Best spots to stay on a summer trip to Toronto

PREPARED BY:

WEBSTER GOVA

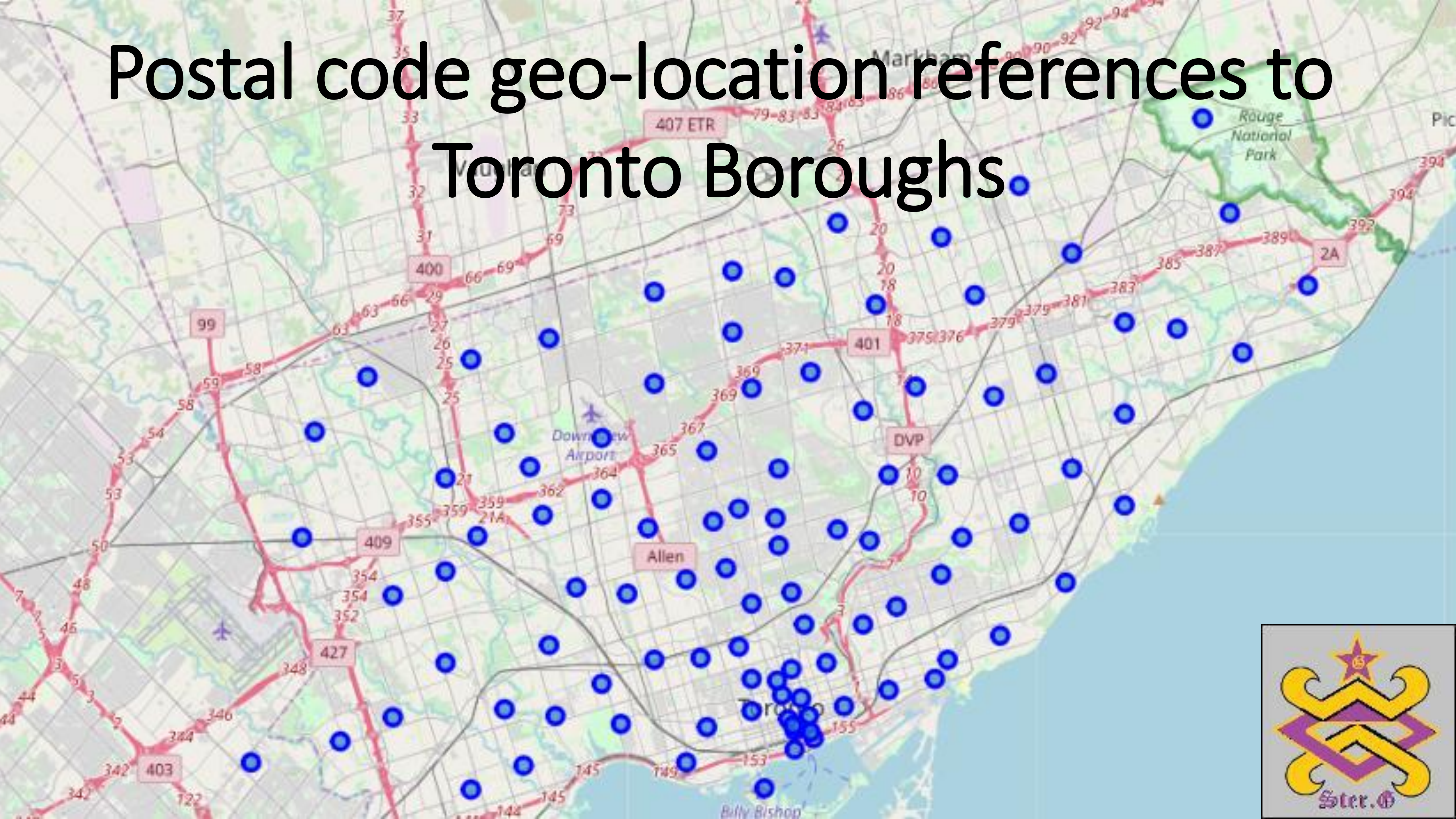
Sterv-AI Consulting Group

Business problem

A travel agent conducted a survey to help them plan a trip for a group of students from Stellenbosch University to visit Toronto. In the survey, the agent asked the students what facilities and activities they'd like to do while in Canada. The results in the survey showed that the students prefer to stay in Central Toronto.

The agent then sent results of the survey to our consulting company (**StervAI**) to analyse the results and recommend boroughs to get the students accommodation in Central Toronto

Postal code geo-location references to Toronto Boroughs



Data requirements

The data required for this project include:

1. Geo-coordinates for the recreational activities, amenities and facilities gathered by the travel agent in the survey (location data for these venues will be obtained from Foursquare *venues* API)
2. Geo-coordinates for borough and neighborhoods in Central Toronto (obtained from Wikipedia Toronto *postal codes*)

Data requirements

Activities, amenities and facilities the students prefer to be in close proximity to include the following venues geo-location matches in FOURSQUARE:

```
activities = ['Neighbourhood',
              'Park', 'Café', 'French Restaurant', 'Other Great Outdoors','Fish & Chips Shop',
              'Gastropub', 'BBQ Joint','Japanese Restaurant','History Museum', 'Grocery Store',
              'Field', 'Liquor Store', 'Coffee Shop', 'Spa', 'Historic Site','Farmers Market',
              'Shopping Mall', 'Juice Bar', 'Castle','Italian Restaurant', 'Indian Restaurant',
              'Greek Restaurant', 'Ice Cream Shop', 'Museum', 'Thai Restaurant','Pizza Place',
              'Gift Shop','Concert Hall', 'Steakhouse', 'Vegetarian / Vegan Restaurant',
              'Breakfast Spot', 'American Restaurant', 'Restaurant','Indie Movie Theater',
              'Dessert Shop','Bar','Sandwich Place', 'Hostel', 'Cocktail Bar','Cupcake Shop',
              'Tapas Restaurant', 'South American Restaurant','Vietnamese Restaurant',
              'Mexican Restaurant','Ramen Restaurant','Egyptian Restaurant','Theater','Movie Theater',
              'Burrito Place','Arts & Crafts Store', 'Taco Place', 'Organic Grocery','Plaza',
              'Beer Bar', 'Whisky Bar','Middle Eastern Restaurant', 'Mediterranean Restaurant',
              'Cosmetics Shop', 'Diner','Seafood Restaurant', 'Asian Restaurant', 'Korean Restaurant'
             ]

venues = venues_toronto_hc[activities].groupby('Neighbourhood').sum()
```

The venues location data will be matched to the Boroughs location data to calculate Neighbourhoods location clusters with the highest proximity to most venues

The best Boroughs location cluster will be identified using a *k-means clustering* model

Methodology

Step 1:

Identify geo-location of Postal Codes and matching them to Boroughs in Central Toronto neighbourhoods

- Geo-location of postal codes in Toronto were scrapped from Wikipedia Postal codes were then matched to Neighbourhoods

- Geospatial data for Toronto was then used to load data on Toronto neighbourhoods

- The postal code and geospatial data were matched and used to create a dataframe with postal code and geo-locations for neighborhoods and boroughs in Toronto
- Create a dataset for Central Toronto Boroughs from the Toronto data

Step 2:

Fetching geo-location data for venues that match venues in Central Toronto ‘Neighborhoods’

•A function to extract geolocation data on venues in Central Toronto neighbourhoods was described as shown below

```
def api_call_4sqr (postal_code_list, neighbourhood_list, lat_list, lng_list, LIMIT = 50000, radius = 10000):
    api = []
    counter = 0
    for postal_code, neighbourhood, lat, lng in zip(postal_code_list, neighbourhood_list, lat_list, lng_list):

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.
            Client_ID, Client_Secret, Version,
            lat, lng, radius, LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']
        api_dict = {}
        api_dict['Postalcode'] = postal_code; api_dict['Neighbourhood(s)'] = neighbourhood;
        api_dict['Latitude'] = lat; api_dict['Longitude'] = lng;
        api_dict['API_calls'] = results;
        api.append(api_dict)
        counter += 1
    return api;
```

```
Toronto_4sqr = api_call_4sqr(list(toronto['Postalcode']),list(toronto['Neighbourhood']),
                             list(toronto['Latitude']),list(toronto['Longitude']),)
```

• Foursquare was used to take advantage of its places database and API of location data with more than 105 million venues worldwide
• This choice was made for ease of scaling up in case the client or similar requests are made by other clients with similar requests for other locations

Methodology

Methodology

Step 3:

Convert venues data fetched from Foursquare into a dataframe to be merged to match the neighborhoods dataframe for Central Toronto

- The geolocation data on venues of interest for the students extracted from Foursquare were merged with the dataframe for geolocation of neighborhoods in Central Toronto
- The category of the venues was the main feature used for matching
- A summary of venue category was conducted to evaluate if there are any neighbourhoods with low numbers of venue categories for exclusion in further
- The next task was to evaluate how close at least 10 venues are to each of the categories to recommend the neighbourhood with more venue categories close to it
- One-hot coding for "Venue Category" was used to identify unique venue categories in the different neighbourhoods
- Numerical assignments of (0) and (1) were made for absent/present respectively
- Total venues per category were then calculated using the numerical assignments from “One-hot coding”

Methodology

Step 4:

Cluster analysis to develop a Machine Learning model to identify neighbourhood closest to more venue categories

- K-Means Clustering was used to develop a machine learning model to identify neighbourhoods (centroids) with the shortest distances to more venue categories around them

```
from sklearn.cluster import KMeans

# Fit the data to the k-means clustering model
kmeans = KMeans(n_clusters = 5, random_state = 0).fit(venues)
```

Methodology

Step 5:

Assign k-means scores to clusters and match the neighbourhoods to their respective clusters. Recommend neighbourhoods based on k-means scores

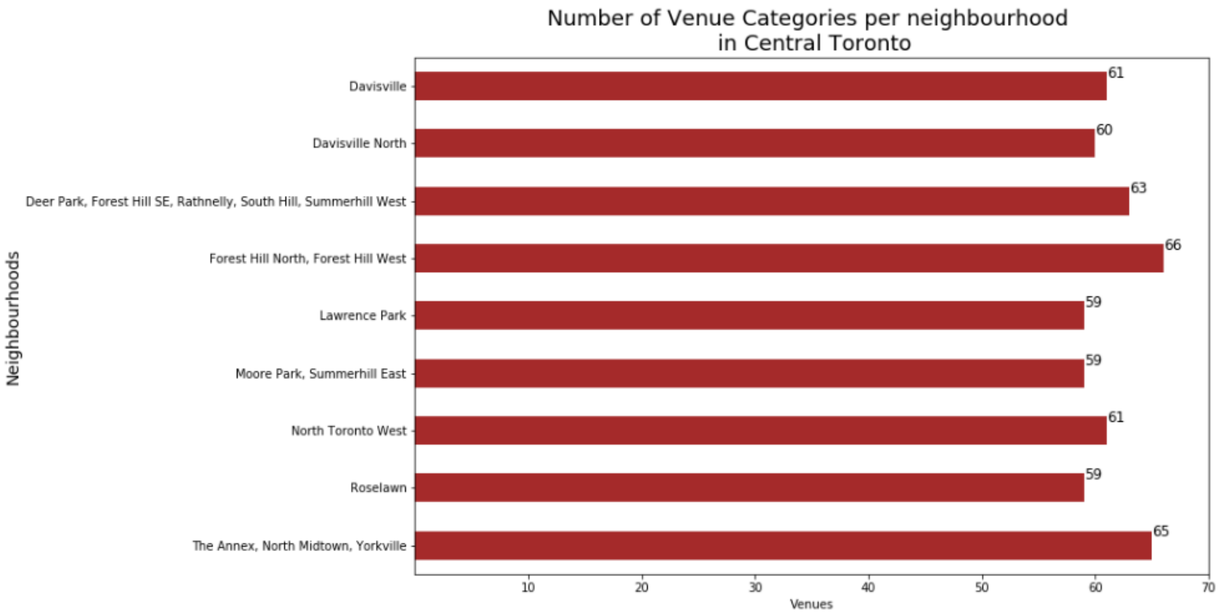
- K-means cluster centers were used to assign total scores to cluster
- Neighbourhoods were then assigned to each cluster using the ‘venues.index’
- Neighbourhoods were then recommended based on total scores for clusters

```
means_df = pd.DataFrame(kmeans.cluster_centers_)
means_df.columns = venues.columns
means_df.index = ['Cluster_1','Cluster_2','Cluster_3','Cluster_4','Cluster_5']
means_df['Sum'] = means_df.sum(axis = 1)
mean = means_df['Sum']
means_df.sort_values(axis = 0, by = ['Sum'], ascending=False).reset_index()
means_df.index.name = 'Cluster'
means_df.head()
```

Step 1:

Identify geo-location of Postal Codes and matching them to Boroughs in Central Toronto neighbourhoods

- Number of neighbourhoods in Central Toronto: **9**
- Highest number of venue categories: **Forest Hill North & West**
- Second highest number of venue categories: **The Annex, North Midtown & Yorkville**
- Third highest number of venue categories: **Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West**

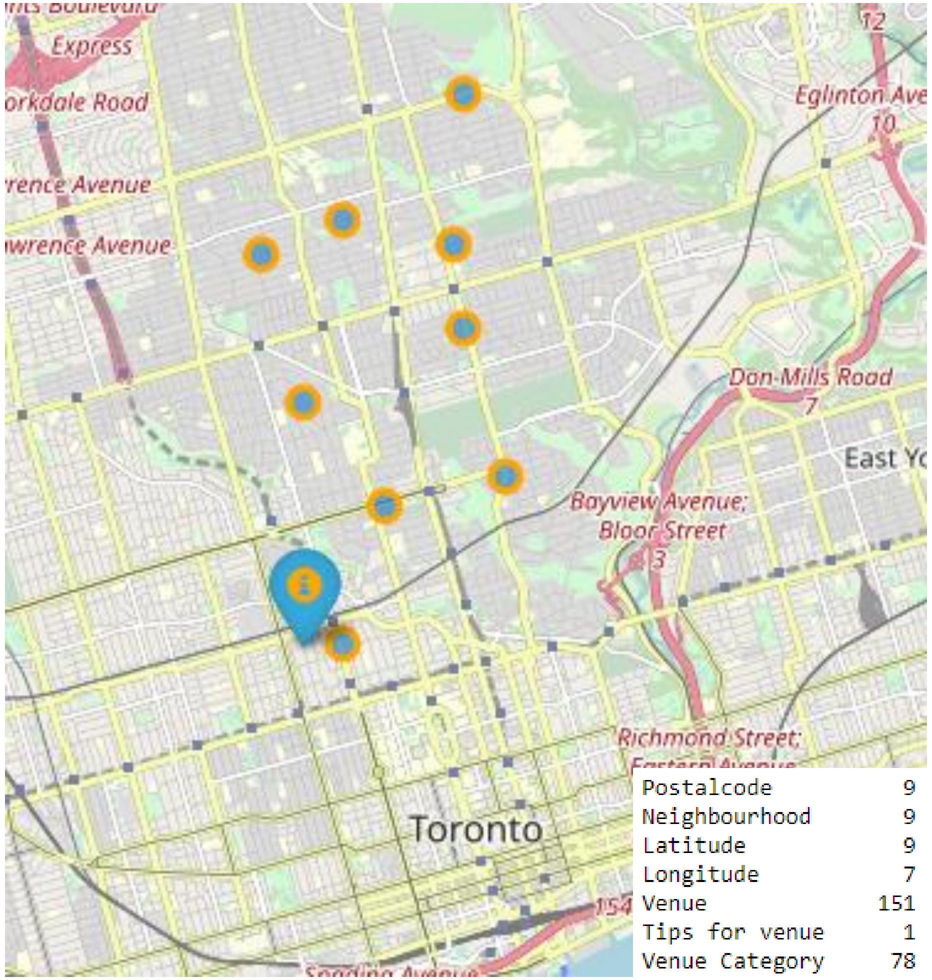


Results

Results

Step 2:

Fetching geo-location data for venues that match venues in Central Toronto ‘Neighborhoods’



Postalcode	9
Neighbourhood	9
Latitude	9
Longitude	7
Venue	151
Tips for venue	1
Venue Category	78

Results

Step 3:

Convert venues data fetched from Foursquare into a dataframe to be merged to match the neighborhoods dataframe for Central Toronto

Neighbourhood	Park	Café	French Restaurant	Other Great Outdoors	Fish & Chips Shop	Gastropub	BBQ Joint	Japanese Restaurant	History Museum	Grocery Store	...	Plaza	Beer Bar	Whisky Bar	Middle Eastern Restaurant
Davisville	6	11	2	1	1	3	2	2	1	3	...	1	0	0	0
Davisville North	7	11	2	1	1	3	2	3	1	3	...	1	0	0	0
Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West	5	10	1	0	1	3	1	2	0	3	...	1	1	1	0

3 rows × 62 columns

Step 4:

Cluster analysis to develop a Machine Learning model to identify and recommend neighbourhoods closest to more venue categories

Results

Cluster	Park	Café	French Restaurant	Other Great Outdoors	Fish & Chips Shop	Gastropub	BBQ Joint	Japanese Restaurant	History Museum	Grocery Store	...	Beer Bar	Whisky Bar	Middle Eastern Restaurant
Cluster_1	5.000000	11.0	2.0	1.0	1.0	3.0	2.000000	2.0	0.0	3.0	...	0.0	0.0	1.0
Cluster_2	2.000000	8.0	2.0	0.0	0.0	1.0	1.000000	1.0	0.0	3.0	...	1.0	1.0	0.0
Cluster_3	8.500000	9.5	1.5	1.0	1.0	3.5	2.000000	3.0	0.5	3.0	...	1.0	1.0	0.5
Cluster_4	7.333333	11.0	2.0	1.0	1.0	3.0	2.333333	3.0	1.0	3.0	...	0.0	0.0	0.0
Cluster_5	5.000000	9.5	1.0	0.5	1.0	3.0	1.500000	2.0	0.0	3.0	...	1.0	1.0	0.5

5 rows × 63 columns

Results

Step 5:

Assign k-means scores to clusters and match the neighbourhoods to their respective clusters. Recommend neighbourhoods based on k-means scores

Cluster	Park	Café	French Restaurant	Other Great Outdoors	Fish & Chips Shop	Mediterranean Restaurant	Cosmetics Shop	Diner	Seafood Restaurant	Asian Restaurant	Korean Restaurant	Sum
Cluster_1	5.000000	11.0	2.0	1.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	87.0
Cluster_2	2.000000	8.0	2.0	0.0	0.0	1.0	1.0	1.0	0.0	2.0	1.0	85.0
Cluster_3	8.500000	9.5	1.5	1.0	1.0	0.0	0.0	0.0	0.5	1.0	0.0	90.0
Cluster_4	7.333333	11.0	2.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	88.0
Cluster_5	5.000000	9.5	1.0	0.5	1.0	0.5	1.0	0.5	0.0	1.0	0.5	89.0

5 rows × 63 columns

- **Best neighbourhood (2)**:
  North Toronto West & Roselawn
- **Second Best neighbourhood (2)**:
  Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West and Forest Hill North, Forest Hill West
- **Third Best neighbourhood (3)**:
  Davisville, Davisvill North & eer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West

The best neighbourhoods are:

```
nb_names1 = nb_summary[nb_summary['Cluster'] == 3]
nb_names1.head()
```

	Neighbourhood	Cluster
6	North Toronto West	3
7	Roselawn	3

The second best neighbourhoods are:

```
nb_names2 = nb_summary[nb_summary['Cluster'] == 5]
nb_names2.head()
```

	Neighbourhood	Cluster
2	Deer Park, Forest Hill SE, Rathnelly, South Hi...	5
3	Forest Hill North, Forest Hill West	5

The third best neighbourhoods are:

```
nb_names3 = nb_summary[nb_summary['Cluster'] == 4]
nb_names3.head()
```

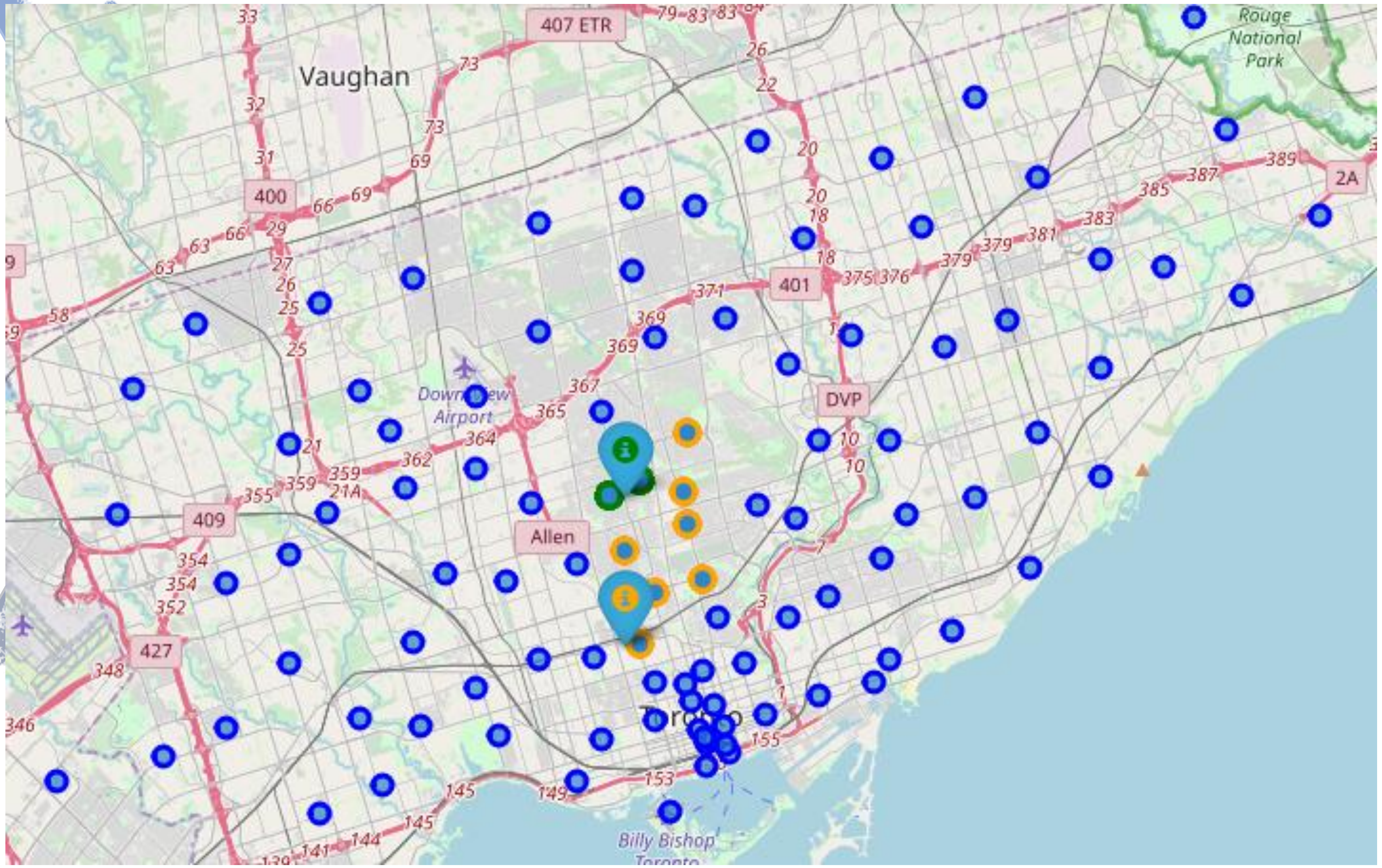
	Neighbourhood	Cluster
0	Davisville	4
1	Davisville North	4
4	Lawrence Park	4

Results

Step 5:

Assign k-means scores to clusters and match the neighbourhoods to their respective clusters. Recommend neighbourhoods based on k-means scores

Green: Best neighbourhood (2): North Toronto West & Roselawn
Orange: Neighbourhoods in Central Toronto
Blue: Neighbourhoods in Toronto



Step 5:

Assign k-means scores to clusters and match the neighbourhoods to their respective clusters. Recommend neighbourhoods based on k-means scores

Results

Discussion

From assignment of neighbourhoods with highest number of venue categories:

- Highest number of venue categories: **Forest Hill North & West**
- Second highest number of venue categories: **The Annex, North Midtown & Yorkville**
- Third highest number of venue categories: **Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West**

From assignment of venue categories to neighbourhoods with shortest distance to k-means centers:

- **Best neighbourhood (2)**:

North Toronto West & Roselawn

- **Second Best neighbourhood (2)**:

Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West and Forest Hill North, Forest Hill West

- **Third Best neighbourhood (3)**:

Davisville, Davisvill North & eer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West

Observations:

- The k-means results are significantly different and more objective
- The 1st & 3rd highest in number of categories were both ranked 2nd best using k-means
- Neighbourhoods not among highest number of venue categories had better k-means scores

Conclusion