Untitled7

August 4, 2019

[1]:
```
import numpy as np # library to handle data in a vectorized manner

import pandas as pd # library for data analsysis
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

import json # library to handle JSON files

!pip install geopy
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors

# import k-means from clustering stage
from sklearn.cluster import KMeans

!conda install -c conda-forge folium=0.5.0 --yes
import folium # map rendering library

print('Libraries imported.')
```

```
Requirement already satisfied: geopy in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (1.20.0)
Requirement already satisfied: geographiclib<2,>=1.49 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from geopy)
(1.49)
Solving environment: done


==> WARNING: A newer version of conda exists. <==
  current version: 4.5.11
  latest version: 4.7.10
```

Please update conda by running

```
$ conda update -n base -c defaults conda
```

All requested packages already installed.

Libraries imported.

[2]:
```
url = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
toronto_list= pd.read_html(url)[0]
```

```
---------------------------------------------------------------------------
ImportError                               Traceback (most recent call last)

<ipython-input-2-f17a33f5ba49> in <module>
      1 url = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
----> 2 toronto_list= pd.read_html(url)[0]


~/conda/envs/python/lib/python3.6/site-packages/pandas/io/html.py in read_html(io,
↪match, flavor, header, index_col, skiprows, attrs, parse_dates, thousands, encoding, decimal,
↪converters, na_values, keep_default_na, displayed_only)
   1103         na_values=na_values,
   1104         keep_default_na=keep_default_na,
-> 1105         displayed_only=displayed_only,
   1106     )


~/conda/envs/python/lib/python3.6/site-packages/pandas/io/html.py in _parse(flavor, io,
↪match, attrs, encoding, displayed_only, **kwargs)
    886     retained = None
    887     for flav in flavor:
--> 888         parser = _parser_dispatch(flav)
    889         p = parser(io, compiled_match, attrs, encoding, displayed_only)
    890


~/conda/envs/python/lib/python3.6/site-packages/pandas/io/html.py in
↪_parser_dispatch(flavor)
    841     else:
    842         if not _HAS_LXML:
--> 843             raise ImportError("lxml not found, please install it")
```

```
844     return _valid_parsers[flavor]
845
```

```
ImportError: lxml not found, please install it
```

[3]:
```
toronto_list.head()
```

```
---------------------------------------------------------------------------

NameError                                 Traceback (most recent call last)

<ipython-input-3-62b0cc7d29a8> in <module>
----> 1 toronto_list.head()

NameError: name 'toronto_list' is not defined
```

[ ]:
```
CanadaData = pd.DataFrame(toronto_list)
```

[ ]:
```
CanadaData.shape
```

[ ]:
```
CanadaData.head()
```

[ ]:
```
CanadaData = CanadaData[CanadaData.Borough != "Not assigned"]
```

[ ]:
```
CanadaData.head()
```

[ ]:
```
CanadaData.shape
```

[ ]:
```
CanadaData.loc[CanadaData.Neighbourhood == 'Not assigned', 'Neighbourhood'] =␣
↪CanadaData.Borough
```

[ ]:
```
CanadaData
```

[ ]:
```
CanadaDataGrouped = CanadaData.groupby(['Postcode', 'Borough'], as_index=False,␣
↪sort=False).agg(','.join)
```

[ ]:
```
CanadaDataGrouped.head()
```

[ ]:
```
CanadaDataGrouped.shape
```