CDDEP Data Science Coding Challenge

You can use Python, R, or SQL to complete the following challenge.

**OBJECTIVES**: Provide a (1) CSV dataset with pipe delimiters, (2) summary statistics, and (3) plots showing the breakdown of racial and income population sizes and life-expectancy for all counties in Louisiana. There’s also a bonus deliverable (4). Turn in a Python, R, or SQL scripts or notebooks that were used to create the deliverable. The specifics for each deliverable are listed below.

# Part A: Data Wrangling

## Deliverable 1: CSV Table:

1. Go to <https://covid19.census.gov/datasets/race-and-ethnicity-county> and download the race and ethnicity population sizes for each US county.
2. Go to <https://www.census.gov/data/experimental-data-products/community-resilience-estimates.html> and download the community resilience estimates for each US county.
3. Now merge the two datasets into the following format with **pipe delimiters** (separators) so that each row corresponds with a county and the columns include the following variables:
   1. Total Population - White alone
   2. Total Population - Black or African American alone
   3. Total Population - American Indian and Alaska Native alone
   4. Total Population - Asian alone
   5. Total Population - Native Hawaiian and Other Pacific Islander alone
   6. Predicted Rate – 3+ Risk factors
   7. Predicted Rate – 1-2+ Risk factors
   8. Predicted Rate – 0 Risk factors

The CSV dataset should be structured as the following:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| GEO\_NAME | GEO\_PARENT\_NAME | FIPS\_CODE | Total Population - White | … | Predicted Rate – 3+ Risk factors | … |
| Autauga County | Alabama | 1001 | 42437 | … | 27.28 | … |
| … | … | … | … | … | … | … |

Note: you might need to map the variable codes with the names using the following data dictionary:

<https://api.census.gov/data/2016/acs/acs5/groups/B02001.html>

<https://www2.census.gov/data/experimental-data-products/community-resilience-estimates/2020/file-layout.pdf>

**BONUS:** Automate the process using a Python, R, or SQL script that extracts the data and formats using the API from the website.

# Part B: Statistical Analysis and Data Visualization

## Deliverable 2: Summary Statistics

1. Create a script or notebook that answers the following questions

* Which **county** has the largest American Indian and Alaska Native population in terms of **percentage** of the total population?
* Which **state** has the smallest Black or African American population in terms of **percentage** of the total population?
* Which **county** has the largest **non-white** **percentage** (i.e. Black, American Indian, Asian, Pacific Islanders)?
* Which counties are the most and least resilient?

## Deliverable 3: Linear Regression

1. Create a script or notebook that generates a scatter plot of all counties where the x-axis measures the **non-white percentage** **of the county population** and the y-axis measures **Predicted Rate of 3+ Risk factors**.
2. Run a linear regression assuming a Gaussian distributed with the following structure with the dataset:
3. Write code that answers the following questions.

* What are the coefficient estimates, standard error, confidence interval and P-value of alpha and beta?
* What is the R2 value?
* Include the linear regression fit and confidence intervals onto the scatter plot in 1.

# BONUS Part: Computational Modeling

## Deliverable: Differential Equation

Solve the following system of differential equation by writing a script or notebook that utilizes a **Backward-stepping (Implicit) Euler Method** where α, β, δ and γ are constants.