IBM APPLIED DATA SCIENCE CAPSTONE PROJECT
Week 5 – report

Identifying the restaurant cuisine trends

By: Hwei Ong

Introduction

The restaurant industry is vibrant in many cities, where changes in demographics and lifestyle are driving the demand for such services. As the economy increases household income, there is potential for increased consumer spending. Nonetheless, with housing, healthcare and common household expenditures consuming majority of this spending, the consumers tend to be conservative in their spending for dine-ins and take-outs. It is not uncommon for food trends to change and drive consumer spending on certain foods. Too many restaurants serving the same cuisine in one neighborhood cluster would mean a highly competitive environment for a newcomer. While a high density of restaurants may signify a strong customer base, it may also mean higher business expenses due to higher rents in a popular area. Starting a new restaurant business requires significant capital and planning upfront. Moreover, it would be more difficult to obtain approval to open a new restaurant from the city for an area that has a concentrated number of restaurants due to congestion concerns. As such, it would be ideal to be able to research the restaurant industry of a city's neighborhoods in order to gain some insights as to what is currently available.

Business Problem / Target Audience

The business problem is how to identify restaurant cuisine trends in the city. The idea is to cluster neighborhoods by cuisine type and gain an understanding on the density of restaurants within the neighborhoods.
The target audience would be:

- investors and/or restauranteurs (i.e. commercial business planning) who are interested in opening a new restaurant;
- policy wonks involved in urban planning as such analytics can guide the zoning of different areas for development; and
- property developers who are interesting in knowing what new property type would be suitable for an area (e.g. a mixed-use residential building vs. a commercial building).

Data Sources and Methods Used

The Canadian cities used in this analysis are Toronto, ON and Calgary, AB. As such, the following data are required:

- A list of neighborhoods and their respective latitude and longitude coordinates for each city – the data required in order to obtain data on the types of venues found in each neighborhood, and for subsequent plotting on the city map during analysis.
- A list of venues found in each neighborhood – the data is required for the machine learning algorithm used to perform clustering on the neighborhoods.

The following Wikipedia pages contain the list of neighborhoods for each city specified below:

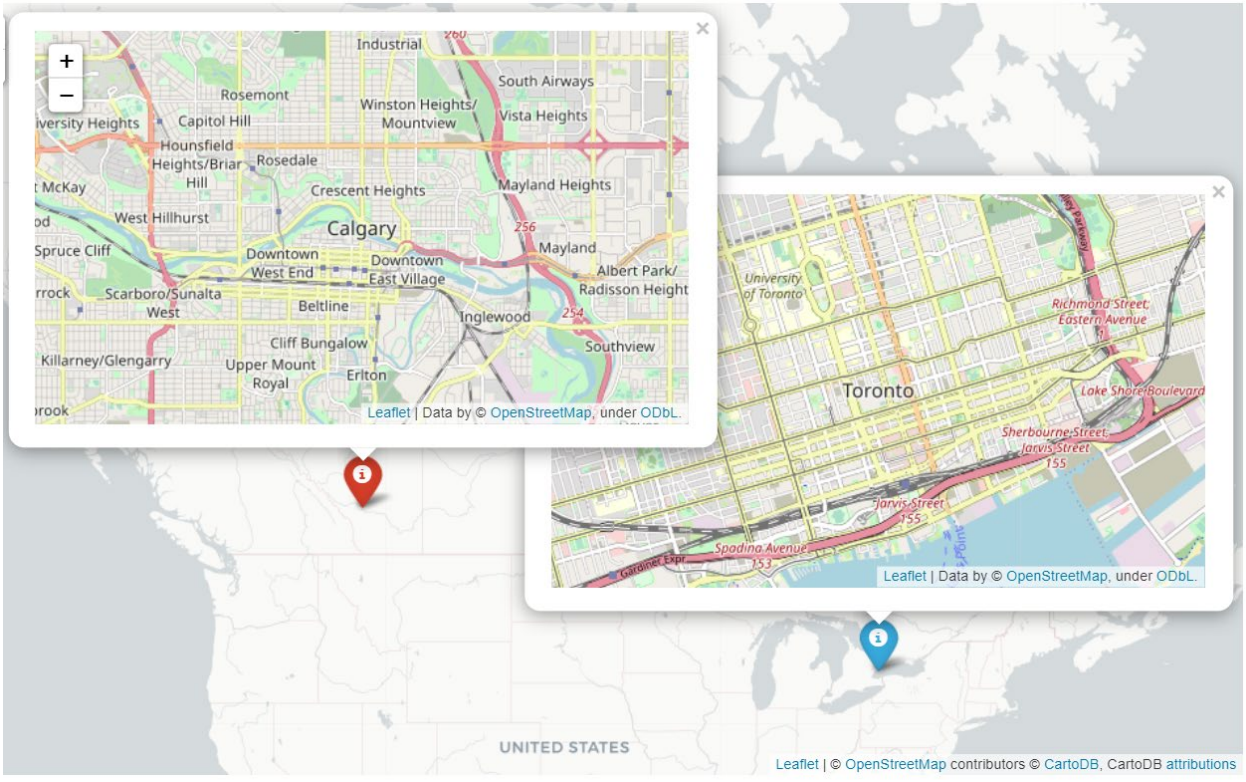
- Toronto (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) [Ref. 1]
- Calgary (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T) [Ref. 2]

The following Python packages will be used in this project:

- Beautiful Soup will be used to design a web scraper that can obtain the information contained on the respective Wikipedia pages listed above.
- The requests and json libraries will be used to work with the data downloaded using Beautiful Soup.
- The numpy library is a dependency in some functions defined within the neighborhood, primarily use to define a range for looping.
- Pandas will be used to structure and clean the downloaded data.
- Geocoder will be used to obtain the coordinates (latitude and longitude) of each neighborhood.
- The unsupervised machine learning algorithm, k-means, from Scikit-learn will be used to cluster the neighborhoods.
- The kneed library will be used to identify the elbow point, which signified the optimal number of clusters for *k*-means.
- Folium will be used to visualize the neighborhoods and clusters for each city.
- The seaborn and matplotlib libraries will be used to visualize the data.

Additionally, the Foursquare API [Ref. 3] will be used to query for venue data using the coordinates of each neighborhood. The venue data contains information on the different types of business found in each neighborhood. From these, we can gain insights into the density of a neighborhood, as well as evaluate the potential customer base and competitors.

Below is a map showing the locations of the different cities covered in this project.



Methodology

(1) Data download, structuring and cleaning

The Python package, Beautiful Soup, was used to create a web scraper that can be used to obtain information from the Wikipedia pages. For each Wikipedia page, the list of neighborhoods is found within a table. Another packaged called pandas was then used to assign the downloaded information to a dataframe and clean the data for subsequent processing. Briefly, the data was downloaded with structured parse tree with various HTML (hyper text markup language) elements and tags contained within. To clean the data, these elements and tags were stripped away and replaced with tabs denoting separations between text in a string. Each row of a HTML table became a string of words separated by tabs and each table turned into a list of word strings. Once this list was structured as a dataframe, a string split function using the tab as a delimiter was used to split the data into different columns within the dataframe. The text in the first dataframe row was used as column labels – any spelling errors and incorrect labels are rectified. Cells within the dataframe with no values, as indicated by the 'Not assigned' text, were removed.

(2) Using the Foursquare API

Neighborhoods sharing the same postal codes were grouped together. The latitude and longitude coordinates of each neighborhood group was required by the Foursquare API in order to query and obtain the venue data for each neighborhood. Only the city of Calgary and Edmonton had these coordinates on their Wikipedia pages, whereas a CSV (comma-separated values) file called 'Geospatial_Coordinates' was obtained separately for Toronto during the labs. For the remaining cities – Montreal and Ottawa, the Geocoder package was used retrieve the coordinates for each neighborhood group found in the dataframe. The coordinates were then used visualized these neighborhoods on a city map using the Folium package. This is to ensure that the coordinates retrieved by Geocoder are correctly mapped out. A Foursquare Developer Account is needed in order to use their API to get the venue data – this can be set up for free on their website (https://foursquare.com/) [Ref. 3]. The client_id, client_secret and access_token from the account were used to create a query to obtain data on the different venues (limited to 100) found within 500 meters of the point coordinates of each neighborhood group. As the results of the Foursquare API query is a JSON (JavaScript Object Notation) file, the data restructured into a pandas dataframe for further processing. The venues were then grouped by neighborhood(s) and data columns with the string 'Restaurant' in their labels will be extracted. This will create a dataframe exclusively for restaurants to be used in *k*-means clustering. The total number of restaurants per neighborhood(s), as well as the total number of different restaurant types, will be obtained using pandas functions.

(3) Plotting, clustering and mapping the neighborhoods

Matplotlib and seaborn will be used to create bar plots that show the sum of different restaurant types per city, as well as the total of all restaurants per neighborhood(s).The *k*-means method from the Scikit-Learn package was used to cluster the neighborhoods, allocating each to the nearest cluster while keeping the summed square errors small. This method is a popular unsupervised machine learning algorithm and has been successfully used previously to cluster neighborhoods in New York City. The optimal number of cluster was determined using the elbow method and the KneeLocator from the kneed library. The top 10 most common type of

restaurants found in each neighborhood can be used to determine the predominant cuisine for each cluster.

Results

The methodology was initially refined using the dataset for Toronto, ON. Then the same methodology was used on Calgary, AB to evaluate the reproducibility of results using a different dataset.

(1) Toronto, ON

The data for all neighborhoods in Downtown Toronto, Central Toronto, East Toronto and West Toronto were used in the analysis. The map in Figure 1 shows these neighborhoods plotted as blue points on the city map of Toronto. These neighborhoods were groups by their postal codes.

Figure 1. Map of Toronto city with the neighborhoods in Downtown Toronto, Central Toronto, East Toronto and West Toronto shown as blue dots using their coordinates.

Aggregating the restaurants by type shows the total number of restaurants for each cuisine in all neighborhoods, as shown by the bar plot below (Figure 2). The predominant cuisine is Italian, followed by Japanese and Sushi (indicated by red arrows). However, there is a large number of venues classified only as Restaurant and a few classified as Theme Restaurant (indicated by purple arrows), which are not useful determining their cuisine types (Figure 2).

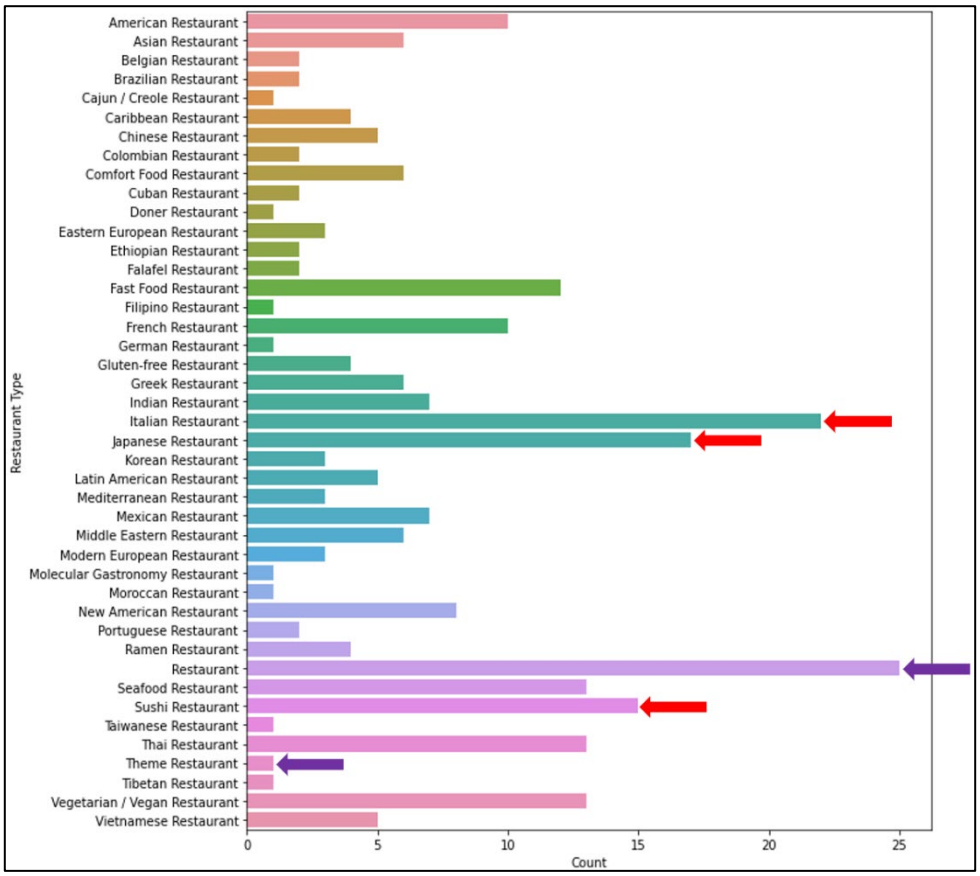


Figure 2. Sum of restaurants for each cuisine in Toronto city for the neighborhoods in Downtown Toronto, Central Toronto, East Toronto and West Toronto. The red arrows show the top three cuisine types, while the purple arrows show the categories with indeterminate cuisines.

Another bar plot (Figure 3) shows the total number of all restaurants for each group of neighborhoods. The red stars show the top two sites with a high number of restaurants – First Canadian Place and Underground City, as well as Church and Wellesley.

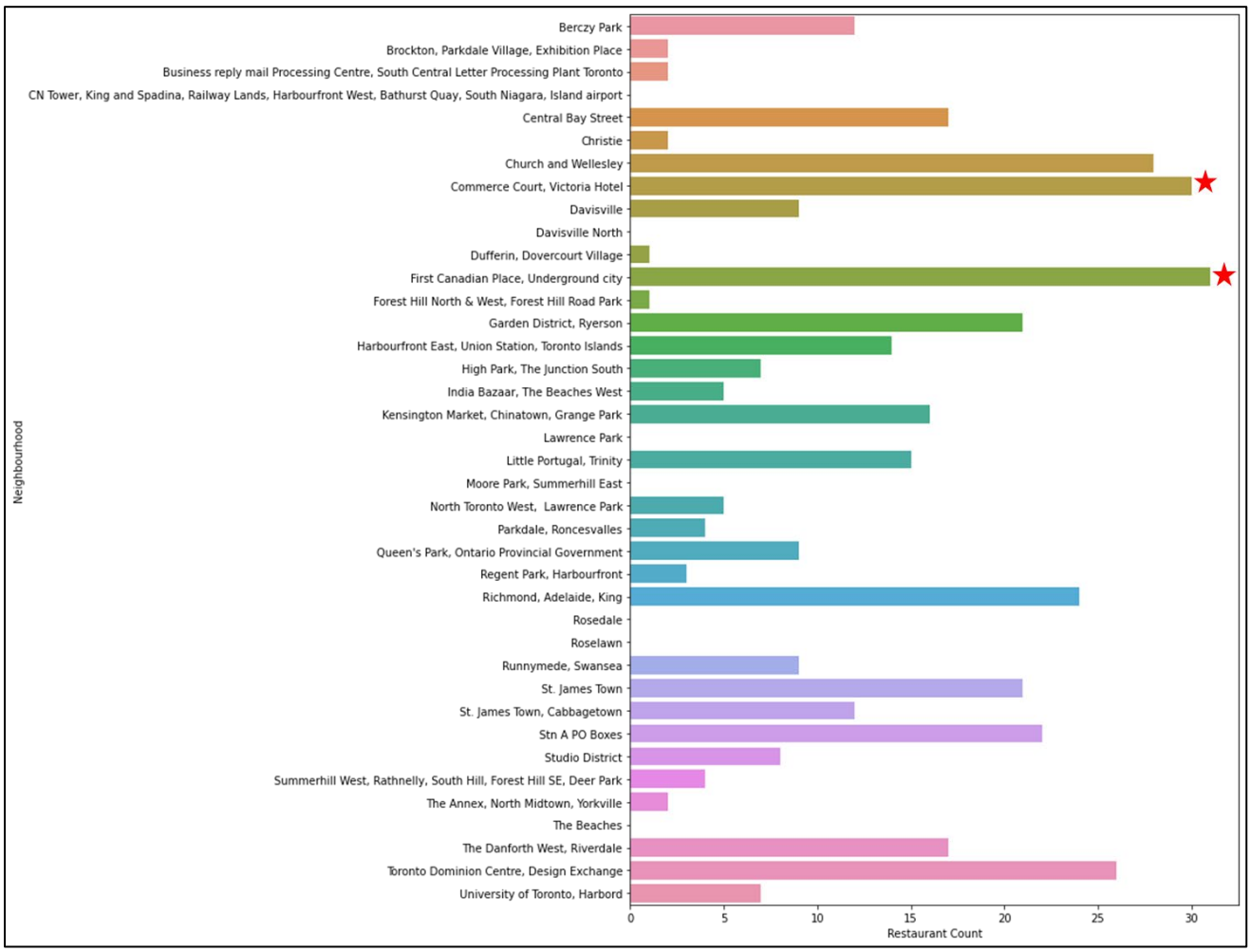


Figure 3. Total number of all restaurants for each group of neighborhood(s) in Downtown Toronto, Central Toronto, East Toronto and West Toronto. The red stars show groups of neighborhoods with the top two highest numbers of restaurants.

The optimal number of clusters was 6, as determined using the elbow method and KneeLocator. Analyzing each cluster reveals the predominant cuisine for each:

- Cluster 1 : Sushi
- Cluster 2 : Vietnamese
- Cluster 3 : Thai
- Cluster 4 : Vietnamese
- Cluster 5 : Greek
- Cluster 6 : equally split between Sushi and Fast Food

This map shows the different clusters on the city map of Toronto, with the legend showing which color for each cluster (Figure 4). The clusters were obtained using the *k*-means method, an unsupervised machine learning algorithm.

Based on the analysis, we can surmise that the predominant cuisine in Toronto is Vietnamese, which came out on top in two clusters.

Figure 4. Map of Toronto city with the *k*-means clusters identified by the algorithm. The legend shows the colors corresponding to each cluster.

(2) Calgary, AB

To ensure reproducibility of the results, the same methodology was tested using a completely dataset for Calgary, AB. The map in Figure 5 shows these neighborhoods plotted as blue points on the city map of Calgary.

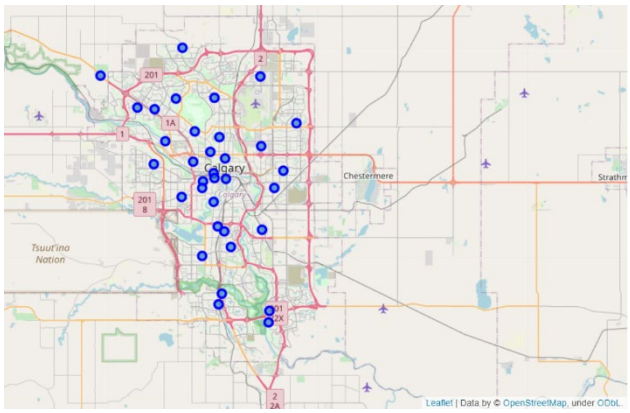


Figure 5. Map of Calgary city with the groups of neighborhood(s) shown as blue dots using their coordinates.

Aggregating the restaurants by type shows the total number of restaurants for each cuisine in all neighborhoods, as shown by the bar plot below (Figure 6). The predominant cuisine is Vietnamese, followed by Middle Eastern, Fast Food and Chinese (indicated by red arrows). Note that there may be an overlap in the cuisines as shown by the different categories. Moreover, one category (called simply 'Restaurant') does not provide any information on their cuisine type (indicated by purple arrow; Figure 6). When comparing the same graphs for Toronto (Figure 2) and Calgary (Figure 6), there is a lower range of cuisines found in Calgary.

Another bar plot (Figure 7) shows the total number of restaurants (irrespective of cuisine type) for each group of neighborhoods. The red stars show the top two sites with a high number of restaurants – Connaught West and Victoria Park, as well as Inglewood, Burnsland, Chinatown, East Victoria Park and Saddlestone. When compared to the same graph for Toronto, (Figure 3), the total number of neighborhoods appears to be lower for Calgary. The caveat here is that the analysis for Toronto covered only four boroughs, while the analysis for Calgary covered all boroughs. Therefore, given the larger area being used in Calgary, it is difficult to directly compare the two cities.

Figure 6 (Right). Sum of restaurants for each cuisine in Calgary. The red arrows show the top three cuisine types, while the purple arrow shows a category with indeterminate cuisines.

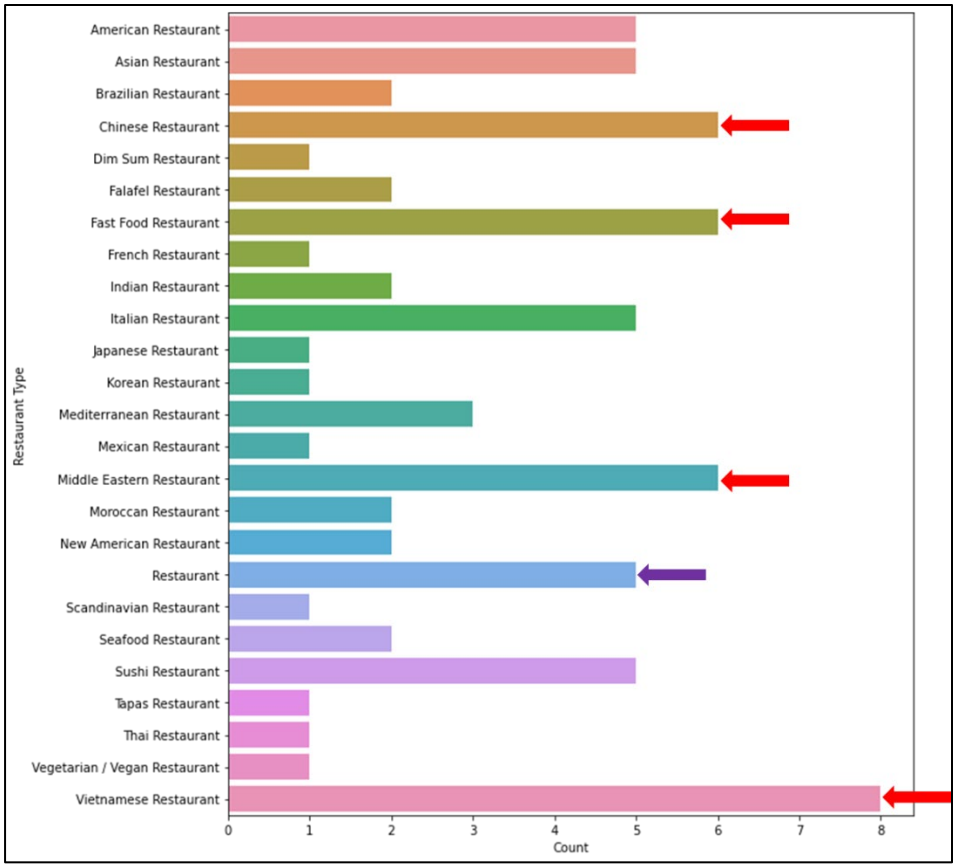
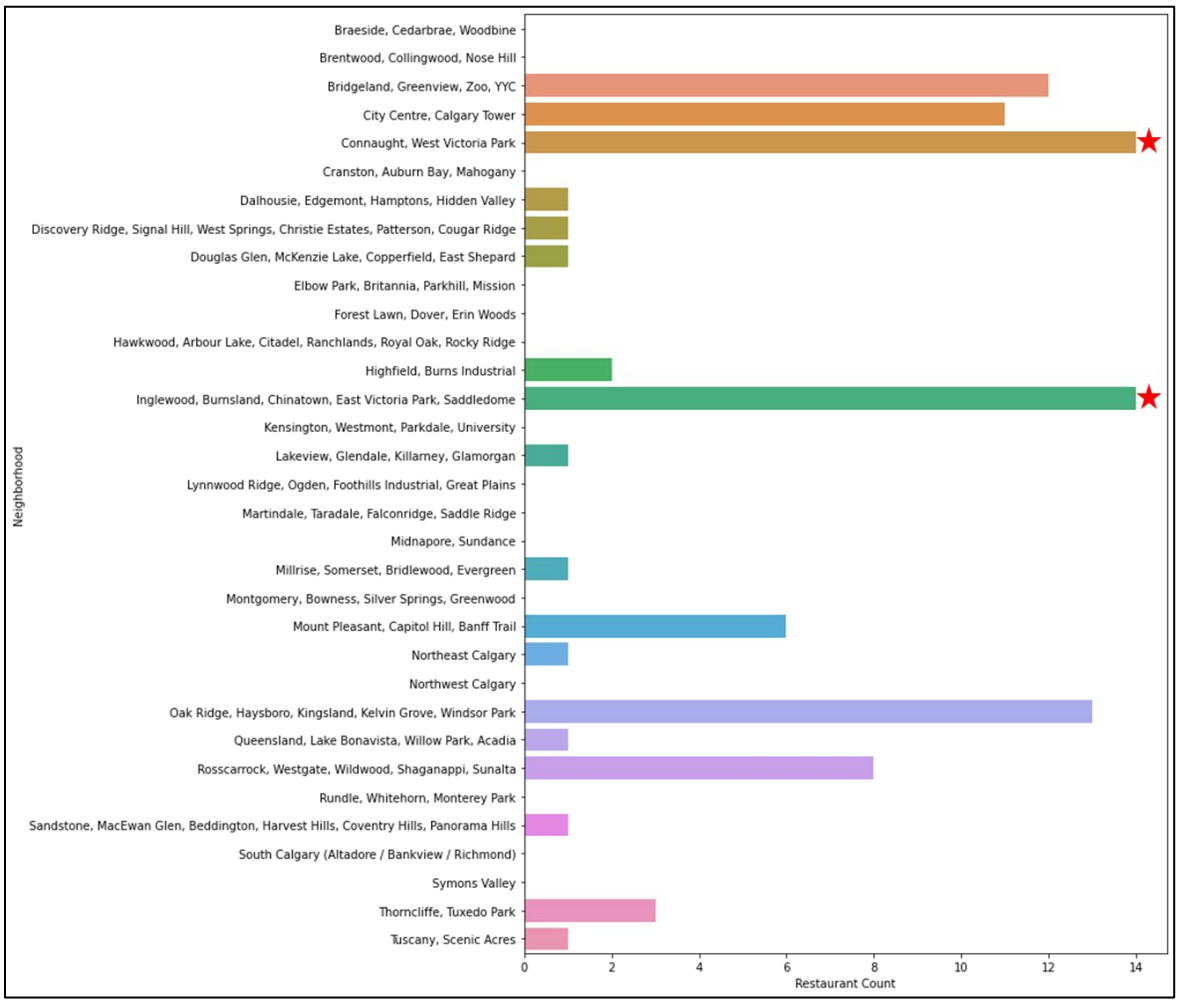


Figure 7 (below). Total number of all restaurants for each group of neighborhood(s) in all boroughs of Calgary. The red stars show groups of neighborhoods with the top two highest numbers of restaurants.



The optimal number of clusters was 6, as determined using the elbow method and KneeLocator. Analyzing each cluster reveals the predominant cuisine for each:

- Cluster 1 : Asian (Vietnamese and Chinese)
- Cluster 2 : Vietnamese
- Cluster 3 : American
- Cluster 4 : Vietnamese
- Cluster 5 : American
- Cluster 6 : Asian

This map (Figure 8) shows the different clusters on the city map of Calgary, with the legend showing which color for each cluster. Based on the analysis, we can surmise that cuisines found in Calgary are predominantly Asian with Vietnamese being the top identifier.

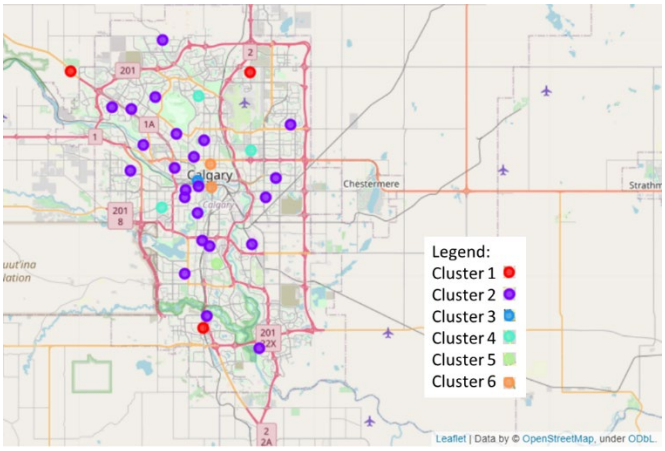


Figure 5. Map of Calgary city with the *k*-means clusters identified by the algorithm. The legend shows the colors corresponding to each cluster.

Discussion

The methodology used in this study has been shown to produce reproducible results using different datasets for the cities of Toronto, ON and Calgary, AB. By aggregating the total number of restaurants per neighborhood(s), as well as by cuisine type, the analysis clearly showed that Toronto has a larger number when compared to Calgary. This suggests that Toronto is a larger city with a more diverse and dense population than Calgary. Cluster analysis showed that Vietnamese cuisine was predominant in both cities. Therefore, a newcomer to the restaurant scene in both cities had a better chance of succeeding by offering a different type of cuisine, preferably one that is not found on the top 10 list of restaurant types. When looking at the total number of all restaurants for the neighborhoods in each city, there are several neighborhoods that appear to be underserved by restaurants. Nonetheless, further analysis is required to address the limitations of the current study, which will improve the accuracy of neighborhood clustering.

While the usefulness of the current study for prospective investors and/or restauranteurs are clear, urban planners and property developers will still benefit, albeit from a different aspect. By looking at the range of cuisine types and also the number of restaurants, urban planners can decide whether to build more infrastructure to support and further development of the area. The type of infrastructure will depend on the neighborhood. For example, a neighborhood with recreational areas may benefit from restaurants with smaller footprints, so as not to disturb the natural ambience. Another option would be dispersed spots that can be licensed to food vendors or food trucks to provide meal options to visitors of these recreational areas. Property developers can look at the concentration of restaurants in different neighborhoods and infer whether the area is better served by building a mixed-use residential or commercial building to increase the customer base. Conversely, neighborhoods with fewer restaurants may be better served by building shopping mall or outlet, which can house both shops and restaurants with ample parking facilities for local residents and visitors.

Study Limitations and Future Research Directions

The Foursquare API data clearly shows several overlapping categories, as well as undefined categories. For example, Japanese and Sushi restaurants can be agglomerated together as sushi is predominantly Japanese (Korean's own version is called gimbap). The main undefined category is of type 'Restaurant', which is not useful as it doesn't indicate the cuisine type. Another example of an undefined category is 'Asian Restaurant'. Given the plethora of cuisines found under this banner, a more discriminate classification is needed. Moreover, the type of restaurant service cannot be gleaned from the current dataset. As described by Ref. 4, there are three primary categories of service styles: quick-service, midscale and upscale. While quick-service style restaurant are often but not always equated with fast food restaurants, it is not certain where the other types of restaurant cuisines fall. The scale of the restaurant will also be useful in segmenting the clusters by size, in addition to cuisine type and service style. As mentioned earlier, a direct comparison between Toronto and Calgary was clouded by the unequal size of the areas covered in the analysis. Moving ahead, being able to normalize the search to area size would enable better comparisons between cities.

Datasets describing the demographics of each city would be useful as well. Key features of these datasets include age, income, education level, race and marital status. These will inform on the type of service styles that would be suitable for a restaurant. For example, Generation Y customers tend to go for fast-food and/or quick-service spots. In contrast, Generation X customers prefer family-style restaurants and depending on their income level, may frequent midscale and upscale restaurants [Ref. 4]. Another feature that may be useful is the development type for each neighborhood, which will determine the scale and service style of a restaurant. For example, a restaurant located in a city may have a smaller footprint due to higher rents, when compared to another located in the suburbs. Instead of a brick and mortar location, a restaurant can be run from a food truck to provide a quick-service style instead [Ref. 4].

Conclusion

The range of data science tools used in the current study has been shown to produce the desired analytics that will help the target audience with their business decisions. Additionally, the methodology has been tested on two different datasets and found to be reproducible. Based on the current results, Vietnamese cuisine is the top trend in both Toronto and Calgary. Therefore, prospective investors and/or restauranteurs should consider other cuisine types in their commercial business planning. Urban planners and property developers will find more opportunities in Calgary for initiating new and improving current developments. Nonetheless, further refinements are required to improve the methodology described herein, in order to obtain more nuanced and accurate analytics.

References

(1) List of Toronto boroughs and neighborhoods. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
(2) List of Calgary boroughs and neighborhoods. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_T)
(3) The Foursquare Developer Account and API (https://foursquare.com/)
(4) "How to Start a Restaurant" – access at the Entrepreneur website (https://www.entrepreneur.com/article/73384)