KausikChattapadhyay_DSC540_Milestone4

DSC 540 Week 9-10 - Milestone 4

Kausik Chattapadhyay

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

Dataset API - I will be fetchng data from https://api.census.gov/data/2019/acs/acs1/profile?get=NAME,gro 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.

In [3]:
```
# Load the necessary 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
```

In [15]:
```
# Reading the API data
apiURl = "https://api.census.gov/data/2019/acs/acs1/profile?get=NAME,group(DP02)&for=county:*"

filename = "ACS2019_state.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)
```

In [21]:
```
county_2019 = pd.read_csv('ACS2019_state.csv', encoding='latin-1')
county_2019.head()
```

Out[21]:

	CountyId	State	County	TotalPop	Men	Women	Hispanic	White	Black	Native	...	Walk	OtherTransp	WorkAtHome	MeanCommute	Employed	PrivateWork	PublicWork	SelfEmployed	FamilyW
0	1001	Alabama	Autauga County	55036	26899	28137	2.7	75.4	18.9	0.3	...	0.6	1.3	2.5	25.8	24112	74.1	20.2	5.6	
1	1003	Alabama	Baldwin County	203360	99527	103833	4.4	83.1	9.5	0.8	...	0.8	1.1	5.6	27.0	89527	80.7	12.9	6.3	
2	1005	Alabama	Barbour County	26201	13976	12225	4.2	45.7	47.8	0.2	...	2.2	1.7	1.3	23.4	8878	74.1	19.1	6.5	
3	1007	Alabama	Bibb County	22580	12251	10329	2.4	74.6	22.0	0.4	...	0.3	1.7	1.5	30.0	8171	76.0	17.4	6.3	
4	1009	Alabama	Blount County	57667	28490	29177	9.0	87.4	1.5	0.3	...	0.4	0.4	2.1	35.0	21380	83.9	11.9	4.0	

5 rows × 37 columns

In [22]:
```
county_2019.describe(include = 'all')
```

Out[22]:

	CountyId	State	County	TotalPop	Men	Women	Hispanic	White	Black	Native	...	Walk	OtherTransp	WorkAtHome	MeanCommute
count	3220.000000	3220	3220	3.220000e+03	3.220000e+03	3.220000e+03	3220.000000	3220.000000	3220.000000	3220.000000	...	3220.000000	3220.000000	3220.000000	3220.000000
unique	NaN	52	1955	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
top	NaN	Texas	Washington County	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
freq	NaN	254	30	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
mean	31393.605280	NaN	NaN	1.007681e+05	4.958781e+04	5.118032e+04	11.296584	74.920186	8.681957	1.768416	...	3.244472	1.598696	4.736894	23.474534
std	16292.078954	NaN	NaN	3.244996e+05	1.593212e+05	1.652164e+05	19.342522	23.056700	14.333571	7.422946	...	3.891510	1.678232	3.073484	5.687241
min	1001.000000	NaN	NaN	7.400000e+01	3.900000e+01	3.500000e+01	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	5.100000
25%	19032.500000	NaN	NaN	1.121350e+04	5.645500e+03	5.553500e+03	2.100000	63.500000	0.600000	0.100000	...	1.400000	0.800000	2.900000	19.600000
50%	30024.000000	NaN	NaN	2.584750e+04	1.287900e+04	1.299350e+04	4.100000	83.600000	2.000000	0.300000	...	2.300000	1.300000	4.100000	23.200000
75%	46105.500000	NaN	NaN	6.660825e+04	3.301725e+04	3.359375e+04	10.000000	92.800000	9.500000	0.600000	...	3.825000	1.900000	5.800000	27.000000
max	72153.000000	NaN	NaN	1.010572e+07	4.979641e+06	5.126081e+06	100.000000	100.000000	86.900000	90.300000	...	59.200000	43.200000	33.000000	45.100000

11 rows × 37 columns

In [23]:
```
# 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
```

Out[23]:

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

In [24]:
```
# Fill missing value in ChildPoverty with zero
county_2019.ChildPoverty.fillna(0)
```

Out[24]:
```
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
```

In [25]:
```
#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()
```

Out[25]:

	CountyId	State	County	Men	Women	White	Black	Native	Hispanic	Asian	Pacific	TotalPop	IncomePerCap	Poverty	ChildPoverty	Employed	SelfEmployed	Unemployment
0	1001	Alabama	Autauga County	26899	28137	75.4	18.9	0.3	2.7	0.9	0.0	55036	27824	13.7	20.1	24112	5.6	5.2
1	1003	Alabama	Baldwin County	99527	103833	83.1	9.5	0.8	4.4	0.7	0.0	203360	29364	11.8	16.1	89527	6.3	5.5
2	1005	Alabama	Barbour County	13976	12225	45.7	47.8	0.2	4.2	0.6	0.0	26201	17561	27.2	44.9	8878	6.5	12.4
3	1007	Alabama	Bibb County	12251	10329	74.6	22.0	0.4	2.4	0.0	0.0	22580	20911	15.2	26.6	8171	6.3	8.2
4	1009	Alabama	Blount County	28490	29177	87.4	1.5	0.3	9.0	0.1	0.0	57667	22021	15.6	25.4	21380	4.0	4.9

In [26]:
```
# 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()
```

Out[26]:

	CountyId	State	County	Men	Women	White	Black	Native	Hispanic	Asian	Pacific	TotalPop	IncomePerCap	Poverty	ChildPoverty	Employed	SelfEmployed	Unemployment	MenPercentage	W
0	1001	Alabama	Autauga County	26899	28137	75.4	18.9	0.3	2.7	0.9	0.0	55036	27824	13.7	20.1	24112	5.6	5.2	48.875282	
1	1003	Alabama	Baldwin County	99527	103833	83.1	9.5	0.8	4.4	0.7	0.0	203360	29364	11.8	16.1	89527	6.3	5.5	48.941286	
2	1005	Alabama	Barbour County	13976	12225	45.7	47.8	0.2	4.2	0.6	0.0	26201	17561	27.2	44.9	8878	6.5	12.4	53.341476	
3	1007	Alabama	Bibb County	12251	10329	74.6	22.0	0.4	2.4	0.0	0.0	22580	20911	15.2	26.6	8171	6.3	8.2	54.255979	
4	1009	Alabama	Blount County	28490	29177	87.4	1.5	0.3	9.0	0.1	0.0	57667	22021	15.6	25.4	21380	4.0	4.9	49.404339	

In [ ]:

Ethical implications of using and transforming the data I am using from APIs/websites and other sources:

Some of the concerns are that as data is broken out of silos and used and transferred, it opens up the risk of hackers getting valuable information. Many consumers are also uneasy about the implications of large corporations gaining access to personal information. This information can be sold to other shady companies and even government organizations who can misuse it. I feel that many of the ethical concerns with APIs boil down to the classic case of a privacy issue.

The age of digital technology is very new and the API economy is even younger. We need more time and research to see whether the use of APIs is ultimately good or bad for mankind. However, we can progress into the future by utilizing the good that APIs bring while setting boundaries to stop the unwanted side effects. A data program would be the first step to balancing APIs and ethics. Customers trust businesses when putting personal data into their hands, and protecting their information will benefit everyone in the equation.

A wide-reaching corporate data program that works together with privacy and data protection laws seems to be the option for the time being. While the state provides regulations, a data program would focus on transparency, allowing consumers to see exactly what data is being collected and stored, and how it is being used.

It would also provide a clear set of terms and conditions of what the business can and can't do with their customer's data. These policies would also amplify the idea of consent; ensuring that customers understand what the business and/or third-party organization intends to do with their data.

Each company's data program should align with its values and overall vision. Moreover, an organization should be well aware of the risks that come with processing large amounts of personal information and have security measures in place to handle breaches and hacking.

A solid data program also creates a culture of trust, transparency, and goodwill which goes a long way when dealing with sensitive things like personal data. As a result, ethical guardrails won't hinder progress; instead, they will help businesses grow. Ensuring privacy and safety to the customers who put their information in your hands will create opportunities, both for yourself and the third-party organizations who deal with them.

In [ ]: