# GGS 366: Lab 4

## General guidelines

**Please submit the .ipynb notebook containing all the codes.** Make sure the notebook is properly formatted by following the instructions below.

* Separate answers to each question using text cells and markdown headings. See example syntax here: [Markdown Guide - Colab](https://colab.research.google.com/notebooks/markdown_guide.ipynb)
* Write the essay questions in your Python notebook using text cells.
* The function body (where appropriate) must include a docstring to provide a general description of the function.
* Each step of the code needs to be commented.
* The code needs to be properly indented and readable. For more on formatting guidelines: [PEP 8 – Style Guide for Python Code | peps.python.org](https://peps.python.org/pep-0008/) and [PEP 257 – Docstring Conventions | peps.python.org](https://peps.python.org/pep-0257/)

You may use GenAI as a supporting tool. However, directly copying code from GenAI will be considered plagiarism and hinder your learning process. This can negatively impact your performance in course exams and, ultimately, your professional success. Therefore, it is in your best interest to thoroughly understand the fundamental concepts and make a genuine effort to solve the problems on your own.

## Question 1 (8 points)

* Go to the [**Mapping Disparities by Social Determinants of Health | CMS Data**](https://data.cms.gov/tools/mapping-disparities-by-social-determinants-of-health)website.
* Select **Year: 2020** and **State: Virginia**, leaving the **County** and **Census Tract** fields blank.
* Download two datasets:
  + **Educational Attainment** as the domain, selecting **Percentage with Only a High School Diploma** as the measure.
  + **Social and Community Context** as the domain, selecting **Percentage of Owner-Occupied Housing Units** as the measure.
* Load both datasets into **Python** using Pandas.
* Rename the “Value” columns:
  + In the **diploma dataset**, rename to **pct\_diploma**.
  + In the **housing dataset**, rename to **pct\_owner**.
* Use the **groupby()** function to summarize these two variables by county and calculate their **median values**.
* Display each summary dataframe.

## Question 2 (7 points)

* Merge the two datasets of the previous question based on the **GEOID** column.
* Count and print the number of census tracts that meet **at least one** of the following criteria:
  + Housing ownership is **more than 30%**.
  + Percentage of people with only a high school diploma is **more than 65%**.
* Count and print the number of census tracts that meet **both** of the following criteria:
  + Housing ownership is **more than 40%**.
  + Percentage of people with only a high school diploma is **more than 55%**

## Question 3 (10 points)

Using the instructions provided in the lecture, load the **ACS 5-Year (2022) data** for all **census tracts in Maryland (FIPS code: 24)** and extract the following variables:

* **B01001\_001E**: Total population
* **B27010\_017E**: Population under **19 years** without health insurance
* **B27010\_033E**: Population **19-34 years** without health insurance
* **B27010\_050E**: Population **35-64 years** without health insurance
* **B27010\_066E**: Population **65+ years** without health insurance

**Tasks:**

* **Rename the variables as appropriate**
* **Create a new column** (total\_no\_insurance) that calculates the total population **without health insurance** by summing the relevant age group columns.
* **Create another column** (pct\_no\_insurance) to calculate the **percentage** of the total population that lacks health insurance.
* **Estimate the quartile values** (25th, 50th, and 75th percentiles) for pct\_no\_insurance.
* **Using these quartile values**, create a new column (insurance\_quartile) to categorize census tracts into **four groups**:
  + **1st Quartile** (Lowest 25%)
  + **2nd Quartile** (25%-50%)
  + **3rd Quartile** (50%-75%)
  + **4th Quartile** (Highest 25%)

## Question 4 (10 points)

Using the instructions provided in the lecture, load the **Maryland block groups** boundaries from the TIGER/Line Shapefiles (hint: block groups shortened as BG in TIGER/Line Shapefiles websites).

* Check the Coordinate Reference System (CRS) of the shapefile.
* Convert the **CRS to EPSG: 3857**. What differences do you observe in the geometries?
* How many **unique counties** are there in Maryland?
* Calculate the **average block group area** for each county.