IBM Data Science Capstone

Battle of the Neighborhoods

Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods (Week 2)

Introduction (Business Problem)

Toronto is the largest metropolitan area in Canada. The city of over six million people is the center of the country's financial commercial efforts. It is growing and projects to attract many people and businesses over the coming decade.

The objective of this study is to identify and cluster the various neighborhoods of Toronto and into similarities for business, dining, entertainment, and housing. The results of this analysis will be applied to individuals and corporations looking to relocate to Toronto.

Our target audience are twofold; individuals/families looking to relocate to a growing, dynamic and relatively safe metropolitan area, and businesses – established or start-up, that wish to leverage the commercial and financial advantages not just of Canada, but North America in general.

Data

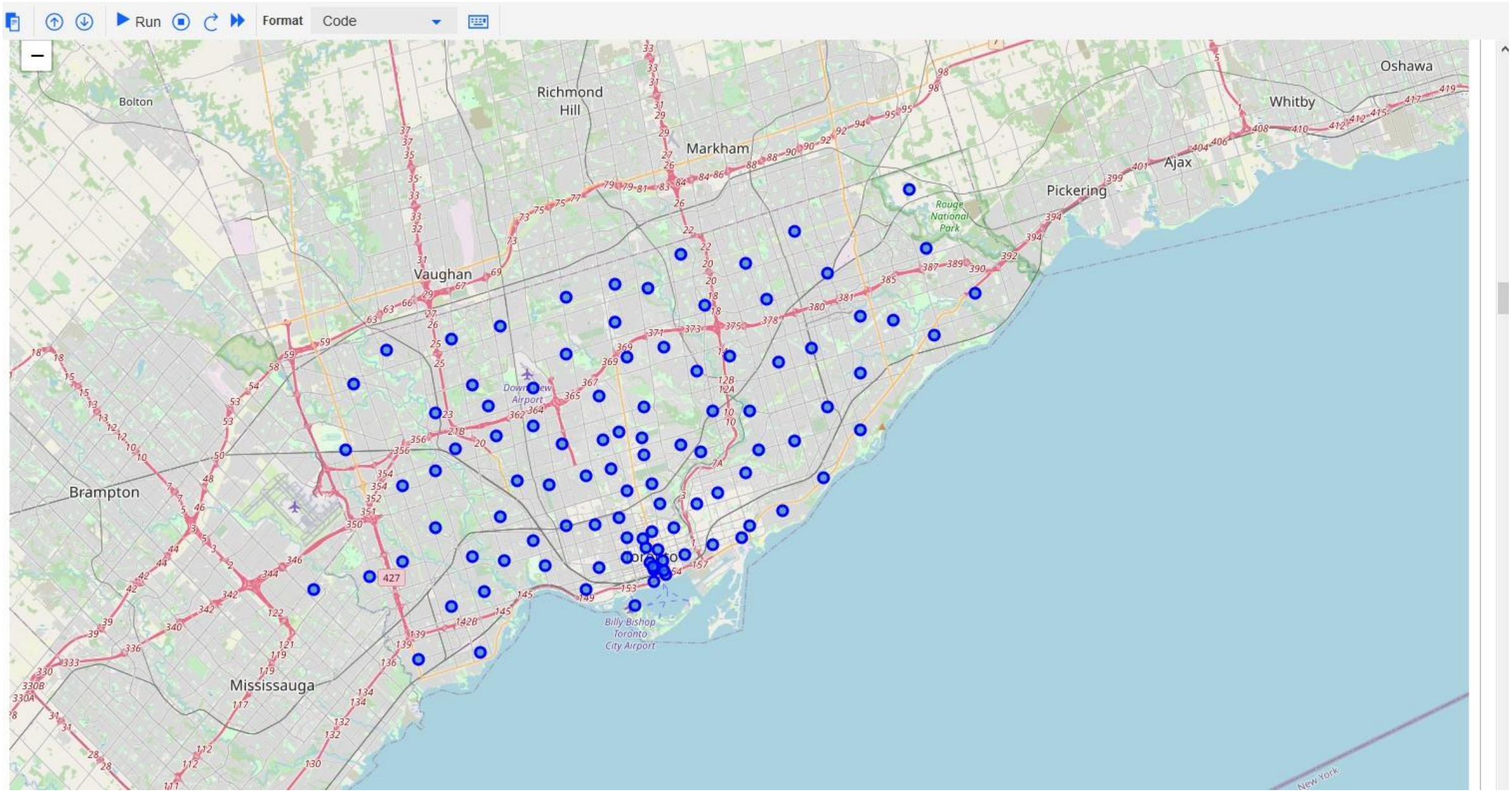
The datasets are compilations of several sites that focus on geospatial and demography. The first database is https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. This provides the individual boroughs and neighborhoods within Toronto as well as the accompanying postal codes. This is the geographical foundation for our study.

Our next data source provides that latitudes and longitudes for Toronto and its constituent neighborhoods, http://cocl.us/Geospatial_data. We will be able to build out our interactive and detailed maps. The data retrieved from Foursquare contains information of venues within a specified distance of the longitude and latitude of the postalcodes. The information obtained per venue as follows:

Neighborhood
Neighborhood Latitude
Neighborhood Longitude
Venue
Name of the venue e.g. the name of a store or restaurant
Venue Latitude
Venue Longitude
Venue Category

Finally, we utilize our Foursquare API to populate venues, categories, businesses and rating to the various concerns within the neighborhoods. These resources will allow us to create an objective, data driven product that will allow potential businesses and people to review the analysis and determine if relocation is the right choice and if so where in Toronto to relocate to.

Metropolitan Toronto



Methodology

Clustering Approach

In order to compare the various venues, housing prices, and school ranking in Toronto we decided to explore neighborhoods, segment them, and group them into clusters to analyze differences and similarities between the neighborhoods. We need to cluster data which in an unsupervised machine learning algorithm: k-means clustering algorithm.

K-means Clustering Coding and Output Sample¶

Utilize Foursquare API to Explore Neighborhoods

Neighborhood_explore.jpg

3.3 Explore the First Neighborhood in the Dataset ¶

[16]:
```
neighborhood_name = df_toronto.loc[0, 'Neighborhood']
print(f"The Neighborhood's name is '{neighborhood_name}'.")
```

```
The Neighborhood's name is 'Malvern, Rouge'.
```

[17]:
```
#get the neighborhood's lat/long values
neighborhood_latitude = df_toronto.loc[0, 'Latitude'] # neighborhood latitude value
neighborhood_longitude = df_toronto.loc[0, 'Longitude'] # neighborhood longitude value

print('Latitude and longitude values of {} are {}, {}.'.format(neighborhood_name,
                                                               neighborhood_latitude,
                                                               neighborhood_longitude))
```

```
Latitude and longitude values of Malvern, Rouge are 43.806686299999996, -79.19435340000001.
```

3.4 The Top 100 Venues within a 500 meter radius

[18]:
```
LIMIT = 100 # limit of number of venues returned by Foursquare API
radius = 500 # define radius
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    neighborhood_latitude,
    neighborhood_longitude,
    radius,
    LIMIT)

# get the result to a json file
results = requests.get(url).json()
```

4.1 K-Means Clustering

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

toronto_grouped_clustering = toronto_grouped.drop('Neighborhood', 1)

# 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]
```

Out[27]: array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int32)

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

toronto_merged = df_toronto

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

toronto_merged.head()
```

Out[28]:

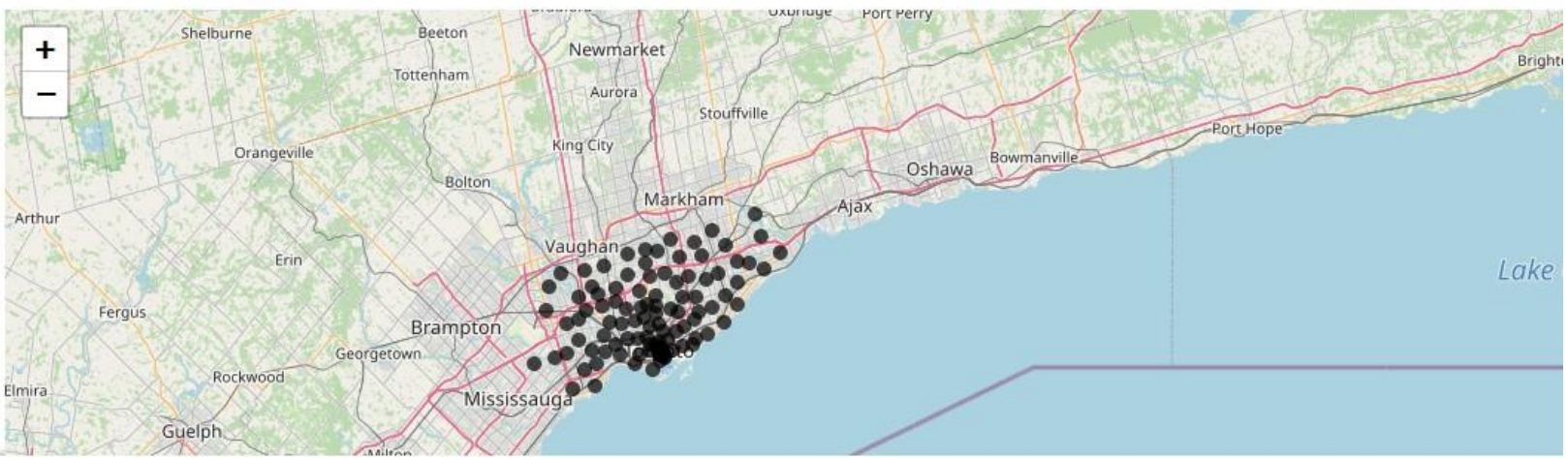
	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Com V
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	1.0	Fast Food Restaurant	Print Shop	Women's Store	Distribution Center	Department Store	Dessert Shop	Dim Sum Restaurant	
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	3.0	Bar	Women's Store	Deli / Bodega	Escape Room	Electronics Store	Eastern European Restaurant	Drugstore	Donut
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	1.0	Rental Car Location	Medical Center	Mexican Restaurant	Intersection	Bank	Restaurant	Electronics Store	Breakfast
3	M1G	Scarborough	Woburn	43.770992	-79.216917	1.0	Coffee Shop	Korean BBQ Restaurant	Women's Store	Dog Run	Dessert Shop	Dim Sum Restaurant	Diner	Discount
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	1.0	Athletics & Sports	Gas Station	Bakery	Bank	Caribbean Restaurant	Thai Restaurant	Fried Chicken Joint	H Resta

Results

```
map_clusters = folium.Map(location=[latitude_x, longitude_y], zoom_start=11)

# set color scheme for the clusters
x = np.arange(kclusters)
colors_array = cm.rainbow(np.linspace(0, 1, kclusters))
rainbow = [colors.rgb2hex(i) for i in colors_array]
print(rainbow)

# add markers to the map
markers_colors = []
for lat, lon, nei, cluster in zip(toronto_merged['Latitude'],
                                  toronto_merged['Longitude'],
                                  toronto_merged['Neighborhood'],
                                  toronto_merged['Cluster Labels']):
    label = folium.Popup(str(nei) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow,
        fill=True,
        fill_color=rainbow,
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

```
['#8000ff', '#00b5eb', '#80ffb4', '#ffb360', '#ff0000']
```

:[42]:

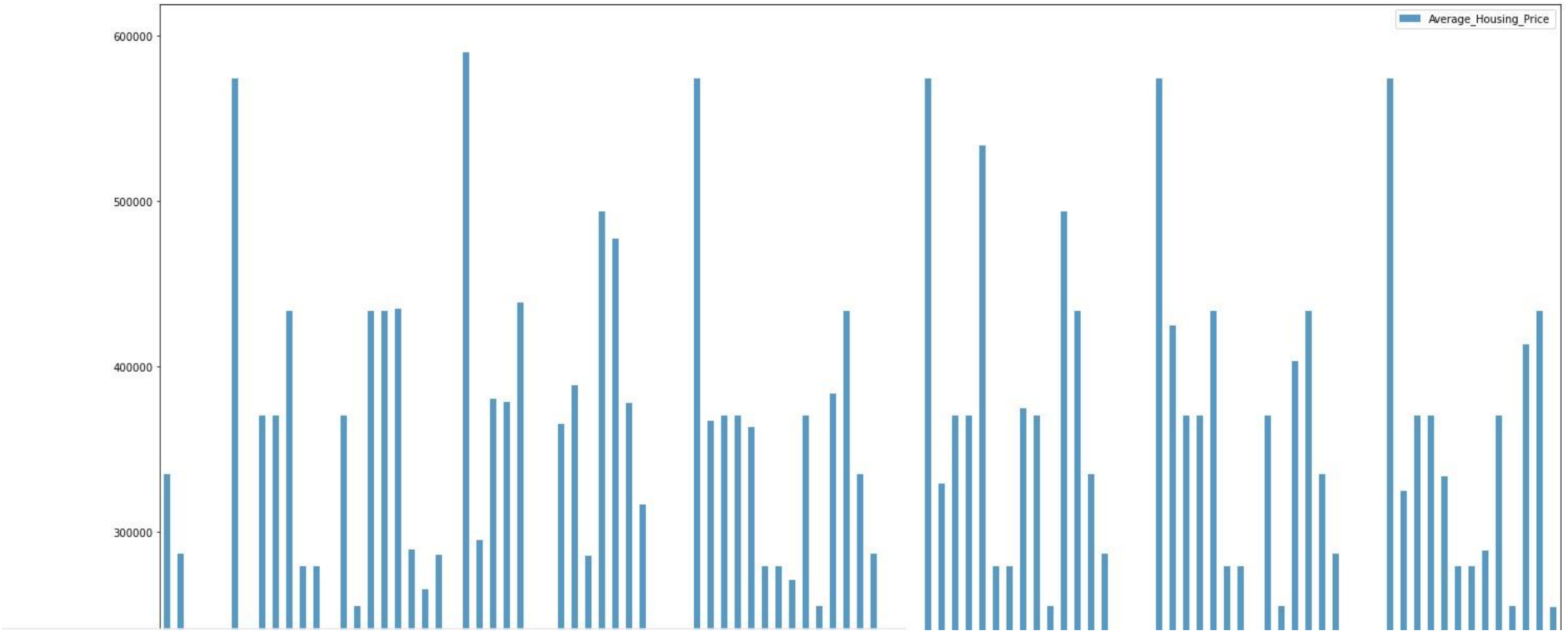


Average Housing Price by Clusters

```
[41]: Toronto_Avg_HousingPrice.set_index('Neighborhood',inplace=True,drop=True)
```

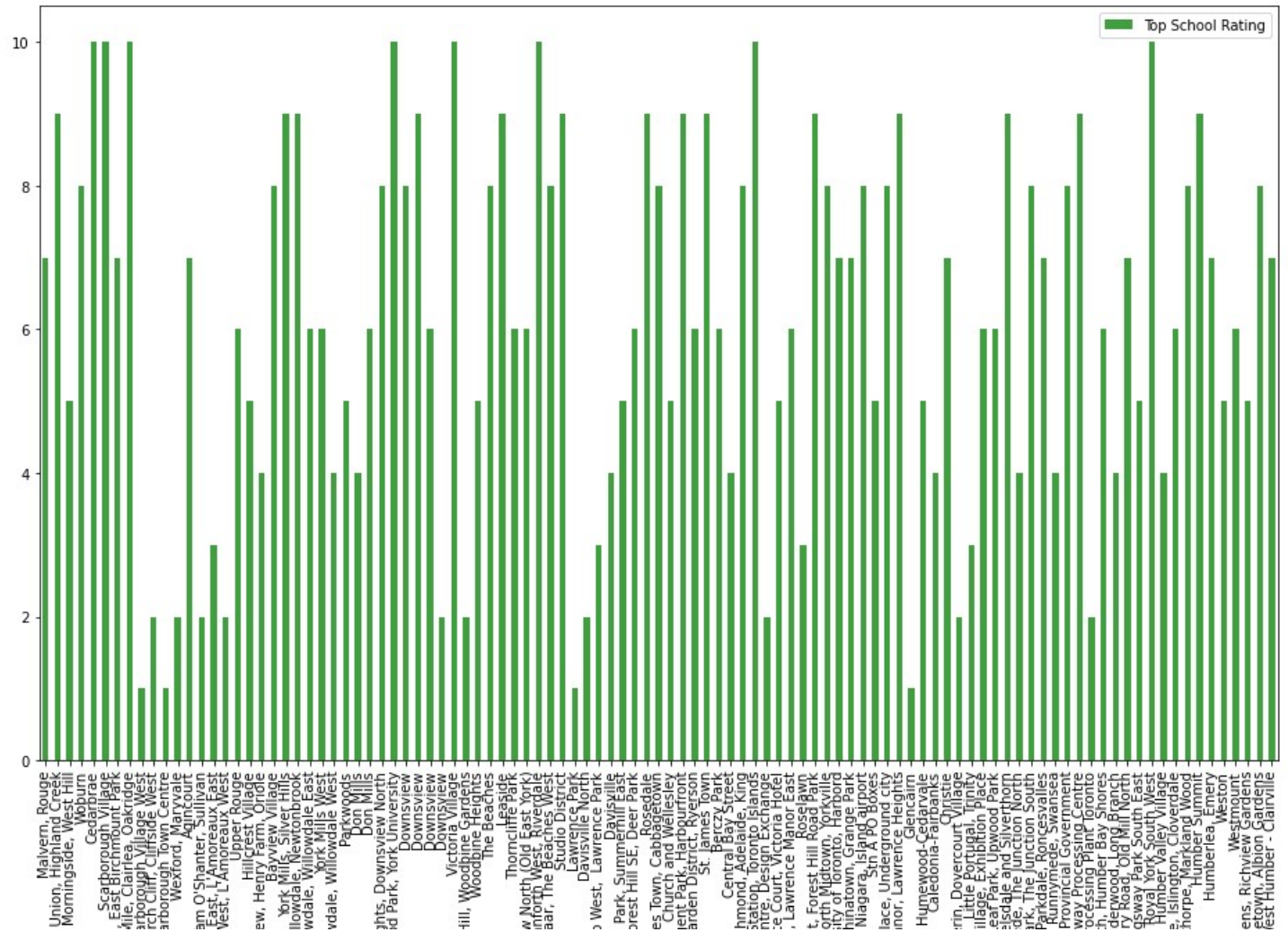
```
[42]: Toronto_Avg_HousingPrice.plot(kind='bar',figsize=(24,18),alpha=0.75)
```

```
Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbcee15db50>
```



School Ratings by Clusters

```
']: Toronto_school_ratings.set_index('Neighborhood',inplace=True,drop=True)

']: Toronto_school_ratings.plot(kind='bar',figsize=(16,10),color='green',alpha=0.75);
```



Discussion

Opportunity/Problem:

The major purpose of this project, is to analyze the various neighborhoods in a new city for the people and/or businesses that are considering relocating. Housing prices and school district ranking are highly influential factors in deciding major life events. The hope is that people can use this analysis to help decide their choices

Sorted list of house in terms of housing prices in a ascending or descending order
Sorted list of schools in terms of location, fees, rating and reviews

Conclusion

Using k-means cluster algorithm I separated Toronto into 5(five) distinctt clusters and for different latitude and longitude from dataset, providing similar neighborhood profiles around them. Individuals can use the outputs of neighborhood analyses to help decide their next actions.

The good part of this output is the data is modular so people at different point of their life can make decisions appropriate to them (people with children will way school ranking while those without will put less weight). This provides the means for people to use data without trying to prescribe end results.

This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision with confidence.