
Creperie	Crêpes à GoGo	43.666609	-79.404061	52	Above Average Recommendations	2
Diner	Fran's	43.661255	-79.383893	167	Very Recommended	1
Doner Restaurant	Otto's Berlin Döner	43.656387	-79.402788	121	Very Recommended	1
Dumpling Restaurant	Dumpling House	43.653860	-79.398558	139	Very Recommended	1
Empanada Restaurant	Jumbo Empanadas	43.654831	-79.402098	51	Above Average Recommendations	2
Fish & Chips Shop	Fresco's Fish & Chips	43.654145	-79.401803	33	Above Average Recommendations	2
Hot Dog Joint	Fancy Franks	43.657480	-79.402733	62	Above Average Recommendations	2
Indian Restaurant	The Host	43.669631	-79.395599	32	Above Average Recommendations	2
Italian Restaurant	Pizzeria Via Mercanti	43.662949	-79.387664	71	Very Recommended	2
Japanese Restaurant	Yasu	43.662837	-79.390613	103	Very Recommended	2
Mexican Restaurant	Torteria San Cosme	43.655832	-79.400545	299	Very Recommended	2
Middle Eastern Restaurant	The Pomegranate	43.656673	-79.406900	47	Above Average Recommendations	2
Persian Restaurant	Sheherzade	43.656659	-79.406969	40	Above Average Recommendations	2
Ramen Restaurant	Sansotei Ramen 三草亭	43.655157	-79.386501	205	Very Recommended	1
Restaurant	Rasa	43.668390	-79.395730	79	Very Recommended	1
Steakhouse	Morton's The Steakhouse	43.669607	-79.394666	36	Above Average Recommendations	2
Sushi Restaurant	Tokyo Sushi	43.665885	-79.386977	32	Above Average Recommendations	2
Vegetarian / Vegan Restaurant	Vegetarian Haven	43.666755	-79.392758	294	Very Recommended	2

Figure 19 List of Recommended Restaurants by Cuisine Type

Conclusion

This report was a fun hypothetical project that allowed me to utilize what I have learnt throughout this program in a real-world scenario! I like that I was able to solve and work with a problem I have ever since I was a freshman. It was a good experience to be able to utilize data analysis tools such as Pandas and Scikit-Learn to a practical problem, and I am glad (in retrospect) to struggle and complete an activity that does not have a clear solution ahead of me. Thank you for taking the time to read this!

Code

https://github.com/ItsMrTurtle/Coursera_Capstone/blob/master/IBM%20Data%20Science%20Final%20Capstone%20Project.ipynb

References

[1] https://www.utoronto.ca/news/u-t-ranked-first-canada-among-world-s-top-20-universities-times-higher-education

[2] https://www.utoronto.ca/about-u-of-t/quick-facts

[3] https://www.utoronto.ca/contacts