Notebook2- Cleaning

March 24, 2020

[1]:
```
from IPython.core.display import display, HTML
display(HTML("<style>.container { width:100% !important; height:100%}</style>"))
```

```
<IPython.core.display.HTML object>
```

1 Cleaning

Based on my approach to build Base Machine learning Model.
I decided to focus on 4 aspects before collecting data and building the base model.

1. Number of people living in every county across United States.
2. Total number of employed/unemployed i.e. civilians labor force in each county. This feature needs to be normalized by dividing it with the total population of the county.
3. Total number of convenient store in each county, again needs to be normalized.
4. Considering the demographics data such as income, poverty percentage, and unemployment rate along side percentage of each race in each of the counties.

[1]:
```
import numpy as np
import pandas as pd
import re
```

[2]:
```
county = pd.read_csv('Supplemental Data - County-Table 1.csv')
print(county.shape)
county.head()
```

```
(3143, 10)
```

[2]:
```
     FIPS     State    County 2010 Census Population   \
0  1001.0   Alabama   Autauga                  54,571
1  1003.0   Alabama   Baldwin                 182,265
2  1005.0   Alabama   Barbour                  27,457
3  1007.0   Alabama      Bibb                  22,915
4  1009.0   Alabama    Blount                  57,322
```

```
  Population Estimate, 2011 Population Estimate, 2012  \
0                    55,255                    55,027
1                   186,653                   190,403
2                    27,326                    27,132
3                    22,736                    22,645
4                    57,707                    57,772

  Population Estimate, 2013 Population Estimate, 2014  \
0                    54,792                    54,977
1                   195,147                   199,745
2                    26,938                    26,763
3                    22,501                    22,511
4                    57,746                    57,621

  Population Estimate, 2015 Population Estimate, 2016
0                    55,035                    55,416
1                   203,690                   208,563
2                    26,270                    25,965
3                    22,561                    22,643
4                    57,676                    57,704
```

[3]:
```python
county = county.dropna()
counties = []
states = []
fips = []
for c in county["County"]:
    counties.append(re.sub(' ', '', c))
for s in county["State"]:
    states.append(re.sub(' ', '', s))
for f in county["FIPS "]:
    fips.append(int(f))
county["County"] = counties
county["State"] = states
county["FIPS "] = fips
county.head()
```

[3]:
```
   FIPS     State   County 2010 Census Population Population Estimate, 2011  \
0  1001  Alabama  Autauga                   54,571                    55,255
1  1003  Alabama  Baldwin                  182,265                   186,653
2  1005  Alabama  Barbour                   27,457                    27,326
3  1007  Alabama     Bibb                   22,915                    22,736
4  1009  Alabama   Blount                   57,322                    57,707

  Population Estimate, 2012 Population Estimate, 2013  \
0                    55,027                    54,792
1                   190,403                   195,147
2                    27,132                    26,938
```

```
3                22,645                22,501
4                57,772                57,746

  Population Estimate, 2014 Population Estimate, 2015  \
0                    54,977                    55,035
1                   199,745                   203,690
2                    26,763                    26,270
3                    22,511                    22,561
4                    57,621                    57,676

  Population Estimate, 2016
0                    55,416
1                   208,563
2                    25,965
3                    22,643
4                    57,704
```

[5]:
```python
demographics = pd.read_csv('acs2015_county_data.csv', encoding='latin-1')
print(demographics.shape)
demographics.head()
```

```
(3220, 37)
```

[5]:
```
   CensusId    State   County  TotalPop    Men  Women  Hispanic  White  Black  \
0      1001  Alabama  Autauga     55221  26745  28476       2.6   75.8   18.5
1      1003  Alabama  Baldwin    195121  95314  99807       4.5   83.1    9.5
2      1005  Alabama  Barbour     26932  14497  12435       4.6   46.2   46.7
3      1007  Alabama     Bibb     22604  12073  10531       2.2   74.5   21.4
4      1009  Alabama   Blount     57710  28512  29198       8.6   87.9    1.5

   Native  ...  Walk  OtherTransp  WorkAtHome  MeanCommute  Employed  \
0     0.4  ...   0.5          1.3         1.8         26.5     23986
1     0.6  ...   1.0          1.4         3.9         26.4     85953
2     0.2  ...   1.8          1.5         1.6         24.1      8597
3     0.4  ...   0.6          1.5         0.7         28.8      8294
4     0.3  ...   0.9          0.4         2.3         34.9     22189

   PrivateWork  PublicWork  SelfEmployed  FamilyWork  Unemployment
0         73.6        20.9           5.5         0.0           7.6
1         81.5        12.3           5.8         0.4           7.5
2         71.8        20.8           7.3         0.1          17.6
3         76.8        16.1           6.7         0.4           8.3
4         82.0        13.5           4.2         0.4           7.7

[5 rows x 37 columns]
```

[6]:
```python
income = pd.read_csv('Unemployment Med HH Inc-Table 1.csv')
print(income.shape)
```

```python
income.head()
```

```
(3275, 52)
```

[6]:

```
  FIPStxt State           Area_name  Rural_urban_continuum_code_2013  \
0       0    US       United States                              NaN
1    1000    AL             Alabama                              NaN
2    1001    AL  Autauga County, AL                              2.0
3    1003    AL  Baldwin County, AL                              3.0
4    1005    AL  Barbour County, AL                              6.0

  Urban_influence_code_2013  Metro_2013 Civilian_labor_force_2007  \
0                      NaN         NaN               152,191,093
1                      NaN         NaN                 2,175,612
2                      2.0         1.0                    24,383
3                      2.0         1.0                    82,659
4                      6.0         0.0                    10,334

 Employed_2007 Unemployed_2007  Unemployment_rate_2007  ...  \
0   145,156,134       7,034,959                     4.6  ...
1     2,089,127          86,485                     4.0  ...
2        23,577             806                     3.3  ...
3        80,099           2,560                     3.1  ...
4         9,684             650                     6.3  ...

 Civilian_labor_force_2016 Employed_2016 Unemployed_2016  \
0              158,921,892   151,183,680       7,738,212
1                2,173,175     2,045,624         127,551
2                   25,918        24,593           1,325
3                   90,500        85,656           4,844
4                    8,402         7,700             702

  Unemployment_rate_2016 Civilian_labor_force_2017 Employed_2017  \
0                    4.9               160,588,515   153,594,100
1                    5.9                   2168444       2073106
2                    5.1                     25909         24908
3                    5.4                     91567         87915
4                    8.4                      8236          7750

 Unemployed_2017  Unemployment_rate_2017 Median_Household_Income_2016  \
0      6,994,415                     4.4                   $57,617.00
1          95338                     4.4                   $46,309.00
2           1001                     3.9                   $54,487.00
3           3652                     4.0                   $56,460.00
4            486                     5.9                   $32,884.00

 Med_HH_Income_Percent_of_State_Total_2016
```

```
0                                     NaN
1                                   100.0
2                                   117.7
3                                   121.9
4                                    71.0

[5 rows x 52 columns]
```

[7]:

```python
income = income.dropna()
print(income.shape)
counties = []
for c in income["Area_name"]:
    counties.append(re.sub(" County, ..", "", c))
income["Area_name"] = counties
income.head()
```

```
(3136, 52)
```

[7]:

```
   FIPStxt State Area_name  Rural_urban_continuum_code_2013  \
2     1001    AL   Autauga                              2.0
3     1003    AL   Baldwin                              3.0
4     1005    AL   Barbour                              6.0
5     1007    AL      Bibb                              1.0
6     1009    AL    Blount                              1.0

   Urban_influence_code_2013  Metro_2013 Civilian_labor_force_2007  \
2                        2.0         1.0                    24,383
3                        2.0         1.0                    82,659
4                        6.0         0.0                    10,334
5                        1.0         1.0                     8,791
6                        1.0         1.0                    26,629

  Employed_2007 Unemployed_2007  Unemployment_rate_2007  ...  \
2        23,577             806                     3.3  ...
3        80,099           2,560                     3.1  ...
4         9,684             650                     6.3  ...
5         8,432             359                     4.1  ...
6        25,780             849                     3.2  ...

  Civilian_labor_force_2016 Employed_2016 Unemployed_2016  \
2                    25,918        24,593           1,325
3                    90,500        85,656           4,844
4                     8,402         7,700             702
5                     8,607         8,050             557
6                    24,576        23,248           1,328

   Unemployment_rate_2016 Civilian_labor_force_2017 Employed_2017  \
2                     5.1                     25909         24908
```

```
3                    5.4                      91567         87915
4                    8.4                       8236          7750
5                    6.5                       8506          8133
6                    5.4                      24494         23509

  Unemployed_2017  Unemployment_rate_2017 Median_Household_Income_2016  \
2            1001                     3.9                   $54,487.00
3            3652                     4.0                   $56,460.00
4             486                     5.9                   $32,884.00
5             373                     4.4                   $43,079.00
6             985                     4.0                   $47,213.00

  Med_HH_Income_Percent_of_State_Total_2016
2                                     117.7
3                                     121.9
4                                      71.0
5                                      93.0
6                                     102.0

[5 rows x 52 columns]
```

[8]:
```
income.columns
```

[8]:
```
Index(['FIPStxt', 'State', 'Area_name', 'Rural_urban_continuum_code_2013',
       'Urban_influence_code_2013', 'Metro_2013', 'Civilian_labor_force_2007',
       'Employed_2007', 'Unemployed_2007', 'Unemployment_rate_2007',
       'Civilian_labor_force_2008', 'Employed_2008', 'Unemployed_2008',
       'Unemployment_rate_2008', 'Civilian_labor_force_2009', 'Employed_2009',
       'Unemployed_2009', 'Unemployment_rate_2009',
       'Civilian_labor_force_2010', 'Employed_2010', 'Unemployed_2010',
       'Unemployment_rate_2010', 'Civilian_labor_force_2011', 'Employed_2011',
       'Unemployed_2011', 'Unemployment_rate_2011',
       'Civilian_labor_force_2012', 'Employed_2012', 'Unemployed_2012',
       'Unemployment_rate_2012', 'Civilian_labor_force_2013', 'Employed_2013',
       'Unemployed_2013', 'Unemployment_rate_2013',
       'Civilian_labor_force_2014', 'Employed_2014', 'Unemployed_2014',
       'Unemployment_rate_2014', 'Civilian_labor_force_2015', 'Employed_2015',
       'Unemployed_2015', 'Unemployment_rate_2015',
       'Civilian_labor_force_2016', 'Employed_2016', 'Unemployed_2016',
       'Unemployment_rate_2016', 'Civilian_labor_force_2017', 'Employed_2017',
       'Unemployed_2017', 'Unemployment_rate_2017',
       'Median_Household_Income_2016',
       'Med_HH_Income_Percent_of_State_Total_2016'],
      dtype='object')
```

[9]:
```
income_cleaned = income.loc[:,['FIPStxt', 'State', 'Area_name',
                               'Civilian_labor_force_2015']]
income_cleaned.head()
```

[9]:
```
  FIPStxt State Area_name Civilian_labor_force_2015
2    1001    AL   Autauga                    25,602
3    1003    AL   Baldwin                    87,705
4    1005    AL   Barbour                     8,609
5    1007    AL      Bibb                     8,572
6    1009    AL    Blount                    24,473
```

[10]:
```
county.columns
```

[10]:
```
Index(['FIPS ', 'State', 'County', '2010 Census Population',
       'Population Estimate, 2011', 'Population Estimate, 2012',
       'Population Estimate, 2013', 'Population Estimate, 2014',
       'Population Estimate, 2015', 'Population Estimate, 2016'],
      dtype='object')
```

[11]:
```
county_cleaned = county.loc[:, ['FIPS ', 'State', 'County', 'Population␣
 ↪Estimate, 2015']]
county_cleaned.head()
```

[11]:
```
  FIPS     State   County Population Estimate, 2015
0  1001  Alabama  Autauga                    55,035
1  1003  Alabama  Baldwin                   203,690
2  1005  Alabama  Barbour                    26,270
3  1007  Alabama     Bibb                    22,561
4  1009  Alabama   Blount                    57,676
```

[12]:
```
demographics.columns
```

[12]:
```
Index(['CensusId', 'State', 'County', 'TotalPop', 'Men', 'Women', 'Hispanic',
       'White', 'Black', 'Native', 'Asian', 'Pacific', 'Citizen', 'Income',
       'IncomeErr', 'IncomePerCap', 'IncomePerCapErr', 'Poverty',
       'ChildPoverty', 'Professional', 'Service', 'Office', 'Construction',
       'Production', 'Drive', 'Carpool', 'Transit', 'Walk', 'OtherTransp',
       'WorkAtHome', 'MeanCommute', 'Employed', 'PrivateWork', 'PublicWork',
       'SelfEmployed', 'FamilyWork', 'Unemployment'],
      dtype='object')
```

[13]:
```
demographics_cleaned = demographics.loc[:, ['State', 'County','Hispanic',
       'White', 'Black', 'Native', 'Asian', 'Pacific', 'Income', 'Poverty',␣
 ↪'Unemployment']]
demographics_cleaned.head()
```

[13]:
```
     State   County  Hispanic  White  Black  Native  Asian  Pacific   Income  \
0  Alabama  Autauga       2.6   75.8   18.5     0.4    1.0      0.0  51281.0
1  Alabama  Baldwin       4.5   83.1    9.5     0.6    0.7      0.0  50254.0
2  Alabama  Barbour       4.6   46.2   46.7     0.2    0.4      0.0  32964.0
3  Alabama     Bibb       2.2   74.5   21.4     0.4    0.1      0.0  38678.0
4  Alabama   Blount       8.6   87.9    1.5     0.3    0.1      0.0  45813.0

  Poverty  Unemployment
0    12.9           7.6
```

```
1     13.4           7.5
2     26.7          17.6
3     16.8           8.3
4     16.7           7.7
```

[14]:
```python
temp = pd.merge(county_cleaned, demographics_cleaned, how = 'inner',␣
↪left_on=['State', 'County'], right_on=['State', 'County'])
temp.head()
```

[14]:
```
   FIPS     State    County Population Estimate, 2015  Hispanic  White  Black  \
0  1001  Alabama  Autauga                      55,035       2.6   75.8   18.5
1  1003  Alabama  Baldwin                     203,690       4.5   83.1    9.5
2  1005  Alabama  Barbour                      26,270       4.6   46.2   46.7
3  1007  Alabama     Bibb                      22,561       2.2   74.5   21.4
4  1009  Alabama   Blount                      57,676       8.6   87.9    1.5

   Native  Asian  Pacific   Income  Poverty  Unemployment
0     0.4    1.0      0.0  51281.0     12.9           7.6
1     0.6    0.7      0.0  50254.0     13.4           7.5
2     0.2    0.4      0.0  32964.0     26.7          17.6
3     0.4    0.1      0.0  38678.0     16.8           8.3
4     0.3    0.1      0.0  45813.0     16.7           7.7
```

[15]:
```python
temp2 = pd.merge(temp, income_cleaned, how='inner', left_on='FIPS ',␣
↪right_on='FIPStxt')

temp2 = temp2.drop(['State_y', 'FIPStxt', 'Area_name'], axis=1).
↪rename(columns={"State_x": "State",
    "Population Estimate, 2015":"Population",
    "Civilian_labor_force_2015":"Civilian Labor Force",
    "FIPS ":"FIPS"})

print(temp2.shape)
temp2.head()
```

```
(2457, 14)
```

[15]:
```
   FIPS    State    County Population  Hispanic  White  Black  Native  Asian  \
0  1001  Alabama  Autauga      55,035       2.6   75.8   18.5     0.4    1.0
1  1003  Alabama  Baldwin     203,690       4.5   83.1    9.5     0.6    0.7
2  1005  Alabama  Barbour      26,270       4.6   46.2   46.7     0.2    0.4
3  1007  Alabama     Bibb      22,561       2.2   74.5   21.4     0.4    0.1
4  1009  Alabama   Blount      57,676       8.6   87.9    1.5     0.3    0.1

   Pacific   Income  Poverty  Unemployment Civilian Labor Force
0      0.0  51281.0     12.9           7.6               25,602
1      0.0  50254.0     13.4           7.5               87,705
2      0.0  32964.0     26.7          17.6                8,609
3      0.0  38678.0     16.8           8.3                8,572
```

```
4      0.0  45813.0     16.7            7.7                 24,473
```

[16]:
```
grocery = pd.read_csv("STORES-Table 1.csv")
print(grocery.columns)
grocery.head()
```

```
Index(['FIPS', 'State', 'County', 'GROC09', 'GROC14', 'PCH_GROC_09_14',
       'GROCPTH09', 'GROCPTH14', 'PCH_GROCPTH_09_14', 'SUPERC09', 'SUPERC14',
       'PCH_SUPERC_09_14', 'SUPERCPTH09', 'SUPERCPTH14', 'PCH_SUPERCPTH_09_14',
       'CONVS09', 'CONVS14', 'PCH_CONVS_09_14', 'CONVSPTH09', 'CONVSPTH14',
       'PCH_CONVSPTH_09_14', 'SPECS09', 'SPECS14', 'PCH_SPECS_09_14',
       'SPECSPTH09', 'SPECSPTH14', 'PCH_SPECSPTH_09_14', 'SNAPS12', 'SNAPS16',
       'PCH_SNAPS_12_16', 'SNAPSPTH12', 'SNAPSPTH16', 'PCH_SNAPSPTH_12_16',
       'WICS08', 'WICS12', 'PCH_WICS_08_12', 'WICSPTH08', 'WICSPTH12',
       'PCH_WICSPTH_08_12'],
      dtype='object')
```

[16]:
```
   FIPS State   County  GROC09  GROC14  PCH_GROC_09_14  GROCPTH09  GROCPTH14  \
0  1001    AL  Autauga       6       4      -33.333333   0.110834   0.072209
1  1003    AL  Baldwin      24      29       20.833333   0.133775   0.144920
2  1005    AL  Barbour       5       5        0.000000   0.180786   0.185963
3  1007    AL     Bibb       6       5      -16.666667   0.261540   0.222163
4  1009    AL   Blount       6       6        0.000000   0.104637   0.103952

   PCH_GROCPTH_09_14  SUPERC09  ...  PCH_SNAPS_12_16  SNAPSPTH12  SNAPSPTH16  \
0         -34.849716         1  ...        12.694878    0.674004    0.760911
1           8.331001         6  ...        43.192771    0.725055    0.949753
2           2.863838         0  ...         0.956938    1.280590    1.354387
3         -15.055985         1  ...        20.512821    0.719122    0.864874
4          -0.654897         1  ...        23.903509    0.657144    0.815946

   PCH_SNAPSPTH_12_16  WICS08  WICS12  PCH_WICS_08_12  WICSPTH08  WICSPTH12  \
0           12.894172       6       5       -16.66667   0.119156   0.090067
1           30.990390      25      27         8.00000   0.141875   0.141517
2            5.762745       6       7        16.66667   0.201099   0.257344
3           20.267995       6       5       -16.66667   0.277919   0.221268
4           24.165470      10       6       -40.00000   0.173028   0.103760

   PCH_WICSPTH_08_12
0         -24.412460
1          -0.252126
2          27.968330
3         -20.383970
4         -40.033200

[5 rows x 39 columns]
```

[17]:
```
grocery_cleaned = grocery.loc[:, ['FIPS', 'GROC14', 'SUPERC14', 'CONVS14',␣
↪'SPECS14']].rename(
    columns={"GROC14":"Grocery Stores",
   "CONVS14":"Convenience Stores", "SUPERC14":"Supercenters", "SPECS14":
↪"Specialty Stores"})
grocery_cleaned['Total Stores'] = grocery_cleaned["Grocery Stores"] \
    + grocery_cleaned["Convenience Stores"] \
    + grocery_cleaned["Specialty Stores"] \
    + grocery_cleaned["Supercenters"]
grocery_cleaned.head()
```

[17]:
```
   FIPS  Grocery Stores  Supercenters  Convenience Stores  Specialty Stores  \
0  1001               4             1                  30                 2
1  1003              29             6                 118                26
2  1005               5             1                  19                 2
3  1007               5             1                  15                 1
4  1009               6             1                  27                 0

   Total Stores
0            37
1           179
2            27
3            22
4            34
```

[18]:
```
final = pd.merge(temp2, grocery_cleaned, how='inner', left_on='FIPS',␣
↪right_on='FIPS').set_index("FIPS")
pops = []
labor = []
for p in final['Population']:
    pops.append(int(re.sub(',', '', p)))
final['Population'] = pops
for l in final['Civilian Labor Force']:
    labor.append(int(re.sub(',', '', l)))
final['Civilian Labor Force'] = labor
print(final.shape)
final.head()
```

```
(2457, 18)
```

[18]:
```
        State   County  Population  Hispanic  White  Black  Native  Asian  \
FIPS
1001  Alabama  Autauga       55035       2.6   75.8   18.5     0.4    1.0
1003  Alabama  Baldwin      203690       4.5   83.1    9.5     0.6    0.7
1005  Alabama  Barbour       26270       4.6   46.2   46.7     0.2    0.4
1007  Alabama     Bibb       22561       2.2   74.5   21.4     0.4    0.1
1009  Alabama   Blount       57676       8.6   87.9    1.5     0.3    0.1
```

```
      Pacific   Income  Poverty  Unemployment  Civilian Labor Force  \
FIPS
1001      0.0  51281.0     12.9           7.6                 25602
1003      0.0  50254.0     13.4           7.5                 87705
1005      0.0  32964.0     26.7          17.6                  8609
1007      0.0  38678.0     16.8           8.3                  8572
1009      0.0  45813.0     16.7           7.7                 24473

      Grocery Stores  Supercenters  Convenience Stores  Specialty Stores  \
FIPS
1001               4             1                  30                 2
1003              29             6                 118                26
1005               5             1                  19                 2
1007               5             1                  15                 1
1009               6             1                  27                 0

      Total Stores
FIPS
1001            37
1003           179
1005            27
1007            22
1009            34
```

[19]:
```
final.to_csv('cleaned_counties.csv')
```