In [ ]:

```
```

In [1]:

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import requests
```

In [2]:

```
url="https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
header = {
  "User-Agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/50.0.2661.75 Safari/5
37.36",
  "X-Requested-With": "XMLHttpRequest"
}

r = requests.get(url, headers=header)

tables = pd.read_html(r.text)

df=pd.DataFrame(tables[0])

# The dataframe will consist of three columns: PostalCode, Borough, and Neighborhood

df.columns=['Postcode','Borough','Neighbourhood']

df.drop([0],axis=0,inplace=True)

df.reset_index()

# Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.

df.drop(df[df['Borough']=="Not assigned"].index,axis=0, inplace=True)

# More than one neighborhood can exist in one postal code area.
# For example, in the table on the Wikipedia page,
# you will notice that M5A is listed twice and has two neighborhoods:
# Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods
# separated with a comma as shown in row 11 in the above table.

df1=df.groupby("Postcode").agg(lambda x:','.join(set(x)))

# If a cell has a borough but a Not assigned neighborhood,
# then the neighborhood will be the same as the borough.
# So for the 9th cell in the table on the Wikipedia page,
# the value of the Borough and the Neighborhood columns will be Queen's Park.

df1.loc[df1['Neighbourhood']=="Not assigned",'Neighbourhood']=df1.loc[df1['Neighbourhood']=="Not assigned",'Borough'
]

df1.shape
```

Out[2]:

```
(103, 2)
```

In [ ]:

```
```