The Battle of Neighborhoods - Coursera Capstone Project

A JOURNEY INTO THE RESTAURANTS OF TORONTO

Introduction

- The goal of this project is to give a recommendation to tourists in Toronto regarding the district of the city in which they could find the higher concentration of specific kinds of restaurants. The target audience of the project are indeed foreign tourists looking for a culinary experience in the city of Toronto.
- After a brief analysis of the venues in Toronto, the work will focus on the restaurants distribution over the territory; in particular, we will classify the neighborhood by clustering them according to the type of restaurants mostly represented in the neighborhood itself.
- Finally, the analysis will focus on the possibility of finding vegetarian restaurants in the city of Toronto.

Data

- We will leverage Foursquare location data and machine learning to address the problem, in particular, Foursquare location data and clustering methods will allow to group the neighbourhoods according to their restaurant venues information.
- In detail, the data will be collected via several CVS file from difference data sources:
  - ✓ via Wikipedia, we will collect the list of neighbourhoods in Toronto (
  - ✓ via Geocoder package, we will address the Geographical location of the neighbourhoods
  - ✓ via Forursquare we will collect the Venue data, and in particular the restaurants in Toronto.

Data acquisition

1. **Toronto Neighborhoods**: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M , the Wiki page of Toronto Neighborhood provided all the information about the postal code, borough and the name of the neighbourhoods of Toronto.

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

2. **Geographical location**: https://cocl.us/Geospatial_data , using the Geocoder Package we obtained the geographical coordinates of the different neighborhoods in Toronto.

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

3. **Venues data:** venues data, and in particolar restaurant ones, have been obtained by Foursquare.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Methodology

Data preparation

After having imported all the relevant informations, dataset were cleaned out and merged toghether.

- First of all:
  1. Neighbourhoods data coming from the Wiki page were cleaned up by deleting rows with not assigned borough
  2. Neighbourhoods with the same postal code were combined in a single row
- Neighbourhoods data were then merged with geographical coordinates based on posta codes, using latitude and longitude collected via Geocoder package.
- Finally, Venues data coming from Foursquare were merged with the previous ones, obtaining the following datatframe

[22]:

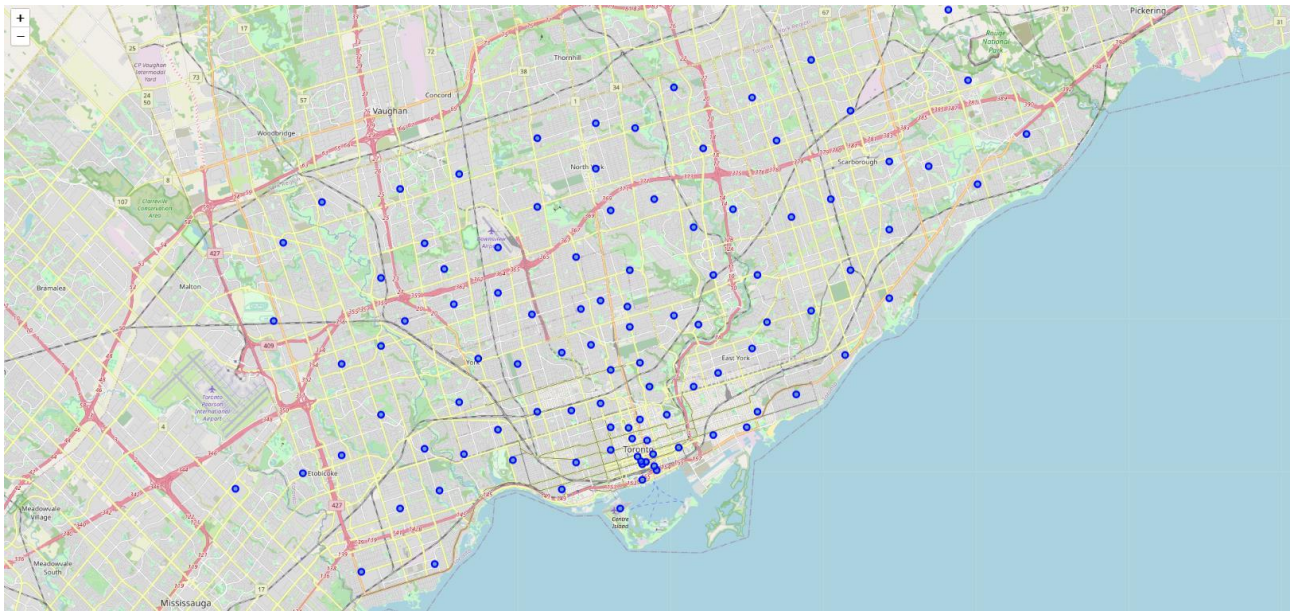
	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Data exploration

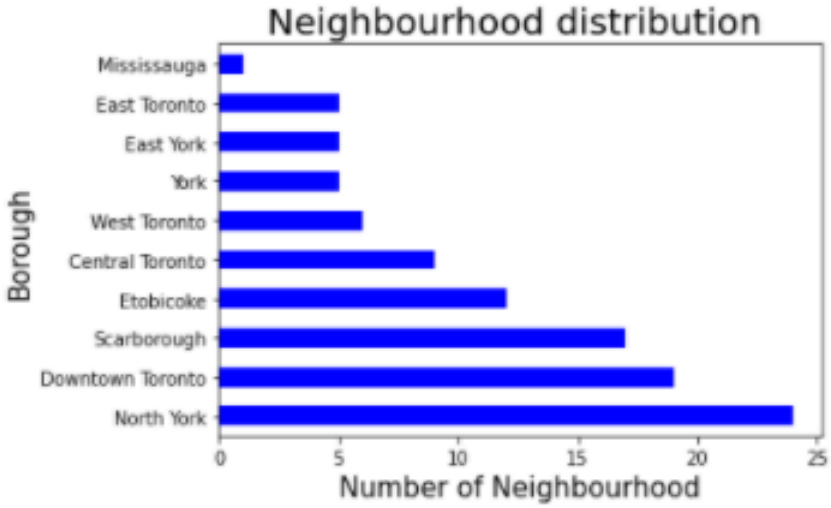
At this point we can start exploring our data set. We will proceed starting from the neighbourhood distribution over the territory. We will then figure out where is the higher concentration of venues and in particular we will focus on restaurants distribution over the neighbourhoods.

Neighbourhoods distribution

In the following map we can observe the neighbourhood distribution per borough in Toronto.



The higher number of neighbourhoods is in the borough of North York followed by Donwntown Toronto. While the smalled borough in terms of neighbourhoods number is Miissauga.

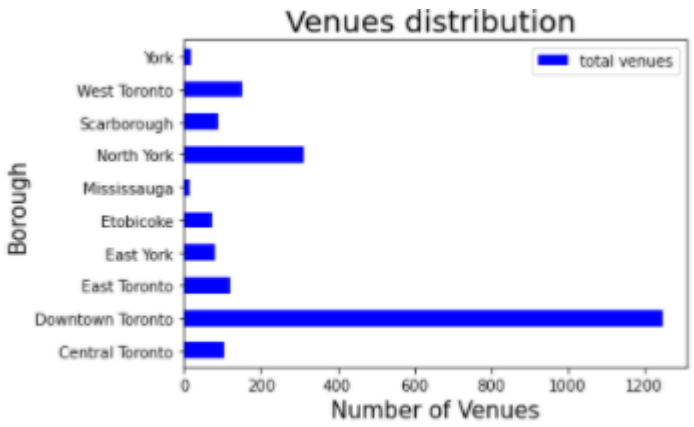


Venues distribution

In Toronto there are 2141 venues. Via onehot encoding we have studied the distribution of venues per neighbourhoods.

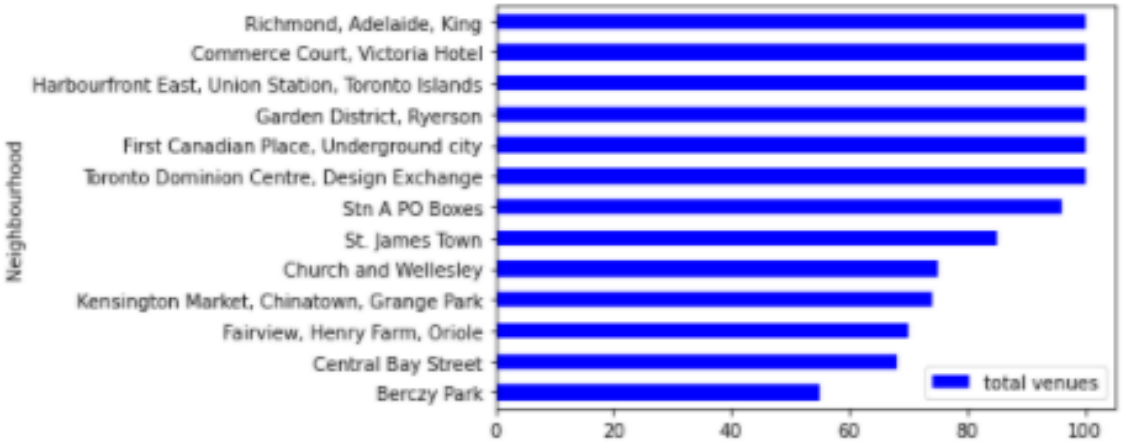
	Neighbourhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	Auto Workshop	BBQ Joint	Baby Store	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Baseball Stadium	Basketball Stadium	Beach	Bed & Breakfast	Beer Bar	Beer Store	Belgian Restaurant	Bike Shop	Bistro	Boat or Ferry	Bookstore	E
0	Agincourt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Alderwood, Long Branch	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Bathurst Manor, Wilson Heights, Downsview North	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	
3	Bayview Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
4	Bedford Park, Lawrence Manor East	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

We observed that most of the venues are mostly concentrated in the Borough of Downtow Toronto (1248), followed by North York (312) and West Toronto (153). Datails are reported in the following graph and table.



	Borough	total venues
1	Downtown Toronto	1248
6	North York	312
8	West Toronto	153
2	East Toronto	119
0	Central Toronto	104
7	Scarborough	90
3	East York	79
4	Etobicoke	74
9	York	20
5	Mississauga	13

While in the following chart we show the total number of venues per neighbourhood, reporting the ones which contain more than 50 venues.



Restaurants distribution

At this point we selected from the venues dataframe, the data related to the Restaurants only.

```
[47]: restaurants_df= toronto_venues[toronto_venues['Venue Category'].str.contains("Restaurant")]

[48]: print('There are {} unique categories or cusines available in Toronto.'.format(len(restaurants_df['Venue Category'].unique())))

      There are 50 unique categories or cusines available in Toronto.
```

We observed that there are 50 different style of cusines in Toronto.

We then repeated the onehot encoding for the Restaurants data and we found that there are 483 restaurants in Toronto.

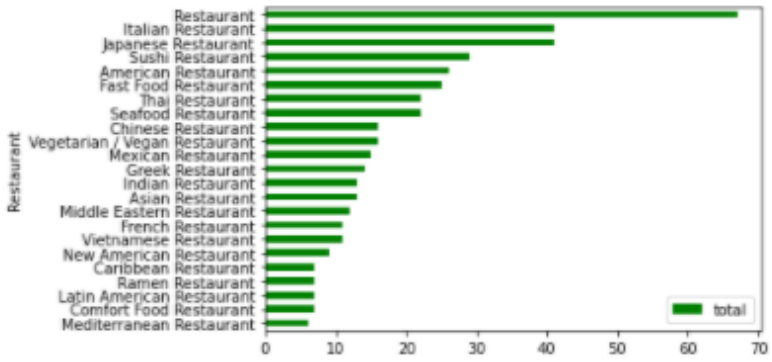
[49]:

	Neighbourhood	Afghan Restaurant	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Colombian Restaurant	Comfort Food Restaurant	Cuban Restaurant	Dim Sum Restaurant	Doner Restaurant	Dumpling Restaurant	Eastern European Restaurant
4	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

[42]:
```
print('There are {} restaurants in Toronto with {} different style of cuisines.'.format(restaurants_onehot.shape[0],(restaurants_onehot.shape[1]-1)))
```

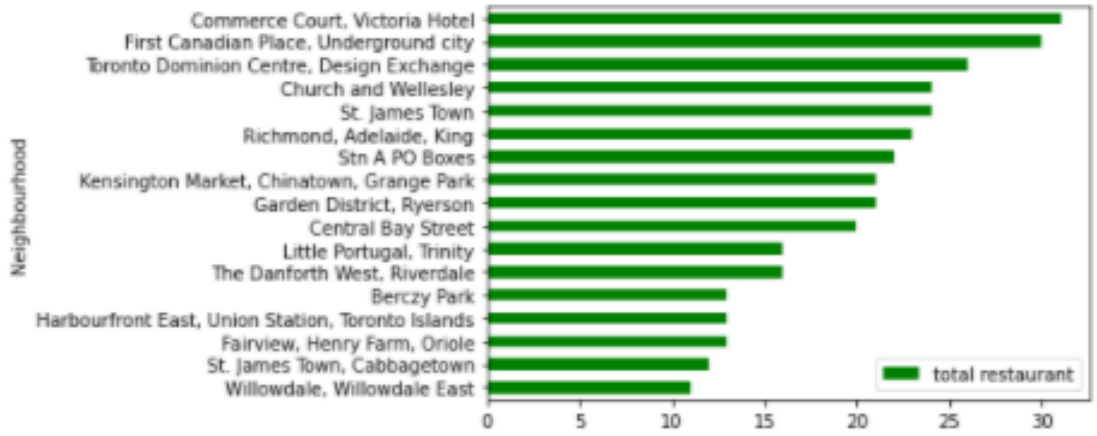
```
There are 483 restaurants in Toronto with 50 different style of cuisines.
```

Below we report the number of restaurants for each category: the most represented cusines are the Japanese and the Italian one, with 41 restaurants in Toronto, followed by Sushi Restaurants (29).



	Restaurant	total
40	Restaurant	67
25	Italian Restaurant	41
26	Japanese Restaurant	41
43	Sushi Restaurant	29
1	American Restaurant	26
17	Fast Food Restaurant	25
45	Thai Restaurant	22
41	Seafood Restaurant	22
48	Vegetarian / Vegan Restaurant	16

Finally, in the follong chart we report the number of restaurants in each neighbourhood, selecting the neighbourhoods with more than 5 restaurants each.



We then proceed in calculating the frequency of occurrence of each restaurant category, which will be used for clustering of neighbourhoods based on the Restaurants distribution.

```
[73]: toronto_restaurant_grouped = restaurants_onehot.groupby('Neighbourhood').mean().reset_index()
toronto_restaurant_grouped.head()
```

[73]:

	Neighbourhood	Afghan Restaurant	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Colombian Restaurant	Comfort Food Restaurant	Cuban Restaurant	Dim Sum Restaurant	Doner Restaurant	Dumpling Restaurant	Eastern European Restaurant	Ethiopian Restaurant	Falafel Restaurant
0	Agincourt	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	Bayview Village	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.50	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	Bedford Park, Lawrence Manor East	0.0	0.111111	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.111111	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Berczy Park	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.076923	0.0	0.0	0.0	0.0	0.076923	0.0	0.0

We observe that in the dataframe there are 67 Restaurants wich do not have any category, since this can be a source of noise for the following clustering procedure, we drop such column from the dataframe.

```
[67]: #Delete the "Restaurant" coiumn since it is not classified
toronto_restaurant_grouped.drop(['Restaurant'],axis=1,inplace=True)

[68]: toronto_restaurant_grouped.shape

[68]: (61, 50)
```

Machine learning

We are now ready to clusterize the Neighbourhoods of Toronto based on the most represented Restaurants.

We prepared the datatset by dropping the neighbourhood column from the toronto_restaurant dataframe:

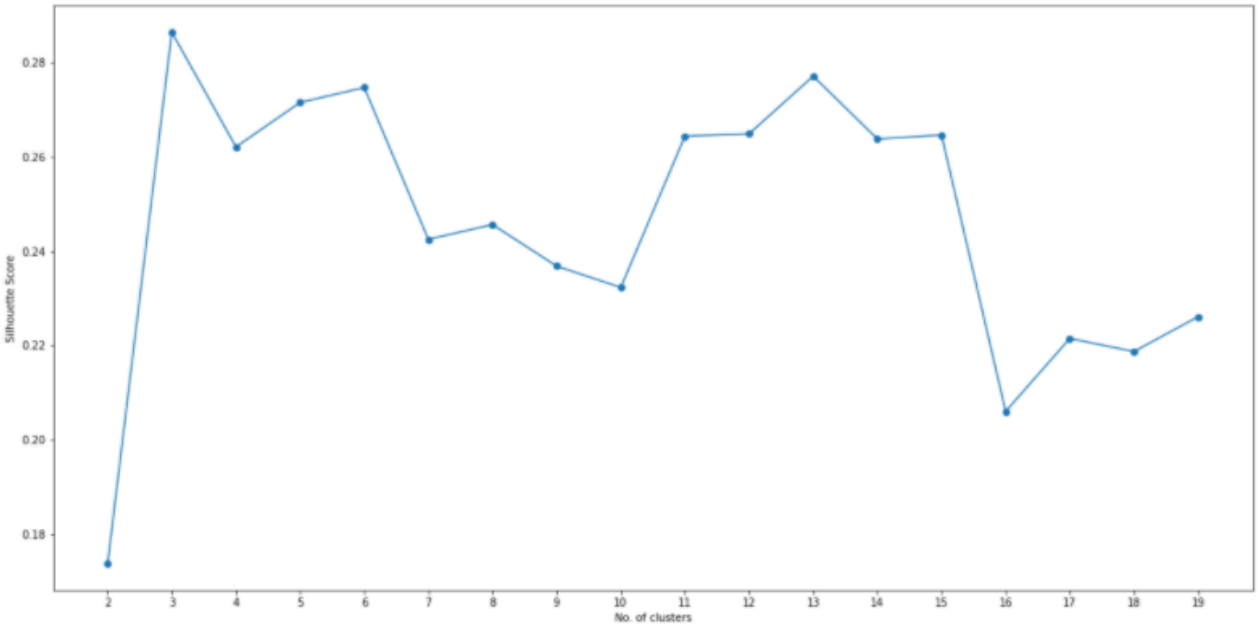
```
[79]: toronto_restaurant_clustering = toronto_restaurant_grouped.drop('Neighbourhood', 1)

[80]: toronto_restaurant_clustering.shape

[80]: (61, 49)
```

We found the best k for the clustering by plotting the silhouette_score:

```
[84]: from sklearn.metrics import silhouette_samples, silhouette_score
def plot(x, y, xlabel, ylabel):
    plt.figure(figsize=(20,10))
    plt.plot(np.arange(2, x), y, 'o-')
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.xticks(np.arange(2, x))
    plt.show()

indices = []
scores = []
max_range = 20

for kclusters in range(2, max_range) :

    # Run k-means clustering
    lct = toronto_restaurant_clustering
    kmeans = KMeans(n_clusters = kclusters, init = 'k-means++', random_state = 0).fit_predict(lct)

    # Gets the score for the clustering operation performed
    score = silhouette_score(lct, kmeans)

    # Appending the index and score to the respective lists
    indices.append(kclusters)
    scores.append(score)


plot(max_range, scores, "No. of clusters", "Silhouette Score")
```

We then obtained the optimal k value from the silhoutte score graph :

```
[101]: opt = np.argmax(scores) + 2 # Finds the optimal value
       opt

[101]: 3
```



We then found the top 10 restaurant for each neighbourhood:

```
[89]: def return_most_common_venues(row, num_top_venues):
          row_categories = row.iloc[1:]
          row_categories_sorted = row_categories.sort_values(ascending=False)

          return row_categories_sorted.index.values[0:num_top_venues]
```

```
[90]: num_top_venues = 10

      indicators = ['st', 'nd', 'rd']

      # create columns according to number of top venues
      columns = ['Neighbourhood']
      for ind in np.arange(num_top_venues):
          try:
              columns.append('{}{} Most Common Restaurant'.format(ind+1, indicators[ind]))
          except:
              columns.append('{}th Most Common Restaurant'.format(ind+1))

      # create a new dataframe
      neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
      neighborhoods_venues_sorted['Neighbourhood'] = toronto_restaurant_grouped['Neighbourhood']

      for ind in np.arange(toronto_restaurant_grouped.shape[0]):
          neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(toronto_restaurant_grouped.iloc[ind, :], num_top_venues)

      neighborhoods_venues_sorted.head()
```

[90]:

	Neighbourhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
0	Agincourt	Latin American Restaurant	Vietnamese Restaurant	Doner Restaurant	Gluten-free Restaurant	German Restaurant	French Restaurant	Filipino Restaurant	Fast Food Restaurant	Falafel Restaurant	Ethiopian Restaurant
1	Bathurst Manor, Wilson Heights, Downsview North	Sushi Restaurant	Chinese Restaurant	Middle Eastern Restaurant	Vietnamese Restaurant	Doner Restaurant	German Restaurant	French Restaurant	Filipino Restaurant	Fast Food Restaurant	Falafel Restaurant
2	Bayview Village	Japanese Restaurant	Chinese Restaurant	Vietnamese Restaurant	Doner Restaurant	Gluten-free Restaurant	German Restaurant	French Restaurant	Filipino Restaurant	Fast Food Restaurant	Falafel Restaurant
3	Bedford Park, Lawrence Manor East	Italian Restaurant	Indian Restaurant	American Restaurant	Thai Restaurant	Sushi Restaurant	Comfort Food Restaurant	Greek Restaurant	Dumpling Restaurant	German Restaurant	French Restaurant
4	Berczy Park	Seafood Restaurant	French Restaurant	Greek Restaurant	Thai Restaurant	Vegetarian / Vegan Restaurant	Sushi Restaurant	Italian Restaurant	Japanese Restaurant	Comfort Food Restaurant	Eastern European Restaurant

And, finally we appended to the top restaurants dataframe the cluster column as well as the neighbourhood and the borough column:

```
[106]: # add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

toronto_merged = toronto_df

# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
toronto_merged = toronto_merged.join(neighborhoods_venues_sorted.set_index('Neighbourhood'), on='Neighbourhood', how='right')

toronto_merged.head() # check the last columns!
```

[106]:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
78	M1S	Scarborough	Agincourt	43.794200	-79.262029	0	Latin American Restaurant	Vietnamese Restaurant	Doner Restaurant	Gluten-free Restaurant	German Restaurant	French Restaurant	Filipino Restaurant	Fast Food Restaurant	Falafel Restaurant	Ethiopian Restaurant
28	M3H	North York	Bathurst Manor, Wilson Heights, Downsview North	43.754328	-79.442259	0	Sushi Restaurant	Middle Eastern Restaurant	Vietnamese Restaurant	Dim Sum Restaurant	German Restaurant	French Restaurant	Filipino Restaurant	Fast Food Restaurant	Falafel Restaurant	Ethiopian Restaurant
39	M2K	North York	Bayview Village	43.786947	-79.385975	0	Japanese Restaurant	Chinese Restaurant	Vietnamese Restaurant	Doner Restaurant	Gluten-free Restaurant	German Restaurant	French Restaurant	Filipino Restaurant	Fast Food Restaurant	Falafel Restaurant
55	M5M	North York	Bedford Park, Lawrence Manor East	43.733283	-79.419750	0	Sushi Restaurant	Italian Restaurant	Indian Restaurant	Comfort Food Restaurant	Greek Restaurant	Japanese Restaurant	American Restaurant	Thai Restaurant	Seafood Restaurant	Cuban Restaurant
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	0	Seafood Restaurant	Indian Restaurant	Comfort Food Restaurant	French Restaurant	Vegetarian / Vegan Restaurant	Italian Restaurant	Japanese Restaurant	Eastern European Restaurant	Greek Restaurant	Thai Restaurant

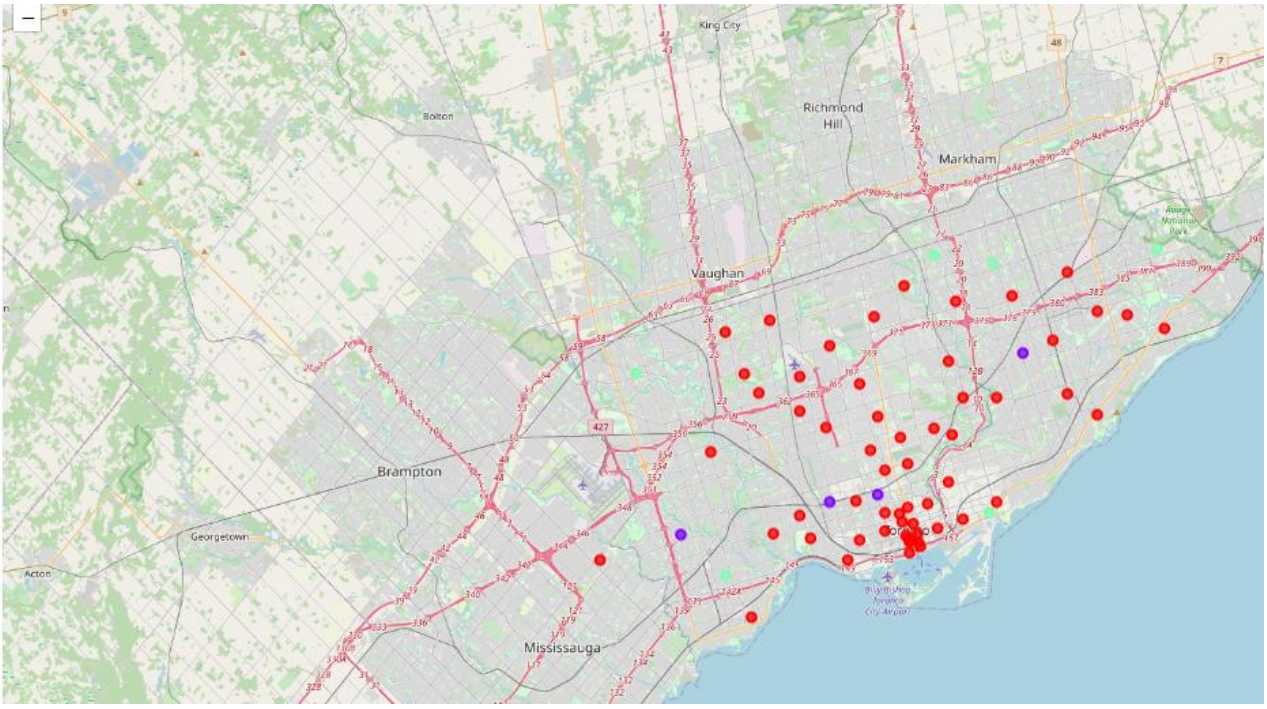
Results

We are now ready to see the results of the clustering of Toronto neighbourhoods. We found three clusters:

- Cluster 0 consists of 58 neighbourhoods,
- Cluster 1 consists of 6 neighbourhoods,
- Cluster 2 consists of 4 neighbourhoods.

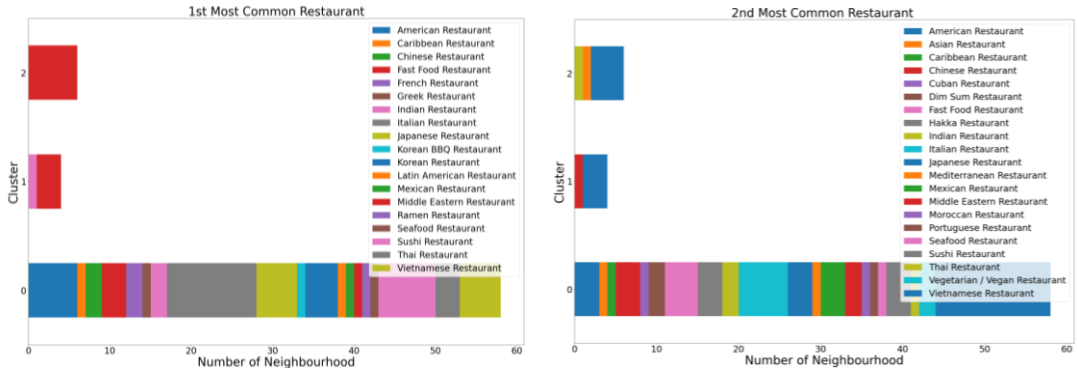
```
[107]: toronto_merged['Cluster Labels'].value_counts()

[107]: 0    58
       2     6
       1     4
       Name: Cluster Labels, dtype: int64
```

The result is then shown in the following graph: neighbourhoods have been coloured depending on the clusters based on restaurants



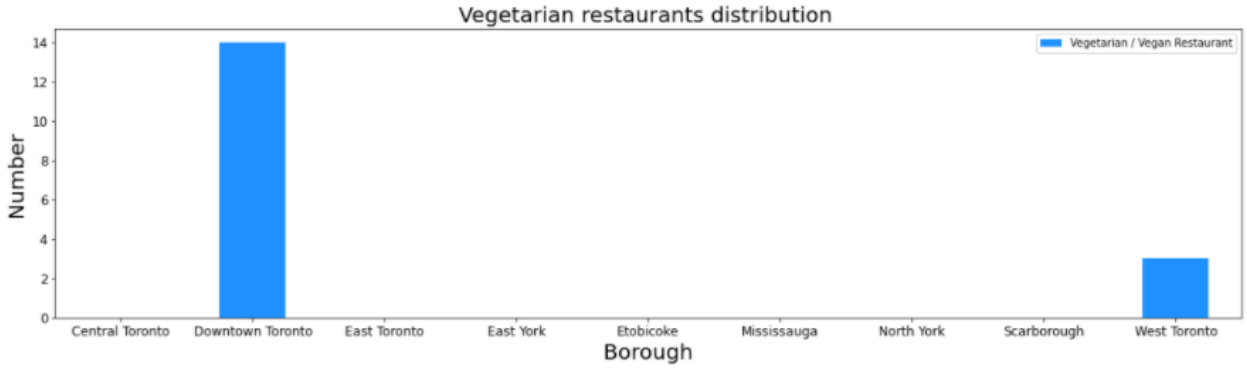
Discussion

We observe that, while for cluster 1 and 2 the 1st most common restaurant is the fast food, in clustr 0 it can range between many options. The 2nd most common restaurant is Vietnamise for all three clusters.

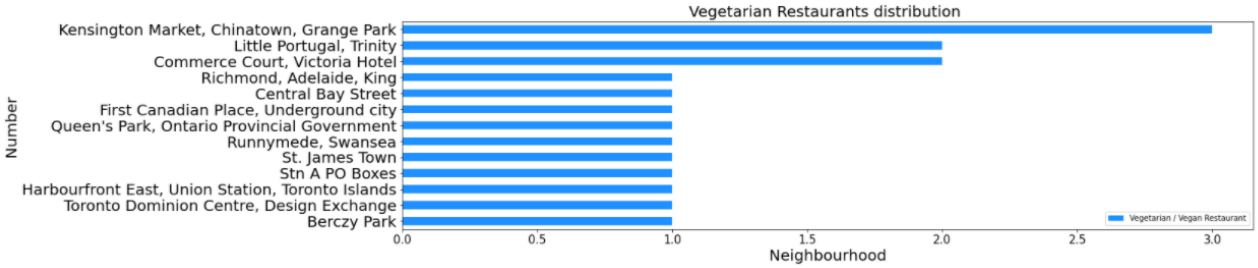


We finally focus our attention on vegetarian restaurants, since the information can be of interest of vegetarian tourists, and they are not widespread in the territory. In particular, vegetarian restaurants are concentrated in the Borough of Downtown Toronto.

```
import matplotlib.pyplot as plt
clr = "dodgerblue"
borough_vegetarian.plot.bar(x='Borough',y='Vegetarian / Vegan Restaurant',figsize=(20,5),color=clr)
plt.title('Vegetarian restaurants distribution ', fontsize=20)
#On x-axis
plt.xlabel('Borough', fontsize = 20)
#On y-axis
plt.ylabel('Number', fontsize=20)
plt.xticks(rotation = 'horizontal', fontsize=12)
plt.yticks(fontsize=12)
plt.show()
```



In the following graph, we reported the neighbourhood distribution of vegetarian restaurants:



Conclusions

- To conclude, in the project we have investigated the restaurant distribution in the neighbourhoods of Toronto. We have classified the neighbourhoods according to the most common restaurants you can find there.
- We used k-means for clustering and we found three different clusters
- Finally, we focused on the possibility of finding vegetarian restaurants in the city of Toronto, and we found they are very concentrated in downtown Toronto.