

Achraf Safsafi

DSC 540

Project_Milestone 4

```
In [1]: import requests

HOST = "https://api.census.gov/data"
year = "2018"
dataset = "acs/acs1/subject"
base_url = "/" + ".join([HOST, year, dataset])
predicates = {}
get_vars = ["NAME", "S0101_C01_015E", "S0101_C01_016E", "S0101_C01_017E", "S0101_C01_018E", "S0101_C01_019E"
]
predicates["get"] = ".join(get_vars)
predicates["for"] = "county:*"
r = requests.get(base_url, params=predicates)

In [2]: #Check the response
print(r.text[0:1000])

[["NAME", "S0101_C01_015E", "S0101_C01_016E", "S0101_C01_017E", "S0101_C01_018E", "S0101_C01_019E", "state", "county"],
["Shelby County, Alabama", "11940", "8412", "6232", "3686", "3025", "01", "117"],
["Talladega County, Alabama", "4531", "4226", "2482", "1910", "1273", "01", "121"],
["Tuscaloosa County, Alabama", "9310", "7501", "5411", "2713", "2537", "01", "125"],
["Pinal County, Arizona", "27844", "27162", "18482", "11173", "6164", "04", "021"],
["Mendocino County, California", "6576", "5744", "2860", "2417", "1475", "06", "045"],
["Orange County, California", "142280", "124254", "80776", "58499", "65417", "06", "059"],
["Sacramento County, California", "70410", "57822", "36056", "25650", "27506", "06", "067"],
["Citrus County, Florida", "16405", "13054", "10321", "7242", "6810", "12", "017"],
["Peoria County, Illinois", "9622", "7828", "4479", "3712", "4560", "17", "143"],
["Hancock County, Indiana", "3932", "4173", "2182", "1487", "1174", "18", "059"],
["Howard County, Indiana", "4473", "4541", "3117", "1423", "2363", "18", "067"],
["Tangipahoa Parish, Louisian

In [3]: # print Column names
r.json()[0]

Out[3]: ['NAME',
'S0101_C01_015E',
'S0101_C01_016E',
'S0101_C01_017E',
'S0101_C01_018E',
'S0101_C01_019E',
'state',
'county']

In [4]: #rename columns
col_names = ["County", "AGE65_69", "AGE70_74", "AGE75_79", "AGE80_84", "AGE85_over", "State_Code", "County_Code"]

In [5]: #Load into Pandas Data Frame
import pandas as pd
df = pd.DataFrame(columns=col_names, data=r.json()[1:])
df.head(10)

Out[5]:
      County  AGE65_69  AGE70_74  AGE75_79  AGE80_84  AGE85_over  State_Code  County_Code
0  Shelby County, Alabama    11940     8412     6232     3686     3025         01         117
1  Talladega County, Alabama    4531     4226     2482     1910     1273         01         121
2  Tuscaloosa County, Alabama    9310     7501     5411     2713     2537         01         125
3    Pinal County, Arizona   27844    27162    18482    11173     6164         04         021
4  Mendocino County, California    6576     5744     2860     2417     1475         06         045
5   Orange County, California  142280   124254    80776    58499    65417         06         059
6  Sacramento County, California   70410    57822    36056    25650    27506         06         067
7    Citrus County, Florida   16405    13054    10321     7242     6810         12         017
8   Peoria County, Illinois    9622     7828     4479     3712     4560         17         143
9   Hancock County, Indiana    3932     4173     2182     1487     1174         18         059

In [6]: df.tail(10)

Out[6]:
      County  AGE65_69  AGE70_74  AGE75_79  AGE80_84  AGE85_over  State_Code  County_Code
828  Hanover County, Virginia    6235     4750     3587     2018     2297         51         085
829  Kitsap County, Washington   16508    13750     9399     4866     3225         53         035
830  Jefferson County, Wisconsin    5089     3763     2754     1758     1281         55         055
831  Outagamie County, Wisconsin    9135     6682     4679     3983     3097         55         087
832   DeKalb County, Alabama     3969     3408     1907     1435     1542         01         049
833    Butte County, California   14949    10320     8903     4207     4441         06         007
834  El Dorado County, California   14742    10686     6861     3685     4453         06         017
835  Imperial County, California    8092     5267     4116     3392     2645         06         025
836   Kern County, California   34857    24449    15794    12300    10676         06         029
837   Marin County, California   18825    14871    11099     7072     6346         06         041

In [7]: #Check the data type
df.dtypes

Out[7]: County      object
AGE65_69      object
AGE70_74      object
AGE75_79      object
AGE80_84      object
AGE85_over    object
State_Code    object
County_Code   object
dtype: object

In [8]: #Count duplicates
dups_County = df.pivot_table(index=['County'], aggfunc='size')
print (dups_County)

County
Ada County, Idaho      1
Adams County, Colorado  1
Adams County, Illinois  1
Adams County, Pennsylvania  1
Aiken County, South Carolina  1
.
York County, Pennsylvania  1
York County, South Carolina  1
York County, Virginia  1
Yuba County, California  1
Yuma County, Arizona  1
Length: 838, dtype: int64

In [9]: # Converting object to Int for age variables
df["AGE65_69"] = df["AGE65_69"].astype(int)
df["AGE70_74"] = df["AGE70_74"].astype(int)
df["AGE75_79"] = df["AGE75_79"].astype(int)
df["AGE80_84"] = df["AGE80_84"].astype(int)
df["AGE85_over"] =df["AGE85_over"].astype(int)

In [10]: #Check the data type
df.dtypes

Out[10]: County      object
AGE65_69      int32
AGE70_74      int32
AGE75_79      int32
AGE80_84      int32
AGE85_over    int32
State_Code    object
County_Code   object
dtype: object

In [11]: #remove whitespace from the beginning and end
df.columns = [x.strip() for x in df.columns]

In [12]: ##Checking for missing values
df.isnull().sum()

Out[12]: County      0
AGE65_69      0
AGE70_74      0
AGE75_79      0
AGE80_84      0
AGE85_over    0
State_Code    0
County_Code    0
dtype: int64

In [13]: #Summary statistic for umeric data
df.describe()

Out[13]:
      AGE65_69  AGE70_74  AGE75_79  AGE80_84  AGE85_over
count  838.000000    838.000000    838.000000    838.000000    838.000000
mean   16883.212411  13261.977327   9188.332936   6057.905728   6310.809069
std    26688.903925  20805.939437  14568.057484  9959.382588  11087.514197
min    2055.000000   1086.000000   1035.000000    436.000000    497.000000
25%    5335.750000   4225.250000   3010.750000   1875.250000   1760.250000
50%    8715.500000   6711.500000   4551.000000   3037.500000   3003.000000
75%   17712.000000  14219.750000   9596.250000   6400.000000   6469.000000
max   450513.000000  334698.000000  232662.000000  165338.000000  192748.000000
```