## DSC-540 Final Project
### Chitramoy Mukherjee-DSC540-T304
### Date : 09/22/2023

### 3 data sources and it's descriptions

1. acs2017_county_data.csv : This data file contains US county level census data for year-2017.This dataset is downloaded from
    kaggle.
2. Wikipedia List of states table contains US state information. Thsi table contains US 50 states information.
3. US Government Data: The US government provides a wide range of public APIs, including data on demographics, economics,
    and crime.(https://www.census.gov/data/developers/data-sets.html)

### Relationship between 3 data sources

US Census Data(acs2017_county_data.csv) and the List of US States table can be linked by geographic location ans state name. We could use the Google Maps API to determine the latitude and longitude for each state in the List of US States table, and then use this information to link the state data to the demographic and economic data in the US Census Data dataset.

### Interpretation and operations on dataset to accomplish future milestones

Based on the state name and it's geographic information, we can merge this 3 datasets after removing the headers from those. After the first step will remove the unwanted columns from the datasets andf then merge those three into one dataset and that dataset could be used to inform policy makers and economic developers about the factors that contribute to population growth. It could also be used to identify states that are at risk of population decline, and to develop targeted interventions to promote population growth in these states.

As a data wrangling project using these datasets would be to create a dataset that maps the demographics and economic factors of each US state to the state's population growth rate. This could be done by linking the US Census Data dataset and the List of US States table, as described above. Once the datasets are linked, we could use statistical analysis to calculate the population growth rate for each state, and then identify correlations between the population growth rate and demographic and economic factors, such as median income, poverty rate, and education levels.

Data Disctionary for acs2017_county_data.csv :

Data columns (total 37 columns):

```
Column No.  Column           Data type     Description
---         ------           -----         -----
 0          Countyid         int64         County identification #
 1          State            object        Name of the state
 2          County           object        Name of the county
 3          TotalPop         int64         Total population
 4          Men              int64         Men count
 5         Women            int64         Women count
 6         Hispanic         float64       % of population that is
Hispanic/Latino
 7         White            float64       % of population that is white
 8         Black            float64       % of population that is black
 9         Native           float64       % of population that is Native
American or Native Alaskan
 10        Asian            float64       % of population that is Asian
 11        Pacific          float64       % of population that is Native
Hawaiian or Pacific Islander
 12        VotingAge          int64       Voting age in days
 13        Income           float64       Median household income ($)
 14        IncomeErr        float64       Median household income error ($)
 15        IncomePerCap     float64       Income per capita ($)
 16        IncomePerCapErr  float64       Income per capita error ($)
 17        Poverty          float64       % under poverty level
 18        ChildPoverty     float64       % of children under poverty level
 19        Professional     float64       % employed in management,
business, science, and arts
 20        Service          float64       % employed in service jobs
 21        Office           float64       % employed in sales and office
jobs
 22        Construction     float64       % employed in natural resources,
construction, and maintenance
 23        Production       float64       % employed in production,
transportation, and material movement
 24        Drive            float64       % commuting alone in a car, van,
or truck
 25        Carpool          float64       % carpooling in a car, van, or
truck
 26        Transit          float64       % commuting on public
transportation
 27        Walk             float64       % walking to work
 28        OtherTransp      float64       % commuting via other means
 29        WorkAtHome       float64       % working at home
 30        MeanCommute      float64       Mean commute time (minutes)
 31        Employed         int64         Number of employed (16+)
 32        PrivateWork      float64       % employed in private industry
 33        PublicWork       float64       % employed in public jobs
 34        SelfEmployed     float64       % self-employed
 35        FamilyWork       float64       % in unpaid family work
 36        Unemployment     float64       Unemployment rate (%)

 List of states Wikipedia Table data dictionary :

 Column No.  Column               Data type     Description
 ---        ------               -----         -----
 1          Postal abbrevation   object        State Name
 2          Cities               object        Major City bypopulation/state
capital
 3          Established          Date          Year state formed
 4          Population           int64         total state population
 5          Total area           int64         Total area
 6          Land area            int64         Total land
 7          Water area           int64         Total water area
```

### Project subject area

Will apply different data wragling techniques on the source data and merge it to perform the analysis.

As a part of this project we will be merging 3 different dataset of differnt type using a common key(state name) and will perform statistical analysis to identify correlations between crime rates and demographic and economic factors, such as poverty, unemployment, and education levels.

### Data Sources:

1. acs2017_county_data.csv (https://www.kaggle.com/code/alawdisoft/us-census-demographic-data/input?select=acs2017_county_data.csv)

2. The US government provides a wide range of public APIs, including data on demographics, economics, and crime. US Census Bureau provides an API for accessing census data. (https://www.census.gov/data/developers/data-sets.html)

3. his Wikipedia table contains a list of all 50 US states, along with their capitals and population. (https://simple.wikipedia.org/wiki/List_of_U.S._states)

### Relationships :

All 3 datasets contain data based on state. The lowest granularity of this 3 dataset data is state name.

### Ethical implications and Challenges :

Ethical implications of using US Census Data for a data wrangling project include:

Privacy: The US Census Data contains personal information about individuals and households. It is important to take steps to protect the privacy of this data, such as anonymizing the data or using differential privacy techniques.

Bias: The US Census Data may be biased in certain ways. For example, it may be more difficult to reach certain populations, such as low-income households or immigrant communities. It is important to be aware of these potential biases and to take steps to mitigate them.

Discrimination: The US Census Data could be used to discriminate against certain groups of people. For example, it could be used to target certain groups with marketing messages or to deny them access to services or opportunities. It is important to use the data in a responsible and ethical way to avoid discrimination.

Use differential privacy techniques: Differential privacy is a set of techniques that can be used to protect the privacy of individuals in a dataset while still allowing for accurate analysis.

Some of the challenges that you might face in a US Census Data project include:

Data quality: The US Census Data is a large and complex dataset. It is important to carefully clean and prepare the data before using it for analysis.

Data complexity: The US Census Data contains a wide range of variables. It is important to understand the meaning of the variables and how they can be used for analysis.

Ethical considerations: As discussed above, there are a number of ethical considerations that must be taken into account when using US Census Data. It is important to design your project in a way that respects the privacy of the data and avoids bias and discrimination.