AbhijitMandal_DSC540_Milestone4

May 22, 2021

0.0.1 DSC 540 Week 9-10 - Milestone 4

Abhijit Mandal

0.0.2 Milestone 4

Perform at least 5 data transformation and/or cleansing steps to your API data. For example:

- Replace Headers
- Format data into a more readable format
- Identify outliers and bad data
- Find duplicates
- Fix casing or inconsistent values
- Conduct Fuzzy Matching

0.0.3 Dataset

API - I will be fetchng data from https://api.census.gov/data/2019/acs/acs1/profile?get=NAME,gr which will provide the demographic info (Age, employment, sex, ethnicity etc) for US at a County level for all the corona virus cases, this information is vital to understand the rate of spread across communities in United States. I will use this data to deep dive into corona cases in US and generate some interesting facts.

0.0.4 Load the necessary libraries.

[60]:
```
#import libraries
import urllib.request, urllib.parse, urllib.error
import json
import requests
import numpy as np
import pandas as pd
#pandasql package allows us to write SQL querry on Pandas DataFrame
import pandasql as psql
import seaborn as sns
import matplotlib.pyplot as plt
```

0.0.5 2. Reading the API data

[61]:
```
apiURl = "https://api.census.gov/data/2019/acs/acs1/profile?
 ↪get=NAME,group(DP02)&for=county:*"

filename = "acs2019_county_data.csv"
chunk_size = 100

response = requests.get(apiURl)

# calling this API and saving it as CSV
with open(filename, 'wb') as fd:
    for chunk in response.iter_content(chunk_size):
        fd.write(chunk)
```

[62]:
```
county_2019 = pd.read_csv('acs2019_county_data.csv', encoding='latin-1')
county_2019.head()
```

[62]:

	CountyId	State	County	TotalPop	Men	Women	Hispanic	\
0	1001	Alabama	Autauga County	55036	26899	28137	2.7	
1	1003	Alabama	Baldwin County	203360	99527	103833	4.4	
2	1005	Alabama	Barbour County	26201	13976	12225	4.2	
3	1007	Alabama	Bibb County	22580	12251	10329	2.4	
4	1009	Alabama	Blount County	57667	28490	29177	9.0	

	White	Black	Native	…	Walk	OtherTransp	WorkAtHome	MeanCommute	\
0	75.4	18.9	0.3	…	0.6	1.3	2.5	25.8	
1	83.1	9.5	0.8	…	0.8	1.1	5.6	27.0	
2	45.7	47.8	0.2	…	2.2	1.7	1.3	23.4	
3	74.6	22.0	0.4	…	0.3	1.7	1.5	30.0	
4	87.4	1.5	0.3	…	0.4	0.4	2.1	35.0	

	Employed	PrivateWork	PublicWork	SelfEmployed	FamilyWork	Unemployment
0	24112	74.1	20.2	5.6	0.1	5.2
1	89527	80.7	12.9	6.3	0.1	5.5
2	8878	74.1	19.1	6.5	0.3	12.4
3	8171	76.0	17.4	6.3	0.3	8.2
4	21380	83.9	11.9	4.0	0.1	4.9

[5 rows x 37 columns]

[63]:
```
county_2019.describe(include = 'all')
```

[63]:

	CountyId	State	County	TotalPop	Men	\
count	3220.000000	3220	3220	3.220000e+03	3.220000e+03	
unique	NaN	52	1955	NaN	NaN	

top	NaN	Texas	Washington County	NaN	NaN	
freq	NaN	254	30	NaN	NaN	
mean	31393.605280	NaN	NaN	1.007681e+05	4.958781e+04	
std	16292.078954	NaN	NaN	3.244996e+05	1.593212e+05	
min	1001.000000	NaN	NaN	7.400000e+01	3.900000e+01	
25%	19032.500000	NaN	NaN	1.121350e+04	5.645500e+03	
50%	30024.000000	NaN	NaN	2.584750e+04	1.287900e+04	
75%	46105.500000	NaN	NaN	6.660825e+04	3.301725e+04	
max	72153.000000	NaN	NaN	1.010572e+07	4.979641e+06	

	Women	Hispanic	White	Black	Native	…	\
count	3.220000e+03	3220.000000	3220.000000	3220.000000	3220.000000	…	
unique	NaN	NaN	NaN	NaN	NaN	…	
top	NaN	NaN	NaN	NaN	NaN	…	
freq	NaN	NaN	NaN	NaN	NaN	…	
mean	5.118032e+04	11.296584	74.920186	8.681957	1.768416	…	
std	1.652164e+05	19.342522	23.056700	14.333571	7.422946	…	
min	3.500000e+01	0.000000	0.000000	0.000000	0.000000	…	
25%	5.553500e+03	2.100000	63.500000	0.600000	0.100000	…	
50%	1.299350e+04	4.100000	83.600000	2.000000	0.300000	…	
75%	3.359375e+04	10.000000	92.800000	9.500000	0.600000	…	
max	5.126081e+06	100.000000	100.000000	86.900000	90.300000	…	

	Walk	OtherTransp	WorkAtHome	MeanCommute	Employed	\
count	3220.000000	3220.000000	3220.000000	3220.000000	3.220000e+03	
unique	NaN	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	NaN	
mean	3.244472	1.598696	4.736894	23.474534	4.709295e+04	
std	3.891510	1.678232	3.073484	5.687241	1.558159e+05	
min	0.000000	0.000000	0.000000	5.100000	3.900000e+01	
25%	1.400000	0.800000	2.900000	19.600000	4.573000e+03	
50%	2.300000	1.300000	4.100000	23.200000	1.061150e+04	
75%	3.825000	1.900000	5.800000	27.000000	2.874725e+04	
max	59.200000	43.200000	33.000000	45.100000	4.805817e+06	

	PrivateWork	PublicWork	SelfEmployed	FamilyWork	Unemployment
count	3220.000000	3220.000000	3220.000000	3220.000000	3220.000000
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	74.863323	17.086118	7.772733	0.278820	6.665590
std	7.647916	6.390868	3.855454	0.448073	3.772612
min	31.100000	4.400000	0.000000	0.000000	0.000000
25%	71.200000	12.700000	5.200000	0.100000	4.475000
50%	76.100000	15.900000	6.800000	0.200000	6.100000
75%	80.200000	19.900000	9.200000	0.300000	8.000000

```
max        88.800000     64.800000      38.000000      8.000000      40.900000

[11 rows x 37 columns]
```

[64]:
```
# Handling Missing Values and Formatting
# We have one missing value in the child poverty column. We fill this with 0.

#Checking missing Data
null_2019 = psql.sqldf("SELECT State, County, TotalPop, Income, IncomeErr,␣
 ↪Poverty, ChildPoverty\
                        FROM county_2019\
                        WHERE ChildPoverty IS NULL")
null_2019
```

[64]:

	State	County	TotalPop	Income	IncomeErr	Poverty	ChildPoverty
0	Hawaii	Kalawao County	86	61750	11280	12.7	None

[65]:
```
# Fill missing value in ChildPoverty with zero
county_2019.ChildPoverty.fillna(0)
```

[65]:
```
0       20.1
1       16.1
2       44.9
3       26.6
4       25.4
        …
3215    49.4
3216    68.2
3217    67.9
3218    62.1
3219    58.2
Name: ChildPoverty, Length: 3220, dtype: float64
```

[66]:
```
#subsetting dataset toget relevant columns
County2019 = county_2019[['CountyId', 'State', 'County', 'Men',␣
 ↪'Women','White', 'Black','Native','Hispanic', 'Asian','Pacific','TotalPop',␣
 ↪'IncomePerCap', 'Poverty', 'ChildPoverty', 'Employed', 'SelfEmployed',␣
 ↪'Unemployment']]
County2019.head()
```

[66]:

	CountyId	State	County	Men	Women	White	Black	Native	\
0	1001	Alabama	Autauga County	26899	28137	75.4	18.9	0.3	
1	1003	Alabama	Baldwin County	99527	103833	83.1	9.5	0.8	
2	1005	Alabama	Barbour County	13976	12225	45.7	47.8	0.2	
3	1007	Alabama	Bibb County	12251	10329	74.6	22.0	0.4	
4	1009	Alabama	Blount County	28490	29177	87.4	1.5	0.3	

```
   Hispanic  Asian  Pacific  TotalPop  IncomePerCap  Poverty  ChildPoverty  \
0       2.7    0.9      0.0     55036         27824     13.7          20.1
1       4.4    0.7      0.0    203360         29364     11.8          16.1
2       4.2    0.6      0.0     26201         17561     27.2          44.9
3       2.4    0.0      0.0     22580         20911     15.2          26.6
4       9.0    0.1      0.0     57667         22021     15.6          25.4

   Employed  SelfEmployed  Unemployment
0     24112           5.6           5.2
1     89527           6.3           5.5
2      8878           6.5          12.4
3      8171           6.3           8.2
4     21380           4.0           4.9
```

[67]:
```
# Adding Calculated column for Men and Women in percentage
pd.options.mode.chained_assignment = None  # default='warn'
County2019['MenPercentage'] = (County2019.Men / County2019.TotalPop)*100
County2019['WomenPercentage'] = (County2019.Women / County2019.TotalPop)*100

County2019.head()
```

[67]:
```
   CountyId    State          County    Men    Women  White  Black  Native  \
0      1001  Alabama  Autauga County  26899   28137   75.4   18.9     0.3
1      1003  Alabama  Baldwin County  99527  103833   83.1    9.5     0.8
2      1005  Alabama  Barbour County  13976   12225   45.7   47.8     0.2
3      1007  Alabama     Bibb County  12251   10329   74.6   22.0     0.4
4      1009  Alabama   Blount County  28490   29177   87.4    1.5     0.3

   Hispanic  Asian  Pacific  TotalPop  IncomePerCap  Poverty  ChildPoverty  \
0       2.7    0.9      0.0     55036         27824     13.7          20.1
1       4.4    0.7      0.0    203360         29364     11.8          16.1
2       4.2    0.6      0.0     26201         17561     27.2          44.9
3       2.4    0.0      0.0     22580         20911     15.2          26.6
4       9.0    0.1      0.0     57667         22021     15.6          25.4

   Employed  SelfEmployed  Unemployment  MenPercentage  WomenPercentage
0     24112           5.6           5.2      48.875282        51.124718
1     89527           6.3           5.5      48.941286        51.058714
2      8878           6.5          12.4      53.341476        46.658524
3      8171           6.3           8.2      54.255979        45.744021
4     21380           4.0           4.9      49.404339        50.595661
```