Optimal Distribution Network for BB Ice Company in North York, Ontario

Coursera IBM Data Science Capstone Project Report

Mayank Panwar

The University of XYZ, ABC

June 2020

123 Publishers

Table of Contents

Introduction

Background

After their huge success in Sydney, an Australian startup “BB Ice Company” decided to expand their business overseas. As one of the founding members of this company was from North York, Ontario, Canada they decided to set up their manufacturing and distribution center there. There is a saying - “Ice is the equivalent to a stove for a chef”. Heat, and Cold, both change every aspect of the final product.

The company makes crystal clear ice which is in demand at many restaurants, bars, pubs, diners, lounges, and nightclubs serving drinks like liquor, cocktails, and mocktails. The temperature affects how the drink feels, by numbing the taste buds. They say that a big cube of perfectly clear, hand-cut ice, is clearly the perfect serve, and won’t overly dilute the drink. Also, cooler temperatures can smooth an average liquor by rounding its edges.

Problem

The company wants to make an optimal distribution network in North York to achieve the timely delivery of Ice pops/cubes/slabs to their potential customers. For this, the stakeholders want to segment all the presumable targets into clusters and then assign each cluster a cold storage truck for better and efficient delivery. In this project, we will cluster the venues based on their GPS coordinates and then will create an information chart for each cluster and map those clusters on North York Map so that the stakeholders can make informed decisions.

Audience

The stakeholders of the company will benefit from this project. The result will help improve their business activities and optimize its distribution network. Other companies that want to optimize their distribution network using similar data analysis and cluster modeling may also be interested in this project.

Data Source

1. We will need a list of all the neighborhoods in North York, Ontario, Canada. The list can be extracted from the Wikipedia Page “List of postal codes of Canada: M” having URL:https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. For web scraping, we will use the BeautifulSoup4 Python package and for data wrangling, we will use the Pandas Python package.

   Example:

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M6A	North York	Lawrence Manor, Lawrence Heights
3	M3B	North York	Don Mills
4	M6B	North York	Glencairn

2. We will also need the GPS coordinates of each neighborhood. For this, we will use the PGeoCoder package that can give the query results based on postal codes.

   Example:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.7545	-79.3300
1	M4A	North York	Victoria Village	43.7276	-79.3148
2	M6A	North York	Lawrence Manor, Lawrence Heights	43.7223	-79.4504
3	M3B	North York	Don Mills	43.7450	-79.3590
4	M6B	North York	Glencairn	43.7081	-79.4479

3. We also need the list of all the pubs, bars, restaurants, nightclubs, diner, and lounge for each neighborhood within the radius of 1.5kms. This data will be extracted from FourSquare Server using FourSquare APIs.

   Example:

	PostalCode	Neighborhood	N_Latitude	N_Longitude	Venue_id	Venue	Venue_Category	V_Latitude	V_Longitude
0	M3A	Parkwoods	43.7545	-79.33	4b8991cbf964a520814232e3	Allwyn's Bakery	Caribbean Restaurant	43.759840	-79.324719
1	M3A	Parkwoods	43.7545	-79.33	4bd4846a6798ef3bd0c5618d	Donalda Golf & Country Club	Golf Course	43.752816	-79.342741
2	M3A	Parkwoods	43.7545	-79.33	4e8d9dcdd5fbbbb6b3003c7b	Brookbanks Park	Park	43.751976	-79.332140
3	M3A	Parkwoods	43.7545	-79.33	4b8ec91af964a520053733e3	Graydon Hall Manor	Event Space	43.763923	-79.342961
4	M3A	Parkwoods	43.7545	-79.33	4b149ea4f964a52029a523e3	Darband Restaurant	Middle Eastern Restaurant	43.755194	-79.348498

4. Features Selection: After data acquisition and cleaning, there were 10845 samples and 9 features in the second dataset containing query results based on the postal code of the neighborhood of North York. Upon examining the venue categories we found that only 6 categories were relevant for our project i.e. Pub, Diner, Bar, Lounge, Restaurants, and Nightclubs. These categories were selected because they can be a potential customer for BB Ice Company. Then this data frame was merged with the original data frame that was scrapped from the Wikipedia page and the final data set contained 2790 samples and 9 features. The relevant features were:

    a. PostalCode
    b. Neighborhood
    c. N_Latitude
    d. N_Longitude
    e. Venue
    f. Venue_Category
    g. V_Latitude
    h. V_Longitude
    i. Cluster_Label

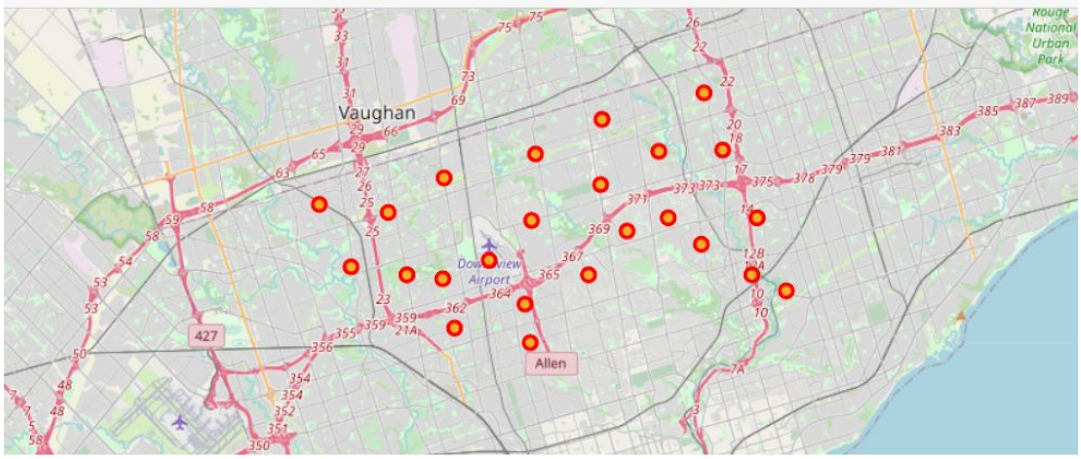
    Other features were dropped from the data set.

Methodology

Exploratory Data Analysis:

After data acquisition and cleaning we used the folium library to visualize all the neighborhoods on the map of North York. GPS coordinates of neighborhoods were used to superimpose their location on the top of the North York map.

Fig 1. Location of all the Neighborhoods in North York

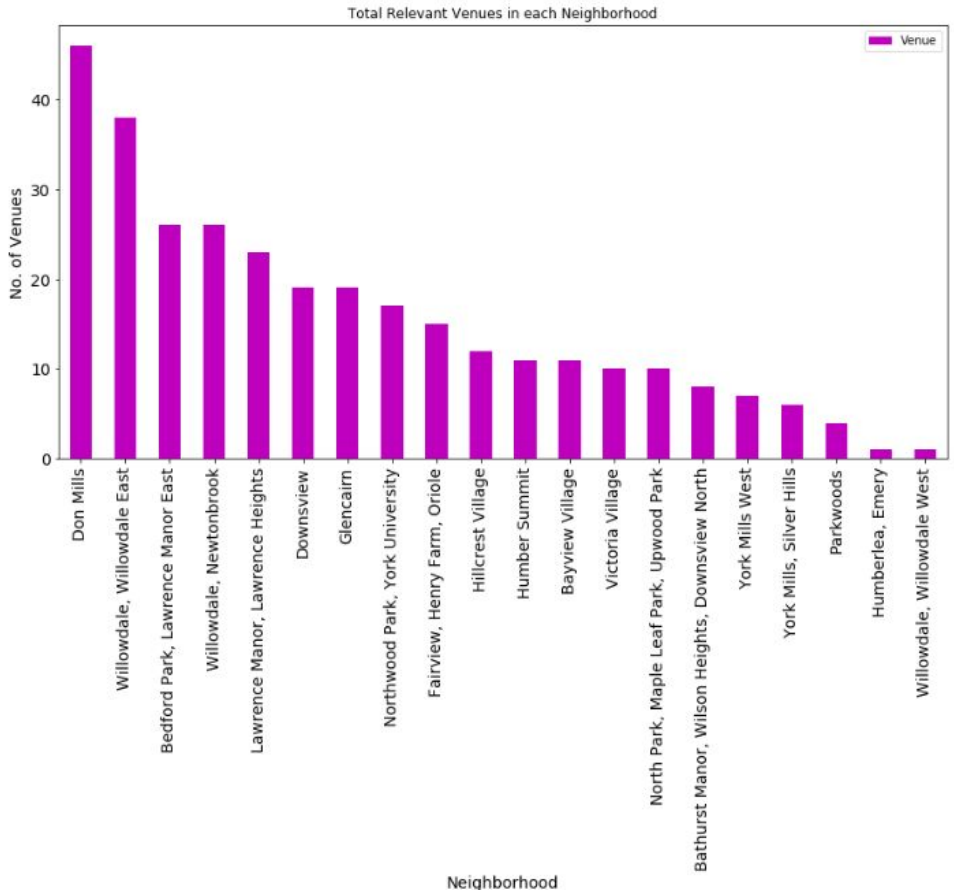


Then FourSquare API was used to explore these neighborhoods. A query was made on the FourSquare server for a list of top 500 venues near each neighborhood within a radius of 1.5kms using their latitude and longitude values. Here is the snippet of the result dataset:

	PostalCode	Neighborhood	N_Latitude	N_Longitude	Venue_id	Venue	Venue_Category	V_Latitude	V_Longitude
0	M3A	Parkwoods	43.7545	-79.33	4b8991cbf964a520814232e3	Allwyn's Bakery	Caribbean Restaurant	43.759840	-79.324719
1	M3A	Parkwoods	43.7545	-79.33	4bd4846a6798ef3bd0c5618d	Donalda Golf & Country Club	Golf Course	43.752816	-79.342741
2	M3A	Parkwoods	43.7545	-79.33	4e8d9dcdd5fbbbb6b3003c7b	Brookbanks Park	Park	43.751976	-79.332140
3	M3A	Parkwoods	43.7545	-79.33	4b8ec91af964a520053733e3	Graydon Hall Manor	Event Space	43.763923	-79.342961
4	M3A	Parkwoods	43.7545	-79.33	4b149ea4f964a52029a523e3	Darband Restaurant	Middle Eastern Restaurant	43.755194	-79.348498

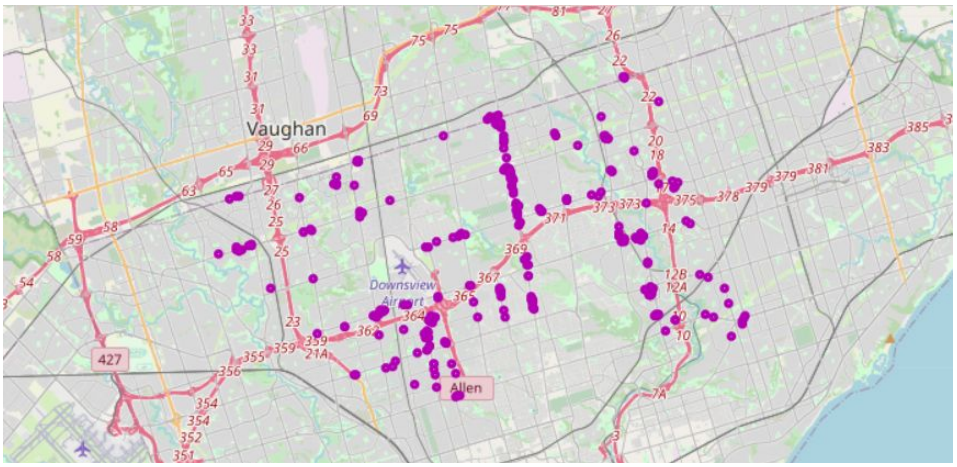
1205 venues were returned after this query out of which we only kept relevant venues and thus the total venues were reduced to 310.

Fig 2. Graph Showing total relevant venues for each neighborhood



The neighborhoods like Done Mills, Willowdale East, Bedford park contains 30+ venues each while Parkwoods, Humberlea, and Willowdale West contain less than 5 venues each. We will also superimpose all the filtered venues on the map of North York for a better Idea of their distribution all over North York.

Fig 3. Distribution of Filtered Venues like Bar, Pubs, Restaurant over North York



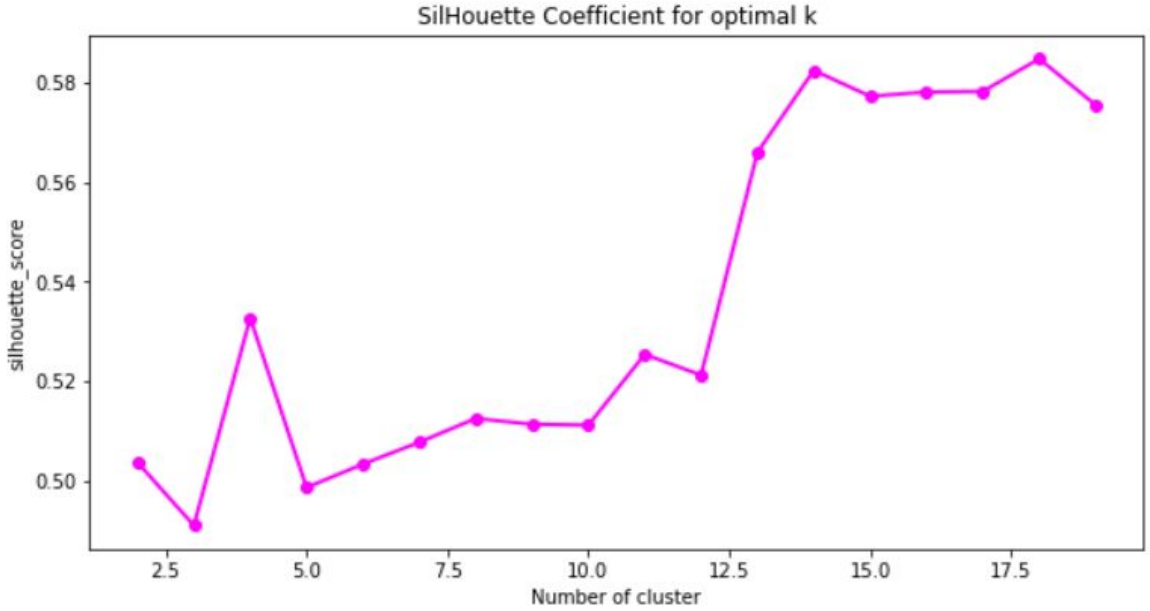
Modeling and Evaluation:

We have segmented these venues into clusters using an unsupervised clustering algorithm called Kmeans clustering. Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster. K-Means algorithm is one of the most common clustering methods of unsupervised learning. We will use the features V_Latitude, V_Longitude for this modeling.

We have made different models using different values of **k** and then to evaluated all the models and selected the best value of k for our model using -

1. Silhouette Coefficient- Silhouette analysis can be used to determine the degree of separation between clusters. We want the Silhouette coefficients to be as big as possible and close to 1 to have a good cluster. Let's check the maximum value of the coefficient using the line graph.

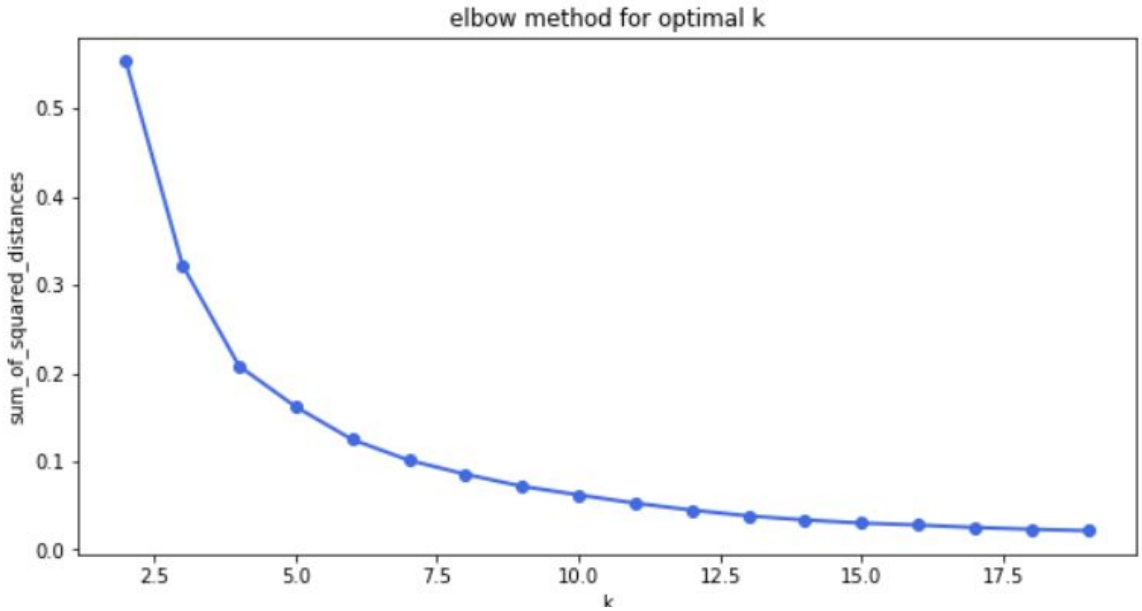
Fig 4. Relationship Between Silhouette Score and Number of Clusters.



Here we can see that k=14 and k=18 are both good candidates for the optimum value of k. As there isn't much difference between the score of both ks we will select the minimum of the two ie **k=14** for cost-efficient distribution.

2. Elbow Method- Elbow method gives us an idea of what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick k at the spot where SSE starts to flatten out and forming an elbow.

Fig 5. Relationship b/w Sum of Squared Distance & Number of Clusters (Elbow Method)



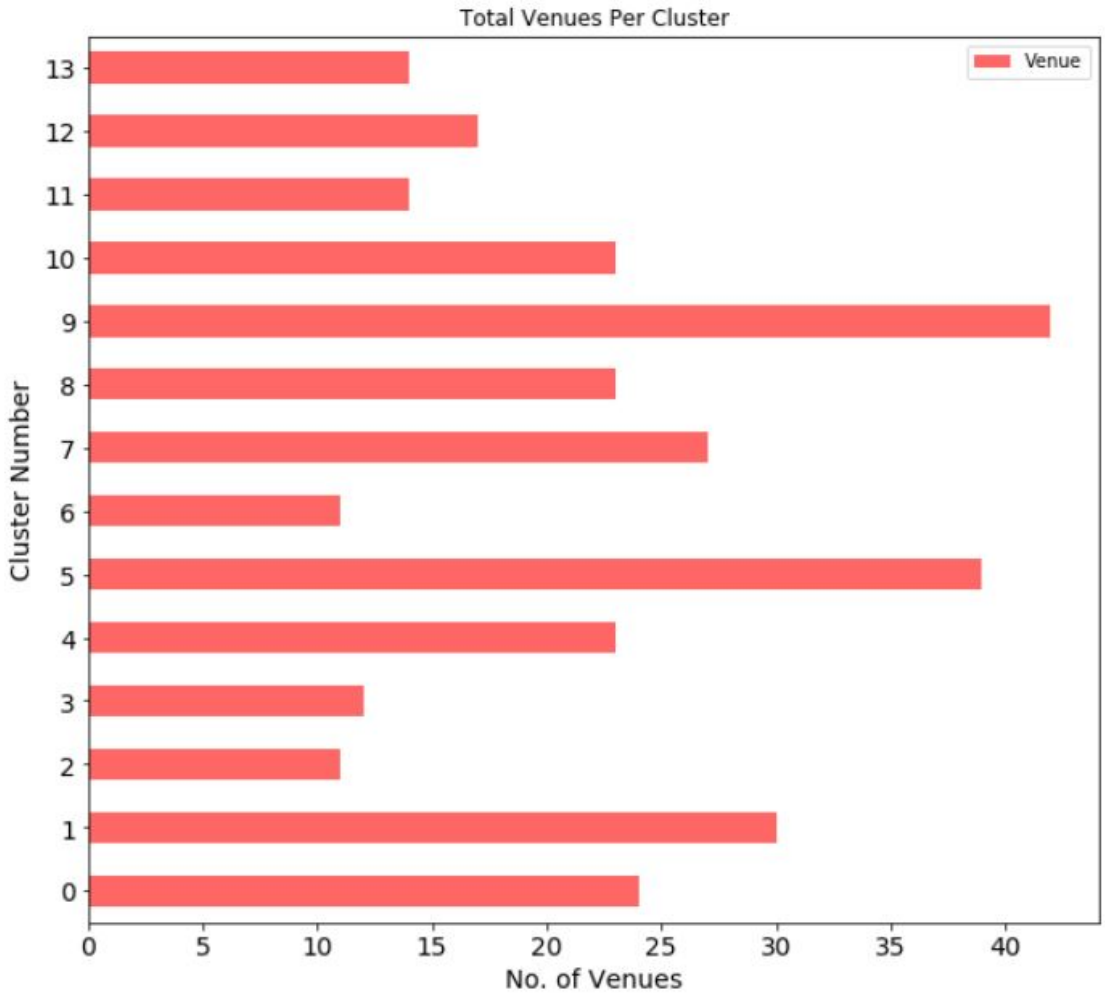
Here the curve is monotonically decreasing and does not show any elbow or has an obvious point where the curve starts flattening out. But looking closely, we can see that after **k=14** the curve slope is almost flat. So, **k=14** is a good choice. So after evaluating all the models using both Elbow Method and Silhouette Analysis we have finalized **k=14** for our Kmeans model. Then after running the Kmeans Model on V_Latitude and V_Longitude we have extracted the cluster labels and then assign these labels to our df_drinks data frame by making a new column named Cluster_Label. Here is the Snippet of the same:

	PostalCode	Neighborhood	N_Latitude	N_Longitude	Venue_id	Venue	Venue_Category	V_Latitude	V_Longitude	Cluster_Label
0	M3A	Parkwoods	43.7545	-79.3300	4b8991cbf964a520814232e3	Allwyn's Bakery	Caribbean Restaurant	43.759840	-79.324719	11
1	M3A	Parkwoods	43.7545	-79.3300	4b149ea4f964a52029a523e3	Darband Restaurant	Middle Eastern Restaurant	43.755194	-79.348498	1
2	M3A	Parkwoods	43.7545	-79.3300	58a8dcaa6119f47b9a94dc05	A&W	Fast Food Restaurant	43.760643	-79.326865	11
3	M3A	Parkwoods	43.7545	-79.3300	4bfc6313d6f2c9b6e9264fc8	China Gourmet	Chinese Restaurant	43.755189	-79.348382	1
4	M4A	Victoria Village	43.7276	-79.3148	550df684498ea2dd2c87bb5a	Jatujak	Thai Restaurant	43.736208	-79.307668	6

Result

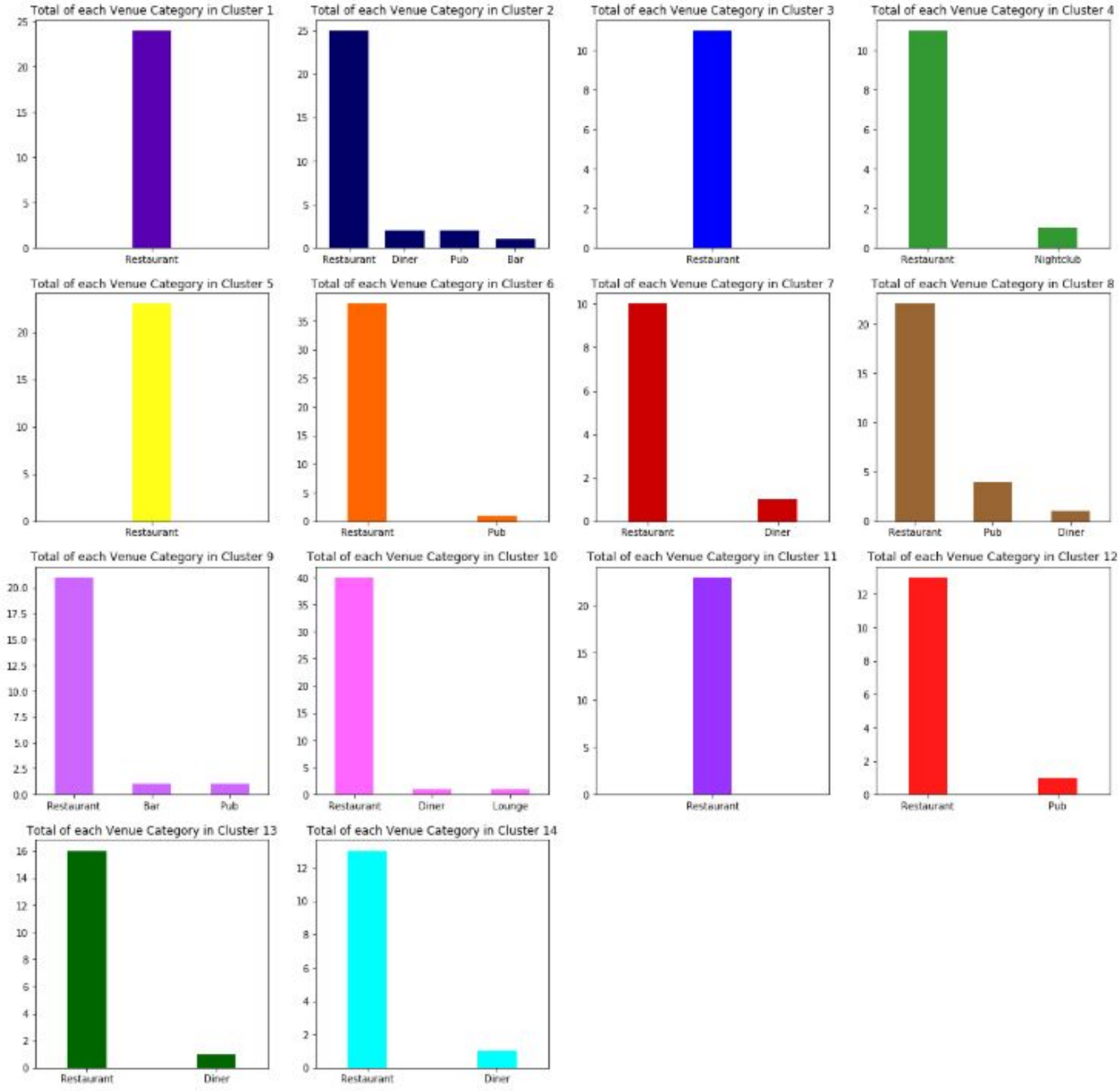
After data analysis and modeling, we have segmented all the relevant venues into 14 clusters. Here is the snippet of the total venues assigned to each cluster:

Fig 6. Total Venues Per Cluster



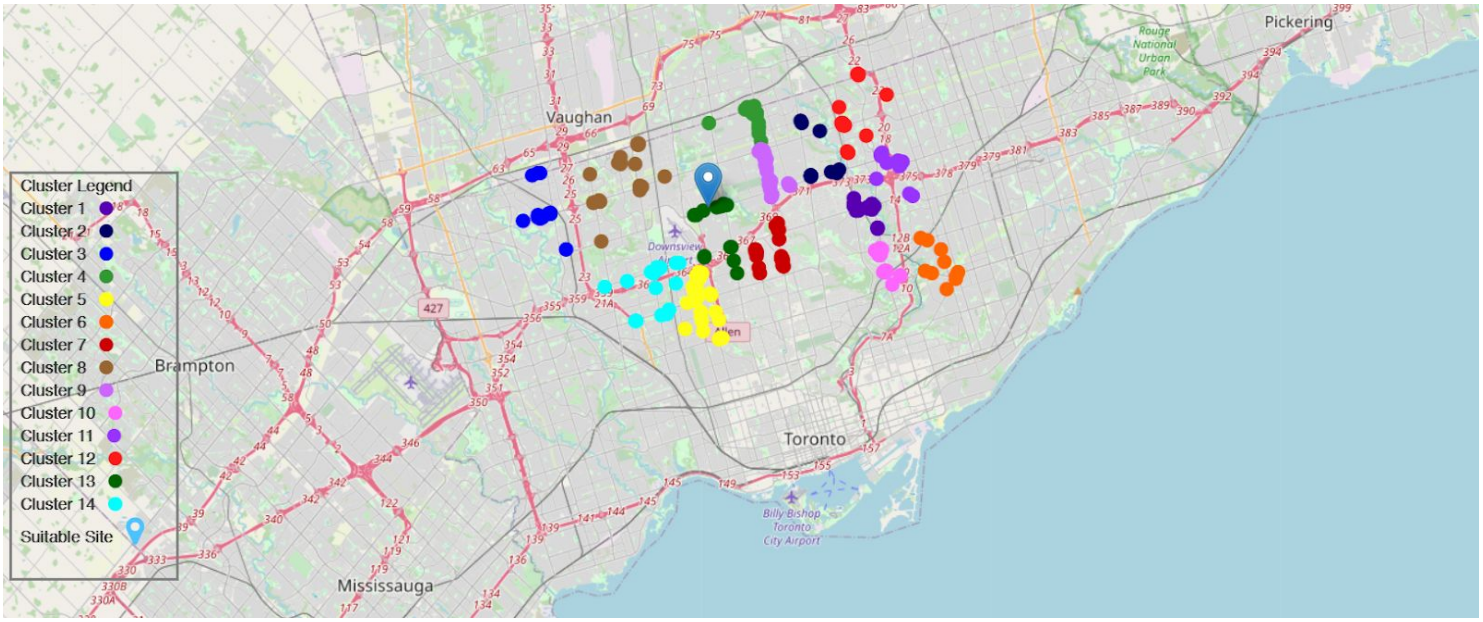
Also, to find out the total of each Venue_Category assigned to each cluster we made a Dataframe grouped by Cluster_Label and extracted each group and then replaced the venue categories having the word 'Restaurant' in it with 'Restaurant' like Chinese Restaurant or India Restaurant will be replaced by Restaurant. Then we made a new DataFrame grouped by Venue Category and found the count for each Category for that cluster. This process was repeated for each cluster label. Here is the collection of graphs visualizing the same :

Fig 7. Total of the unique Venue Categories for Cluster 1 to 14



Finally, we have superimposed these clusters on the Map of North York. Different color is assigned to each cluster.

Fig 8. Visualization of Venue Clusters on North York Map



Discussion

The startup, BB ice company has proved their idea in Sydney and now are looking for expanding their business in North York, Ontario, Canada. But as the area is new to them a proper strategy needs to be formulated to list all the potential customers, make optimum distribution networks and decide the best site to set up their manufacturing and distribution hub. Making a proper strategy will give them a competitive edge and will help them make efficient and timely deliveries. As the temperature and texture of the ice are very important to make the drinks and experience smooth it is very important that timely delivery of ice cubes or pops are done.

There are a total of 20 probable neighborhoods and a total of 310 probable customers distributed all over North York. There is a complexity due to such large numbers so to segment these customers/venues efficiently unsupervised Machine Learning can be used. There are different clustering approaches that can be tried to segment the probable customers so that one cold storage truck can be assigned to each segment.

We used the Kmeans algorithm with k=14 for this analysis. Then we assigned each venue a cluster label. We also used exploratory data analysis to see how many venues were assigned to each cluster. The result was satisfactory and this information can be used by the company to formulate their strategies to set up their business with an efficient distribution network in North York.

Finally, the analysis was ended by visualizing data using various types of graphs and clustering information on the map of North York.

Conclusion

Using the Kmeans algorithm as part of this clustering study when we evaluated the models using the Elbow method and Silhouette Coefficient, we found the optimum value of k to be 14. So to form the optimal distribution center in North York the stakeholders can segment the venues and make 14 zones for efficient delivery. Also as they can choose a location near coordinates 43.7615° N, 79.4111° W to set up their manufacturing plant and distribution warehouse as these coordinates are at the center of North York and all the distribution zones would be easily accessible from here.

Also, it would be best to assign each distribution zone a cold storage truck for the timely and efficient delivery of Ice cubes, pops, and slabs.

Not only for BB Ice company but such a model can be beneficial to many other companies that want to make their distribution network more efficient using similar data analysis and modeling.

References:

1. Wikipedia “List of postal codes of Canada: M”
2. Pgeocode
3. FourSquare API
4. Coursera - IBM Data Science