Data-driven Recipe to New Cuisine Business at Toronto, ON, Canada

Gilberto da Silva Guimarães Filho

April 25, 2021

1. Introduction and Business Problem

1.1 Problem Background

Toronto is Canada's biggest city, and its miles a global chief in business, finance, technology, leisure, and culture. A massive wide variety of immigrants from all around the world make Toronto one of the maximum multicultural towns withinside the globe. Toronto is one of the maximum livable towns withinside the international as proven through diverse statistics, reviews, and worldwide rankings.

A quote of Arthur C. Nielsen, Market Researcher & Founder of ACNielsen in Data Tells a Story. Are you Listening? says: "The price of light is less than the cost of darkness."

Toronto's diversity certainly poses several challenges in terms of data-driven decision-making. Fortunately, this is a challenge that can become an exclusive opportunity, precisely because of the diversity that Toronto presents. This project was conceived hypothetically to serve people or companies interested in opening their restaurants in Toronto. In this sense, we will analyze several neighborhoods in Toronto to see if there is any pattern of behavior in terms of preferences for specific types of cuisine.

1.2 Problem Description

A restaurant, eating place or an eatery, is a commercial enterprise that prepares and serves food and drink to customers. Meals are commonly served and eaten at the premises; however, many eating places additionally provide take-out and meals shipping services. Toronto cuisine business is recognized for its diversity and competitiveness.

Toronto's diversity guarantees that their meals producers stay at the leading edge of culinary developments and product developments, making the place not anything brief of a “Foodie Paradise,” supplying flavors and fares from throughout the world.

On the one hand, this competitiveness and diversity in the sector can pose serious challenges for those less experienced in the restaurant sector, on the other hand, diversity allows more entrepreneurs to settle in Toronto.

In this scenario of high competitiveness and diversity, making business data-driven decisions dramatically increases the chances of new business initiatives success at this field.

1.3 Interest

Stablishing a new local for a business could be a challenge task, for this reason some groups of customers or set of people may be more interested on it:

- Businessmen who desire to make investments or get insights on restaurant field in Toronto. This evaluation may be a manual to begin or enlarge eating places on Toronto's neighborhoods.
- Freelancers who like to have their very own eating place as an aspect business.
- Data science professionals who are interested in the topic.

2. Data acquisition and cleaning

2.1 Data sources

To have a look at the above said, following data assets could be used:

- Wikipedia's "List of Postal Code of Canada: M": this is a website that give us data about boroughs and neighborhoods with postal codes from Toronto, ON, Canada.
- Foursquare API: this tool allows interested people get location data and other resources by RESTfull API calls. Using this API, I will get some data like categories of business in Toronto and specificities of cuisine entrepreneurship at this city. Beyond this the Foursquare API will be used to get locations and some Toronto’s restaurants data.
- For geolocation data: Geopy, a Python library, will be used. This library is especially good in retrieving longitude and latitude data. But, if necessary, some additional tools will be used as well.
- If geolocation data Python packages does not work well for our purposes, Coursera's Toronto geospatial data will be used to get latitude and longitude at Toronto by neighborhood's postal codes.

2.2 Data cleaning and exploratory data analysis

After web scraping some geospatial data of Toronto from Wikipedia website, the data will be cleaned and then Exploratory Data Analysis will be done by data wrangling. K-means clustering machine learning algorithm will be used to try to find some insights about cuisine entrepreneurship at Toronto.

There are different methods to gather data. These two methods will be used because were studied in this certification process. Beside that Python libraries, packages and community have a lot of material for a faster process. As an example, if you wish to learn more complicated cases of web scraping using the BeautifulSoup Python package, you can start your journey here: http://beautiful-soup-4.readthedocs.io/en/latest/

In this work it will be used some hints form Cousera's course, like hints for web scraping. Pandas, another Python library, or the BeautifulSoup package, or any other way of transforming the data can be used as well.

2.3 Feature selection

In this work the focus will be on international cuisine. Therefore, during data processing, the selection of features will be made, hypothetically, with the choices of restaurants of interest by the customer. In this way, we will restrict our analysis according to its objective, namely: to provide insights for new entrepreneurs in the international restaurant industry in Toronto, ON, Canada.

Kept features	Dropped features
'Caribbean Restaurant', 'Chinese Restaurant', 'Indian Restaurant', 'African Restaurant', 'Filipino Restaurant', 'Greek Restaurant', 'German Restaurant', 'Japanese Restaurant', 'Italian Restaurant', 'Mexican Restaurant', 'Korean BBQ Restaurant', 'French Restaurant', 'South Indian Restaurant', 'Vegetarian / Vegan Restaurant', 'Thai Restaurant', 'Asian Restaurant', 'Hakka Restaurant', 'Halal Restaurant', 'Sushi Restaurant', 'Korean Restaurant', 'Middle Eastern Restaurant', 'Sri Lankan Restaurant', 'Falafel Restaurant', 'American Restaurant', 'Bangladeshi Restaurant', 'Pakistani Restaurant',	'Coffee Shop', 'Sandwich Place', 'Fast Food Restaurant', 'Fried Chicken Joint', 'Restaurant', 'Bakery', 'Burger Joint', 'Breakfast Spot', 'Fish & Chips Shop', 'Café', 'BBQ Joint', 'Pizza Place', 'Shawarma Place', 'Donut Shop', 'Grocery Store', 'Salad Place', 'Sports Bar', 'Juice Bar', 'Snack Place', 'Wings Joint', 'Ice Cream Shop', 'Food Truck', 'Deli / Bodega', 'Steakhouse', 'Diner', 'Tea Room', 'Gastropub', 'Cajun / Creole Restaurant', 'Harbor / Marina', 'Bistro', 'Hot Dog Joint', 'Bubble Tea Shop', 'Noodle House', 'Food Court', 'Internet Cafe', 'Event Space', 'Bar', 'Frozen Yogurt Shop', 'Burrito Place', 'Chaat Place', 'Pub', 'Cafeteria',

'Vietnamese Restaurant', 'North Indian Restaurant', 'Kebab Restaurant', 'Afghan Restaurant', 'Seafood Restaurant', 'Iraqi Restaurant', 'Cantonese Restaurant', 'Hong Kong Restaurant', 'Malay Restaurant', 'Taiwanese Restaurant', 'Dumpling Restaurant', 'Szechuan Restaurant', 'Ramen Restaurant', 'Udon Restaurant', 'Mediterranean Restaurant', 'Dim Sum Restaurant', 'Eastern European Restaurant', 'New American Restaurant', 'South American Restaurant', 'Persian Restaurant', 'Doner Restaurant', 'Tapas Restaurant', 'Cuban Restaurant', 'Brazilian Restaurant', 'Modern European Restaurant', 'Caucasian Restaurant', 'Latin American Restaurant', 'Turkish Restaurant', 'Portuguese Restaurant', 'Russian Restaurant', 'Bosnian Restaurant', 'Moroccan Restaurant', 'Ethiopian Restaurant', 'Venezuelan Restaurant', 'Egyptian Restaurant', 'Syrian Restaurant', 'Tibetan Restaurant', 'Belgian Restaurant', 'Kosher Restaurant', 'Lebanese Restaurant', 'Dutch Restaurant', 'Peruvian Restaurant', 'Colombian Restaurant', 'Empanada Restaurant', 'Indian Chinese Restaurant', 'Argentinian Restaurant', 'Hungarian Restaurant', 'Burmese Restaurant', 'Polish Restaurant', 'Scandinavian Restaurant', 'Swiss Restaurant', 'Cajun / Creole Restaurant'	'Buffet', 'Food Stand', 'Cha Chaan Teng', 'Bagel Shop', 'Pastry Shop', 'Cupcake Shop', 'Butcher', 'Gourmet Shop', 'Lounge', 'Factory', 'Office', 'Irish Pub', 'Hookah Bar', 'Beer Bar', 'Brewery', 'Wine Bar', 'Gelato Shop', 'Comfort Food Restaurant', 'Soup Place', 'Food & Drink Shop', 'Supermarket', 'Poke Place', 'Convenience Store', 'Gay Bar', 'Chocolate Shop', 'Market', 'Gluten-free Restaurant', 'Bowling Alley', 'Taco Place', 'Hotel', 'Cheese Shop', 'Pie Shop', 'Shopping Mall', 'Comic Shop', 'Theme Restaurant', 'Dessert Shop', 'Food', 'Hotpot Restaurant', 'Poutine Place'

Table 1 – Feature selection

3. Methodology

3.1 Web Scraping

Initially, a brief study of the space organization of Toronto, ON, Canada was made. From these data, it was realized the need to segment neighborhoods in regions of proximity. After this step of defining data that would be scraped from the web, this scraping was made from the following URL: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

For data scraping was used BeautifulSoup, a module from the Python bs4 library. Then, the structure of the downloaded data was understood, and the data was organized in a list of dictionaries containing the names of Toronto boroughs, neighborhoods, and postal codes.

After the data were downloaded and properly adjusted, the need for a listing of latitude and longitude by postal code was realized to carry out the necessary analyzes. This data was obtained through a link provided by Coursera: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv

The web scraping result was a data frame with 103 rows (entries) and 5 columns (features). The figure 1 bellow shows the result of this step:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
...	...	...	...	...	...
98	M9N	York	Weston	43.706876	-79.518188
99	M9P	Etobicoke	Westmount	43.696319	-79.532242
100	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.688905	-79.554724
101	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437
102	M9W	Etobicoke Northwest	Clairville, Humberwood, Woodbine Downs, West H...	43.706748	-79.594054

103 rows × 5 columns

Figure 1 – Data frame with geospatial data from Toronto, ON, Canada.

With these data, it was possible to plot a Toronto's chart containing all the neighborhoods that should be clustered according to the most common restaurants in these regions. This map can be viewed at next page (figure 2).

To interpret this map and its data it is important to note the following structure:

1. The last line of data is the borough to which the neighborhood marked on the map belongs.
2. The line (or lines) above borough is(are) the neighborhood(s) of Toronto.

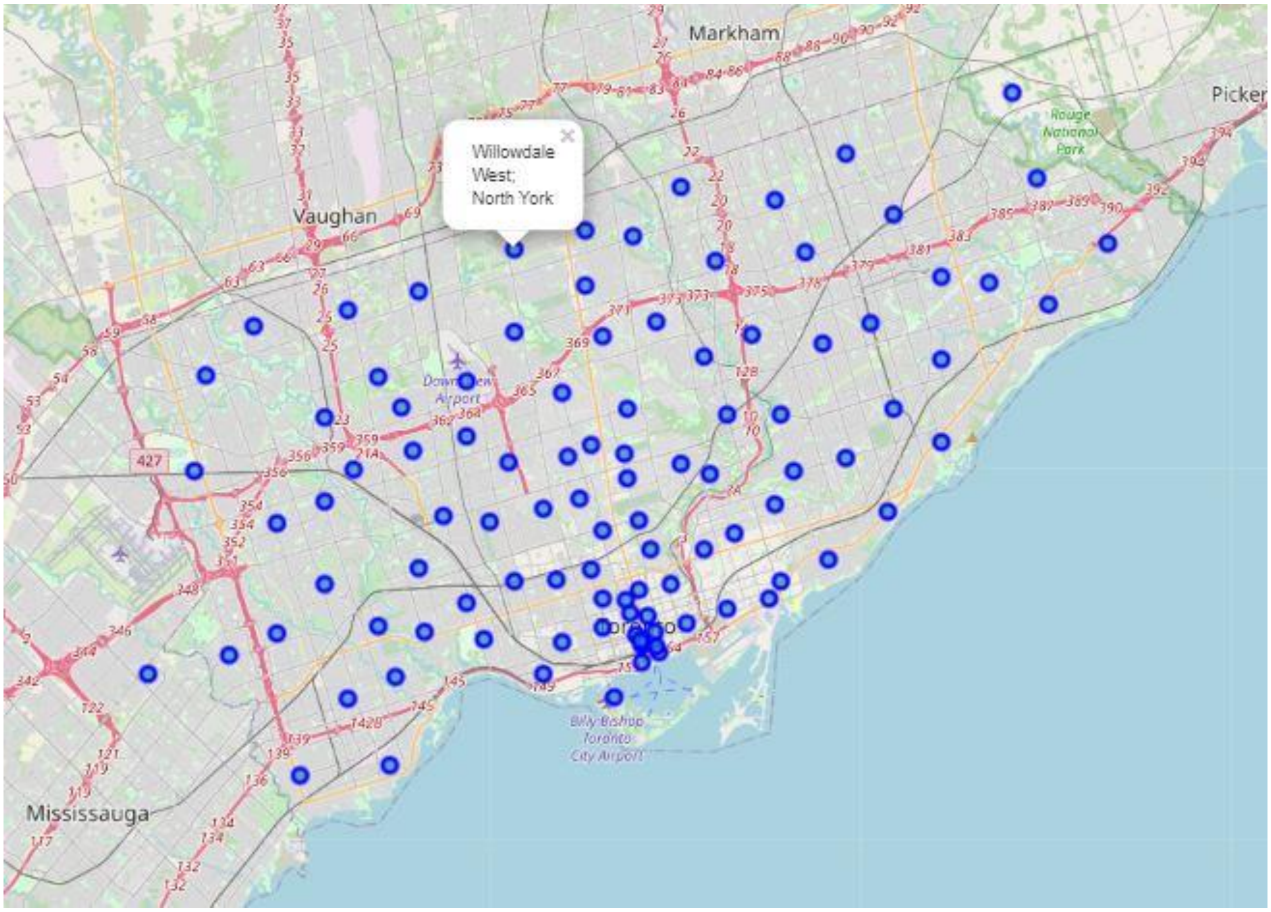

Figure 2 – Map from Toronto, ON, Canada with its neighborhoods.

3.2 Foursquare API, Pandas and Pickle

This API provided by Foursquare was used to collect data for the purpose of this project, i.e., to gather data like:

1. Business segment categories provided by Foursquare.
2. Sub-categories made available by Foursquare in the food category.

After knowing the available categories and sub-categories, a test with a neighborhood was carried out so that it was possible to make the inference of the analyzes that could be made from then on.

It was observed that the request returned some redundancy like two fields for name, for example. At the next page, figure 3 represents this issue. It can be observed that there are two names for the same place: ‘Fast Food Restaurant’, which means the category name and the name of establishment: ‘Wendy’s’.

To overcome this a function was defined and it loops through the neighborhoods of Toronto and creates an API request with a radius of 1,000 meters with a limit of 500 maximum nearby venues that should be returned. Then the function would get the relevant data necessary to the analysis for each nearby venue.


```
{'meta': {'code': 200, 'requestId': '6084734e75bcec2ea4e2725b'},
 'response': {'confident': False,
  'venues': [{'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/fastfood_',
       'suffix': '.png'},
      'id': '4bf58dd8d48988d16e941735',
      'name': 'Fast Food Restaurant',
      'pluralName': 'Fast Food Restaurants',
      'primary': True,
      'shortName': 'Fast Food'}],
    'hasPerk': False,
    'id': '4bb6b9446edc76b0d771311c',
    'location': {'cc': 'CA',
     'city': 'Toronto',
     'country': 'Canada',
     'crossStreet': 'Morningside & Sheppard',
     'distance': 387,
     'formattedAddress': ['Toronto ON', 'Canada'],
     'labeledLatLngs': [{'label': 'display',
       'lat': 43.80744841934756,
       'lng': -79.19905558052072}],
     'lat': 43.80744841934756,
     'lng': -79.19905558052072,
     'state': 'ON'},
    'name': 'Wendy’s',
    'referralId': 'v-1619293006'}]}}
```


Figure 3 – Redundancy in the name fields from returned data.

After being collected and adjusted, the data necessary for the analysis were organized in serialized data frames using the Python pickle and pandas libraries. The serialization of the data frames was done to improve their processing in future analyzes.

4. Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and look at data and summarize their maximum essential characteristics. This is a regular step in the process of getting information from data and regularly use visualization methods.

This step permits to discover some information to get some answers you need, making it much less hard for data scientists to discover patterns, spot anomalies, check a hypothesis, or check assumptions.

4.1 Important data from toronto_venues data frame

The merged data frame called toronto_venues have all the required information for this analysis. This data frame has 4,282 rows, which means the venues scattered throughout neighborhoods. It also has 7 columns that show data like neighborhoods, neighborhood latitude and longitude, venue, venue longitude, latitude, and category.

From this data frame it was possible to know what the unique categories were returned; these categories represent the boundaries of this study. The most occurring unique categories are:

Venue Category	
Coffee Shop	600
Pizza Place	266
Café	244
Restaurant	205
Bakery	201
Fast Food Restaurant	185
Chinese Restaurant	159
Sandwich Place	136
Italian Restaurant	108
Caribbean Restaurant	100
Bubble Tea Shop	94
Indian Restaurant	87
Sushi Restaurant	84
Grocery Store	82
Burger Joint	81
Fried Chicken Joint	72
Breakfast Spot	69
Asian Restaurant	68
Middle Eastern Restaurant	63
Ice Cream Shop	62
Japanese Restaurant	58
Thai Restaurant	52
Vietnamese Restaurant	51
BBQ Joint	50
Mexican Restaurant	49
Dessert Shop	47
Deli / Bodega	43
Diner	41
American Restaurant	36
Korean Restaurant	35

Figure 4 – Top 30 occurring unique venues at Toronto.

4.2 Data Cleaning

The customer's choice of restaurants was hypothetically assumed in this analysis. From the client's list of interests, the other categories were removed from this analysis. First, a list of all the unique categories that were available was made, and then the hypothetical customer chose the ones that most interested him.

That done, the other categories that were not of interest to the study were discarded. The result of this processing was a data frame composed of 77 unique restaurant categories, 1,165 unique restaurants and with no missing values, as can be seen in the next figure.

```
# filter specialized categories

toronto_venues = toronto_venues[toronto_venues['Venue Category'].isin(client_interest_food_venues)].reset_index()
toronto_venues.head(5)
```

	index	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	7	Malvern, Rouge	43.806686	-79.194353	Bella's Lechon	43.801291	-79.198378	Filipino Restaurant
1	8	Malvern, Rouge	43.806686	-79.194353	Charley's West Indian Foods Ltd	43.800581	-79.199309	Caribbean Restaurant
2	10	Malvern, Rouge	43.806686	-79.194353	Mr. Greek	43.799853	-79.198234	Greek Restaurant
3	15	Malvern, Rouge	43.806686	-79.194353	Just Soups	43.801286	-79.198802	Caribbean Restaurant
4	16	Malvern, Rouge	43.806686	-79.194353	Charley's Exotic Cuisine	43.800982	-79.200233	Chinese Restaurant

```
# print unique venues categories

print(f'Number of unique restaurant categories selected by client at Toronto, ON, Canada: {len(toronto_venues["Venue Category"].unique())}')
```

```
Number of unique restaurant categories selected by client at Toronto, ON, Canada: 77
```

```
# print unique venues

print(f'Number of unique restaurants in categories selected by client at Toronto, ON, Canada: {len(toronto_venues["Venue"].unique())}')
```

```
Number of unique restaurants in categories selected by client at Toronto, ON, Canada: 1165
```

```
# check the quantity of missing values

print(f'There is(are) {toronto_venues.isnull().sum().sum()} missing value(s) in this dataframe')
```

```
There is(are) 0 missing value(s) in this dataframe
```

Figure 5 – toronto_venues data frame and some important data

4.3 Insights from communities

Now, every community will be analyzed separately to recognize the maximum restaurant types inside its 1,000 meters of vicinity. The one hot encoding technique is a feature of python pandas library that converts the explicit variables (which are venue Category) right into a shape that might be supplied to ML algorithms to do a higher task in prediction. This technique was applied and then some statistical results were printed, and the top 10 occurrences of restaurants was found at this data set. The result can be verified bellow.

	count	mean	std	min	25%	50%	75%	max
Chinese Restaurant	101.0	1.574257	2.113511	0.0	0.0	1.0	2.0	13.0
Greek Restaurant	101.0	0.336634	1.328738	0.0	0.0	0.0	0.0	11.0
Middle Eastern Restaurant	101.0	0.623762	1.255798	0.0	0.0	0.0	1.0	10.0
Caribbean Restaurant	101.0	0.990099	1.705843	0.0	0.0	0.0	1.0	10.0
Korean Restaurant	101.0	0.346535	1.117458	0.0	0.0	0.0	0.0	10.0
Asian Restaurant	101.0	0.673267	1.040278	0.0	0.0	0.0	1.0	7.0
Vietnamese Restaurant	101.0	0.504950	0.996231	0.0	0.0	0.0	1.0	6.0
Indian Restaurant	101.0	0.861386	1.249237	0.0	0.0	0.0	1.0	6.0
Italian Restaurant	101.0	1.069307	1.358362	0.0	0.0	1.0	1.0	6.0
Sushi Restaurant	101.0	0.831683	1.068357	0.0	0.0	0.0	1.0	5.0

Figure 6 – top 10 occurrences of restaurants founded.

4.4 Data Visualization of Top 10 occurrences of restaurants at Toronto

From the information above we know the top 10 restaurants occurrences at Toronto. Now, through plotting a bar chart, it will be found the neighborhoods where these restaurants are distributed.

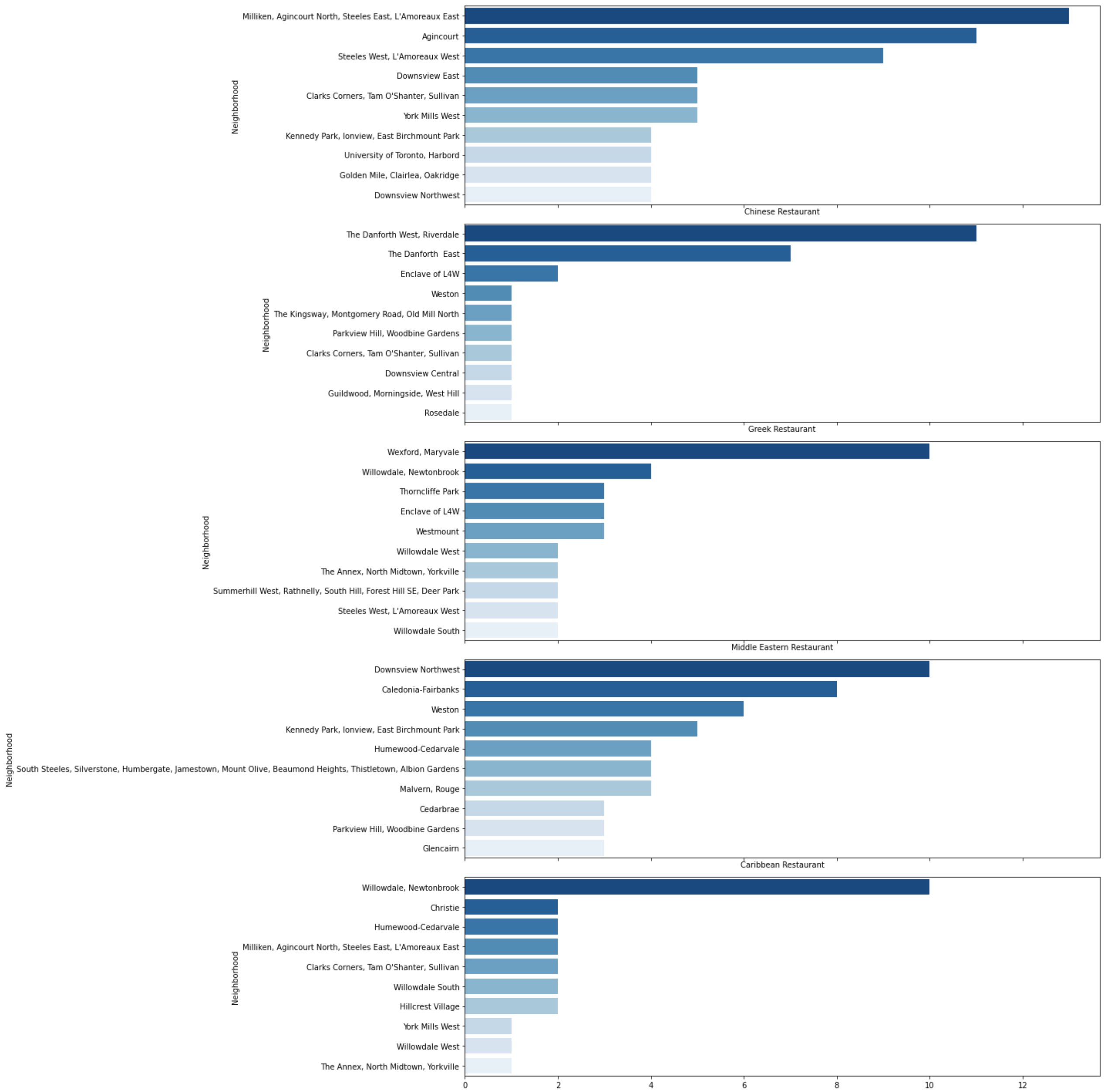

Figure 7 – top 10 occurrences of restaurants spread by neighborhoods.

5. Clusterization Modeling

5.1 K-means

K-means is one of the most broadly used unsupervised clustering methods. The K-means set of rules clusters the records handy with the aid of using seeking to separate samples into K organizations of identical variance, minimizing a criterion called the inertia or within-cluster sum-of-squares. This set of rules calls for the variety of clusters to be specified. It scales nicely to big variety of samples and has been used throughout a big variety of software regions in lots of one-of-a-kind fields.

The k-means set of rules divides a fixed of N samples (saved in a records matrix X) into K disjoint clusters C, every defined with the aid of using the suggest μj of the samples withinside the cluster. The means are generally known as the cluster “centroids”.

K-means set of rules falls into the circle of relatives of unsupervised clusterization algorithms/methods. For this own circle of relatives of models, the studies desires to have handy a dataset with a few observations without the want of getting additionally the labels/instructions of the observations. Unsupervised clusterization research how structures can infer a feature to explain a hidden shape from unlabeled records.

5.1.1 Elbow and Silhouette Methods

The Elbow Method calculates the sum of squared distances of samples to their closest cluster middle for unique values of ‘k’. The most appropriate range of clusters is the value after which there may be no great lower values withinside the sum of squared distances.

Sometimes Elbow technique does not provide the desired result, which occurred on this case. As, there may be a slow lower withinside the sum of squared distances, optimum quantity of clusters can't be determined. To counter this, other technique may be implemented: the silhouette method. According to Wikipedia, “the silhouette method measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).” This definition can be found at: https://en.wikipedia.org/wiki/Silhouette_(clustering)

At this study 9 was chosen as the best number for k-means clusterization. This value as assumed after analyzing Silhouette method, because Elbow technique does not show a good result. Even silhouette score was bellowing the minimum expected of 0.3, but for this hypothetical study the focus will be on the methodology by itself. For better results further analysis must be done and some paths inside this technique should be reviewed, like check the need of standardization or even try other techniques, like DBSCAN, for example.

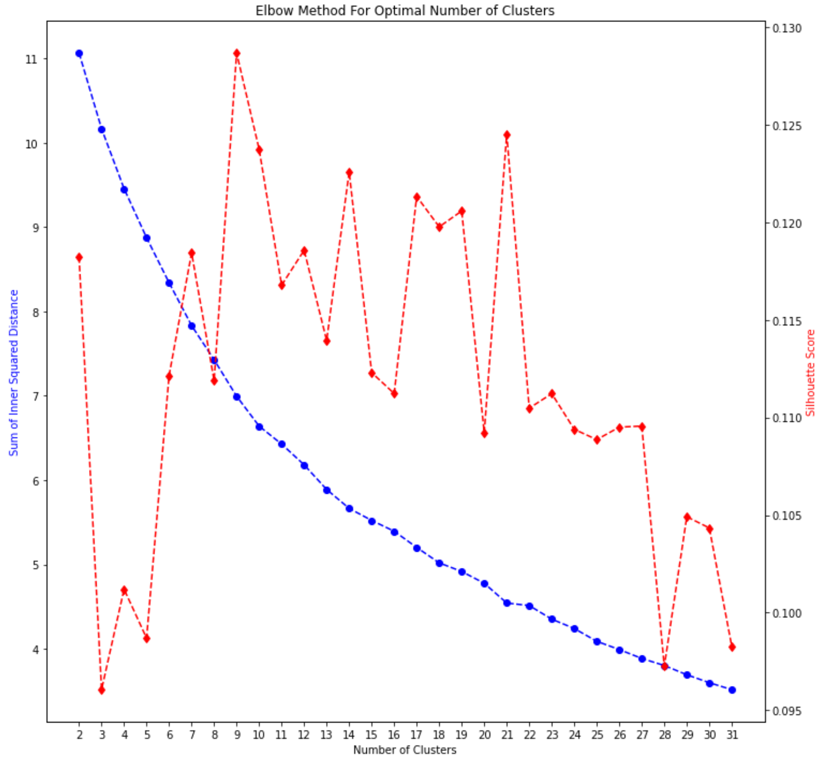

Figure 8 – Elbow and Silhouette methods to find the optimal number of clusters.

5.1.2 K-means clusterization results: neighborhoods segmentation

After defining the optimal number of clusters, the results of the clustering were delivered by the KMeans class of the Python sklearn library. As soon as the clustering is done, it is possible to plot the graph that shows how the geographic layout of the neighborhoods was in relation to the relative characteristics of the types of restaurants present. See the next figure.

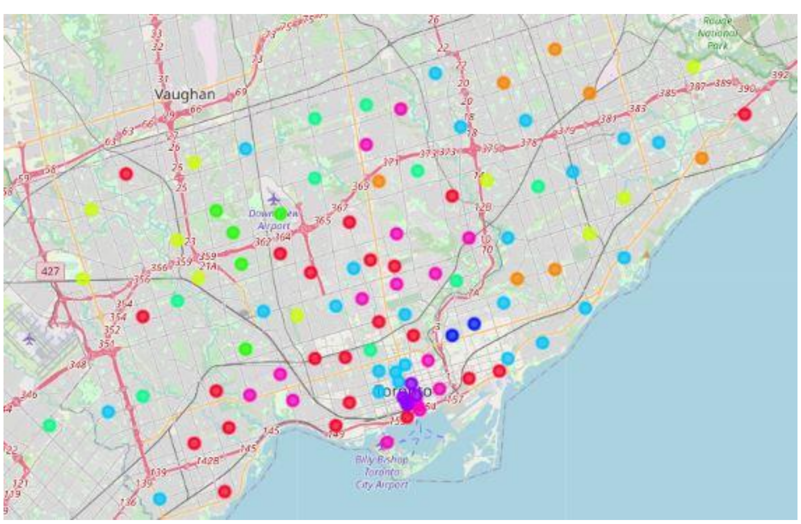

Figure 9 – Geographical results of k-means clusterization method.

6. Clusterization Results

6.1.1 K-means clusterization results: Cluster 1

Indian Restaurant holds the greatest accountability for this cluster with 5 occurrences in First Most Common Venue across different neighborhoods followed by Asian Restaurant with 4 occurences. The Asiatic food sounds to be predominant in this cluster and many of the neighborhoods from this cluster are in Scarborough, Toronto.

Cluster 1 – Most boroughs occurrences	Cluster 1 – Most restaurants occurrences
Scarborough 5 East Toronto 2 Etobicoke 1 North York 1 Central Toronto 1 Downtown Toronto 1 Name: Borough, dtype: int64 --	Indian Restaurant 5 Asian Restaurant 3 Caribbean Restaurant 2 Japanese Restaurant 1 Name: 1st Top Common Venue, dtype: int64 -- Asian Restaurant 4 Indian Restaurant 2 Caribbean Restaurant 2 Japanese Restaurant 1 Vietnamese Restaurant 1 South Indian Restaurant 1 Name: 2nd Top Common Venue, dtype: int64 -- Indian Restaurant 2 Thai Restaurant 2 French Restaurant 1 Caribbean Restaurant 1 Egyptian Restaurant 1 Vegetarian / Vegan Restaurant 1 Asian Restaurant 1 Afghan Restaurant 1 Filipino Restaurant 1 Name: 3rd Top Common Venue, dtype: int64 --

Table 2 – Cluster 1

6.1.2 K-means clusterization results: Cluster 2

Chinese Restaurant holds the greatest accountability for this cluster with 6 occurrences in First Most Common Venue across different neighborhoods followed by Vietnamese Restaurant with 4 occurrences. The attention point for this cluster is the Caribbean Restaurant, it appears in third place in terms of occurrence in First Most Common Venue for neighborhoods in this cluster. The Chinese and Vietnamese migration in this cluster sounds to be great, maybe there is a market opportunity or saturation.

Cluster 2 – Most boroughs occurrences	Cluster 2 – Most restaurants occurrences
North York 6 York 4 Downtown Toronto 3 Etobicoke Northwest 1 Etobicoke 1 Scarborough 1 Queen's Park 1 East York 1 Name: Borough, dtype: int64 --	Chinese Restaurant 6 Vietnamese Restaurant 4 Caribbean Restaurant 2 Thai Restaurant 1 Latin American Restaurant 1 Eastern European Restaurant 1 Asian Restaurant 1 African Restaurant 1 Portuguese Restaurant 1 Name: 1st Top Common Venue, dtype: int64 -- Vietnamese Restaurant 3 Latin American Restaurant 2 Caribbean Restaurant 2 Italian Restaurant 2 Japanese Restaurant 2 Chinese Restaurant 2 Korean Restaurant 1 French Restaurant 1 Scandinavian Restaurant 1 Mexican Restaurant 1 African Restaurant 1 Name: 2nd Top Common Venue, dtype: int64 -- Caribbean Restaurant 3 Halal Restaurant 2 Portuguese Restaurant 2 Chinese Restaurant 2 Turkish Restaurant 1 French Restaurant 1 Vietnamese Restaurant 1 Cantonese Restaurant 1 Italian Restaurant 1 Japanese Restaurant 1 Latin American Restaurant 1 Thai Restaurant 1 Afghan Restaurant 1 Name: 3rd Top Common Venue, dtype: int64 --

Table 3 – Cluster 2

6.1.3 K-means clusterization results: Cluster 3

In this cluster we can observe the massive presence of restaurants specialized in Middle Eastern cuisine with 4 occurrences in First Most Common Venue across different neighborhoods. Then we can also observe a strong presence of Asian cuisine, especially Chinese and Japanese.

Cluster 3 – Most boroughs occurrences	Cluster 3 – Most restaurants occurrences
North York 5 Etobicoke 2 East York 1 Mississauga 1 Central Toronto 1 Scarborough 1 Name: Borough, dtype: int64 --	Middle Eastern Restaurant 4 Chinese Restaurant 2 Korean Restaurant 1 French Restaurant 1 Indian Restaurant 1 Sushi Restaurant 1 Mexican Restaurant 1 Name: 1st Top Common Venue, dtype: int64 -- Middle Eastern Restaurant 4 Mediterranean Restaurant 2 Sushi Restaurant 2 Japanese Restaurant 1 African Restaurant 1 Thai Restaurant 1 Name: 2nd Top Common Venue, dtype: int64 -- Chinese Restaurant 3 Caribbean Restaurant 2 Dumpling Restaurant 1 Middle Eastern Restaurant 1 Japanese Restaurant 1 Vietnamese Restaurant 1 Sushi Restaurant 1 Greek Restaurant 1 Name: 3rd Top Common Venue, dtype: int64 --

Table 4 – Cluster 3

6.1.4 K-means clusterization results: Cluster 4

Cluster 4 – Most boroughs occurrences	Cluster 4 – Most restaurants occurrences
North York 5 Etobicoke 4 West Toronto 3 Central Toronto 3 Downtown Toronto 2 Scarborough 1 East Toronto Business 1 East Toronto 1 Name: Borough, dtype: int64 --	Italian Restaurant 18 Asian Restaurant 1 Japanese Restaurant 1 Name: 1st Top Common Venue, dtype: int64 -- Sushi Restaurant 3 Chinese Restaurant 3 Vietnamese Restaurant 2 Italian Restaurant 2 Latin American Restaurant 1 German Restaurant 1 Polish Restaurant 1 American Restaurant 1 Middle Eastern Restaurant 1 Japanese Restaurant 1 Caribbean Restaurant 1 Falafel Restaurant 1 Mexican Restaurant 1 New American Restaurant 1 Name: 2nd Top Common Venue, dtype: int64 -- Japanese Restaurant 3 Indian Restaurant 3 Mexican Restaurant 2 Thai Restaurant 2 Ramen Restaurant 1 French Restaurant 1 Vegetarian / Vegan Restaurant 1 Korean Restaurant 1 Caribbean Restaurant 1 Falafel Restaurant 1 Tapas Restaurant 1 Afghan Restaurant 1 Dim Sum Restaurant 1 African Restaurant 1 Name: 3rd Top Common Venue, dtype: int64 --

Table 5 – Cluster 4

Italian Restaurant holds a massive accountability for cluster 4 with 18 occurrences followed in First Most Common Venue across different neighborhoods in this cluster. Despite the predominance of Italian cuisine in this cluster, we can see that Chinese and Vietnamese cuisine are also a relevant presence and maybe a business opportunity.

6.1.5 K-means clusterization results: Cluster 5

Japanese Restaurant holds the greatest accountability for this cluster with 9 occurrences in First Most Common Venue across different neighborhoods. This finding is reinforced by the 6 occurrences of Japanese restaurants when we look at the ranking of Second Most Common Venue from this cluster. Despite the predominance of Japanese cuisine in this clustered set of neighborhoods, we must be aware of Italian and Indian cuisines, because when we look at Third Most Common Venue in this cluster Italian cuisine has 5 occurrences and Indian cuisine has 4 occurrences.

Cluster 5 – Most boroughs occurrences	Cluster 5 – Most restaurants occurrences
Downtown Toronto 6 Central Toronto 4 West Toronto 3 North York 3 Etobicoke 2 East York 1 Name: Borough, dtype: int64 --	Sushi Restaurant 9 Italian Restaurant 2 Asian Restaurant 2 Mediterranean Restaurant 1 Japanese Restaurant 1 Indian Restaurant 1 Falafel Restaurant 1 American Restaurant 1 Chinese Restaurant 1 Name: 1st Top Common Venue, dtype: int64 -- Sushi Restaurant 6 Japanese Restaurant 4 Italian Restaurant 2 Chinese Restaurant 2 Mexican Restaurant 2 Asian Restaurant 2 Indian Restaurant 1 Name: 2nd Top Common Venue, dtype: int64 -- Italian Restaurant 5 Indian Restaurant 4 Sushi Restaurant 2 Thai Restaurant 1 Middle Eastern Restaurant 1 Ramen Restaurant 1 French Restaurant 1 North Indian Restaurant 1 Vietnamese Restaurant 1 Asian Restaurant 1 American Restaurant 1 Name: 3rd Top Common Venue, dtype: int64 --

Table 6 – Cluster 5

6.1.6 K-means clusterization results: Cluster 6

This cluster is dominated by Caribbean cuisine with 5 occurrences in the First Most Common Venue in this cluster. Chinese cuisine, on the other hand, has 4 occurrences in the Second Most Common Venue and this represents the first place in this segmentation.

Cluster 6 – Most boroughs occurrences	Cluster 6 – Most restaurants occurrences
North York 3 Scarborough 2 York 1 Name: Borough, dtype: int64 --	Caribbean Restaurant 5 Italian Restaurant 1 Name: 1st Top Common Venue, dtype: int64 -- Chinese Restaurant 4 Greek Restaurant 1 Caribbean Restaurant 1 Name: 2nd Top Common Venue, dtype: int64 -- Vietnamese Restaurant 2 Afghan Restaurant 1 Sri Lankan Restaurant 1 German Restaurant 1 African Restaurant 1 Name: 3rd Top Common Venue, dtype: int64 --

Table 7 – Cluster 6

6.1.7 K-means clusterization results: Cluster 7

Chinese Restaurant holds the greatest accountability for this cluster with 7 occurrences in First Most Common Venue across different neighborhoods. The attention point for this cluster is the Indian, Caribbean, and Mexican cuisines. There are 3 Indian Restaurants in the Second Most Common Venue and 2 Caribbean Restaurants, but when Third Most Common Venue is taken in consideration Mexican and Caribbean cuisines appear with 2 occurrences each.

Cluster 7 – Most boroughs occurrences	Cluster 7 – Most restaurants occurrences
Scarborough 6 East York 1 North York 1 Name: Borough, dtype: int64 --	Chinese Restaurant 7 American Restaurant 1 Name: 1st Top Common Venue, dtype: int64 -- Indian Restaurant 3 Caribbean Restaurant 2 Thai Restaurant 1 Vietnamese Restaurant 1 Dumpling Restaurant 1 Name: 2nd Top Common Venue, dtype: int64 -- Mexican Restaurant 2 Caribbean Restaurant 2 Chinese Restaurant 1 Asian Restaurant 1 Japanese Restaurant 1 Indian Restaurant 1 Name: 3rd Top Common Venue, dtype: int64 --

Table 8 – Cluster 7

6.1.8 K-means clusterization results: Cluster 8

A Ramen restaurant denotes a preference for Japanese cuisine in the First and Second Most Common places in this cluster. The point of attention of this cluster is the Belgian cuisine that appears at second place in the list of Second Most Common places in the cluster and at the first place in the ranking of Third Most Common Places.

Cluster 8 – Most boroughs occurrences	Cluster 8 – Most restaurants occurrences
Downtown Toronto 5 Downtown Toronto Stn A 1 Name: Borough, dtype: int64 --	Ramen Restaurant 3 Sushi Restaurant 2 Mexican Restaurant 1 Name: 1st Top Common Venue, dtype: int64 -- Ramen Restaurant 2 Belgian Restaurant 2 Chinese Restaurant 1 Japanese Restaurant 1 Name: 2nd Top Common Venue, dtype: int64 -- Belgian Restaurant 2 Chinese Restaurant 1 Mexican Restaurant 1 Asian Restaurant 1 American Restaurant 1 Name: 3rd Top Common Venue, dtype: int64 --

Table 9 – Cluster 8

6.1.9 K-means clusterization results: Cluster 9

Despite the predominance of Greek cuisine in this cluster, it is not possible to reach any conclusion if there is a real predominance of Greek cuisine over other cuisines in this cluster due to the little occurrence that can be observed.

Cluster 9 – Most boroughs occurrences	Cluster 9 – Most restaurants occurrences
East Toronto 1 East York/East Toronto 1 Name: Borough, dtype: int64 --	Greek Restaurant 2 Name: 1st Top Common Venue, dtype: int64 -- Mediterranean Restaurant 1 Thai Restaurant 1 Name: 2nd Top Common Venue, dtype: int64 -- Chinese Restaurant 1 Ramen Restaurant 1 Name: 3rd Top Common Venue, dtype: int64 --

Table 10 – Cluster 9

7. Conclusion

The application of the k-means clustering algorithm to a multidimensional data set can generate interesting results that can be selected in business decision making.

In this study, it is possible to observe that boroughs (and even neighborhoods) at the city of Toronto have a great diversity in terms of different types of cuisines that are present within a radius of 1,000 meters of their latitude and longitude of reference.

Such regions were segmented into 9 groups and, after the application of the k-means algorithm, it was possible to observe some interesting information that needs further analysis or the use of other algorithms for a better verification of patterns.

8. Complementary Analysis

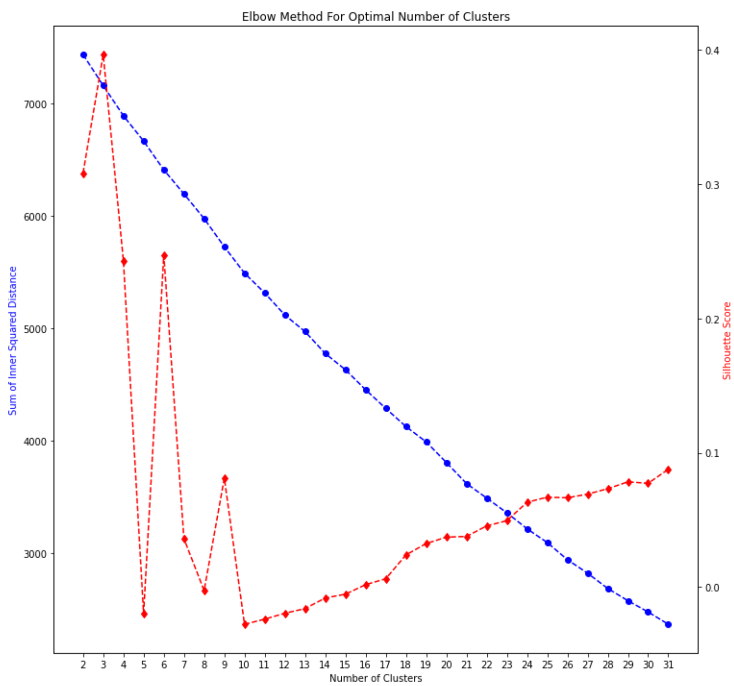

Figure 10 – Elbow and Silhouette methods to find the optimal number of clusters after standardization.

After standardizing the data frame that shows the count of restaurants present in each neighborhood, the elbow and silhouette methods were applied to identify the optimal number of clusters. As can be seen, although we have a good silhouette score when the number of clusters is 3, the inner squared sum of the distances is quite high. For this reason, it is concluded that the standardization of data from the mentioned data frame did not improve this model. The results of this study can and should be improved by other clustering algorithms and new more up-to-dated databases that can be used to improve this initial study.