Import modules

In [1]:

```
import pandas as pd
import csv
import os
import arcpy
from urllib.request import urlopen
from bs4 import BeautifulSoup
import requests
import zipfile
```

Download 2019 March Student Data

In [2]:

```
# Read the 2019 March table in the website and save it as csv file
url = 'https://studyinthestates.dhs.gov/sevis-data-mapping-tool/march-2019-sevis-data-mappi
table = pd.read_html(url)[0]
table.to_csv('2019temp.csv')
```

In [3]:

```
# Read the csv file, rename the columns, and drop the unnecessary columns
df = pd.read_csv('2019temp.csv')
df.columns = ['Country of Citizenship','Continent','Region','# of Active Students','Male','
dfstate = df.drop(labels=['Country of Citizenship','Continent','Region','# of Active Studer
```

In [4]:

```
# Sum up the student number by state, add the columns name and save it back as a new csv fi
data=dfstate.sum()
data.to_csv('2019temp.csv')
data = pd.read_csv('2019temp.csv')
data.columns = ['2019State','2019Student']
data.to_csv(r'D:\2021-spring\ArcGIS\Project\2019Mar.csv')
```

Download 2020 September Student Data

In [5]:

```
# Read the 2020 September table in the website and save it as csv file
url = 'https://studyinthestates.dhs.gov/sevis-data-mapping-tool/september-2020-sevis-data-m
table = pd.read_html(url)[0]
table.to_csv('2020temp.csv')
```

In [6]:

```
# Read the csv file, and drop the unnecessary columns
df = pd.read_csv('2020temp.csv')
df.columns = ['list','Country of Citizenship','Continent','Region','# of Active Students','
dfstate = df.drop(labels=['list','Country of Citizenship','Continent','Region','# of Active
```

In [7]:

```
# Sum up the student number by state, add the columns name and save it back as a new csv fi
data=dfstate.sum()
data.to_csv('2020temp.csv')
data = pd.read_csv('2020temp.csv')
data.columns = ['2020State','2020Student']
data.to_csv(r'D:\2021-spring\ArcGIS\Project\2020Sep.csv')
```

Download State Boundaries Data

In [8]:

```
# Use urlopen to get the data and save it as a variable
response = urlopen("https://www.census.gov/geographies/mapping-files/time-series/geo/carto-
html = response.read()
```

In [9]:

```
# Build the BeautifulSoup item from the website data
parser = BeautifulSoup(html.decode("utf-8"), "html.parser")
```

In [10]:

```
# Find the class for downloading the shapefile
target = parser.find_all('a', class_="uscb-layout-align-start-start", name_='cb_2018_us_sta
```

In [11]:

```
# Find the class for downloading the shapefile
target = parser.find_all("a", href="//www2.census.gov/geo/tiger/GENZ2018/shp/cb_2018_us_sta
```

In [12]:

```
# Get the downloading link from the class above
for link in target:
    url = 'http:'+link.get('href')
    print(url)
```

```
http://www2.census.gov/geo/tiger/GENZ2018/shp/cb_2018_us_state_500k.zip (htt
p://www2.census.gov/geo/tiger/GENZ2018/shp/cb_2018_us_state_500k.zip)
```

In [13]:

```
# Get the data and save it as a zip file
r = requests.get(url, allow_redirects=True)
open('stateboundary.zip', 'wb').write(r.content)
```

Out[13]:

```
0
```

In [14]:

```
# unzip the file
with zipfile.ZipFile('stateboundary.zip', 'r') as tempzip:
    tempzip.extractall(r'D:\2021-spring\ArcGIS\Project\stateboundary')
```

```
---------------------------------------------------------------------------
BadZipFile                                Traceback (most recent call last)
<ipython-input-14-a04908f2fea8> in <module>
      1 # unzip the file
----> 2 with zipfile.ZipFile('stateboundary.zip', 'r') as tempzip:
      3     tempzip.extractall(r'D:\2021-spring\ArcGIS\Project\stateboundar
y')

C:\Program Files\ArcGIS\Pro\bin\Python\envs\arcgispro-py3\lib\zipfile.py in
__init__(self, file, mode, compression, allowZip64, compresslevel)
   1256         try:
   1257             if mode == 'r':
-> 1258                 self._RealGetContents()
   1259             elif mode in ('w', 'x'):
   1260                 # set the modified flag so central directory gets wr
itten

C:\Program Files\ArcGIS\Pro\bin\Python\envs\arcgispro-py3\lib\zipfile.py in
_RealGetContents(self)
   1323             raise BadZipFile("File is not a zip file")
   1324         if not endrec:
-> 1325             raise BadZipFile("File is not a zip file")
   1326         if self.debug > 1:
   1327             print(endrec)

BadZipFile: File is not a zip file
```

Arcpy - Join Data

In [15]:

```
# Set up the Arcpy environment
arcpy.env.workspace = r'D:\2021-spring\ArcGIS\Project'
aprx = arcpy.mp.ArcGISProject(r'D:\2021-spring\ArcGIS\Project\Final Project\Final Project.a
```

In [16]:

```
# Join 2019 March Student Number
arcpy.management.JoinField(r'\stateboundary\cb_2018_us_state_500k.shp','STUSPS','2019Mar.cs
```

Out[16]:

Output

D:\2021-spring\ArcGIS\Project\\stateboundary\cb_2018_us_state_500k.shp

Messages

Start Time: 2021年5月3日 下午 03:21:10

Succeeded at 2021年5月3日 下午 03:21:10 (Elapsed Time: 0.79 seconds)

In [17]:

```
# Join 2020 September Student Number
arcpy.management.JoinField(r'\stateboundary\cb_2018_us_state_500k.shp','STUSPS','2020Sep.cs
```

Out[17]:

Output

D:\2021-spring\ArcGIS\Project\\stateboundary\cb_2018_us_state_500k.shp

Messages

Start Time: 2021年5月3日 下午 03:21:11

Succeeded at 2021年5月3日 下午 03:21:12 (Elapsed Time: 0.93 seconds)

In [18]:

```
# Join State Population
arcpy.management.JoinField(r'stateboundary\cb_2018_us_state_500k.shp','NAME','Population.cs
```

Out[18]:

Output

D:\2021-spring\ArcGIS\Project\stateboundary\cb_2018_us_state_500k.shp

Messages

Start Time: 2021年5月3日 下午 03:21:13

Succeeded at 2021年5月3日 下午 03:21:14 (Elapsed Time: 0.78 seconds)

In [19]:

```
arcpy.management.JoinField(r'stateboundary\cb_2018_us_state_500k.shp','STUSPS','CovidCase.c
```

Out[19]:

Output

D:\2021-spring\ArcGIS\Project\stateboundary\cb_2018_us_state_500k.shp

Messages

Start Time: 2021年5月3日 下午 03:21:16

Succeeded at 2021年5月3日 下午 03:21:17 (Elapsed Time: 0.91 seconds)

Arcpy - Calculate Data

In [20]:

```
arcpy.management.CalculateField('stateboundary\cb_2018_us_state_500k.shp','StuVar','!2019St
```

Out[20]:

Output

D:\2021-spring\ArcGIS\Project\stateboundary\cb_2018_us_state_500k.shp

Messages

Start Time: 2021年5月3日 下午 03:22:21
Adding StuVar to cb_2018_us_state_500k...
Succeeded at 2021年5月3日 下午 03:22:21 (Elapsed Time: 0.03 seconds)

In [21]:

```
arcpy.management.CalculateField('stateboundary\cb_2018_us_state_500k.shp','StuVarPer','!Stu
```

Out[21]:

Output

D:\2021-spring\ArcGIS\Project\stateboundary\cb_2018_us_state_500k.shp

Messages

Start Time: 2021年5月3日 下午 03:22:22
Adding StuVarPer to cb_2018_us_state_500k...
Succeeded at 2021年5月3日 下午 03:22:22 (Elapsed Time: 0.03 seconds)

In [22]:

```
rcpy.management.CalculateField('stateboundary\cb_2018_us_state_500k.shp','CasePer','!CaseNum
```

Out[22]:

Output

D:\2021-spring\ArcGIS\Project\stateboundary\cb_2018_us_state_500k.shp

Messages

Start Time: 2021年5月3日 下午 03:22:25
Adding CasePer to cb_2018_us_state_500k...
Succeeded at 2021年5月3日 下午 03:22:26 (Elapsed Time: 0.04 seconds)

Arcpy - Analyze Data

In [23]:

```
(r'stateboundary\cb_2018_us_state_500k.shp','2019Studen','CONTINUOUS', r'Analysis\StuVar.shp
```

Out[23]:

Output

id	value
0	D:\2021-spring\ArcGIS\Project\Analysis\StuVar.shp
1	

Messages

```
Start Time: 2021年5月3日 下午 03:22:30
WARNING 001605: Distances for Geographic Coordinates (degrees, minutes, seconds) are
analyzed using Chordal Distances in meters.

--------------- Summary of GLR Results [Model Type: Continuous (Gaussian/OLS)] --------------
------------------------------------------------------------------------------------------
Variable Coefficient [a] StdError t-Statistic Probability [b] Robust_SE Robust_t Robust_Pr [b]
Intercept -365.023368 280.509513 -1.301287 0.198790 326.623114 -1.117567 0.268791
2020STUDEN 1.314007 0.009315 141.068629 0.000000* 0.028670 45.831556 0.000000*
------------------------------------------------------------------------------------------



--------------------------------------------- GLR Diagnostics ---------------------------------------------
------------------------------------------------------------------------------------------------------------
Input Features: cb_2018_us_state_500k.shp Dependent Variable: 2019STUDEN
Number of Observations: 55 Akaike's Information Criterion (AICc) [d]: 981.429403
Multiple R-Squared [d]: 0.997344 Adjusted R-Squared [d]: 0.997294
Joint F-Statistic [e]: 19900.358134 Prob(>F), (1,53) degrees of freedom: 0.000000*
Joint Wald Statistic [e]: 2100.531524 Prob(>chi-squared), (1) degrees of freedom: 0.000000*
Koenker (BP) Statistic [f]: 35.794498 Prob(>chi-squared), (1) degrees of freedom: 0.000000*
Jarque-Bera Statistic [g]: 117.379678 Prob(>chi-squared), (2) degrees of freedom:
0.000000*
------------------------------------------------------------------------------------------------------------



Notes on Interpretation
* An asterisk next to a number indicates a statistically significant p-value (p < 0.01).
[a] Coefficient: Represents the strength and type of relationship between each explanatory
variable and the dependent variable.
[b] Probability and Robust Probability (Robust_Pr): Asterisk (*) indicates a coefficient is
statistically significant (p < 0.01); if the Koenker (BP) Statistic [f] is statistically significant, use
the Robust Probability column (Robust_Pr) to determine coefficient significance.
[c] Variance Inflation Factor (VIF): Large Variance Inflation Factor (VIF) values (> 7.5)
indicate redundancy among explanatory variables.
[d] R-Squared and Akaike's Information Criterion (AICc): Measures of model fit/performance.
[e] Joint F and Wald Statistics: Asterisk (*) indicates overall model significance (p < 0.01); if
the Koenker (BP) Statistic [f] is statistically significant, use the Wald Statistic to determine
```

```
overall model significance.
[f] Koenker (BP) Statistic: When this test is statistically significant (p < 0.01), the relationships
modeled are not consistent (either due to non-stationarity or heteroskedasticity). You should
rely on the Robust Probabilities (Robust_Pr) to determine coefficient significance and on the
Wald Statistic to determine overall model significance.
[g] Jarque-Bera Statistic: When this test is statistically significant (p < 0.01) model
predictions are biased (the residuals are not normally distributed).

Succeeded at 2021年5月3日 下午 03:22:31 (Elapsed Time: 0.77 seconds)
```

In [24]:

```
on(r'stateboundary\cb_2018_us_state_500k.shp','StuVarPer','CONTINUOUS', r'Analysis\StuVar_Ca
```

Out[24]:

Output

id	value
0	D:\2021-spring\ArcGIS\Project\Analysis\StuVar_Case.shp
1	

Messages

```
Start Time: 2021年5月3日 下午 03:22:31
WARNING 001605: Distances for Geographic Coordinates (degrees, minutes, seconds) are
analyzed using Chordal Distances in meters.

------------ Summary of GLR Results [Model Type: Continuous (Gaussian/OLS)] ------------
-------------------------------------------------------------------------------------
Variable Coefficient [a] StdError t-Statistic Probability [b] Robust_SE Robust_t Robust_Pr [b]
Intercept 0.072758 0.025476 2.855978 0.006114* 0.014436 5.040220 0.000006*
CASEPER 0.002683 0.015004 0.178790 0.858785 0.011440 0.234491 0.815509
-------------------------------------------------------------------------------------



--------------------------------------------- GLR Diagnostics ---------------------------------------------
-----------------------------------------------------------------------------------------------------------
Input Features: cb_2018_us_state_500k.shp Dependent Variable: STUVARPER
Number of Observations: 55 Akaike's Information Criterion (AICc) [d]: -118.041679
Multiple R-Squared [d]: 0.000603 Adjusted R-Squared [d]: -0.018254
Joint F-Statistic [e]: 0.031966 Prob(>F), (1,53) degrees of freedom: 0.984446
Joint Wald Statistic [e]: 0.054986 Prob(>chi-squared), (1) degrees of freedom: 0.814604
Koenker (BP) Statistic [f]: 0.498354 Prob(>chi-squared), (1) degrees of freedom: 0.480224
Jarque-Bera Statistic [g]: 1054.015893 Prob(>chi-squared), (2) degrees of freedom:
0.000000*
-----------------------------------------------------------------------------------------------------------



Notes on Interpretation
* An asterisk next to a number indicates a statistically significant p-value (p < 0.01).
[a] Coefficient: Represents the strength and type of relationship between each explanatory
variable and the dependent variable.
[b] Probability and Robust Probability (Robust_Pr): Asterisk (*) indicates a coefficient is
statistically significant (p < 0.01); if the Koenker (BP) Statistic [f] is statistically significant, use
the Robust Probability column (Robust_Pr) to determine coefficient significance.
[c] Variance Inflation Factor (VIF): Large Variance Inflation Factor (VIF) values (> 7.5)
indicate redundancy among explanatory variables.
[d] R-Squared and Akaike's Information Criterion (AICc): Measures of model fit/performance.
[e] Joint F and Wald Statistics: Asterisk (*) indicates overall model significance (p < 0.01); if
the Koenker (BP) Statistic [f] is statistically significant, use the Wald Statistic to determine
```

```
overall model significance.
[f] Koenker (BP) Statistic: When this test is statistically significant (p < 0.01), the relationships
modeled are not consistent (either due to non-stationarity or heteroskedasticity). You should
rely on the Robust Probabilities (Robust_Pr) to determine coefficient significance and on the
Wald Statistic to determine overall model significance.
[g] Jarque-Bera Statistic: When this test is statistically significant (p < 0.01) model
predictions are biased (the residuals are not normally distributed).

Succeeded at 2021年5月3日 下午 03:22:32 (Elapsed Time: 0.68 seconds)
```

In [25]:

```
stateboundary\cb_2018_us_state_500k.shp','CasePer','CONTINUOUS', r'Analysis\Case_StuVar.shp'
```

Out[25]:

Output

id	value
0	D:\2021-spring\ArcGIS\Project\Analysis\Case_StuVar.shp
1	

Messages

```
Start Time: 2021年5月3日 下午 03:22:34
WARNING 001605: Distances for Geographic Coordinates (degrees, minutes, seconds) are
analyzed using Chordal Distances in meters.

------------- Summary of GLR Results [Model Type: Continuous (Gaussian/OLS)] ------------
----------------------------------------------------------------------------------------
Variable Coefficient [a] StdError t-Statistic Probability [b] Robust_SE Robust_t Robust_Pr [b]
Intercept 1.523167 0.137686 11.062618 0.000000* 0.129123 11.796222 0.000000*
STUVARPER 0.224701 1.256789 0.178790 0.858785 0.851307 0.263949 0.792845
----------------------------------------------------------------------------------------



-------------------------------------------- GLR Diagnostics -------------------------------------------
--------------------------------------------------------------------------------------------------------
Input Features: cb_2018_us_state_500k.shp Dependent Variable: CASEPER
Number of Observations: 55 Akaike's Information Criterion (AICc) [d]: 125.499120
Multiple R-Squared [d]: 0.000603 Adjusted R-Squared [d]: -0.018254
Joint F-Statistic [e]: 0.031966 Prob(>F), (1,53) degrees of freedom: 0.984446
Joint Wald Statistic [e]: 0.069669 Prob(>chi-squared), (1) degrees of freedom: 0.791820
Koenker (BP) Statistic [f]: 1.082274 Prob(>chi-squared), (1) degrees of freedom: 0.298189
Jarque-Bera Statistic [g]: 0.811695 Prob(>chi-squared), (2) degrees of freedom: 0.666412
--------------------------------------------------------------------------------------------------------



Notes on Interpretation
* An asterisk next to a number indicates a statistically significant p-value (p < 0.01).
[a] Coefficient: Represents the strength and type of relationship between each explanatory
variable and the dependent variable.
[b] Probability and Robust Probability (Robust_Pr): Asterisk (*) indicates a coefficient is
statistically significant (p < 0.01); if the Koenker (BP) Statistic [f] is statistically significant, use
the Robust Probability column (Robust_Pr) to determine coefficient significance.
[c] Variance Inflation Factor (VIF): Large Variance Inflation Factor (VIF) values (> 7.5)
indicate redundancy among explanatory variables.
[d] R-Squared and Akaike's Information Criterion (AICc): Measures of model fit/performance.
[e] Joint F and Wald Statistics: Asterisk (*) indicates overall model significance (p < 0.01); if
the Koenker (BP) Statistic [f] is statistically significant, use the Wald Statistic to determine
overall model significance.
```

```
[f] Koenker (BP) Statistic: When this test is statistically significant (p < 0.01), the relationships
modeled are not consistent (either due to non-stationarity or heteroskedasticity). You should
rely on the Robust Probabilities (Robust_Pr) to determine coefficient significance and on the
Wald Statistic to determine overall model significance.
[g] Jarque-Bera Statistic: When this test is statistically significant (p < 0.01) model
predictions are biased (the residuals are not normally distributed).

Succeeded at 2021年5月3日 下午 03:22:34 (Elapsed Time: 0.67 seconds)
```

In [26]:

```
arcpy.stats.GeneralizedLinearRegression(r'stateboundary\cb_2018_us_state_500k.shp','StuVar'
```

Out[26]:

Output

id	value
0	D:\2021-spring\ArcGIS\Project\Analysis\StuVar_Pop.shp
1	

Messages

```
Start Time: 2021年5月3日 下午 03:22:38
WARNING 001605: Distances for Geographic Coordinates (degrees, minutes, seconds) are
analyzed using Chordal Distances in meters.

--------------- Summary of GLR Results [Model Type: Continuous (Gaussian/OLS)] --------------
---------------------------------------------------------------------------------------------
Variable Coefficient [a] StdError t-Statistic Probability [b] Robust_SE Robust_t Robust_Pr [b]
Intercept -1323.163754 635.607366 -2.081731 0.042212* 752.354843 -1.758696 0.084404
POPULATION 0.001016 0.000068 14.977289 0.000000* 0.000173 5.859826 0.000000*
---------------------------------------------------------------------------------------------



--------------------------------------------- GLR Diagnostics ---------------------------------------------
----------------------------------------------------------------------------------------------------------
Input Features: cb_2018_us_state_500k.shp Dependent Variable: STUVAR
Number of Observations: 55 Akaike's Information Criterion (AICc) [d]: 1061.512897
Multiple R-Squared [d]: 0.808884 Adjusted R-Squared [d]: 0.805279
Joint F-Statistic [e]: 224.319177 Prob(>F), (1,53) degrees of freedom: 0.000000*
Joint Wald Statistic [e]: 34.337557 Prob(>chi-squared), (1) degrees of freedom: 0.000000*
Koenker (BP) Statistic [f]: 38.223716 Prob(>chi-squared), (1) degrees of freedom: 0.000000*
Jarque-Bera Statistic [g]: 31.309207 Prob(>chi-squared), (2) degrees of freedom: 0.000000*
----------------------------------------------------------------------------------------------------------



Notes on Interpretation
* An asterisk next to a number indicates a statistically significant p-value (p < 0.01).
[a] Coefficient: Represents the strength and type of relationship between each explanatory
variable and the dependent variable.
[b] Probability and Robust Probability (Robust_Pr): Asterisk (*) indicates a coefficient is
statistically significant (p < 0.01); if the Koenker (BP) Statistic [f] is statistically significant, use
the Robust Probability column (Robust_Pr) to determine coefficient significance.
[c] Variance Inflation Factor (VIF): Large Variance Inflation Factor (VIF) values (> 7.5)
indicate redundancy among explanatory variables.
[d] R-Squared and Akaike's Information Criterion (AICc): Measures of model fit/performance.
[e] Joint F and Wald Statistics: Asterisk (*) indicates overall model significance (p < 0.01); if
the Koenker (BP) Statistic [f] is statistically significant, use the Wald Statistic to determine
overall model significance.
```

```
[f] Koenker (BP) Statistic: When this test is statistically significant (p < 0.01), the relationships
modeled are not consistent (either due to non-stationarity or heteroskedasticity). You should
rely on the Robust Probabilities (Robust_Pr) to determine coefficient significance and on the
Wald Statistic to determine overall model significance.
[g] Jarque-Bera Statistic: When this test is statistically significant (p < 0.01) model
predictions are biased (the residuals are not normally distributed).

Succeeded at 2021年5月3日 下午 03:22:39 (Elapsed Time: 0.68 seconds)
```

In [ ]: