CS 2316 Final Project:

Exploring the Effect of Income & Population on COVID cases in Georgia Counties

By Rakshanda Khan

Topic: Exploring the Effect of Income & Population on COVID cases in Georgia Counties

- Do Income & Population have an effect on COVID cases in Georgia Counties?
    - What kind of an effect do they have? (directly/inversely proportional, weak/strong effect)
    - Which of the two - Income or Population - has a greater effect?

Why did I pick this topic?

- COVID-19 continues to be something that affects the everyday lives of several people all over the world.
- It's a unique event and so I was interested in exploring how such a unique event might be affected by factors such as income & population.
- I heard different experiences & stories from friends in different counties in Georgia; with some of them being more worried about COVID in their county than some others from a different country- so I was curious about getting more objective insights regarding this.

DATA COLLECTION PROCESS

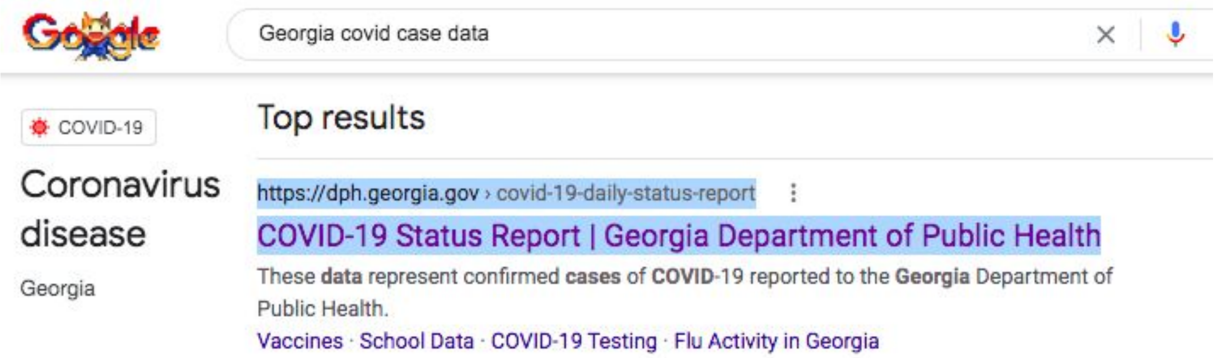
How did I find my datasets?

- My topic of interest had 3 clear variables/factors:

  Income, Population, COVID cases
- Specific to Georgia Counties
- 3 dataset requirements: downloaded, web-scrape, web-based API/JSON

Downloaded Dataset: CSV file from Georgia DPH

- Google Searched “Georgia covid case data”



Georgia covid case data

COVID-19

Coronavirus disease

Georgia

Top results

https://dph.georgia.gov › covid-19-daily-status-report

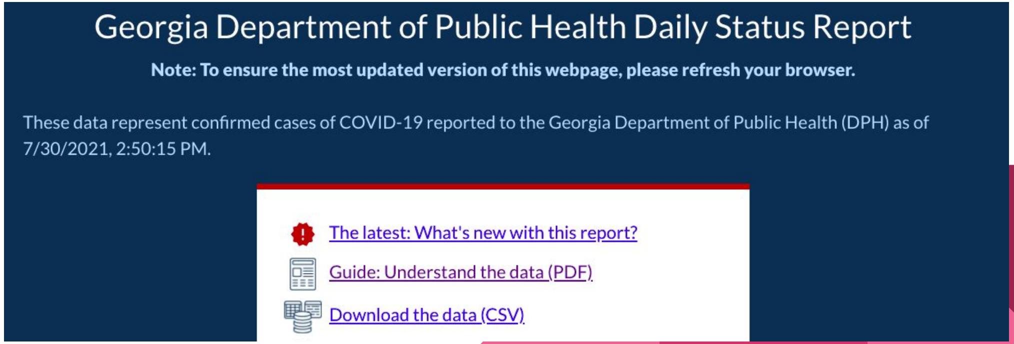
COVID-19 Status Report | Georgia Department of Public Health

These **data** represent confirmed **cases** of **COVID**-19 reported to the **Georgia** Department of Public Health.

Vaccines · School Data · COVID-19 Testing · Flu Activity in Georgia

- Downloaded the CSV Zip file

  > county_cases.csv



Georgia Department of Public Health Daily Status Report

Note: To ensure the most updated version of this webpage, please refresh your browser.

These data represent confirmed cases of COVID-19 reported to the Georgia Department of Public Health (DPH) as of 7/30/2021, 2:50:15 PM.

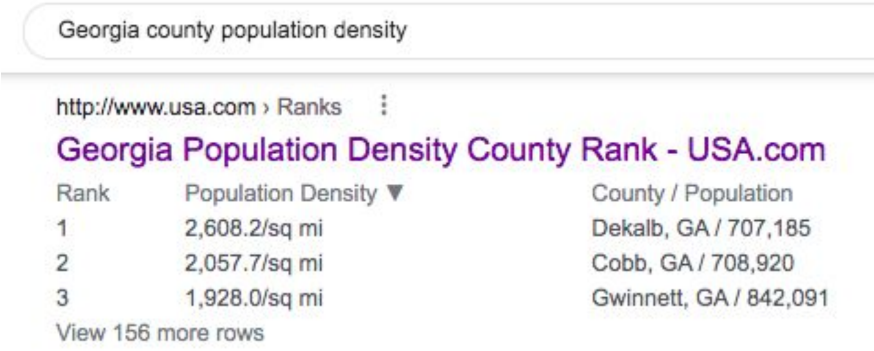
The latest: What's new with this report?

Guide: Understand the data (PDF)

Download the data (CSV)

Web Requirement #1: www.usa.com/rank/georgia-state--population-density--county-rank.htm

- Google Searched “Georgia county population density”



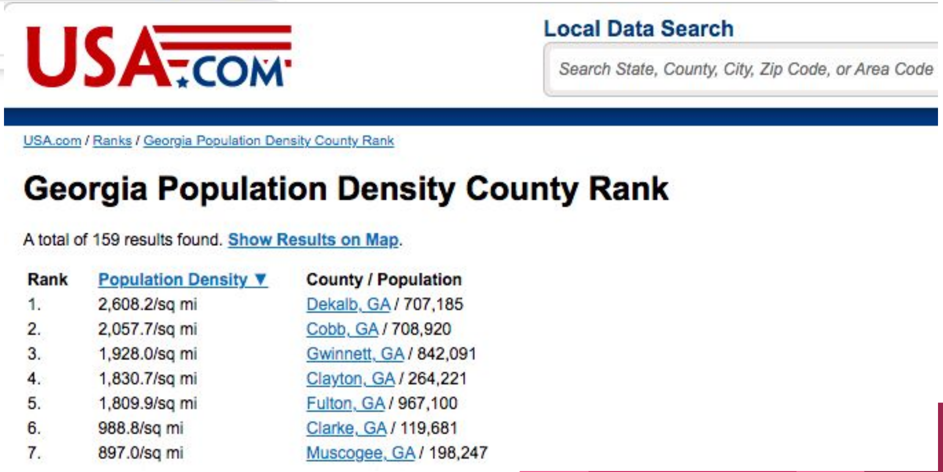
Georgia county population density

http://www.usa.com › Ranks

Georgia Population Density County Rank - USA.com

Rank	Population Density ▼	County / Population
1	2,608.2/sq mi	Dekalb, GA / 707,185
2	2,057.7/sq mi	Cobb, GA / 708,920
3	1,928.0/sq mi	Gwinnett, GA / 842,091

View 156 more rows



USA.com / Ranks / Georgia Population Density County Rank

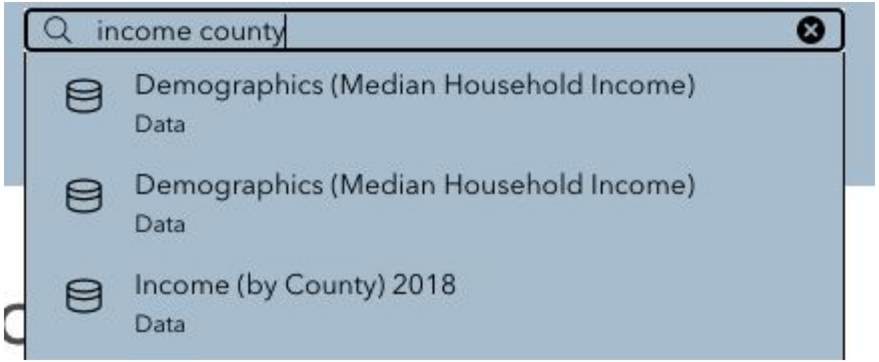
Georgia Population Density County Rank

A total of 159 results found. Show Results on Map.

Rank	Population Density ▼	County / Population
1.	2,608.2/sq mi	Dekalb, GA / 707,185
2.	2,057.7/sq mi	Cobb, GA / 708,920
3.	1,928.0/sq mi	Gwinnett, GA / 842,091
4.	1,830.7/sq mi	Clayton, GA / 264,221
5.	1,809.9/sq mi	Fulton, GA / 967,100
6.	988.8/sq mi	Clarke, GA / 119,681
7.	897.0/sq mi	Muscogee, GA / 198,247

Web Requirement #2: ARC 2018 Income API/JSON

- Already familiar with the ARC Website
- Searched for “income county” on their website
- Initially came across 2017 version > updated to 2018 version after Phase 2
- They had a link ready to access:

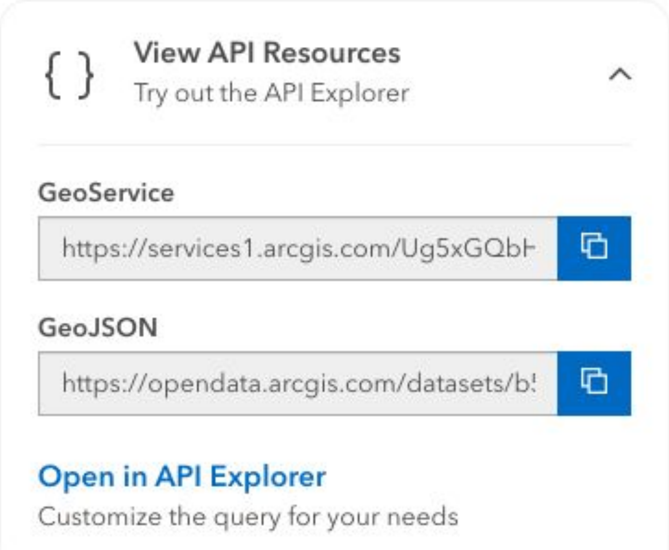


income county

Demographics (Median Household Income)
Data

Demographics (Median Household Income)
Data

Income (by County) 2018
Data



View API Resources
Try out the API Explorer

GeoService
https://services1.arcgis.com/Ug5xGQbH

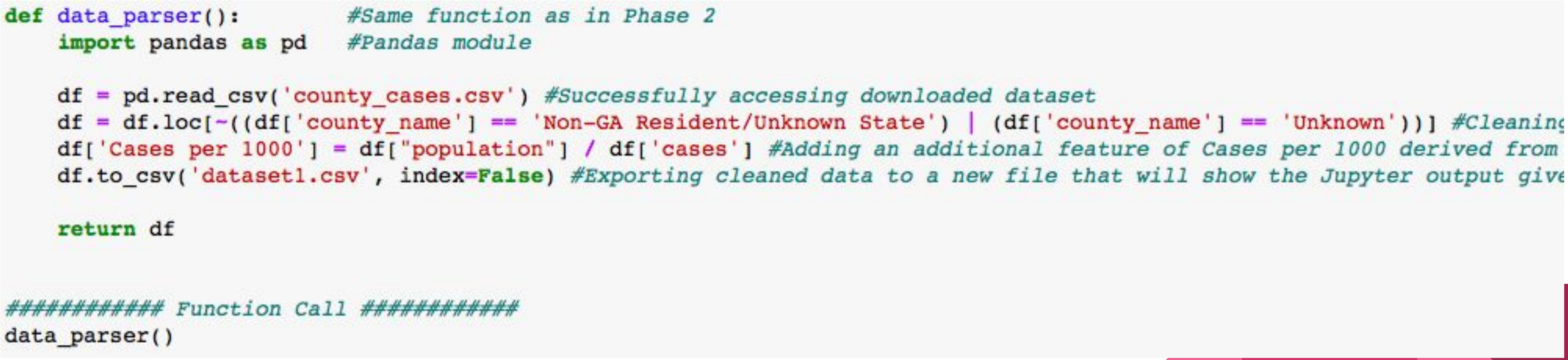
GeoJSON
https://opendata.arcgis.com/datasets/b!

Open in API Explorer
Customize the query for your needs

DATA CLEANING PROCESS

Downloaded dataset:

- Modules: **Pandas**
- Inconsistency: Removing rows that had county_name values of Non-GA Resident /Unknown State using **df.loc** to select other rows
- Added an additional feature of Cases per 1000 into a new column



```
def data_parser():        #Same function as in Phase 2
    import pandas as pd   #Pandas module

    df = pd.read_csv('county_cases.csv') #Successfully accessing downloaded dataset
    df = df.loc[~((df['county_name'] == 'Non-GA Resident/Unknown State') | (df['county_name'] == 'Unknown'))] #Cleaning
    df['Cases per 1000'] = df["population"] / df['cases'] #Adding an additional feature of Cases per 1000 derived from
    df.to_csv('dataset1.csv', index=False) #Exporting cleaned data to a new file that will show the Jupyter output give

    return df


############ Function Call ############
data_parser()
```

Web Requirement #1:

Modules used:

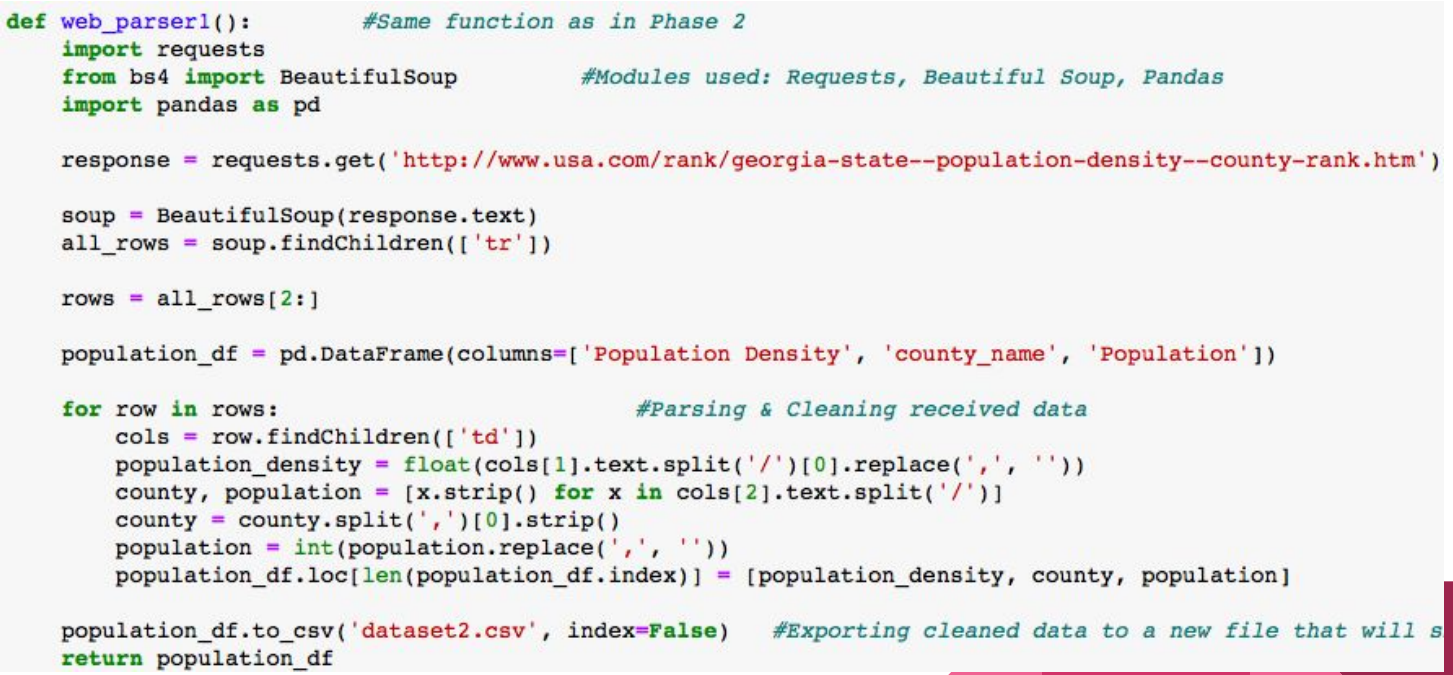
Pandas,

Requests,

BeautifulSoup

soup.findChildren

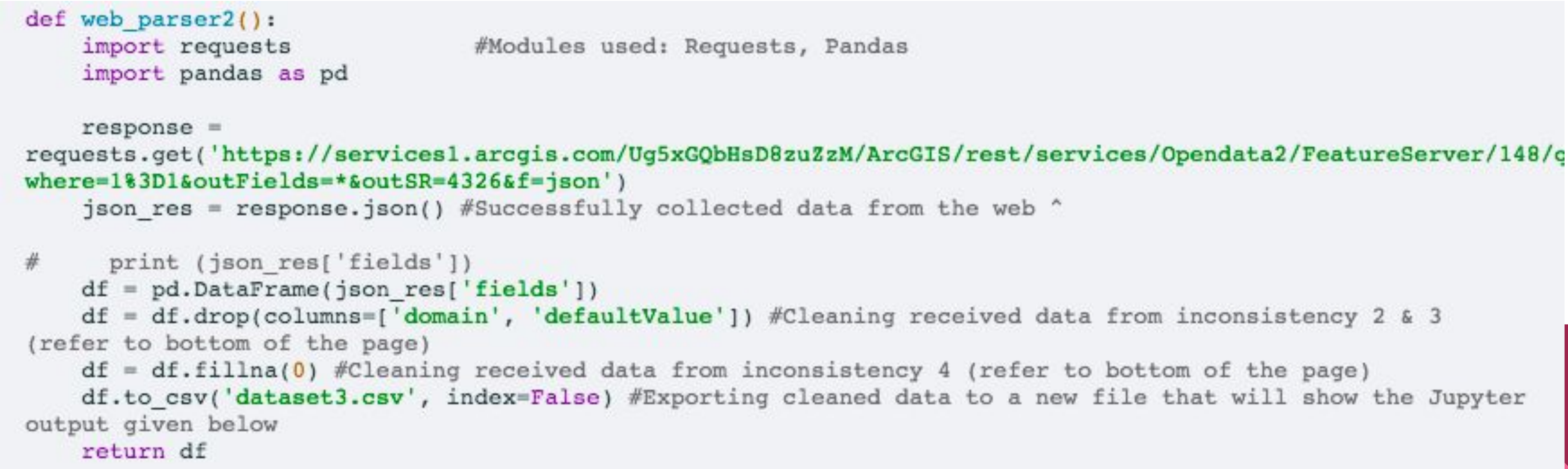
For-loop



```
def web_parser1():          #Same function as in Phase 2
    import requests
    from bs4 import BeautifulSoup         #Modules used: Requests, Beautiful Soup, Pandas
    import pandas as pd

    response = requests.get('http://www.usa.com/rank/georgia-state--population-density--county-rank.htm')

    soup = BeautifulSoup(response.text)
    all_rows = soup.findChildren(['tr'])

    rows = all_rows[2:]

    population_df = pd.DataFrame(columns=['Population Density', 'county_name', 'Population'])

    for row in rows:                               #Parsing & Cleaning received data
        cols = row.findChildren(['td'])
        population_density = float(cols[1].text.split('/')[0].replace(',', ''))
        county, population = [x.strip() for x in cols[2].text.split('/')]
        county = county.split(',')[0].strip()
        population = int(population.replace(',', ''))
        population_df.loc[len(population_df.index)] = [population_density, county, population]

    population_df.to_csv('dataset2.csv', index=False)   #Exporting cleaned data to a new file that will s
    return population_df
```

Web requirement #2 (old during phase 2):

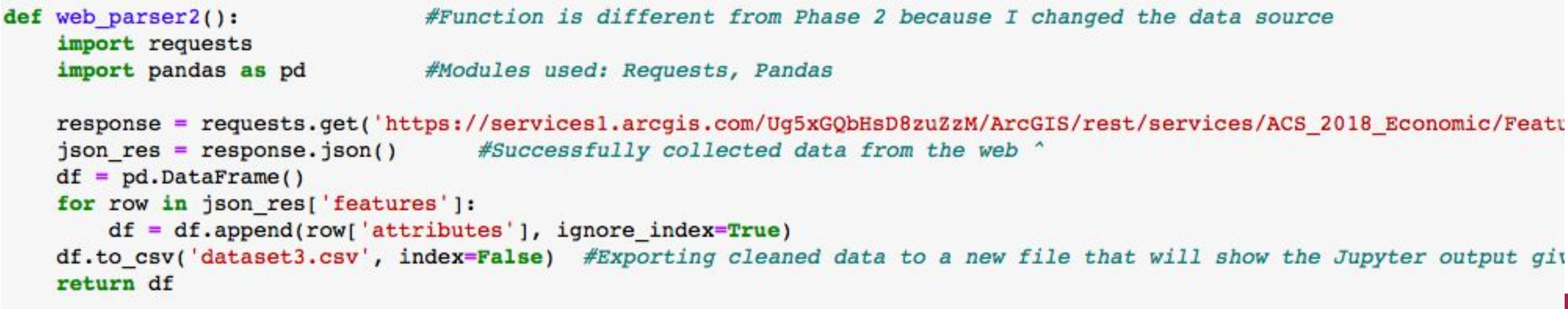
- Modules: **Requests, Pandas**
- Removing inconsistencies using **df.drop** (drop columns with None values) and **df.fillna(0)** (replace NaN values with zero)



```
def web_parser2():
    import requests                    #Modules used: Requests, Pandas
    import pandas as pd

    response =
requests.get('https://services1.arcgis.com/Ug5xGQbHsD8zuZzM/ArcGIS/rest/services/Opendata2/FeatureServer/148/q
where=1%3D1&outFields=*&outSR=4326&f=json')
    json_res = response.json() #Successfully collected data from the web ^

#     print (json_res['fields'])
    df = pd.DataFrame(json_res['fields'])
    df = df.drop(columns=['domain', 'defaultValue']) #Cleaning received data from inconsistency 2 & 3
(refer to bottom of the page)
    df = df.fillna(0) #Cleaning received data from inconsistency 4 (refer to bottom of the page)
    df.to_csv('dataset3.csv', index=False) #Exporting cleaned data to a new file that will show the Jupyter
output given below
    return df
```

Web requirement #2 (new updated after Phase 2):

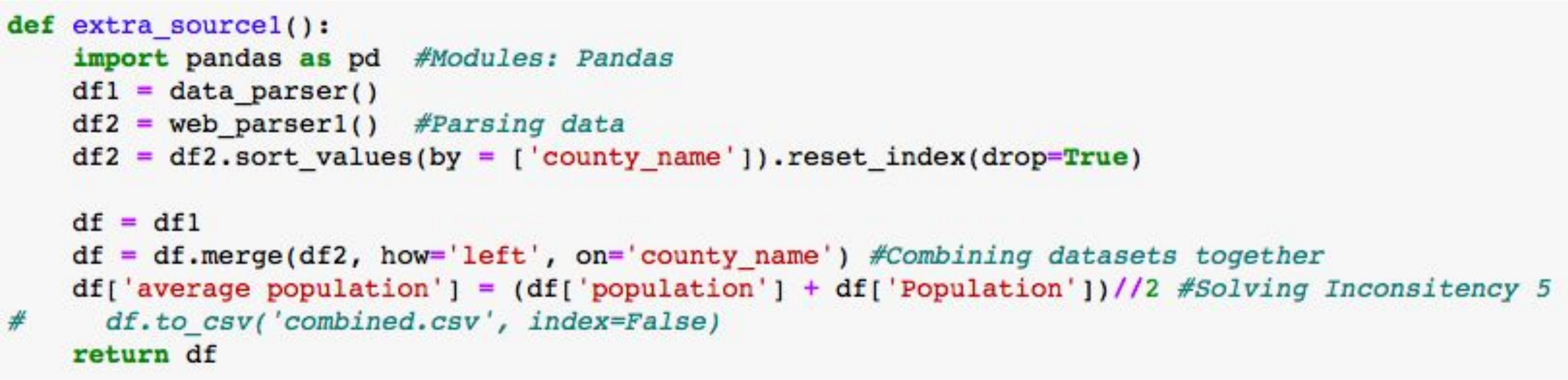
- Modules used: **Requests, Pandas**
- **For-loop, df.append**



```
def web_parser2():                  #Function is different from Phase 2 because I changed the data source
    import requests
    import pandas as pd             #Modules used: Requests, Pandas

    response = requests.get('https://services1.arcgis.com/Ug5xGQbHsD8zuZzM/ArcGIS/rest/services/ACS_2018_Economic/Featu
    json_res = response.json()          #Successfully collected data from the web ^
    df = pd.DataFrame()
    for row in json_res['features']:
        df = df.append(row['attributes'], ignore_index=True)
    df.to_csv('dataset3.csv', index=False)  #Exporting cleaned data to a new file that will show the Jupyter output gi
    return df
```

Additional parsing/cleaning function 1:

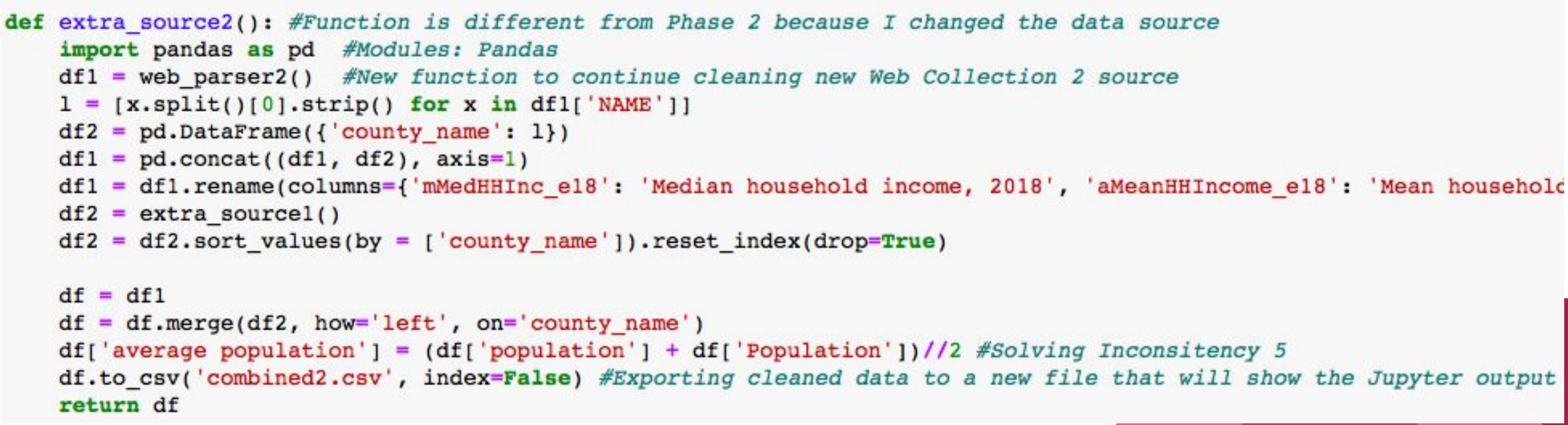
- Combined datasets using **df.merge**
- For the same attribute, the datasets had different values, so we solved this logical inconsistency by taking the average of the 2 values
- **df.sort_values(by=)**



```
def extra_source1():
    import pandas as pd  #Modules: Pandas
    df1 = data_parser()
    df2 = web_parser1()  #Parsing data
    df2 = df2.sort_values(by = ['county_name']).reset_index(drop=True)

    df = df1
    df = df.merge(df2, how='left', on='county_name') #Combining datasets together
    df['average population'] = (df['population'] + df['Population'])//2 #Solving Inconsitency 5
#     df.to_csv('combined.csv', index=False)
    return df
```

Additional parsing/cleaning function 2:

- Similar to last function with **df.merge, df.sort_values(by=)** & solving logical inconsistency
- Further cleaning with **pd.concat** & **df.rename**

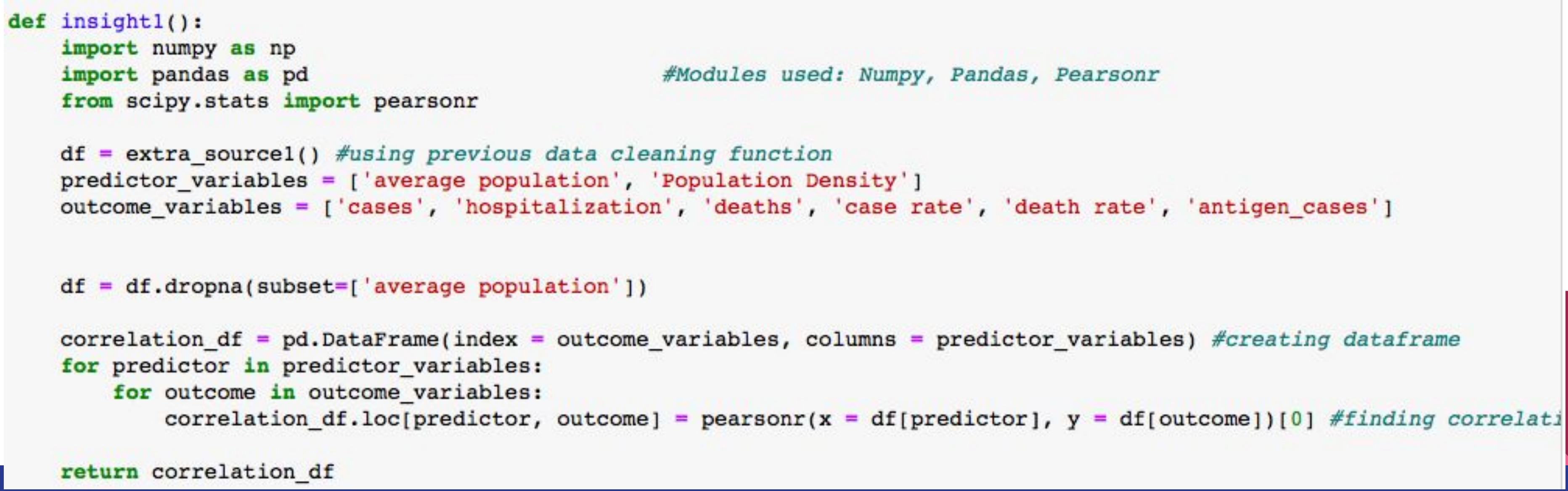


```python
def extra_source2(): #Function is different from Phase 2 because I changed the data source
    import pandas as pd  #Modules: Pandas
    df1 = web_parser2()  #New function to continue cleaning new Web Collection 2 source
    l = [x.split()[0].strip() for x in df1['NAME']]
    df2 = pd.DataFrame({'county_name': l})
    df1 = pd.concat((df1, df2), axis=1)
    df1 = df1.rename(columns={'mMedHHInc_e18': 'Median household income, 2018', 'aMeanHHIncome_e18': 'Mean househol
    df2 = extra_source1()
    df2 = df2.sort_values(by = ['county_name']).reset_index(drop=True)

    df = df1
    df = df.merge(df2, how='left', on='county_name')
    df['average population'] = (df['population'] + df['Population'])//2 #Solving Inconsistency 5
    df.to_csv('combined2.csv', index=False) #Exporting cleaned data to a new file that will show the Jupyter output
    return df
```

DATA ANALYSIS (INSIGHTS)

Insight 1: Correlation between COVID case variables & Population Variables

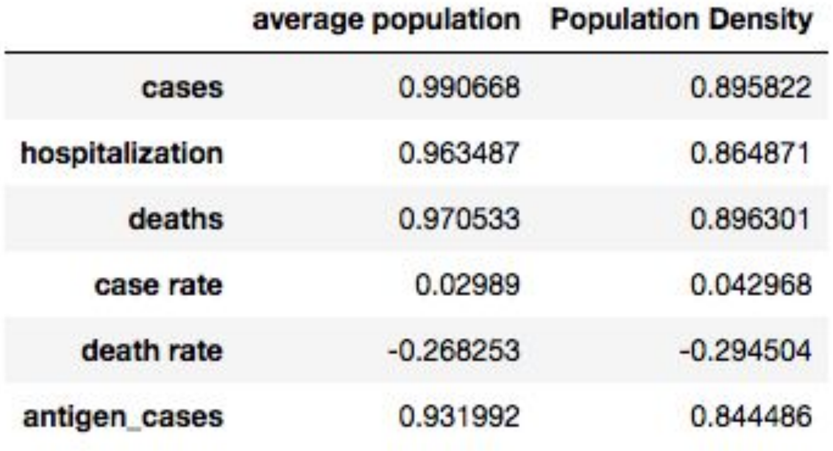
- Modules: **Numpy, Pandas, Pearsonr**
- Determines the **linear correlations** between different COVID variables & **2 population variables** and produces a dataframe with these correlation values.



```
def insight1():
    import numpy as np
    import pandas as pd                              #Modules used: Numpy, Pandas, Pearsonr
    from scipy.stats import pearsonr

    df = extra_source1() #using previous data cleaning function
    predictor_variables = ['average population', 'Population Density']
    outcome_variables = ['cases', 'hospitalization', 'deaths', 'case rate', 'death rate', 'antigen_cases']


    df = df.dropna(subset=['average population'])

    correlation_df = pd.DataFrame(index = outcome_variables, columns = predictor_variables) #creating dataframe
    for predictor in predictor_variables:
        for outcome in outcome_variables:
            correlation_df.loc[predictor, outcome] = pearsonr(x = df[predictor], y = df[outcome])[0] #finding correlati

    return correlation_df
```

Insight 1: Correlation between COVID case variables & Population Variables

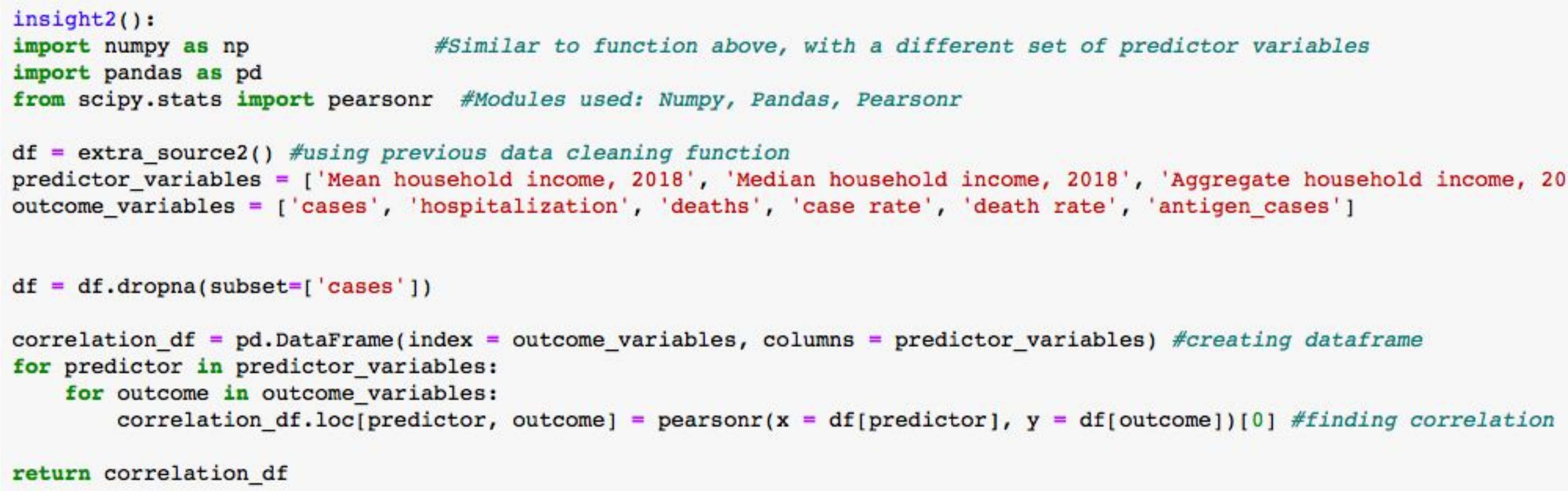
- Majority of COVID variables have a very strong correlation with both the population variables, with the exception of case rate & death rate which show little to no correlation with the population variables.
- Among the population variables, average population consistently shows a higher/stronger correlation than population density.
- This insight function is necessary to help see the effect population has on COVID cases. From seeing the data the function produces, our overall insight is that population levels have a extremely strong effect on COVID-19 cases for counties in Georgia.



	average population	Population Density
cases	0.990668	0.895822
hospitalization	0.963487	0.864871
deaths	0.970533	0.896301
case rate	0.02989	0.042968
death rate	-0.268253	-0.294504
antigen_cases	0.931992	0.844486

Insight 2: Correlation between COVID case variables & Income Variables

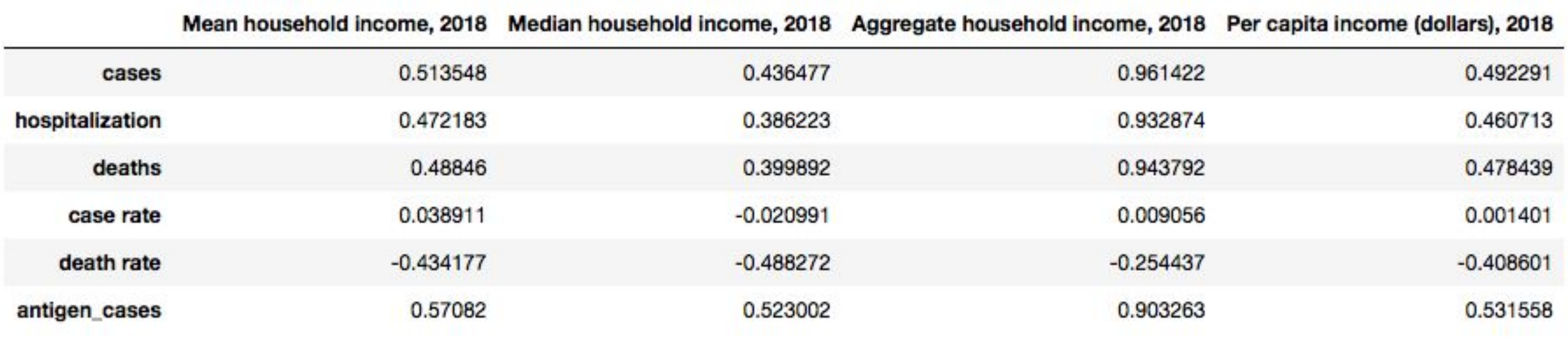
- Modules: **Numpy, Pandas, Pearsonr**
- Determines the **linear correlations** between different COVID variables & different **income variables** and produces a dataframe with these correlation values.



```
insight2():
import numpy as np              #Similar to function above, with a different set of predictor variables
import pandas as pd
from scipy.stats import pearsonr  #Modules used: Numpy, Pandas, Pearsonr

df = extra_source2() #using previous data cleaning function
predictor_variables = ['Mean household income, 2018', 'Median household income, 2018', 'Aggregate household income, 20
outcome_variables = ['cases', 'hospitalization', 'deaths', 'case rate', 'death rate', 'antigen_cases']


df = df.dropna(subset=['cases'])

correlation_df = pd.DataFrame(index = outcome_variables, columns = predictor_variables) #creating dataframe
for predictor in predictor_variables:
    for outcome in outcome_variables:
        correlation_df.loc[predictor, outcome] = pearsonr(x = df[predictor], y = df[outcome])[0] #finding correlation

return correlation_df
```

Insight 2: Correlation between COVID case variables & Income Variables

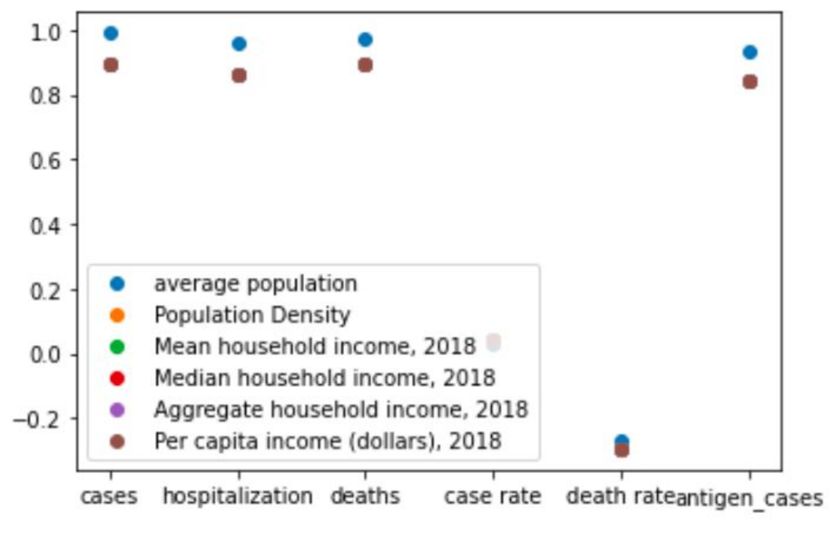
- Overall, income variables show a lower/weaker correlation to COVID variables, compared to the previous insight function's population variables.
- However, majority of income variables still show a moderate correlation to COVID variables, with the exception of case rate which shows no correlation.
- Moreover, in the last insight function, the population variables had little to no correlation to the death_rate variable, however this time we see majority of the income variables have a low to moderate correlation with death_rate.
- In addition, Aggregate household income is the income variable with the overall highest/strongest correlation to COVID variables; with correlations comparable to that of population variables.
- This insight function is necessary to help see the effect income levels have on COVID cases. From seeing the data the function produces, our overall insight is that income levels have a moderately strong effect on COVID-19 cases for counties in Georgia. The overall effect of income levels on COVID cases seems to be less than the effect of population levels, however income levels have a larger effect on the COVID death rate than population.



	Mean household income, 2018	Median household income, 2018	Aggregate household income, 2018	Per capita income (dollars), 2018
cases	0.513548	0.436477	0.961422	0.492291
hospitalization	0.472183	0.386223	0.932874	0.460713
deaths	0.48846	0.399892	0.943792	0.478439
case rate	0.038911	-0.020991	0.009056	0.001401
death rate	-0.434177	-0.488272	-0.254437	-0.408601
antigen_cases	0.57082	0.523002	0.903263	0.531558

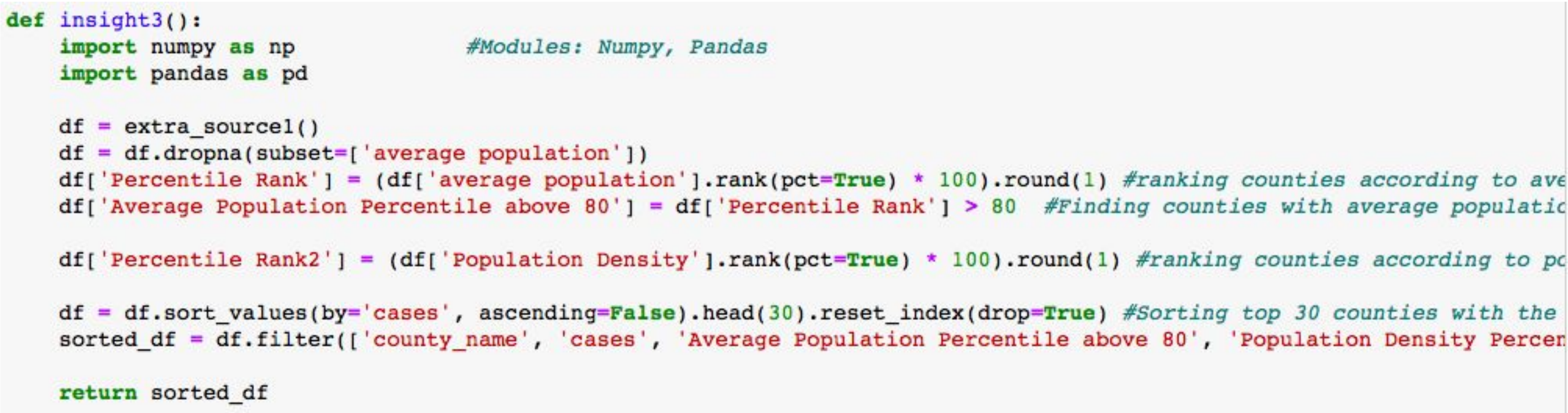
VISUALIZATION #1

- Scatterplot that visualizes some data from our insight functions 1 & 2. COVID variables on the x-axis, correlation coefficient values on the y-axis, and data points for average population & per capita income variables.
- It allows us to easily compare the correlation values for average population & per capita income for the different COVID variables to see if the population variable/income variable has a larger correlation/effect & if this correlation is positive (directly proportional) or negative (inversely proportional).
- From this visualization, we can easily tell that both income & population levels have a strong correlation/effect to the number of COVID cases, with average population having a slightly stronger correlation/effect.
- Case rate & death rate are the only COVID variables that are an exception to this, with the former having no correlation with both income & population variables, and the latter having a weak correlation to per capita income.



Insight 3: Sorting COVID cases & Ranking Population Percentile

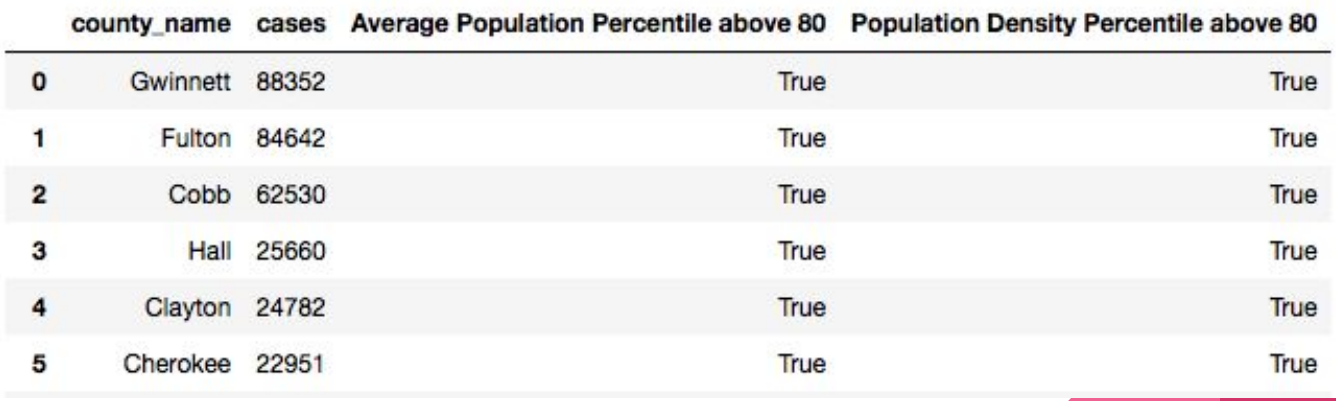
- Modules: **Numpy, Pandas**
- Sorts the top 30 counties with the most COVID cases & also produces a bool of True/False to indicate whether that county also has a population level above the 80th percentile for population variables, average population & population density.



```
def insight3():
    import numpy as np             #Modules: Numpy, Pandas
    import pandas as pd

    df = extra_source1()
    df = df.dropna(subset=['average population'])
    df['Percentile Rank'] = (df['average population'].rank(pct=True) * 100).round(1) #ranking counties according to ave
    df['Average Population Percentile above 80'] = df['Percentile Rank'] > 80  #Finding counties with average populatio

    df['Percentile Rank2'] = (df['Population Density'].rank(pct=True) * 100).round(1) #ranking counties according to po

    df = df.sort_values(by='cases', ascending=False).head(30).reset_index(drop=True) #Sorting top 30 counties with the
    sorted_df = df.filter(['county_name', 'cases', 'Average Population Percentile above 80', 'Population Density Percen

    return sorted_df
```

Insight 3: Sorting COVID cases & Ranking Population Percentile

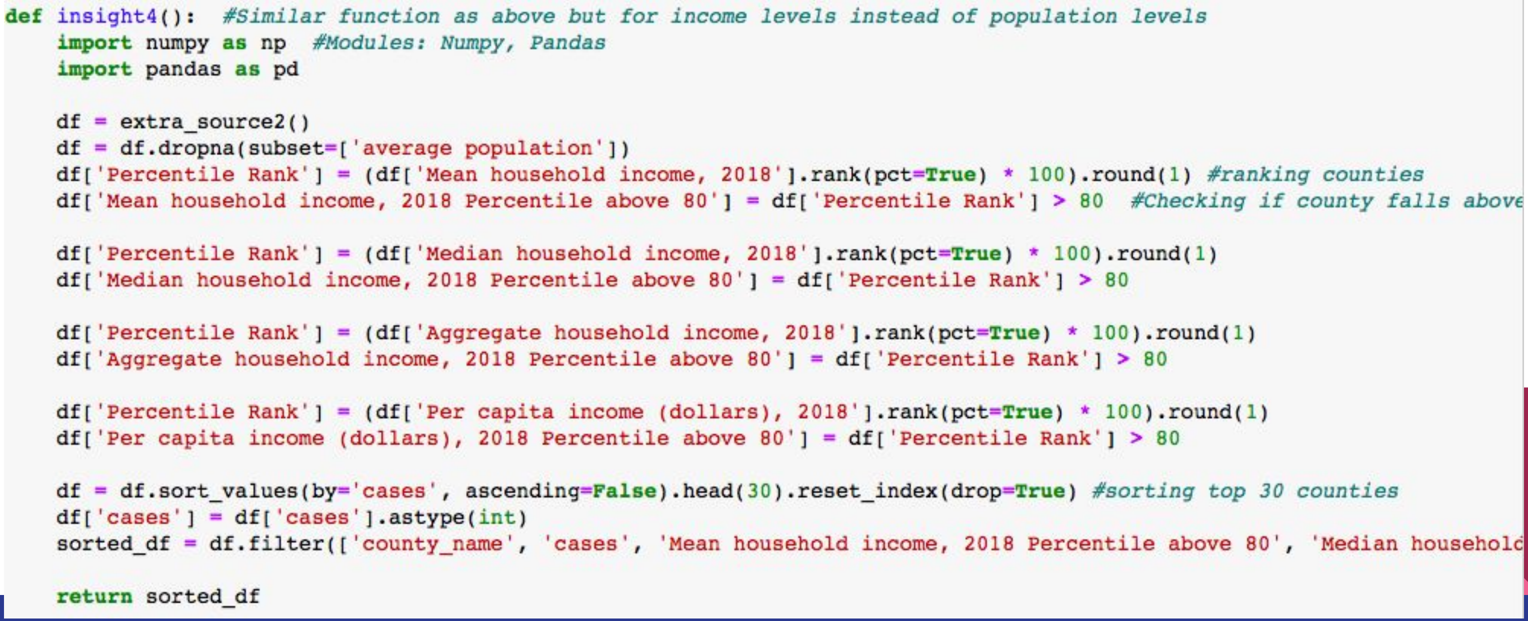
- Almost all the counties out of the top 30 counties for most COVID cases, also fall above the 80th percentile for both average population & population density; with the exception of 3 counties that were ranked in the lower 10.
- This insight function provides data that is further evidence to indicate that population levels have a strong effect on the number of COVID cases for the counties in Georgia.



	county_name	cases	Average Population Percentile above 80	Population Density Percentile above 80
0	Gwinnett	88352	True	True
1	Fulton	84642	True	True
2	Cobb	62530	True	True
3	Hall	25660	True	True
4	Clayton	24782	True	True
5	Cherokee	22951	True	True

Insight 4: Sorting COVID cases & Ranking Income Percentile

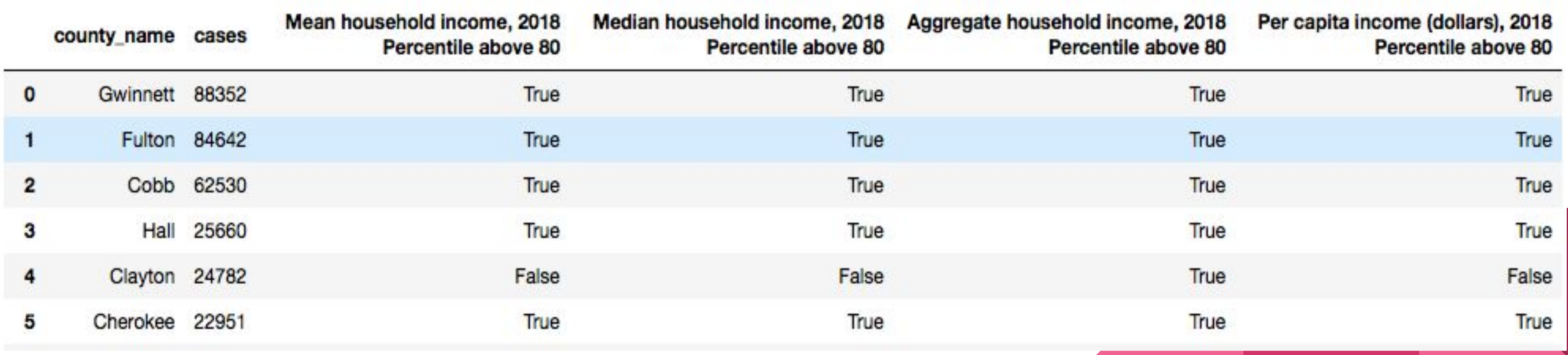
- Modules: **Numpy, Pandas**
- Sorts the top 30 counties with the most COVID cases & also produces a bool of True/False to indicate whether that county also has a income level above the 80th percentile for income variables.



```
def insight4():  #Similar function as above but for income levels instead of population levels
    import numpy as np  #Modules: Numpy, Pandas
    import pandas as pd

    df = extra_source2()
    df = df.dropna(subset=['average population'])
    df['Percentile Rank'] = (df['Mean household income, 2018'].rank(pct=True) * 100).round(1) #ranking counties
    df['Mean household income, 2018 Percentile above 80'] = df['Percentile Rank'] > 80  #Checking if county falls above

    df['Percentile Rank'] = (df['Median household income, 2018'].rank(pct=True) * 100).round(1)
    df['Median household income, 2018 Percentile above 80'] = df['Percentile Rank'] > 80

    df['Percentile Rank'] = (df['Aggregate household income, 2018'].rank(pct=True) * 100).round(1)
    df['Aggregate household income, 2018 Percentile above 80'] = df['Percentile Rank'] > 80

    df['Percentile Rank'] = (df['Per capita income (dollars), 2018'].rank(pct=True) * 100).round(1)
    df['Per capita income (dollars), 2018 Percentile above 80'] = df['Percentile Rank'] > 80

    df = df.sort_values(by='cases', ascending=False).head(30).reset_index(drop=True) #sorting top 30 counties
    df['cases'] = df['cases'].astype(int)
    sorted_df = df.filter(['county_name', 'cases', 'Mean household income, 2018 Percentile above 80', 'Median household

    return sorted_df
```

Insight 4: Sorting COVID cases & Ranking Income Percentile

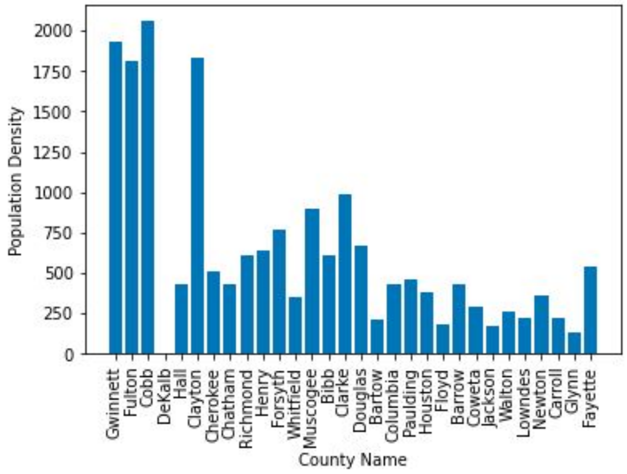
- For all income variables, the majority of top 30 counties (atleast 15 or more counties), also fall above the 80th percentile for their income level.
- This provides further evidence that income levels have a moderately strong effect on the number of COVID cases in Georgia counties- however this effect is less than the effect population levels have, with the exception of Aggregate household income which seems to have almost the same effect as population variables.

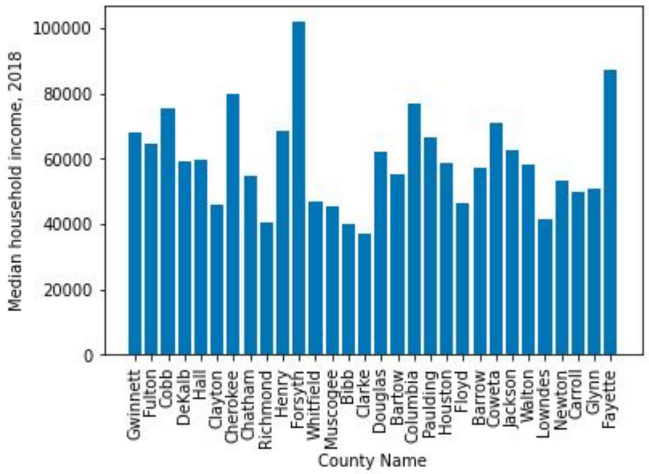


	county_name	cases	Mean household income, 2018 Percentile above 80	Median household income, 2018 Percentile above 80	Aggregate household income, 2018 Percentile above 80	Per capita income (dollars), 2018 Percentile above 80
0	Gwinnett	88352	True	True	True	True
1	Fulton	84642	True	True	True	True
2	Cobb	62530	True	True	True	True
3	Hall	25660	True	True	True	True
4	Clayton	24782	False	False	True	False
5	Cherokee	22951	True	True	True	True

VISUALIZATION #2

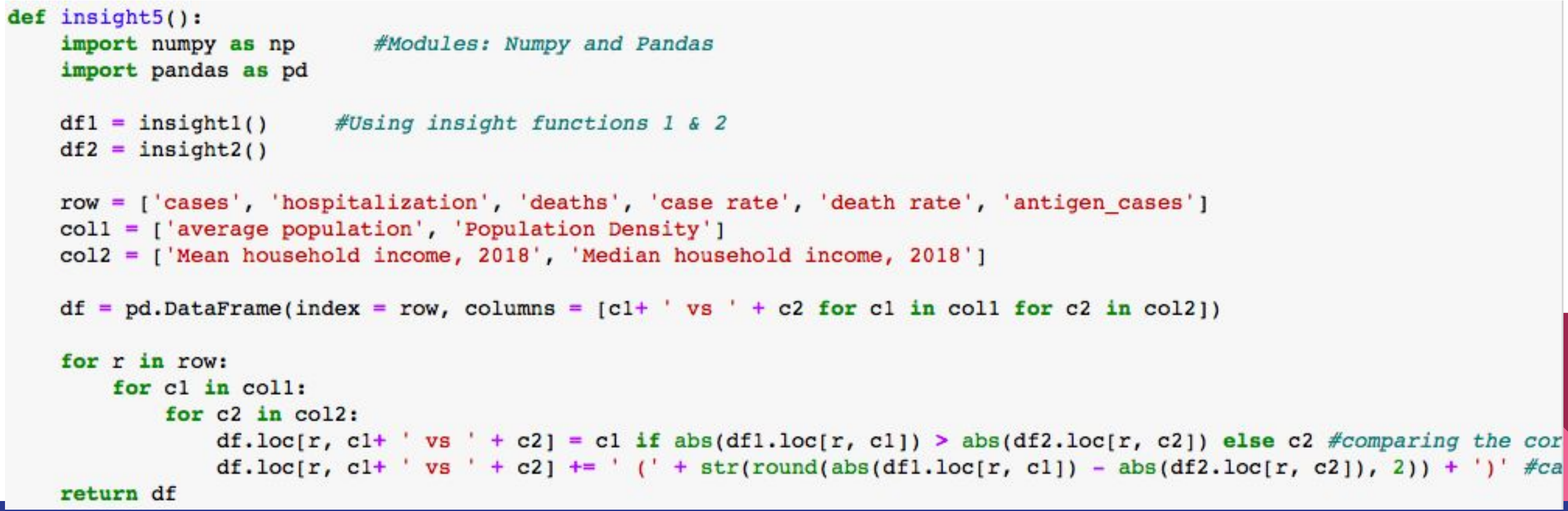
- The visualizations are bar graphs showing the population density & median household income levels (y-axis) for the top 30 counties with the most COVID cases (x-axis). The visualization is somewhat similar to our insight functions 3 & 4.
- From the bar graphs, there doesn't seem to be any overall obvious trend/pattern.
- However, we can see the top 3 counties for COVID cases are also among the top 4 counties in terms of population density by a large margin- hence providing some evidence to support population density being directly proportional to number of COVID cases.
- Whereas, we can see that the top 30 counties for COVID cases generally fall in a similar range of median household income; with the exception of a few outliers.





Insight 5: Population Correlations vs. Income Correlations

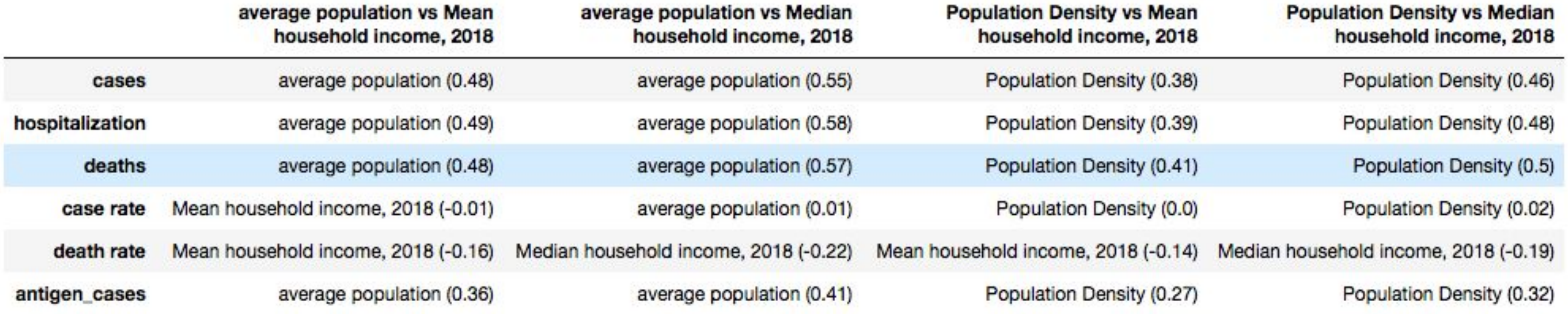
- Modules: **Numpy, Pandas**
- Directly compares the correlation values found from insight functions 1 & 2, to see if population variables or income variables have a larger effect on COVID variables. It also shows the difference in their correlation values to help us understand if the difference is significant enough to draw a conclusion or not.



```
def insight5():
    import numpy as np      #Modules: Numpy and Pandas
    import pandas as pd

    df1 = insight1()      #Using insight functions 1 & 2
    df2 = insight2()

    row = ['cases', 'hospitalization', 'deaths', 'case rate', 'death rate', 'antigen_cases']
    col1 = ['average population', 'Population Density']
    col2 = ['Mean household income, 2018', 'Median household income, 2018']

    df = pd.DataFrame(index = row, columns = [c1+ ' vs ' + c2 for c1 in col1 for c2 in col2])

    for r in row:
        for c1 in col1:
            for c2 in col2:
                df.loc[r, c1+ ' vs ' + c2] = c1 if abs(df1.loc[r, c1]) > abs(df2.loc[r, c2]) else c2 #comparing the cor
                df.loc[r, c1+ ' vs ' + c2] += ' (' + str(round(abs(df1.loc[r, c1]) - abs(df2.loc[r, c2]), 2)) + ')' #ca
    return df
```

Insight 5: Population Correlations vs. Income Correlations

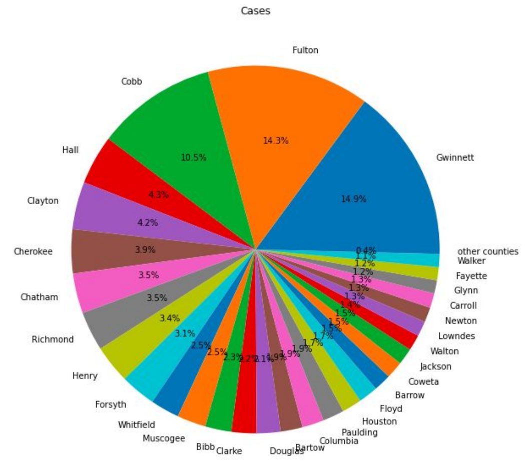
- For almost all the COVID variables, population variables have a significantly larger correlation & hence larger effect than income variables, with the exception of case rate in which neither population nor income seem to have a correlation with, and finally death rate, which is the only COVID variable in which income variables have a slightly larger correlation & hence slightly larger effect than population variables.

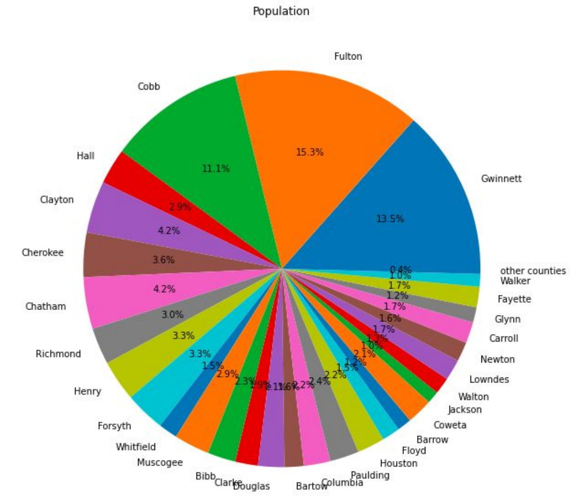


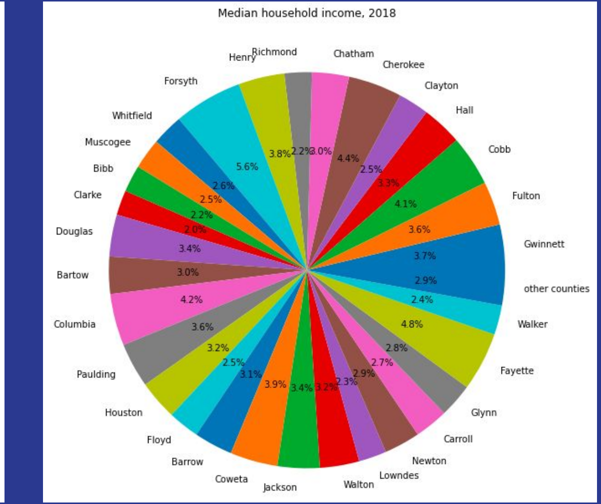
	average population vs Mean household income, 2018	average population vs Median household income, 2018	Population Density vs Mean household income, 2018	Population Density vs Median household income, 2018
cases	average population (0.48)	average population (0.55)	Population Density (0.38)	Population Density (0.46)
hospitalization	average population (0.49)	average population (0.58)	Population Density (0.39)	Population Density (0.48)
deaths	average population (0.48)	average population (0.57)	Population Density (0.41)	Population Density (0.5)
case rate	Mean household income, 2018 (-0.01)	average population (0.01)	Population Density (0.0)	Population Density (0.02)
death rate	Mean household income, 2018 (-0.16)	Median household income, 2018 (-0.22)	Mean household income, 2018 (-0.14)	Median household income, 2018 (-0.19)
antigen_cases	average population (0.36)	average population (0.41)	Population Density (0.27)	Population Density (0.32)

VISUALIZATION #3

- Pie chart 1 shows the percent of COVID cases each county from the top 30 counties contributes to the total number of covid cases in all Georgia counties. There's a category of ‘other counties’ to represent the percent contribution from counties below the top 30. Pie chart 2 shows the percent population each county from the top 30 countries contributes to the total population of all Georgia counties. Pie chart 3 shows the percent income each county contributes to the total income of all counties.
- From this visualization, we can see that the top 3 counties for COVID cases, are also the top 3 for population & among the top 4 for income- hence indicating population & income effect the number of COVID cases. Furthermore, we see that the pie chart for cases is more similar to the pie chart for population compared to the pie chart for income- hence indicating population has a larger effect on the number of COVID cases, than income.







Overall Results & Conclusion

- Do Income & Population have an effect on COVID cases in Georgia Counties? **YES**
    - What kind of an effect do they have?
    - **Population variables have a very strong correlation to most COVID case variables & are directly proportional to them (with the exception of Case rate & death rate).**
    - **Income variables have a low-moderate correlation to most COVID case variables (with the exception of Case rate). Case rate has no correlation to either.**
    - Which of the two - Income or Population - has a greater effect?
    - **Since population variables have a stronger correlation, we can conclude population has a greater effect overall. The only exception was with the death rate, in which, Income variables have more of an effect than Population variables, although this effect is much lower (lower correlation) than the effect Population**
    - **Had on other COVID case variables.**