coursera project 3

June 14, 2019

Segmentation and Clustering
Import Libraries

In [1]:
```
import numpy as np

import pandas as pd
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

import json

from geopy.geocoders import Nominatim

import requests
from pandas.io.json import json_normalize

import matplotlib.cm as cm
import matplotlib.colors as colors

from sklearn.cluster import KMeans

import folium

print('Libraries imported.')
```

```
Libraries imported.
```

Obtain Dataset from Toronto

In [2]:
```
url='https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
df=pd.read_html(url, header=0)[0]
df.head()
```

Out[2]:

	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods

3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront

Ignore cells with Borough as "Not assigned" by deleting and resetting Index

```
In [4]: df = df[df.Borough != 'Not assigned']
        df.reset_index(inplace = True)
        df.drop('index', axis=1,inplace=True)
        df.head()
```

Out[4]:

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Harbourfront
3	M5A	Downtown Toronto	Regent Park
4	M6A	North York	Lawrence Heights

Linking cells based on borough and postcode for Neighborhood column

```
In [5]: df_group = df.groupby(['Postcode','Borough'])['Neighbourhood'].apply(lambda x:', '.join(x))
        df_group2 = pd.DataFrame(df_group)
        df_group2.reset_index(inplace = True)
        df_group2.head()
```

Out[5]:

	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Using the csv file for the dataframe

```
In [6]: import pandas as pd
        Toronto_df = pd.read_csv("https://cocl.us/Geospatial_data")
        Toronto_df.rename(columns={'Postal Code':'Postcode'}, inplace=True)
        Toronto_df.head()
```

Out[6]:

	Postcode	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

```
In [7]: df_inner = pd.merge(Toronto_df, df_group2, on='Postcode', how='inner')
        df_inner = df_inner[['Postcode','Borough','Neighbourhood','Latitude','Longitude']]
        df_inner.head()
```

Out[7]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	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

Clustering

Delete all Boroughs excluding the ones that contains 'Toronto'

In [8]:
```
Torontodf = df_inner[df_inner.Borough.str.contains("Toronto")]
Torontodf
```

Out[8]:

	Postcode	Borough
37	M4E	East Toronto
41	M4K	East Toronto
42	M4L	East Toronto
43	M4M	East Toronto
44	M4N	Central Toronto
45	M4P	Central Toronto
46	M4R	Central Toronto
47	M4S	Central Toronto
48	M4T	Central Toronto
49	M4V	Central Toronto
50	M4W	Downtown Toronto
51	M4X	Downtown Toronto
52	M4Y	Downtown Toronto
53	M5A	Downtown Toronto
54	M5B	Downtown Toronto
55	M5C	Downtown Toronto
56	M5E	Downtown Toronto
57	M5G	Downtown Toronto
58	M5H	Downtown Toronto
59	M5J	Downtown Toronto
60	M5K	Downtown Toronto
61	M5L	Downtown Toronto
63	M5N	Central Toronto
64	M5P	Central Toronto
65	M5R	Central Toronto
66	M5S	Downtown Toronto
67	M5T	Downtown Toronto

68	M5V	Downtown Toronto	
69	M5W	Downtown Toronto	
70	M5X	Downtown Toronto	
75	M6G	Downtown Toronto	
76	M6H	West Toronto	
77	M6J	West Toronto	
78	M6K	West Toronto	
82	M6P	West Toronto	
83	M6R	West Toronto	
84	M6S	West Toronto	
87	M7Y	East Toronto	

	Neighbourhood	Latitude	Longitude
37	The Beaches	43.676357	-79.293031
41	The Danforth West, Riverdale	43.679557	-79.352188
42	The Beaches West, India Bazaar	43.668999	-79.315572
43	Studio District	43.659526	-79.340923
44	Lawrence Park	43.728020	-79.388790
45	Davisville North	43.712751	-79.390197
46	North Toronto West	43.715383	-79.405678
47	Davisville	43.704324	-79.388790
48	Moore Park, Summerhill East	43.689574	-79.383160
49	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686412	-79.400049
50	Rosedale	43.679563	-79.377529
51	Cabbagetown, St. James Town	43.667967	-79.367675
52	Church and Wellesley	43.665860	-79.383160
53	Harbourfront, Regent Park	43.654260	-79.360636
54	Ryerson, Garden District	43.657162	-79.378937
55	St. James Town	43.651494	-79.375418
56	Berczy Park	43.644771	-79.373306
57	Central Bay Street	43.657952	-79.387383
58	Adelaide, King, Richmond	43.650571	-79.384568
59	Harbourfront East, Toronto Islands, Union Station	43.640816	-79.381752
60	Design Exchange, Toronto Dominion Centre	43.647177	-79.381576
61	Commerce Court, Victoria Hotel	43.648198	-79.379817
63	Roselawn	43.711695	-79.416936
64	Forest Hill North, Forest Hill West	43.696948	-79.411307
65	The Annex, North Midtown, Yorkville	43.672710	-79.405678
66	Harbord, University of Toronto	43.662696	-79.400049
67	Chinatown, Grange Park, Kensington Market	43.653206	-79.400049
68	CN Tower, Bathurst Quay, Island airport, Harbo...	43.628947	-79.394420
69	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846
70	First Canadian Place, Underground city	43.648429	-79.382280
75	Christie	43.669542	-79.422564
76	Dovercourt Village, Dufferin	43.669005	-79.442259
77	Little Portugal, Trinity	43.647927	-79.419750
78	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191
82	High Park, The Junction South	43.661608	-79.464763

83	Parkdale, Roncesvalles	43.648960	-79.456325
84	Runnymede, Swansea	43.651571	-79.484450
87	Business Reply Mail Processing Centre 969 Eastern	43.662744	-79.321558

Hot Coding

In [9]:
```
Toronto_onehot = pd.get_dummies(Torontodf[['Borough']], prefix="", prefix_sep="")

Toronto_onehot['Neighbourhood'] = Torontodf['Neighbourhood']

fixed_columns = [Toronto_onehot.columns[-1]] + list(Toronto_onehot.columns[:-1])
Toronto_onehot = Toronto_onehot[fixed_columns]

Toronto_onehot.head()
```

Out[9]:

	Neighbourhood	Central Toronto	Downtown Toronto	East Toronto	West Toronto
37	The Beaches	0	0	1	0
41	The Danforth West, Riverdale	0	0	1	0
42	The Beaches West, India Bazaar	0	0	1	0
43	Studio District	0	0	1	0
44	Lawrence Park	1	0	0	0

In [10]:
```
Toronto_grouped = Toronto_onehot.groupby('Neighbourhood').mean().reset_index()
Toronto_grouped.head()
```

Out[10]:

	Neighbourhood	Central Toronto	Downtown Toronto	East Toronto	West Toronto
0	Adelaide, King, Richmond	0	1	0	0
1	Berczy Park	0	1	0	0
2	Brockton, Exhibition Place, Parkdale Village	0	0	0	1
3	Business Reply Mail Processing Centre 969 Eastern	0	0	1	0
4	CN Tower, Bathurst Quay, Island airport, Harbo...	0	1	0	0

In [11]:
```
from sklearn.cluster import KMeans
```

In [12]:
```
kclusters = 4
```

```
Toronto_grouped_clustering = Toronto_grouped.drop('Neighbourhood', 1)

kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(Toronto_grouped_clustering)

kmeans.labels_[0:10]
```

Out[12]: array([0, 0, 2, 3, 0, 0, 0, 0, 0, 0], dtype=int32)

Merging clusters with the dataframe for Toronto

```
In [13]: Torontodf.insert(0, 'Cluster Labels', kmeans.labels_)
         Torontodf.head()
```

Out[13]:

	Cluster Labels	Postcode	Borough	Neighbourhood	\
37	0	M4E	East Toronto	The Beaches	
41	0	M4K	East Toronto	The Danforth West, Riverdale	
42	2	M4L	East Toronto	The Beaches West, India Bazaar	
43	3	M4M	East Toronto	Studio District	
44	0	M4N	Central Toronto	Lawrence Park	

	Latitude	Longitude
37	43.676357	-79.293031
41	43.679557	-79.352188
42	43.668999	-79.315572
43	43.659526	-79.340923
44	43.728020	-79.388790

Creating Map

```
In [ ]: import folium
        map_clusters = folium.Map(location=[latitude, longitude], zoom_start=11)

        x = np.arange(kclusters)
        ys = [i + x + (i*x)**2 for i in range(kclusters)]
        colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
        rainbow = [colors.rgb2hex(i) for i in colors_array]

        # add markers to the map
        markers_colors = []
        for lat, lon, poi, cluster in zip(Torontodf['Latitude'], Torontodf['Longitude'], Torontodf['Neighbourhood'], To
            label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
            folium.CircleMarker(
                [lat, lon],
                radius=5,
                popup=label,
                color=rainbow[cluster-1],
                fill=True,
                fill_color=rainbow[cluster-1],
                fill_opacity=0.7).add_to(map_clusters)
```

```
map_clusters
```

In [16]:
```
address = 'Toronto, CA'

geolocator = Nominatim(user_agent="T_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))
```

The geograpical coordinate of Toronto are 43.653963, -79.387207.

Cluster Examination

In [17]:
```
Torontodf.loc[Torontodf['Cluster Labels'] == 0, Torontodf.columns[[1] + list(range(5, Torontodf.shape[1])
```

Out[17]:

	Postcode	Longitude
37	M4E	-79.293031
41	M4K	-79.352188
44	M4N	-79.388790
45	M4P	-79.390197
46	M4R	-79.405678
47	M4S	-79.388790
48	M4T	-79.383160
49	M4V	-79.400049
50	M4W	-79.377529
54	M5B	-79.378937
56	M5E	-79.373306
58	M5H	-79.384568
59	M5J	-79.381752
60	M5K	-79.381576
68	M5V	-79.394420
75	M6G	-79.422564
76	M6H	-79.442259
77	M6J	-79.419750

In [18]:
```
Torontodf.loc[Torontodf['Cluster Labels'] == 1, Torontodf.columns[[1] + list(range(5, Torontodf.shape[1])
```

Out[18]:

	Postcode	Longitude
51	M4X	-79.367675
52	M4Y	-79.383160
53	M5A	-79.360636
57	M5G	-79.387383
63	M5N	-79.416936
65	M5R	-79.405678
66	M5S	-79.400049
69	M5W	-79.374846
82	M6P	-79.464763

In [19]:
```
Torontodf.loc[Torontodf['Cluster Labels'] == 2, Torontodf.columns[[1] + list(range(5, Torontodf.shape[1])
```

Out[19]:

	Postcode	Longitude
42	M4L	-79.315572
55	M5C	-79.375418
61	M5L	-79.379817
64	M5P	-79.411307
67	M5T	-79.400049
70	M5X	-79.382280

In [20]:
```
Torontodf.loc[Torontodf['Cluster Labels'] == 3, Torontodf.columns[[1] + list(range(5, Torontodf.shape[1])
```

Out[20]:

	Postcode	Longitude
43	M4M	-79.340923
78	M6K	-79.428191
83	M6R	-79.456325
84	M6S	-79.484450
87	M7Y	-79.321558

In [ ]: