Capstone Project- The Battle of Neighborhoods

Introduction/Business Problem

Toronto is the largest city if Canada and home of many touristic attractions. The city is full of museums, theatres, art galleries and festival events.

The Toronto City Culture Association heard about the capabilities of Data Science and were interested in a working with us to create a recommendation system for city visitors, based on where they planned to stay.

As such, they contacted us, proposing the project with a list of requirements.

Requirements

- The system should be capable of get the list of all cultural places in Toronto, worth visiting (e.g. Museums);

- Be able to rate the places according to typical user experience

- Get the top 3 cultural places

- Cluster the selected places by Neighbourhood

- Display the results Graphically on a map

- Create bar graphs analysing the results

Data

In order to be able to carry out this project we need data from a few sources:

1) Wikipedia website - we need the information regarding the Boroughs and Neighbourhoods of Toronto - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Toronto - FSAs [edit]

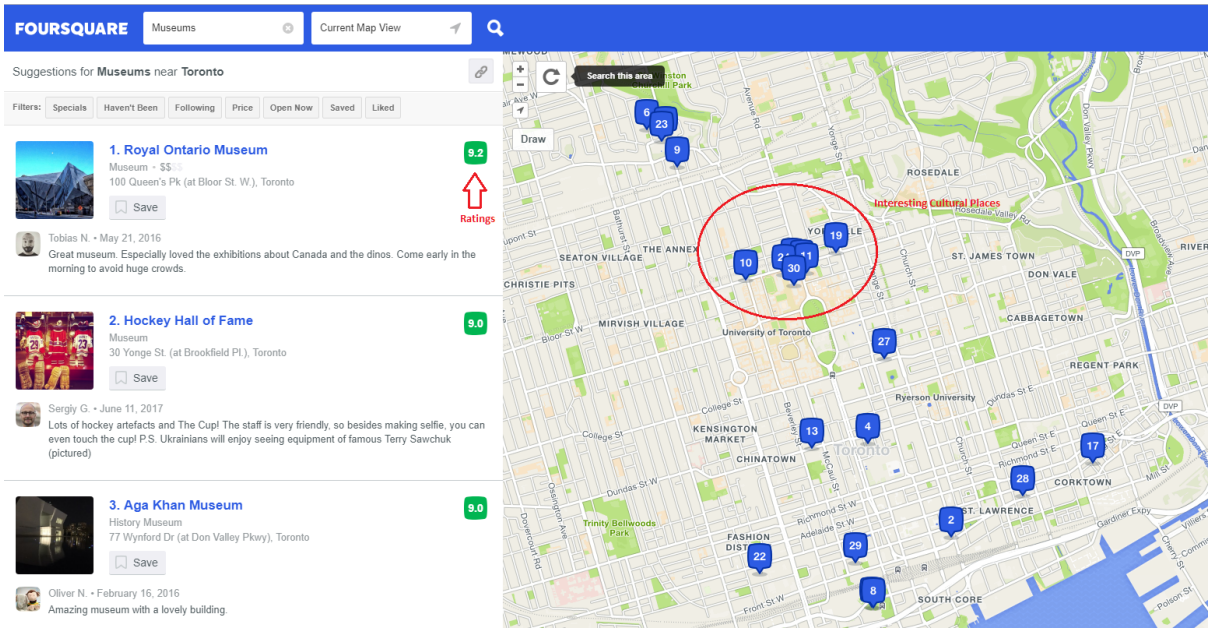
Note: There are no rural FSAs in Toronto, hence no postal codes start with M0.

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M5A	Downtown Toronto	Regent Park
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor
M7A	Queen's Park	Not assigned
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue
M1B	Scarborough	Rouge
M1B	Scarborough	Malvern
M2B	Not assigned	Not assigned
M3B	North York	Don Mills North
M4B	East York	Woodbine Gardens

2) Foursquare website:

- from here we can extract the information about the list of all Cultural venues of Toronto

- we can also see the popularity of each venue to create our top list



3) Geospatial_Coordinates.csv file which contains the coordinates for each Toronto PostCode

	A	B	C
1	Postal Code	Latitude	Longitude
2	M1B	43.8066863	-79.1943534
3	M1C	43.7845351	-79.1604971
4	M1E	43.7635726	-79.1887115
5	M1G	43.7709921	-79.2169174
6	M1H	43.773136	-79.2394761
7	M1J	43.7447342	-79.2394761
8	M1K	43.7279292	-79.2620294
9	M1L	43.7111117	-79.2845772
10	M1M	43.716316	-79.2394761
11	M1N	43.692657	-79.2648481
12	M1P	43.7574096	-79.273304
13	M1R	43.7500715	-79.2958491
14	M1S	43.7942003	-79.2620294
15	M1T	43.7816375	-79.3043021
16	M1V	43.8152522	-79.2845772
17	M1W	43.7995252	-79.3183887

Methodology

The approach to achieve this work was fundamentally the following:

1 – Getting the data

2 – Cleaning and preparing the data sets

3 – Create Machine learning Clusters

4 – Plot graphs for graphical analysis

5 – Plot bar charts to make conclusions

1 – Getting the data

Getting the Postcodes from Wikipedia:

Open file and transform it on a soup object (bs4.BeautifulSoup)

[7]:
```
response = requests.get('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M')
soup = BeautifulSoup(response.text,'lxml')
```

[8]:
```
myTable = soup.find('table')
```

Now that we have the table, let's loop through the rows and create a list for each cell

[9]:
```
table_list = []

for row in myTable.tbody.find_all('tr'):
    row_list = []
    for cell in row.find_all('td'):
        #print(cell.text.strip())
        row_list.append(cell.text.strip())

    #Make sure to ignore entries that haven't got a borough
    if len(row_list) > 0 and row_list[1] != 'Not assigned':
        #Define "Not assigned" neighbourhoods as their borough name
        if row_list[2] == 'Not assigned':
            row_list[2] = row_list[1]

        table_list.append(row_list)
```

Loading the coordinates from csv:

Next, read from csv and get latitude/longitude

[13]:
```
lat_log_df = pd.read_csv('Geospatial_Coordinates.csv')
```

[14]:
```
lat_log_df.head()
```

[14]:

	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

Finally, get the data from the Avenues Foursquare API:

Let's create a function to get all venues from all the neighborhoods in Toronto in a 500 miles radius

[25]:
```
def getNearbyVenues(names, latitudes, longitudes, radius=500):
    LIMIT = 100
    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

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

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

2 – Cleaning and preparing the data sets

Creating the first dataframe from Wikipedia:

Now that we have the data, just create a dataframe and populate it

[11]:
```
# define the dataframe columns
column_names = ['PostalCode', 'Borough', 'Neighborhood']

# instantiate the dataframe
ca_postcodes_df = pd.DataFrame(columns=column_names)


for item in unique_postcode_table_list:
    PostalCode = item[0]
    Borough = item[1]
    Neighborhood = item[2]

    ca_postcodes_df = ca_postcodes_df.append({'PostalCode': PostalCode,
                                              'Borough': Borough,
                                              'Neighborhood': Neighborhood}, ignore_index=True)

ca_postcodes_df.head()
```

[11]:

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront, Regent Park
3	M6A	North York	Lawrence Heights, Lawrence Manor
4	M7A	Queen's Park	Queen's Park

The dataframe with geo coordinates:

Next, read from csv and get latitude/longitude

[13]:
```
lat_log_df = pd.read_csv('Geospatial_Coordinates.csv')
```

[14]:
```
lat_log_df.head()
```

[14]:

	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

… and merging them both:

Merge both dataframes

```
[15]: lat_log_df_new = lat_log_df.copy()
#lat_log_df_new['Postal Code'] = lat_log_df_new['PostalCode']
lat_log_df_new.rename(columns={'Postal Code':'PostalCode'}, inplace=True)
#lat_log_df_new
df_pc_lat_long = pd.merge(ca_postcodes_df, lat_log_df_new, how='inner', on = 'PostalCode')
df_pc_lat_long.head()
```

[15]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494

Finally get the dataframe with Cultural Avenues to work with:

Slice by Arts and Museums categories only

```
[82]: #toronto_cultural_venues = toronto_venues[toronto_venues['Venue Category'].str.contains('art',  case=False)] & toronto_venues[toronto_venues['Venue Category'].str.contains('museum',  case=False)]

#toronto_cultural_venues = toronto_venues.loc[toronto_venues['Venue Category'] == 'Bakery'] | toronto_venues.loc[toronto_venues['Venue Category'] == 'Spa']
toronto_cultural_venues = toronto_venues.loc[(toronto_venues['Venue Category'].str.contains('Museum',  case=False)) |
                                             (toronto_venues['Venue Category'].str.contains('Art',  case=False)) &
                                             (~toronto_venues['Venue Category'].str.contains('Store',  case=False)) &
                                             (~toronto_venues['Venue Category'].str.contains('Dojo',  case=False)) |
                                             (toronto_venues['Venue Category'].str.contains('music',  case=False)) |
                                             (toronto_venues['Venue Category'].str.contains('event',  case=False)) |
                                             (toronto_venues['Venue Category'].str.contains('theatre',  case=False))




                                            ]
toronto_cultural_venues
#df.loc[(df['column_name'] == some_value) & df['other_column'].isin(some_values)]
```

[82]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
19	Harbourfront, Regent Park	43.654260	-79.360636	Young Centre for the Performing Arts	43.650825	-79.357593	Performing Arts Venue
30	Harbourfront, Regent Park	43.654260	-79.360636	Berkeley Church	43.655123	-79.365873	Event Space
38	Harbourfront, Regent Park	43.654260	-79.360636	Arta Gallery	43.650022	-79.361222	Art Gallery
70	Ryerson, Garden District	43.657162	-79.378937	Jazz Bistro	43.655678	-79.379276	Music Venue
79	Ryerson, Garden District	43.657162	-79.378937	Ryerson Image Centre	43.657523	-79.379460	Art Gallery
163	St. James Town	43.651494	-79.375418	Club 120	43.652100	-79.375522	Performing Arts Venue
191	St. James Town	43.651494	-79.375418	St. Lawrence Market Plaza	43.649169	-79.372330	Art Gallery
255	Berczy Park	43.644771	-79.373306	Hockey Hall Of Fame (Hockey Hall of Fame)	43.646974	-79.377323	Museum
283	Berczy Park	43.644771	-79.373306	St. Lawrence Market Plaza	43.649169	-79.372330	Art Gallery
320	Central Bay Street	43.657952	-79.387383	Textile Museum of Canada	43.654396	-79.386500	Art Museum
482	Adelaide, King, Richmond	43.650571	-79.384568	Textile Museum of Canada	43.654396	-79.386500	Art Museum
488	Adelaide, King, Richmond	43.650571	-79.384568	Design Exchange	43.647972	-79.380104	Art Gallery

3 – Create Machine learning Clusters

Creating clusters with Avenues and the Neighbourhoods:

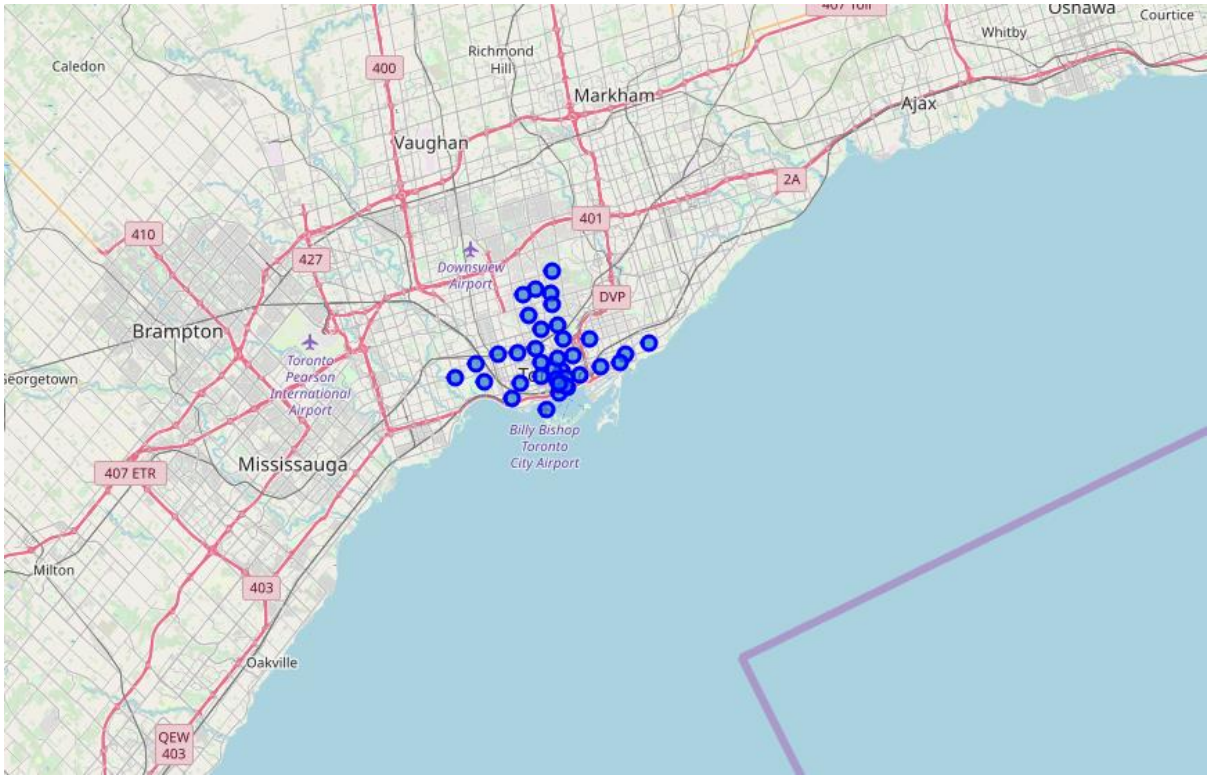
Cluster Neighborhoods

[149]:
```
# set number of clusters
kclusters = 5

toronto_grouped_clustering = toronto_grouped.drop('Neighborhood', 1)
#print(toronto_grouped_clustering)
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_#[0:10]
```

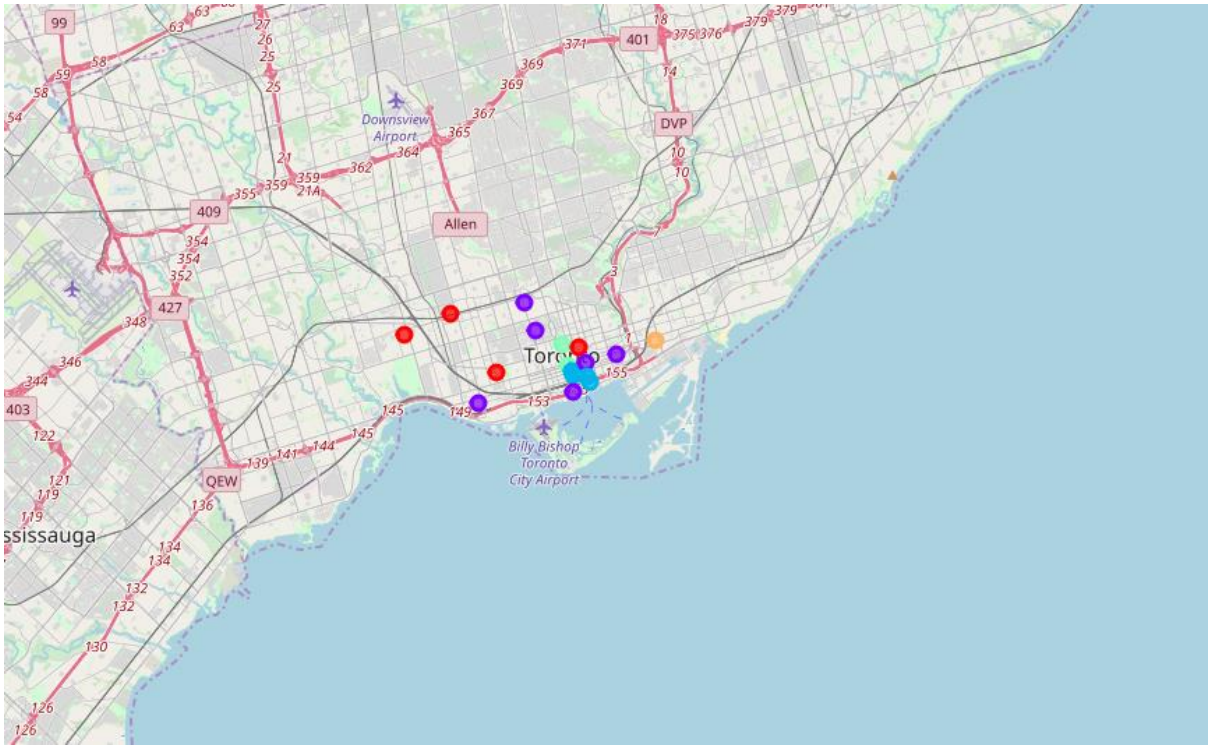
[149]:
```
array([3, 2, 1, 3, 2, 2, 0, 2, 1, 1, 1, 0, 0, 0, 1, 2, 4, 1], dtype=int32)
```

4 – Plot graphs for graphical analysis

Plot and see the neighbourhoods we loaded:

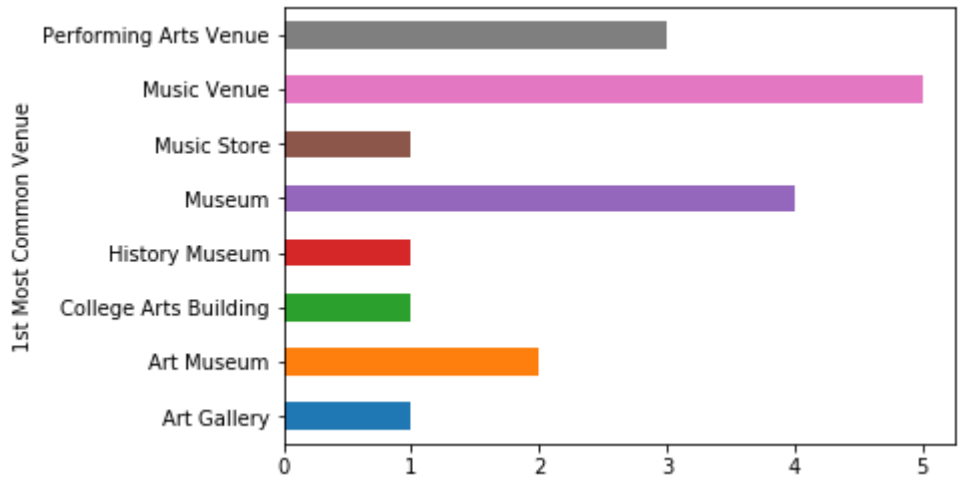


Plot and see the clusters of avenues (5 per cluster):

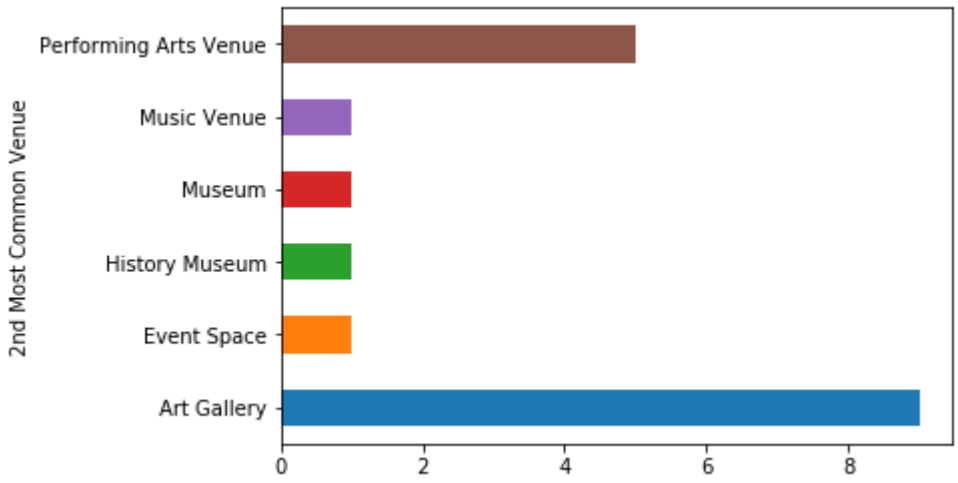


5 – Plot bar charts to make conclusions

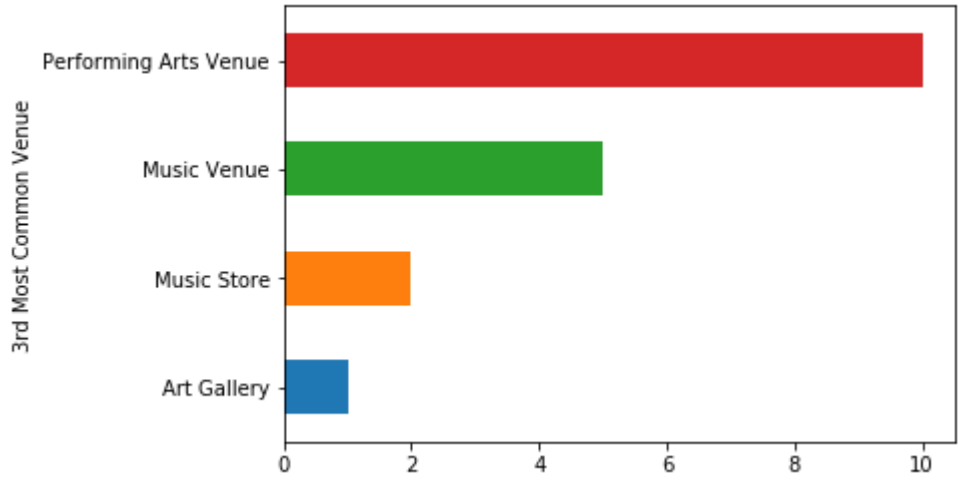
The most sought-after Avenue across all neighbourhoods:



The second sought-after Avenue across all neighbourhoods:



The third sought-after Avenue across all neighbourhoods:



Results

From the combination of this data set we can conclude:

- The density of Avenue clusters indicated by the graph chart;
- The most sought-after Avenue across all neighbourhoods is Music Avenue
- The second sought-after Avenue across all neighbourhoods is Art Gallery
- The third sought-after Avenue across all neighbourhoods is Performing Arts Venue
- The most common avenue by Neighbourhood given by the following table:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	M5H	Downtown Toronto	Adelaide, King, Richmond	43.650571	-79.384568	3	Art Museum	Art Gallery	Performing Arts Venue
1	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	2	Museum	Art Gallery	Performing Arts Venue
2	M6K	West Toronto	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191	1	Performing Arts Venue	Music Venue	Music Store
3	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	3	Art Museum	Performing Arts Venue	Music Venue
4	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817	2	Museum	Art Gallery	Performing Arts Venue
5	M5K	Downtown Toronto	Design Exchange, Toronto Dominion Centre	43.647177	-79.381576	2	Museum	Art Gallery	Performing Arts Venue
6	M6H	West Toronto	Dovercourt Village, Dufferin	43.669005	-79.442259	0	Music Venue	Art Gallery	Performing Arts Venue
7	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.382280	2	Museum	Art Gallery	Performing Arts Venue
8	M5S	Downtown Toronto	Harbord, University of Toronto	43.662696	-79.400049	1	College Arts Building	Performing Arts Venue	Music Venue
9	M5J	Downtown Toronto	Harbourfront East, Toronto Islands, Union Station	43.640816	-79.381752	1	Music Venue	History Museum	Performing Arts Venue
10	M5A	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636	1	Performing Arts Venue	Event Space	Art Gallery
11	M6P	West Toronto	High Park, The Junction South	43.661608	-79.464763	0	Music Venue	Performing Arts Venue	Music Store
12	M6J	West Toronto	Little Portugal, Trinity	43.647927	-79.419750	0	Music Venue	Art Gallery	Performing Arts Venue
13	M5B	Downtown Toronto	Ryerson, Garden District	43.657162	-79.378937	0	Music Venue	Art Gallery	Performing Arts Venue
14	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	1	Performing Arts Venue	Art Gallery	Music Venue
15	M5W	Downtown Toronto	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	2	Art Gallery	Museum	Performing Arts Venue
16	M4M	East Toronto	Studio District	43.659526	-79.340923	4	Music Store	Performing Arts Venue	Music Venue
17	M5R	Central Toronto	The Annex, North Midtown, Yorkville	43.672710	-79.405678	1	History Museum	Performing Arts Venue	Music Venue

Discussion

We can clearly see that the most overall rated cultural place is "Performing Arts Venue" followed by "Music Avenue" when combining all the data. Therefore, our advice to the stakeholders would be to increase this particular Avenue type if possible and make sure to also increase the budget on marketing the other cultural avenues as well to help promoting them for the city.

Conclusion

The analysis of this data is rather limited because we were to use Foursquare. The basis for the ranking positions is based on this platform and are given by users directly. It’s possible to have different results if the data set were other than Foursquare, for example, official data provided by some Government division.

The Clustering function can be changed and will yield different clusters. We assumed 5, but stakeholders can change this at will.

Was interesting to note that Museums, are popular but not as much as one might think.

Given the time allowed and the limited data, this was an interesting project that could very well benefit the Tourism agencies of Toronto!