Data Collection

July 26, 2020

1 Data Collection

[5]:

```
import json
import sys
sys.path.append('/home/nbuser/library/')
import pandas as pd
import requests
```

API Call First step - Get URL from documentation Second step - look at the required parameters Third step - Catch when status code does not equal to 200

[44]:

```
def get_C18pop(x,y):

    url = 'https://koordinates.com/services/query/v1/vector.json'

    params = {
        'key' : 'aa35290d615e43c4ad41b17c7a08881e',
        'layer' : 104612,
        'x' : x,
        'y' : y
    }

    response = requests.get(url, params=params)

    if response.status_code != 200:
        return response.status_code

    C18pop = response.
 ↪json()['vectorQuery']['layers']['104612']['features'][0]['properties']['C18_CURPop']
    return C18pop
```

2 Putting C18pop into the CSV

[69]:

```
datafile = pd.read_csv('Dataset.csv')
sample = datafile.head()
```

2.1 Testing with sample first

[55]:
```
sample
```

[55]:
```
   Bedrooms  Bathrooms                                    Address Land area  \
0         5        3.0  106 Lawrence Crescent Hill Park, Auckland       714
1         5        3.0               8 Corsica Way Karaka, Auckland       564
2         6        4.0     243 Harbourside Drive Karaka, Auckland       626
3         2        1.0  2/30 Hardington Street Onehunga, Auckland        65
4         3        1.0     59 Israel Avenue Clover Park, Auckland       601

        CV   Latitude   Longitude      SA1  0-19 years  20-29 years  \
0   960000 -37.012920  174.904069  7009770          48           27
1  1250000 -37.063672  174.922912  7009991          42           18
2  1250000 -37.063580  174.924044  7009991          42           18
3   740000 -36.912996  174.787425  7007871          42            6
4   630000 -36.979037  174.892612  7008902          93           27

   30-39 years  40-49 years  50-59 years  60+ years      Suburbs
0           24           21           24         21     Manurewa
1           12           21           15         30       Karaka
2           12           21           15         30       Karaka
3           21           21           12         15     Onehunga
4           33           30           21         33  Clover Park
```

[56]:
```
sample['C18_CURPop'] = sample.apply(lambda row: get_C18pop(row['Longitude'],␣
 ↪row['Latitude']), axis = 1)
```

```
/home/nbuser/anaconda3_501/lib/python3.6/site-packages/ipykernel/__main__.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/indexing.html#indexing-view-versus-copy
  if __name__ == '__main__':
```

[57]:
```
sample
```

[57]:
```
   Bedrooms  Bathrooms                                    Address Land area  \
0         5        3.0  106 Lawrence Crescent Hill Park, Auckland       714
1         5        3.0               8 Corsica Way Karaka, Auckland       564
2         6        4.0     243 Harbourside Drive Karaka, Auckland       626
3         2        1.0  2/30 Hardington Street Onehunga, Auckland        65
4         3        1.0     59 Israel Avenue Clover Park, Auckland       601

        CV   Latitude   Longitude      SA1  0-19 years  20-29 years  \
0   960000 -37.012920  174.904069  7009770          48           27
1  1250000 -37.063672  174.922912  7009991          42           18
```

```
2  1250000 -37.063580  174.924044  7009991          42           18
3   740000 -36.912996  174.787425  7007871          42            6
4   630000 -36.979037  174.892612  7008902          93           27

   30-39 years  40-49 years  50-59 years  60+ years      Suburbs  C18_CURPop
0           24           21           24         21     Manurewa         174
1           12           21           15         30       Karaka         129
2           12           21           15         30       Karaka         129
3           21           21           12         15     Onehunga         120
4           33           30           21         33  Clover Park         231
```

2.2 Converting in Datafile

[71]:

```
datafile['C18_CURPop'] = datafile.apply(lambda row: get_C18pop(row['Longitude'],␣
 ↪row['Latitude']), axis = 1)
```

[72]:

```
datafile.head()
```

[72]:

```
   Bedrooms  Bathrooms                                    Address Land area  \
0         5        3.0  106 Lawrence Crescent Hill Park, Auckland       714
1         5        3.0             8 Corsica Way Karaka, Auckland       564
2         6        4.0     243 Harbourside Drive Karaka, Auckland       626
3         2        1.0  2/30 Hardington Street Onehunga, Auckland        65
4         3        1.0     59 Israel Avenue Clover Park, Auckland       601

        CV   Latitude   Longitude      SA1  0-19 years  20-29 years  \
0   960000 -37.012920  174.904069  7009770          48           27
1  1250000 -37.063672  174.922912  7009991          42           18
2  1250000 -37.063580  174.924044  7009991          42           18
3   740000 -36.912996  174.787425  7007871          42            6
4   630000 -36.979037  174.892612  7008902          93           27

   30-39 years  40-49 years  50-59 years  60+ years      Suburbs  C18_CURPop
0           24           21           24         21     Manurewa         174
1           12           21           15         30       Karaka         129
2           12           21           15         30       Karaka         129
3           21           21           12         15     Onehunga         120
4           33           30           21         33  Clover Park         231
```

Writing out to csv file

[73]:

```
datafile.to_csv('Dataset_C18pop.csv', index=False)
```

3 Land Area Clean Up

[86]:
```
# df_1 represent Dataset_C18pop
df_1 = pd.read_csv('Dataset_C18pop.csv')
sample = df_1.head()
```

[75]:
```
sample
```

[75]:
```
   Bedrooms  Bathrooms                                    Address Land area  \
0         5        3.0  106 Lawrence Crescent Hill Park, Auckland       714
1         5        3.0              8 Corsica Way Karaka, Auckland       564
2         6        4.0      243 Harbourside Drive Karaka, Auckland       626
3         2        1.0  2/30 Hardington Street Onehunga, Auckland        65
4         3        1.0      59 Israel Avenue Clover Park, Auckland       601

        CV   Latitude   Longitude      SA1  0-19 years  20-29 years  \
0   960000 -37.012920  174.904069  7009770          48           27
1  1250000 -37.063672  174.922912  7009991          42           18
2  1250000 -37.063580  174.924044  7009991          42           18
3   740000 -36.912996  174.787425  7007871          42            6
4   630000 -36.979037  174.892612  7008902          93           27

   30-39 years  40-49 years  50-59 years  60+ years      Suburbs  C18_CURPop
0           24           21           24         21     Manurewa         174
1           12           21           15         30       Karaka         129
2           12           21           15         30       Karaka         129
3           21           21           12         15     Onehunga         120
4           33           30           21         33  Clover Park         231
```

[87]:
```
df_1["Land area"] = df_1["Land area"].str.extract('(\d+)').astype(float)
df_1.describe()
```

[87]:
```
          Bedrooms    Bathrooms     Land area            CV     Latitude  \
count  1051.000000  1049.000000   1051.000000  1.051000e+03  1051.000000
mean      3.777355     2.073403    856.989534  1.387521e+06   -36.893715
std       1.169412     0.992985   1588.156219  1.182939e+06     0.130100
min       1.000000     1.000000     40.000000  2.700000e+05   -37.265021
25%       3.000000     1.000000    321.000000  7.800000e+05   -36.950565
50%       4.000000     2.000000    571.000000  1.080000e+06   -36.893132
75%       4.000000     3.000000    825.000000  1.600000e+06   -36.855789
max      17.000000     8.000000  22240.000000  1.800000e+07   -36.177655

         Longitude           SA1   0-19 years  20-29 years  30-39 years  \
count  1051.000000  1.051000e+03  1051.000000  1051.000000  1051.000000
mean    174.799325  7.006319e+06    47.549001    28.963844    27.042816
std       0.119538  2.591262e+03    24.692205    21.037441    17.975408
min     174.317078  7.001130e+06     0.000000     0.000000     0.000000
25%     174.720779  7.004416e+06    33.000000    15.000000    15.000000
50%     174.798575  7.006325e+06    45.000000    24.000000    24.000000
```

```
75%      174.880944  7.008384e+06    57.000000    36.000000    33.000000
max      175.492424  7.011028e+06   201.000000   270.000000   177.000000

       40-49 years  50-59 years   60+ years   C18_CURPop
count  1051.000000  1051.000000  1051.000000  1051.000000
mean     24.125595    22.615604    29.360609   179.914367
std      10.942770    10.210578    21.805031    71.059280
min       0.000000     0.000000     0.000000     3.000000
25%      18.000000    15.000000    18.000000   138.000000
50%      24.000000    21.000000    27.000000   174.000000
75%      30.000000    27.000000    36.000000   210.000000
max     114.000000    90.000000   483.000000   789.000000
```

[88]:
```
df_1.to_csv('Dataset_C18pop.csv', index=False)
```

4 Add dev index

[99]:
```
# df_1 represents Dataset_C18pop
# df_2 represents DevIndex
df_1 = pd.read_csv('Dataset_C18pop.csv')
df_2 = pd.read_csv('DevIndex.csv')
df_1.head()
```

[99]:
```
   Bedrooms  Bathrooms                                   Address  Land area  \
0         5        3.0  106 Lawrence Crescent Hill Park, Auckland      714.0
1         5        3.0              8 Corsica Way Karaka, Auckland      564.0
2         6        4.0      243 Harbourside Drive Karaka, Auckland      626.0
3         2        1.0  2/30 Hardington Street Onehunga, Auckland       65.0
4         3        1.0      59 Israel Avenue Clover Park, Auckland      601.0

        CV   Latitude   Longitude      SA1  0-19 years  20-29 years  \
0   960000 -37.012920  174.904069  7009770          48           27
1  1250000 -37.063672  174.922912  7009991          42           18
2  1250000 -37.063580  174.924044  7009991          42           18
3   740000 -36.912996  174.787425  7007871          42            6
4   630000 -36.979037  174.892612  7008902          93           27

   30-39 years  40-49 years  50-59 years  60+ years      Suburbs  C18_CURPop
0           24           21           24         21     Manurewa         174
1           12           21           15         30       Karaka         129
2           12           21           15         30       Karaka         129
3           21           21           12         15     Onehunga         120
4           33           30           21         33  Clover Park         231
```

[100]:
```
# rename the key
df_2 = df_2.rename({'SA12018_code':'SA1','NZDep2018':'DepIndex'}, axis=1)
df_2.head()
```

[100]:
```
       SA1  DepIndex  NZDep2018_Score  URPopnSA1_2018  SA22018_code  \
0  7000000      10.0           1245.0             141        100100   
1  7000001      10.0           1245.0             114        100100   
2  7000002       NaN              NaN               0        100300   
3  7000003      10.0           1207.0             225        100100   
4  7000004       9.0           1093.0             138        100100   

                SA22018_name  
0                 North Cape  
1                 North Cape  
2  Inlets Far North District  
3                 North Cape  
4                 North Cape  
```

[102]:
```
df_1 = df_1.merge(df_2[['SA1', 'DepIndex']], on='SA1', how='left')
df_1.head()
```

[102]:
```
   Bedrooms  Bathrooms                                  Address  Land area  \
0         5        3.0  106 Lawrence Crescent Hill Park, Auckland      714.0   
1         5        3.0               8 Corsica Way Karaka, Auckland      564.0   
2         6        4.0     243 Harbourside Drive Karaka, Auckland      626.0   
3         2        1.0  2/30 Hardington Street Onehunga, Auckland       65.0   
4         3        1.0     59 Israel Avenue Clover Park, Auckland      601.0   

        CV   Latitude   Longitude      SA1  0-19 years  20-29 years  \
0   960000 -37.012920  174.904069  7009770          48           27   
1  1250000 -37.063672  174.922912  7009991          42           18   
2  1250000 -37.063580  174.924044  7009991          42           18   
3   740000 -36.912996  174.787425  7007871          42            6   
4   630000 -36.979037  174.892612  7008902          93           27   

   30-39 years  40-49 years  50-59 years  60+ years      Suburbs  C18_CURPop  \
0           24           21           24         21     Manurewa         174   
1           12           21           15         30       Karaka         129   
2           12           21           15         30       Karaka         129   
3           21           21           12         15     Onehunga         120   
4           33           30           21         33  Clover Park         231   

   DepIndex  
0       6.0  
1       1.0  
2       1.0  
3       2.0  
4       9.0  
```

[103]:
```
df_1.to_csv('Dataset_C18pop.csv', index=False)
```