EPBI5208 Data Management and Analysis
Final Project

Due Date: December 4, 2024
Data Source: National Health and Nutrition Examination Survey (NHANES), August 2021 –August 2023 (link)

Instructions:
You may use **R**, **SAS**, or both for this project. Include annotated code for each step, with clear explanations of major steps. The project should be submitted as a single word or pdf document.
Clarity and readability are crucial for this project. Ensure that your code, results, discussions, and any outputs are well-organized and easy to follow. Your project will be evaluated based on how well you present and explain each step, so make sure that each part of your work is clear and logically structured. Projects lacking clarity and organization may not be evaluated fully.

Variables:
The 24 variables for this project are listed below:

- Age (RIDAGEYR)
- Gender (RIAGENDR)
- Race/Ethnicity (RIDRETH1)
- Education Level (DMDEDUC2)
- Ratio of family income to poverty (INDFMPIR)
- Health Insurance Coverage (HIQ011)
- Body Mass Index (BMI) (BMXBMI)
- Smoking Status (SMQ040)
- Alcohol Use in Past 12 Months (ALQ121)
- Physical Activity Level (PAD680)
- Hours of Sleep (SLD012)
- Hypertension Diagnosis (BPQ020)
- Diabetes Diagnosis (DIQ010)
- Cancer Diagnosis (MCQ220)
- Coronary Disease Diagnosis (MCQ160C)
- Stroke History (MCQ160F)
- Total Cholesterol (LBXTC)
- Blood Glucose (LBXGLU)
- Prescription for Blood Pressure Medication (BPQ150)
- Prescription for Diabetes Medication (DIQ050)
- Prescription for Cholesterol Medication (BPQ101D)
- Self-Reported AIDS Test Status (HSQ590)
- Depression Score (PHQ-9) (DPQ020)
- Quality of Sleep (DPQ030)

Project Sections

1. **Introduction (5 points)**

   Provide a brief overview of the NHANES dataset and the project's objectives.

2. **Initial Exploration and Data Cleaning (40 points)**

   a. Download and import the NHANES (8/2021 – 8/2023) dataset. Load only the specified 24 variables. Verify each variable's data type (e.g., numerical, categorical) and adjust as necessary. *(5 points)*

   b. Perform an initial exploration to understand the dataset. Check for and document any missing data in each variable. Summarize the percentage of missing data by variable and consider the potential impact of missing data on future analyses. *(10 points)*

   c. Review and document the dataset's structure, noting aspects such as the number of observations, variable types, and overall completeness. *(5 points)*

   d. Conduct checks to confirm that the data is logically consistent (e.g., values fall within expected ranges for continuous variables and categories are valid for categorical variables) and make sure that you fixed any inconsistencies. *(10 points)*

   e. Record key observations and any immediate decisions or adjustments made to prepare the dataset for cleaning and recoding. *(5 points)*

3. **Variable Recoding (20 points)**

   a. Choose 6 variables for recoding: *(15 points)*

      i. Two variables from "Quantitative to Categorical": For example, recode BMI into categories ("Underweight," "Normal Weight," "Overweight," "Obese").

      ii. Two variables from "Quantitative to Binary": For example, creating a "Senior" (≥65) vs. "Non-Senior" (<65) indicator from the age variable.

      iii. Two variables from "Categorical to Binary": Simplify a categorical variable like education level into two categories (e.g., "High School or Less" vs. "More than High School").

   b. Briefly document and justify each recoding decision, specifying the criteria or cutoffs used. *(5 points)*

4. **Descriptive Analysis and Visualization (25 points)**

   a. Calculate descriptive statistics for the variables, including the recoded ones. Pay attention to the accuracy and the consistency of the variable types and the descriptives, i.e. frequencies for categorical and binary variables, and descriptives such as mean or median for quantitative variables. *(10 points)*

   b. Generate visualizations for at least three of the original and three of the recoded variables to illustrate the distribution of your cleaned data. *(10 points)*

   c. Report key results and any notable trends. *(5 points)*

5. **Statistical Analysis (10 points)**

   Select one quantitative variable and one of the recoded binary or categorical variables and then perform a basic comparison (e.g., t-test, chi-square test) to examine differences between groups. Briefly interpret the result of your comparison test and discuss any insights related to public health. *(10 points)*

6. **Reflections (5 points)**

Discuss any challenges faced and reflections on using R or SAS for this project.