2019

IBM Applied Data Science Capstone Report

EVALUATING RESTAURANT LOCATIONS BY TORONTO WELLBEING HEALTHY FOOD INDICES AND VENUE AVERAGE HEALTH RATINGS

JIM HUEBNER, PH.D. MBA, CPA, MBSS

Contents

Introduction 2
Business Problem 2
Audience & Stakeholders 2
Data 2
Data Sources 2
Data Gathering & Cleansing 4
Data Feature Selections 6
Methodology 6
Data Preparation Choices 7
Choropleth Neighbourhood Maps 7
Neighbourhood HFI Choropleth 7
Neighbourhood Venue Average Health Ratings (VAHR) 9
Histograms of Wellbeing Toronto Neighbourhoods 10
Clustering of Neighbourhoods 12
Results 16
Choropleth Maps 16
Histograms 16
Clusters 16
Neighbourhoods and Venues 17
Discussion 18
Gaps among Neighbourhoods' HFI and VAHR 18
Market Penetration Strategy 19
Population/Demographics 19
Number of Venues 19
Top Venues 19
Neighbourhood Topographics 19
Recommendations 20
Business Problem 20
Future Directions 20
Conclusion 21

Introduction

This project seeks to identify health-conscious neighbourhoods in Toronto that may be under-served by healthy quick-serve restaurants (QSRs) and sit-in food options.

Business Problem

Demand for healthy food is growing. Healthy food chains are emerging to attempt to meet the demand for low-calorie ingredients and fresh produce.[1]

The challenge for the restaurant industry is balancing a variety of demand factors that sometimes compete, for example genuinely healthy and organic foods and low cost.[2]

Selections for new venue locations must target and effectively service the demand in health-conscious neighbourhoods. This project seeks to explore data insights specifically to identify restaurant locations in health-conscious neighbourhoods in Toronto that may be under-served by existing venues, and what are the factors that lead to such insights.

Audience & Stakeholders

The audience for this project is restauranteurs, investors and anyone servicing the QSR and sit-in food segments. These stakeholders have a vested interest in responding to market demand by making investment and operations decisions based on data insights. This project explores some of those data insights specific to venue location.

Data

This section describes the data sourced for this project, as well as the data cleansing and preparation for subsequent exploration.

Data Sources

This project sources and integrates data from Wellbeing Toronto as well as Foursquare data. This section describes each of these data sources and provides examples of the data.

Wellbeing Toronto data

Wellbeing Toronto (WbTo) provides neighbourhood-level datasets about Toronto services, facilities, and well-being. The datasets are segmented by community indicators grouped under 11 categories, including demographics, civics, and health as examples. While primarily a mapping application, and underlying datasets are downloadable.

[1] Garfield, Leanna (2017). "American fast food as we know it is dying — and healthier chains may be replacing it." Business Insider, November 2017. Available at https://www.businessinsider.com/future-of-fast-food-healthy-affordable-2017-11.

[2] Hardy, Kevin (2017). "9 Fast Food Trends for 2018." QSR, December 2017. Available at https://www.qsrmagazine.com/exclusives/9-fast-food-trends-2018.

For this project, the dataset of interest includes the Healthy Food Index (HFI) and Total Population indicators for each of Toronto's 140 neighbourhoods. These data are downloadable as a csv file and include the neighbourhood name and number. Wellbeing Toronto is publicly accessible at https://map.toronto.ca/wellbeing.

Wellbeing Toronto data sample

Following is a sample of the imported data showing the populations and HFI for each neighbourhood:

Index	Neighbourhood	Total Population	Healthy Food Index
0	West Humber-Clairville	33312.0	23.82
1	Mount Olive-Silverstone-Jamestown	32954.0	37.57
2	Thistletown-Beaumond Heights	10360.0	42.26
3	Rexdale-Kipling	10529.0	23.31
4	Elms-Old Rexdale	9456.0	24.71

Foursquare data

Foursquare provides a mobile app that allows users to search for near-by venues and see information and reviews. Users also feed information back to Foursquare both passively, as the app tracks users' locations, and actively as users enter venue names, locations, and reviews.

Since 2009, users have provided Foursquare with location data on over 105 million venues, with "over 75 million tips from local experts." As one of the largest sources of location-based venue data, the company describes itself as "a is a technology company that uses location intelligence to build meaningful consumer experiences and business solutions."[3]

This project will access Foursquare venue data for the selected "Wellbeing Toronto" neighbourhoods, specifically food venues. The Foursquare venue data will particularly seek to identify food venues that are categorized as *most healthy* and *least healthy*. These data will then be used for subsequent comparison and categorization to provide insight to the business problem.

The Foursquare venue data are accessible via application programming interface (API). A free developer account is used to access the data from https://developer.foursquare.com/places-api.

Foursquare data sample

Following is a sample of the imported data showing particularly the venues (by name) and the respective venue categories for each neighbourhood.

[3] Sourced from https://foursquare.com/about

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rosedale	43.679563	-79.377529	Rosedale Park	43.682328	-79.378934	Playground
1	Rosedale	43.679563	-79.377529	Whitney Park	43.682036	-79.373788	Park
2	Rosedale	43.679563	-79.377529	Alex Murray Parkette	43.678300	-79.382773	Park
3	Rosedale	43.679563	-79.377529	Milkman's Lane	43.676352	-79.373842	Trail
4	Cabbagetown, St. James Town	43.667967	-79.367675	Cranberries	43.667843	-79.369407	Diner

Data Gathering & Cleansing

Load Toronto Health Food Index data from Wellbeing Toronto

The HFI data from Wellbeing Toronto were downloaded as comma separated values (csv) and saved locally for subsequent access. Non-essential data columns were removed. The data included neighbourhood names but no geocodes.

Get HFI Neighbourhood Geocodes

Initial tests in obtaining geocodes of the Toronto Wellbeing data indicated that many neighbourhoods were not recognized by geocoding service (e.g. Nominatim[4]). This was due to the hyphenation of the neighbourhood names as shown in the sample data above.

Further testing revealed that the non-hyphenated segments were recognized at a higher rate than the hyphenated neighbourhood names. This led to a four-step process to obtain geocodes for the HFI data as following:

1) Split the neighbourhood names: each hyphenated segment was copied to its own row in the data set along with the corresponding population, neighbourhood id, and HFI. This resulted in 206 rows.

2) Get geocodes: the free Nominatim service of the openstreetmap project was used.

3) Eliminate neighbourhoods without geocodes: Rows in the dataset for which Nominatim could not find geocodes were deleted. 193 rows remained in the dataset.

4) Eliminate duplicate neighbourhoods: Only one of the rows sharing the same neighbourhood Id was retained, while the duplicated rows were retained. This ensures that WbTo neighbourhoods

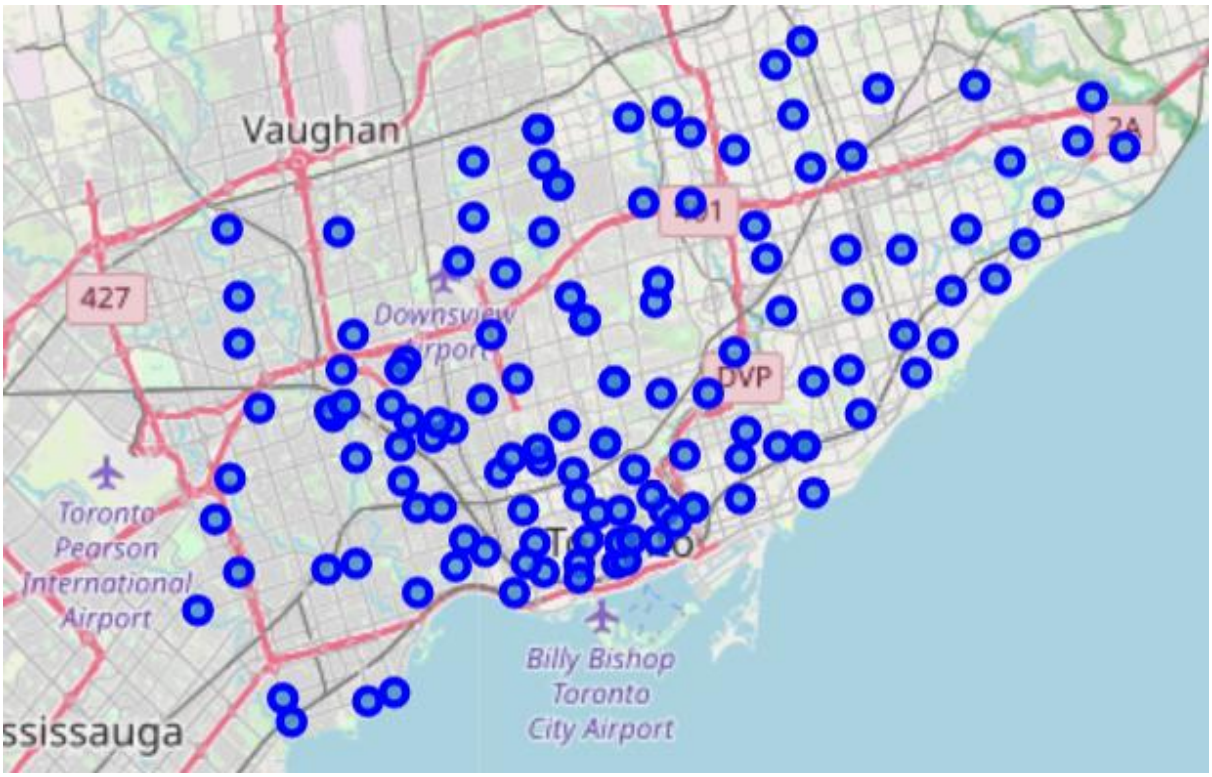
[4] https://nominatim.openstreetmap.org/

are included only once in the resulting dataset. 136 of the original neighbourhoods remained in the dataset, which is less than a 3% loss.

The resulting geocoded WbTo dataset ("dfh4") was saved to a tabular (csv) file which could then be accessed in subsequent notebook sessions without having to re-run the process. This dataset formed the basis for obtaining the Foursquare food venues.

View Toronto HFI Neighbourhood map

The following is a map of the Toronto Wellbeing neighbourhoods after geocoding.



This map shows the distribution of the neighbourhoods, and helps validate the geocoding process. This map also shows the broad distribution across Toronto, as well as some grouping toward the downtown (center bottom) and east of the Toronto airport.

Foursquare Food Venues and Health Ratings

The Foursquare application programming interface (API) was accessed to obtain the food venues in the resulting Wellbeing Toronto neighbourhoods from above. Also, each venue was given a health rating depending on the type of venue.

The process required as described as follows:

1) Obtain all venues and venue categories within 500 metres of each neighbourhood centroid: This resulted in 3171 venues*[5]. Further, Foursquare associates each venue with a venue category, from which the food venues can by identified. The venues data comprised 287 unique categories.

[5] Note that actual number of venues in the workbook may vary depending on the run date.

2) Extract Food categories:
   The unique categories were extracted and exported to a csv file. Then the non-food categories were eliminated from the category list in the csv file, resulting in 115* food categories.

3) Assign health ratings to food categories:
   For purposes of this project, it was necessary to identify the healthy food venues. Indicators of healthy cuisine are generally available[6], but do not directly map to the Foursquare food categories. Because of this subjectivity, the rating was simply assigned on a 3-point scale to avoid any illusion of accuracy. A value of 3 was assigned to healthy food venues, including Greek, Japanese, and health food categories; a value of 2 to moderate food venues, for example, other Asian, European, and South/Central American food categories that offer some healthy food choices; and a value of 1 to American, burger or fast and dessert food categories. These ratings were entered manually against the 114* food categories in the csv file. The file was then uploaded for use in this project as "4Scategories_food.csv."

4) Merge food category ratings dataset with Foursquare venue data.
   The 3171* Foursquare venue dataset was merged with the food categories originally exported from the Foursquare data that were reduced to 114* food categories. An inner join on the food categories value resulted in the removal of all non-food venues and 1762* remaining food venues.

5) Merge food venue, categories and ratings dataset with Wellbeing Toronto data
   The final step in data gathering and cleansing is merging the Foursquare venue data with the Wellbeing Toronto data.

Data Feature Selections

The final composite neighbourhood dataset includes the Foursquare venue data, with over 1700 food venues across 114* unique categories with health ratings (1-3), along with the Wellbeing Toronto data including neighbourhood populations and HFI.

These data will be used in subsequent data analysis and exploration.

Methodology

This section describes the data exploration, inferences, data exploration, and machine learnings that were conducted and how they relate to the original business problem of gaining data insights specifically to identify restaurant locations in health-conscious neighbourhoods in Toronto that may be under-served by existing venues.

[6] For example, Butler, N, RD, LD, and Schaefer, A. (2016) "*10 Healthiest Cuisines*" Healthline.com, Feb. 10, 2016. Available at https://www.healthline.com/health/food-nutrition/travel-healthiest-cuisines.

The methodology includes data preparation, choropleth mapping, histogram plots, neighbourhood cluster analysis, as well as the choices and considerations within the methods.

Data Preparation Choices

The data preparation described above focused on collating the Wellbeing Toronto HFI data in comparison with the available food venues. This then provides for analysing and exploring the data related to the original business problem, namely, to identify restaurant locations in health-conscious neighbourhoods in Toronto that may be under-served by existing venues, and the factors that lead to such insights.

One of the fundamental choices is the data perspective, and the inherent data preparation choices. Fundamentally this involves choosing whether to anchor the data analysis on the HFI as the independent variable, or to anchor the data analysis on the venue category health rating. Several approaches are identified in the methodology including visual comparisons, plots, and discussion about the data and business factors related to this choice, which is then resolved in the "Selecting the clustering approach" section below.

The next section will describe the processes of data preparation and visual exploration of the HFI data in comparison to the available venues by health rating.

Choropleth Neighbourhood Maps

The first exploration of the data presents the two choropleth maps for comparison. The first map presents the Toronto neighbourhoods by HFI data, while the second presents the neighbour venue average health ratings.

Choropleth mapping requires an overlay definition of the objects being mapped, in this case the Toronto neighbourhoods. This definition is in the "geojson" file format. An internet scan produced several available geojson options. A particular github repository made available both the geojson format, as well a shape file option[7]

Neighbourhood HFI Choropleth

Preparing the HFI choropleth map required several steps, including preparing the mapping data, exploring the data, and creating the map.

Preparing the Map Data

The HFI map required extraction of the 115* unique neighbourhood names with their respective healthy food indices. The collated venue dataset from above was trimmed to just the two required columns, and grouped by the neighbourhood names.

Exploring the Data

The first rows of the resulting HFI mapping dataset are shown below:

	Neighbourhood	Healthy Food Index
0	Agincourt North	38.09

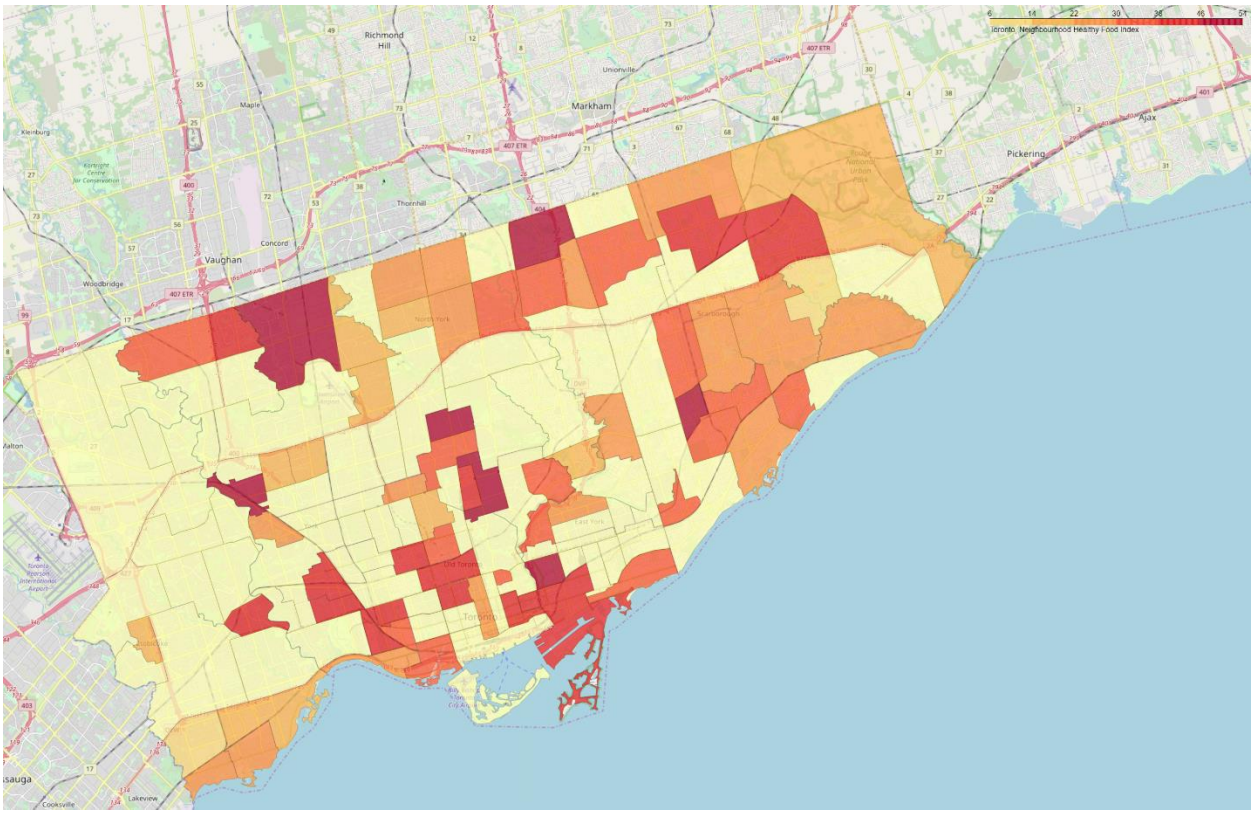
[7] Niehgbourhoods.js file available at https://github.com/adamw523/toronto-geojson

	Neighbourhood	Healthy Food Index
1	Agincourt South	35.26
2	Alderwood	11.41
3	Annex	38.85
4	Bathurst Manor	18.80

The 115* rows of the dataset align with the targeted number of neighbourhoods. Further the maximum HFI value is 53.48, while the minimum is 6.67. This provides an excellent range with which to differentiate between neighbourhoods.

Mapping the HFI data

The HFI data were mapped as shown below.



This choropleth map illustrates the HFI scores of the Wellbeing Toronto communities. The darker areas indicate the healthiest ranked neighbourhoods, while the lightest areas indicate the least healthy.

The choropleth map illustrates that the majority of neighbourhoods have little access to healthy food sources (low HFI). The high HFI neighbourhoods appear to be grouped across the northern and eastern neighbourhoods, and also distributed loosely outward from Toronto's core (middle bottom region of the map).

Further observations and comparisons will be discussed in the Results section below.

Neighbourhood Venue Average Health Ratings (VAHR)

Similar to the HFI map above, preparing the neighbourhood venue average health rated choropleth map required several steps, including preparing the mapping data, exploring the data, and creating the map.

Preparing the Map Data

The HFI map required extraction of the 115* unique neighbourhood names with their respective venue average health ratings (VAHR). The collated venue dataset from above was trimmed to just the two required columns, namely neighbourhood names and venue health ratings. The dataset was then grouped by the neighbourhood names while the health ratings for each venue were averaged for each neighbourhood. This averaging resulted in the VAHR.

Exploring the Data

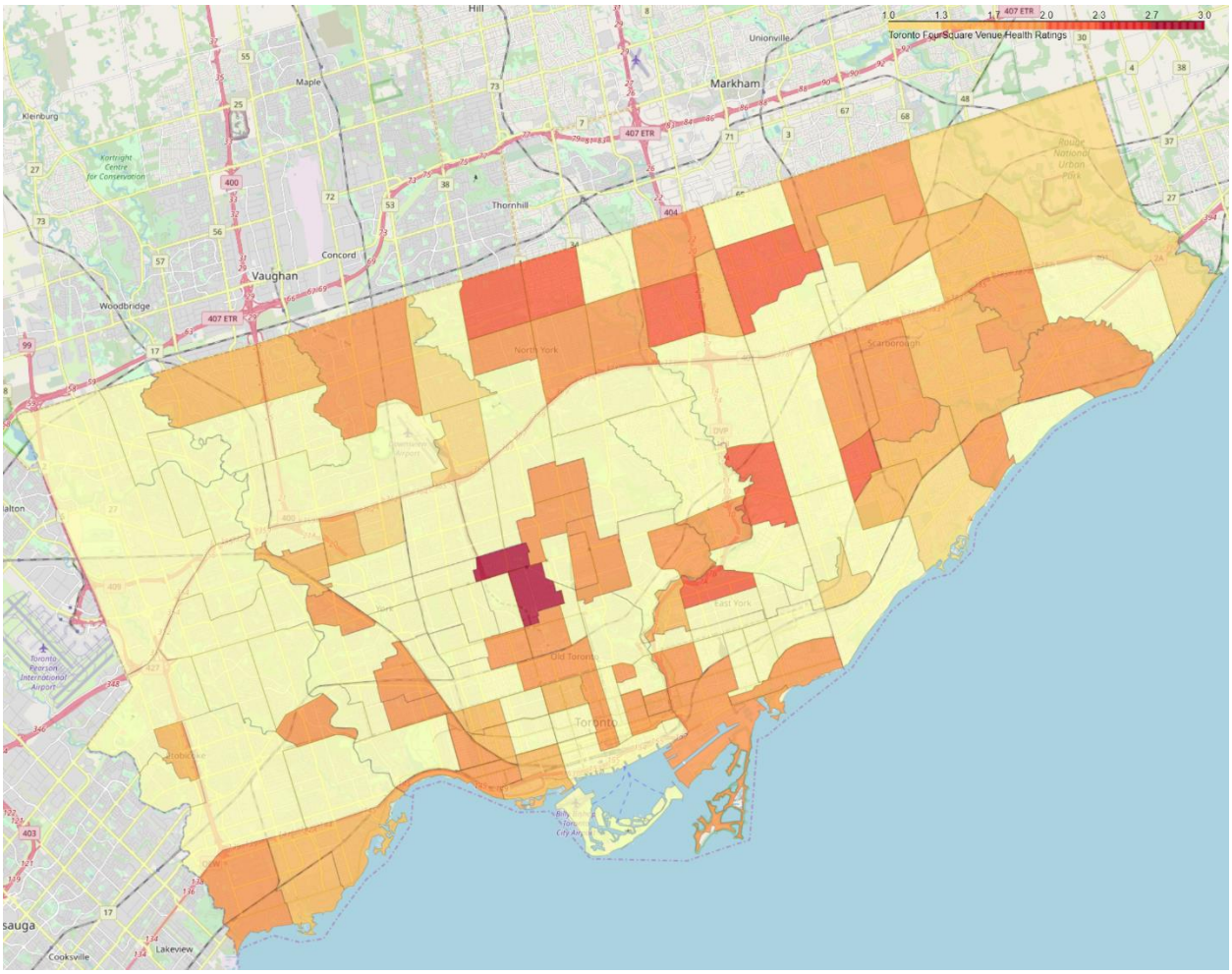
The first rows of the resulting HFI mapping dataset are shown below:

	Neighbourhood	rating
0	Agincourt North	1.578947
1	Agincourt South	2.000000
2	Alderwood	1.666667
3	Annex	1.774194
4	Bathurst Manor	1.000000

The 115* rows of the dataset align with the targeted number of neighbourhoods. Further the maximum HFI value is 3 while the minimum is 1. This range is expected, since the rating scale applied was 1 to 3.

Mapping the VAHR data

The VAHR data were mapped as shown below.



This choropleth map illustrates the calculated VAHR scores of the Wellbeing Toronto communities. The darker areas indicate the neighbourhoods with the healthiest average ranked venues, while the lightest areas indicate the least healthy venues.

The choropleth map illustrates that the majority of neighbourhoods have little access to healthy food sources (low HFI). The high HAVR neighbourhoods appear to be grouped across the northern and eastern neighbourhoods, and also distributed loosely outward from Toronto's core (middle bottom region of the map).

Further observations and comparisons will be discussed in the Results section below.

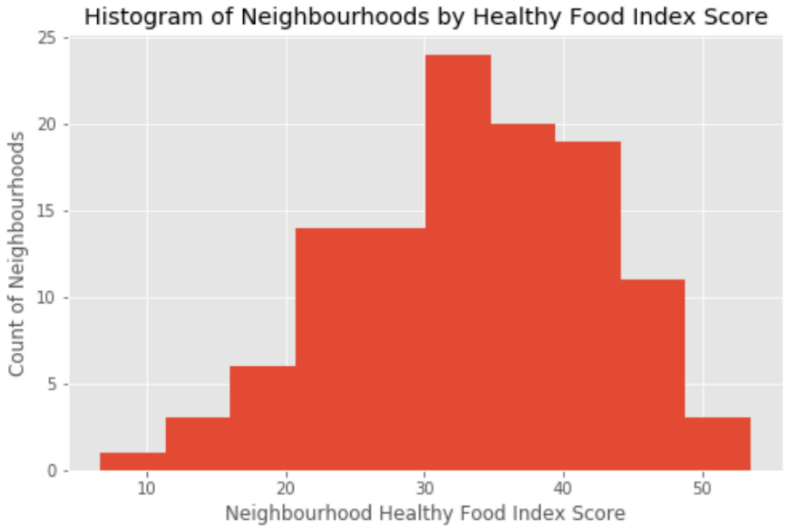
Histograms of Wellbeing Toronto Neighbourhoods

This exploration of the data presents the two histograms for comparison. The first histogram presents the Toronto neighbourhoods by HFI data, while the second presents the neighbour venue average health ratings.

The histograms utilize the same datasets used by the choropleth maps. Therefore no additional data preparation was required.

HFI Histogram

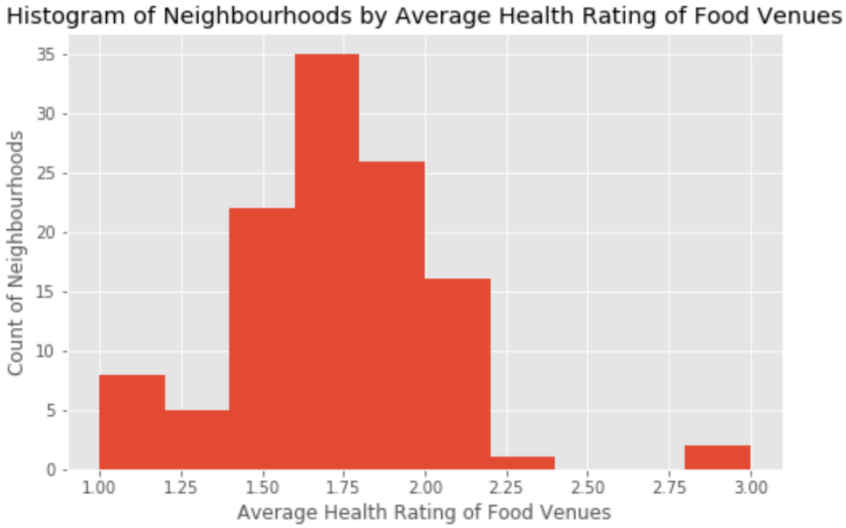
The following histogram illustrates the count of Wellbeing Toronto communities across range of HFI scores.



The histogram illustrates that the neighbourhood counts are greatest in the HFI scores between 30 and 45. Further observations and comparisons will be discussed in the Results section below.

VAHR Histogram

The following histogram illustrates the count of Wellbeing Toronto communities across range of HFI scores.



The histogram illustrates that the neighbourhood counts are greatest in the VAHR scores between 1.4 and 2, or below the median. Further observations and comparisons will be discussed in the Results section below.

Clustering of Neighbourhoods

The venue data will now be explored for similar distributions of venue types by selected neighbourhoods. The clusters will be generated using the k-means algorithm, and displayed on the map of Toronto.

In preparation, we will prepare the clustering dataset. However, two approaches are possible. Clustering can occur on healthy neighbourhoods, or on neighbourhoods with healthy venues. The following sections describe the two approaches, followed by the preparation of the two datasets, the process used in selecting between the two approaches, the final preparation of the chosen dataset, the mapped clusters, and finally reviewing the clustered datasets.

Two Approaches

Clustering can occur following one of two approaches, namely, healthy neighbourhoods, or neighbourhoods with healthy venues. The healthy neighbourhoods approach examines neighbourhoods with high HFI. The choropleth map allocated HFI across 6 groups, the top 3 being 46-54, 38-46, and 30-38. An initial examination of neighbourhoods with HFI greater than 30 resulted in a large dataset of venues. A second attempt targeting neighbourhoods with HFI above 38 resulted in a large but more manageable dataset.

The second approach examines neighbourhoods with healthy food venues, namely those falling in a food category with a health rating equal to 3.

Prepare Comparison Datasets

The two approaches identified above required preparation of two distinct datasets.

The preparation of the healthy neighbourhoods dataset required the removal of all venues in neighbourhoods with HFI less than 38 from the collated Wellbeing Toronto dataset prepared in the data gathering and cleansing section above. The healthy neighbourhoods dataset resulted in 744* venues across 38* neighbourhoods.

The preparation of the healthy food venues dataset also required building on the Wellbeing Toronto dataset prepared above by removing all venues in neighbourhoods where venue categories equalled 1 or 2. The healthy food venues dataset resulted in 244* venues across 56* neighbourhoods.

Selecting the Clustering Approach

Choosing one of these datasets for clustering comparison is both a data and business decision.

From a data perspective, the clustering comparisons would benefit from the greater number of neighbourhoods offered by the second dataset over the first (58 : 38), while still providing a sufficient number of venues for valid clustering.

The business perspective however is more significant. Choosing to compare clusters based on only healthy food venues (rating = 3) targets competitive businesses. This follows the principle of The Central

Place Theory[8], which suggests the economic advantage of establishing a venue near competitors, particularly in larger urban centers. At this same time, the cluster comparisons can factor in the neighbourhood population and HFI indicators to highlight neighbourhoods of greater market opportunity. This point will be discussed further in the conclusions.

Considering both the data and business perspectives, the choice for cluster comparison will focus on the second dataset, namely the neighbourhoods with only healthy food venues (rating = 3).

Prepare the Cluster Dataset

Preparing the chosen dataset for clustering, namely the health food venues dataset, requires several steps, including:

- encoding the venue categories into columns
- group by neighbourhood
- select 10 most common venues
- create clusters
- merge cluster labels

The following table lists the top five rows of the resulting cluster dataset.

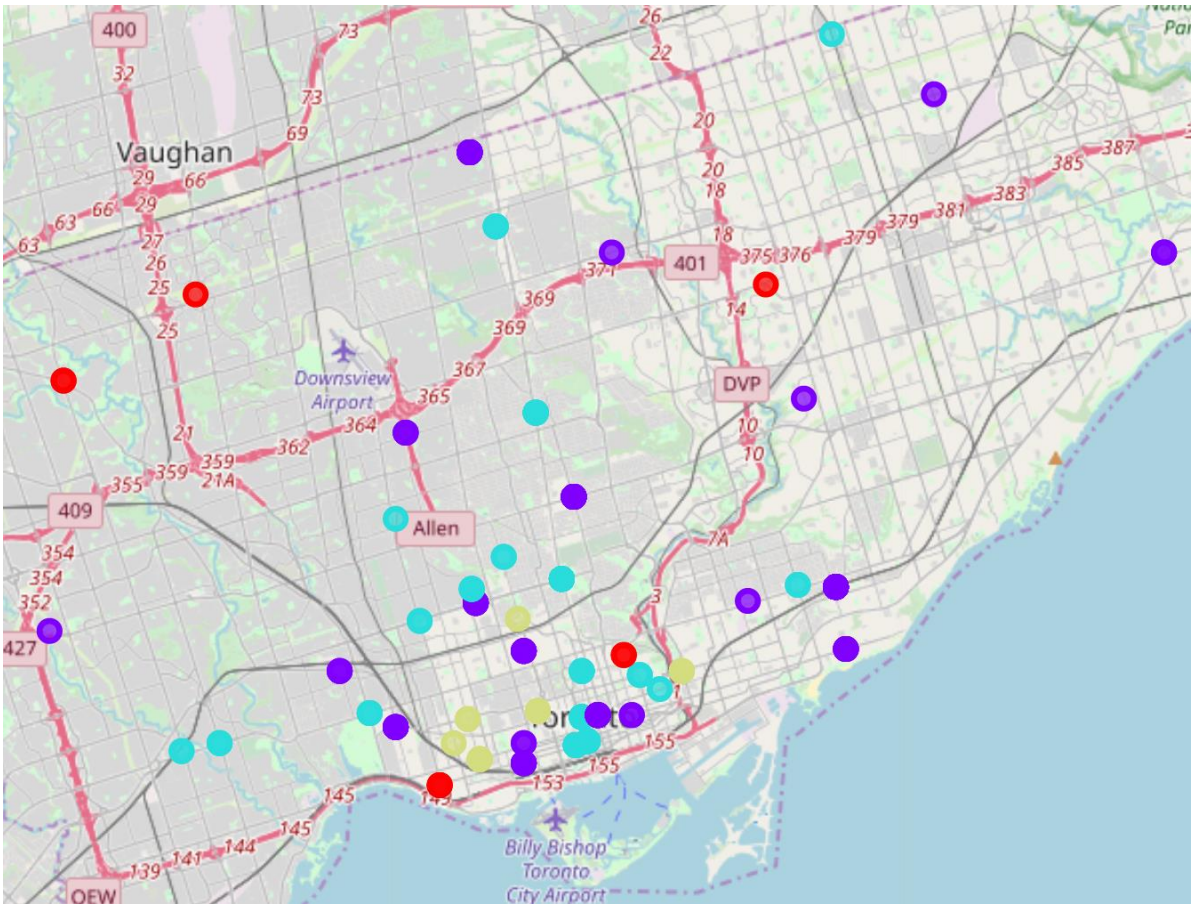
	Neighbourhood	Total Population	Healthy Food Index	Lat	Long	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Co
0	Mount Olive	32954.0	37.57	43.653963	-79.387207	Sushi Restaurant	2	Sushi Restaurant	Vegetarian / Vegan Restaurant	I E Rest
1	Mount Olive	32954.0	37.57	43.653963	-79.387207	Sushi Restaurant	2	Sushi Restaurant	Vegetarian / Vegan Restaurant	I E Rest
2	Mount Olive	32954.0	37.57	43.653963	-79.387207	Sushi Restaurant	2	Sushi Restaurant	Vegetarian / Vegan Restaurant	I E Rest
3	Mount Olive	32954.0	37.57	43.653963	-79.387207	Sushi Restaurant	2	Sushi Restaurant	Vegetarian / Vegan Restaurant	I E Rest
4	Mount Olive	32954.0	37.57	43.653963	-79.387207	Tea Room	2	Sushi Restaurant	Vegetarian / Vegan Restaurant	I E Rest

Due to space limitations, the above table shows only the first two of ten “most common venue” columns of the cluster dataset. This dataset is used to create the following map.

[8] Steif, Ken. 2013. "Why Do Certain Retail Stores Cluster Together?", Planetizen, Oct. 2013. Available at https://www.planetizen.com/node/65765.

Mapping the Clusters

The following map illustrates the neighbourhood clusters on the map of Toronto.



The four different clusters are indicated by different colours, illustrating neighbourhoods with similar distributions of venue types for all the healthy venues.

The following section will review the resulting clustering data.

Examine the Clusters

This section briefly examines the resulting data clusters of neighbourhoods with similar distributions of venue types. These clusters are formed only the healthy venues, where the category healthy rating equals 3 as described above.

The four clusters are listed by individual venue, and then also by neighbourhood group, resulting in two listings for each cluster.

The first listing is shown by the following example of the data cluster details for Cluster One.

```
Shape is:  (9, 17)
```

	Neighbourhood	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
9	Thistletown	Caribbean Restaurant	0	Caribbean Restaurant	Vegetarian / Vegan Restaurant	Tea Room	Eastern European Restaurant	Fish Market	Fruit & Vegetable Store	Gluten-free Restaurant	Gourmet Shop	Greek Restaurant	Health Food Store
10	Thistletown	Caribbean Restaurant	0	Caribbean Restaurant	Vegetarian / Vegan Restaurant	Tea Room	Eastern European Restaurant	Fish Market	Fruit & Vegetable Store	Gluten-free Restaurant	Gourmet Shop	Greek Restaurant	Health Food Store
35	York University Heights	Caribbean Restaurant	0	Caribbean Restaurant	Vegetarian / Vegan Restaurant	Tea Room	Eastern European Restaurant	Fish Market	Fruit & Vegetable Store	Gluten-free Restaurant	Gourmet Shop	Greek Restaurant	Health Food Store
53	Parkwoods	Caribbean Restaurant	0	Caribbean Restaurant	Vegetarian / Vegan Restaurant	Tea Room	Eastern European Restaurant	Fish Market	Fruit & Vegetable Store	Gluten-free Restaurant	Gourmet Shop	Greek Restaurant	Health Food Store
104	North St.James Town	Caribbean Restaurant	0	Caribbean Restaurant	Market	Vegetarian / Vegan Restaurant	Health Food Store	Eastern European Restaurant	Fish Market	Fruit & Vegetable Store	Gluten-free Restaurant	Gourmet Shop	Greek Restaurant

The Cluster Detail results lists the cluster pattern in each row following the top 10 venues for each cluster pattern, along with the neighbourhood name and cluster label. The Results section below summarizes and discusses these results in more detail.

The second listing is shown by the following example of the data cluster neighbourhood groups for Cluster One.

			Lat
Neighbourhood	**Total Population**	**Healthy Food Index**	
North St.James Town	**18615.0**	**34.03**	2
Parkwoods	**34805.0**	**33.60**	1
South Parkdale	**21849.0**	**26.84**	3
Thistletown	**10360.0**	**42.26**	2
York University Heights	**27593.0**	**47.00**	1

The Cluster neighbourhood groups listing provides the total population, HFI, and count of cluster patterns attributed to the neighbourhood. These two listings are provided for each of the four clusters.

The Results section below summarizes and discusses these results in more detail.

Results

This results section provides an overview of the outcomes of the methodology and their relevance to the original problem of identifying restaurant locations in health-conscious neighbourhoods in Toronto that may be under-served by existing venues, and the related factors.

Choropleth Maps

The two choropleth maps illustrate significant differences between the neighbourhoods' Healthy Food Index scores and the venues' average health ratings. The Health Food Index map indicates more neighbourhoods at the healthier end of the scale (darker areas) compared to the average health ratings of the venues in those neighbourhoods.

This comparison may generally indicate that most Toronto neighbourhoods are underserved by healthy food venues compared to other food venues. More specifically, there appears to be opportunity for further data exploration of the differential, such as the clustering machine learning technique described above.

Histograms

The two histograms affirm the differences between the neighbourhoods' Healthy Food Index Scores and the venues' average health ratings observed in the choropleth maps. The Health Food Index histogram illustrates skewing of neighbourhoods' Healthy Food Index more toward the upper range (healthier) compared to the average health ratings of the food venues in those neighbourhoods shown in the second histogram.

Clusters

The clustering machine learning technique provides additional insights into the neighbourhood data. The chosen approach was to examine the cluster patterns of all venues which were manually rated by category as healthy (rating = 3). The four resulting clusters were examined by cluster pattern detail, as well as neighbourhood grouping properties.

The following table summarizes the results of each cluster, including cluster label (number), number of neighbourhoods, number of venues, the top 3 most common venues, population, and HFI.

Cluster #->	**One**	**Two**	**Three**	**Four**
Neighbourhoods	5	21	23	7
Venues	9	101	107	27
Most common venues	Caribbean/Greek Vegetarian/vegan tea Room	Juice /smoothie/tea Middle Eastern Sushi	Sushi Salad/Vegetarian/vegan Greek/Caribbean	Vegetarian Tea/Juice/smoothie Health food/organic
Max Population	34805	31180	65913	31340
Min Population	10360	7804	7865	11785
Max HFI	47	50	52	47
Min HFI	27	20	18	32
HFI range	20	30	34	15
Max cluster patterns per neighbourhood	3	9	10	13

Table 1: Summary of Cluster Results

This table highlights key differences between the clusters which are discussed next.

Neighbourhoods and Venues

Clusters one and four represent a lower number of neighbourhoods and venues compared to clusters two and three. These results may provide guidance in restaurant location choice depending on a penetration strategy. The Central Place Theory referenced earlier might suggest locating restaurants with the greatest number of venues in a neighbourhood, and pursue demand based on largest populations.

Population and Venues

Population patterns across the clusters one, two and four are moderately consistent. The maximum populations for these three clusters range within 3625, or about ±5.5% of the average maximum. The maximum population for cluster three, however, is more than double the average maximum of the other three clusters. Further examination of the populations of the 23 neighbourhoods in cluster three reveals four neighbourhoods with populations above the average maximum population of the other three clusters. A market penetration strategy seeking to maximize opportunity based on population along might favour these four neighbourhoods in cluster three.

Another strategy may consider the density factor of venues to population in each neighbourhood. A simple calculation of neighbourhood population divided by number of venues would provide the objective selection options. This strategy would favour the larger ratios as the most favourable location options for new venues.

HFI Min/Max & Range

The HFI values vary considerably across the four clusters. The HFI maximum values are fairly consistent, varying only 5 points, from 47 to 52, across the four clusters. This indicates high availability of healthy food choices within each cluster.

However, the HFI minimum values show considerably greater variation, varying 19 points, from 15 to 34, across the four clusters. This would indicate that some neighbourhoods within cluster four, particularly, and in cluster 20 are significantly underserved by healthy food choices.

These conditions would favour a market penetration strategy seeking underserved markets, in contrast to a market penetration strategy seeking to locate among competitors as suggested by the Central Place theory discussed earlier.

Most Common Venues

Finally, the most common venues across the four show some variety as observed in table 1 above. The top three most common venues share some similar venue categories, but none share exactly the same pattern. For example, Caribbean/Greek cuisine is ranked first in cluster one, and third in cluster three, while not appearing in the top three of clusters two and four. Similarly, juice and smoothie bars rank first and second in clusters two and four respectively, while cluster one includes only tea rooms ranked third. Notably, only cluster four includes health food and organic grocery outlets among its top three venue categories.

The venue data – both the top three venue categories in the summary table above, as well as the top ten given in the detail charts in the workbook – offer valuable insights into market penetration options. Depending on the type of venue and market penetration strategy, entrepreneurs can choose to align with direct competitors operating in the same category (e.g. sushi venues), align with indirect

competitors in similar categories (e.g. Asian but not sushi venues), or to opt for neighbourhoods with few direct competitors (e.g. cuisines other than Asian).

These results collectively offer insights to market opportunities and factors to consider based on entrepreneurs' selected market penetration strategy. The following section summarizes the observations and discusses recommendations.

Discussion

The Results presented above describe how the analyses performed above lead to various insights and market opportunities. This section summarizes those insights and the decision factors for restauranteurs and stakeholders servicing the quick server restaurant (QSR) and sit-in food segments.

Gaps among Neighbourhoods' HFI and VAHR

The maps and histograms illustrate significant gaps among neighbourhoods in both the healthy food indices (HFI) scores and Venue Average Health Ratings (VAHR), highlighting different types of opportunities for restauranteurs. High HFI scores may offer entrepreneurs the opportunity to identify neighbourhoods with higher demand for healthy food choices. These differentials are illustrated in the following table.

HFI	low VAHR	high VAHR
high	- high demand, low supply - opportunities for new entrants - broad range of venue category options	- high demand, high supply - select or low opportunities for new entrants - need to carefully target venue category choices
low	- low demand, low supply - low opportunities for new entrants - choose venue categories to drive demand	- low demand, high supply - very low opportunities for new entrants - very narrow range of venue category options

Legend
Good opportunities
Moderate opportunities
Poor opportunities

Table 2: HFI-VAHR Matrix

This table highlights opportunities by HFI as an indicator of demand, and VAHR as an indicator of supply. Each quadrant offers different levels of opportunities in relation to the type of venue category offered by the new restauranteur as shown above.

Thus, the opportunities highlighted by the maps and histograms illustrate both HFI and VAHR as important factors in determining opportunities by neighbourhood for new healthy food locations.

However, these factors are mitigated by other important factors, as discussed next.

Market Penetration Strategy

As mentioned above, the choice of market penetration strategy will impact how opportunities are identified across the other factors. As already mentioned, for example, the Central Place Theory suggests that new entrants may benefit from competitors. Alternatively, selecting a high HFI/low VAHR neighbourhood may be more attractive for entrants more averse to competition. As another example, a restauranteur with a unique offering may target high population areas regardless of the number and category types of competing venues.

Population/Demographics

Population density is a factor of entrant choice as a possible indicator of market demand. As identified above, cluster three offers neighbourhoods with higher maximum populations, which may offer more attractive opportunities for new entrants depending on other factors, including also the number of venues.

Number of Venues

Number of venues is an important choice factor as a possible indicator of supply. The choropleth maps illustrate a lower overall supply (VAHR) in many Toronto neighbourhoods compared to demand (HFI). However, population to venue count ratios may be a key indicator for entrants looking at market opportunity in terms of venue density. High ratios are more attractive to this restauranteur.

Top Venues

Depending on the market penetration strategy of the restauranteur, the existing vendor categories as indicated by the clustering maps are a key decision factor for new restaurant locations. The cluster details as described above are useful for positioning a new entrant offering in relation to existing top venues. One entrant strategy may choose to locate in neighbourhoods among the same venue categories, as suggested by the Central Place theory. Another strategy may seek to complete similar but different venue categories, or alternatively identify complementary categories that are different from the restauranteur's offering but known to attract customers of similar taste.

Further, the four clusters offer different choice options as identified in Table 1 above that can be useful as a decision tool for different market penetration strategies. Overall, the top venues by neighbourhood identified by the clustering is a key tool particularly after market strategy has been selected, and other factors considered as presented above in order to narrow the neighbourhood options.

Neighbourhood Topographics

Other decision factors that are beyond the scope of this project may also impact restauranteur location choices. One such factor is the neighbourhood topographics, including demographics that would describe and account for population density, age distribution, access to public transportation, types of road networks, public amenities, attractions, and so on. All of these would contribution to and impact the restauranteur's decision process in selecting new venue locations, although are beyond the scope of this project.

Recommendations

The analysis conducted in this project highlights various decision factors surrounding the insights gained by the data analysis and machine learning. These factors together provide a framework in considering locations for restauranteurs and stakeholders seeking to establish new venues or even re-invigorate an existing brand.

The data analyses, visualizations, and machine learning in this project suggest how the HFI data and venue data, particularly the VAHR, can be considered as part of a decision framework for locating new venues. One key contribution is the composition of Wellbeing Toronto data in combination with venue data. The construction of a health index for venue categories provides an important comparative indicator to potentially highlight demand and supply by Toronto neighbourhood.

This framework does not result in specific recommendations for all cases, but rather works together to result in recommendations depending on the entrant's venue type, risk tolerance, and market strategy as discussed. Numerous examples have been already cited of how the various decision factors highlighted above contribute to a decision framework for several different scenarios.

Business Problem

This decision framework, particularly the identification of factors resulting from the data analysis, provides guidance for restauranteurs and stakeholders serving the QSR and eat-in food industry.

The data analysis and framework highlight how various factors can be implemented to select for new venue locations that target and effectively service the demand in health-conscious neighbourhoods.

Further, a more simplistic application the data analyses is for existing brands to map existing venues onto both the HFI and VAHR choropleth maps. This would serve to indicate how effectively existing venues are located within health conscious neighbourhoods depending on the brand's market strategy.

Future Directions

The preceding discussions have highlighted, in addition to factors and insights resulting from the data analysis, other factors beyond the scope of this project.

One such factor is neighbourhood topograhics/demographics, that describes population density, age distribution, access to public transportation, types of road networks, public amenities, attractions, geography and so on. The addition of information resources pertinent to these neighbourhood dimensions can aid additional analysis and visualization to complement this project. These additional analyses would advance this research and aid in restaurant location and re-invigoration selection.

Another factor to consider is the assignment of health ratings to the food categories. For purposes of this project, this was conducted somewhat subjectively based on readily available health cuisine data as cited. However, additional research and a more rigorous may enhance the objectivity around this rating scale and further help to validate the decision framework.

A future consideration for building on these analyses is the development of a model for assessing marketing penetration strategy given the existing decision factors and analyses given above. This would be a significant undertaking based on existing and new research. However, such a model could

complement this decision framework by advancing the sophistication of the decision process to add prescriptive analytics that includes the scope of complexity surround marketing penetration strategy.

These recommendations describe how the decision framework and factors developed in this project can help guide restauranteurs and stakeholders in locating new venues within Toronto neighbourhoods, or in evaluating strategic options for revitalizing existing venues.

Conclusion

The restaurant industry is challenged with balancing a variety of demand factors that sometimes compete, for example genuinely healthy and organic foods and low cost. Restauranteurs and other stakeholders can use these data analyses and insights presented in this project to explore how to effectively target new venue locations and effectively service the demand in health-conscious neighbourhoods.

This project has first exploited neighbourhood wellness data from Wellbeing Toronto merged with Foursquare venue data and manually applied health ratings for food venue categories. These data provided the basis for several visualizations including two choropleth maps illustrating the healthy food index (HFI) and the derived venue average health rating (VAHR) by neighbourhood respectively. Additionally, two histograms also compared the HFI and VAHR distributions across neighbours, providing visual affirmation of the gap between the HFI and VAHR levels.

Building on these data, the cluster analysis identified the geographic distribution of Toronto venues by similarities in venue categories. The four resulting clusters were mapped to illustrate the neighbour distributions, and then listed and grouped by neighbourhood to provide detail decision data of the top venue types by cluster-neighbourhood.

The methodology section detailed the steps performed in the data analyses and machine learning, and also described the data and business factors in selecting between two analytical approaches and preparation of the dataset for the clustering machine learning technique.

These analyses and data insights together provide a decision framework for restauranteurs evaluating new or existing venue locations. The results section identified numerous decision factors, and illustrated their implementation across several marketing penetration strategies which included a demand/supply matrix table.

Finally, the discussion section detailed the decision factors made visible by the data analysis and machine learnings that contribute to the decision framework and how they are implemented as a decision process leading to business recommendations for the original business problem. Several future directions and model enhancements were discussed.

In summary, this project explored data insights that contribute to the identification of restaurant locations in health-conscious neighbourhoods in Toronto that may be under-served by existing venues, and identified factors that lead to such insights. These insights and factors may also be useful for evaluating existing venue locations with respect to neighbourhood HFIs and VAHRs.