Segmenting and Clustering Neighbourhoods in Toronto

In [1]:
```
import pandas as pd
import numpy as np
import requests
```

In [8]:
```
wiki = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_
M'
wiki_page = requests.get(wiki)

df = pd.read_html(wiki_page.content, header = 0)[0]
df = df[wiki_raw.Neighbourhood != 'Not assigned']
df.reset_index(inplace = True)
df.head()
```

Out[8]:

	index	Postal Code	Borough	Neighbourhood
0	2	M3A	North York	Parkwoods
1	3	M4A	North York	Victoria Village
2	4	M5A	Downtown Toronto	Regent Park, Harbourfront
3	5	M6A	North York	Lawrence Manor, Lawrence Heights
4	6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

In [9]:
```
df.groupby(['Postal Code']).first()
```

Out[9]:

Postal Code	index	Borough	Neighbourhood

Postal Code	index	Borough	Neighbourhood
M1B	9	Scarborough	Malvern, Rouge
M1C	18	Scarborough	Rouge Hill, Port Union, Highland Creek
M1E	27	Scarborough	Guildwood, Morningside, West Hill
M1G	36	Scarborough	Woburn
M1H	45	Scarborough	Cedarbrae
...	...	...	...
M9N	98	York	Weston
M9P	107	Etobicoke	Westmount
M9R	116	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...
M9V	143	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...
M9W	152	Etobicoke	Northwest, West Humber - Clairville

103 rows × 3 columns

In [11]:
```
len(df['Postal Code'].unique())
```

Out[11]: 103

In [12]:
```
df[df['Borough'] == 'Not assigned']
```

Out[12]:

index	Postal Code	Borough	Neighbourhood

In [13]:
```
df.shape
```

Out[13]: (103, 4)

Part 2

In [14]:
```
pip install geocoder
```

```
Collecting geocoder
  Downloading geocoder-1.38.1-py2.py3-none-any.whl (98 kB)
Requirement already satisfied: click in c:\users\abhishek\anaconda3\lib
\site-packages (from geocoder) (7.1.2)
Requirement already satisfied: six in c:\users\abhishek\anaconda3\lib\s
ite-packages (from geocoder) (1.15.0)
Requirement already satisfied: future in c:\users\abhishek\anaconda3\li
b\site-packages (from geocoder) (0.18.2)
Requirement already satisfied: requests in c:\users\abhishek\anaconda3
\lib\site-packages (from geocoder) (2.24.0)
Collecting ratelim
  Downloading ratelim-0.1.6-py2.py3-none-any.whl (4.0 kB)
Requirement already satisfied: chardet<4,>=3.0.2 in c:\users\abhishek\a
naconda3\lib\site-packages (from requests->geocoder) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1
in c:\users\abhishek\anaconda3\lib\site-packages (from requests->geocod
er) (1.25.9)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\abhishek
\anaconda3\lib\site-packages (from requests->geocoder) (2020.6.20)
Requirement already satisfied: idna<3,>=2.5 in c:\users\abhishek\anacon
da3\lib\site-packages (from requests->geocoder) (2.10)
Requirement already satisfied: decorator in c:\users\abhishek\anaconda3
\lib\site-packages (from ratelim->geocoder) (4.4.2)
Installing collected packages: ratelim, geocoder
Successfully installed geocoder-1.38.1 ratelim-0.1.6
Note: you may need to restart the kernel to use updated packages.
```

In [15]:
```
import geocoder
```

In [16]:
```
url = 'http://cocl.us/Geospatial_data'
```

In [17]:
```
df_geo = pd.read_csv(url)
df_geo.head()
```

Out[17]:

	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

In [18]:
```
df_geo.dtypes
```

Out[18]:
```
Postal Code     object
Latitude       float64
Longitude      float64
dtype: object
```

In [19]:
```
df.dtypes
```

Out[19]:
```
index             int64
Postal Code      object
Borough          object
Neighbourhood    object
dtype: object
```

In [21]:
```
df.shape
```

Out[21]:
```
(103, 4)
```

In [22]:
```
df_geo.shape
```

Out[22]:
```
(103, 3)
```

In [23]:
```
df = df.join(df_geo.set_index('Postal Code'), on='Postal Code')
df
```

Out[23]:

	index	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	2	M3A	North York	Parkwoods	43.753259	-79.329656
1	3	M4A	North York	Victoria Village	43.725882	-79.315572
2	4	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	5	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
...	...	...	...	...	...	...
98	160	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.653654	-79.506944
99	165	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
100	168	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.662744	-79.321558
101	169	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509
102	178	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.628841	-79.520999

103 rows × 6 columns

In [24]:
```
df = df.reset_index()
df.drop(['index'], axis = 'columns', inplace = True)
df = df.set_index('level_0')
df.head()
```

Out[24]:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
level_0					
0	M3A	North York	Parkwoods	43.753259	-79.329656

level_0	Postal Code	Borough	Neighbourhood	Latitude	Longitude
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

In [25]:
```
df = df.rename(index = {'level_0' : 'index'})
```

In [26]:
```
df.index.name = 'index'
```

In [27]:
```
df.head()
```

Out[27]:

index	Postal Code	Borough	Neighbourhood	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	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

In [28]:
```
df.shape
```

Out[28]:
```
(103, 5)
```

In [ ]: