Our Question: How do socioeconomic and demographic factors influence self-reported health outcomes, mental health, and access to preventive care?

- Will be composed of answering smaller questions:
    - Self-reported rates of depression/anxiety across age groups (MENTHLTH)
    - Income level and health status relationship (INCOME2, GENHLTH)
    - Physical activity variations by demographics (EXERANY2, PASTAE1)
    - Chronic conditions correlation with socioeconomic factors (DIABETE3, CVDINFR4, ASTHMA3)
    - Health insurance status and preventive care access (HLTHPLN1)

1. What is an observation in your study?

- Each observation corresponds to a survey respondent in the BRFSS dataset (1993-2023). The dataset includes self-reported data on various health behaviors, conditions, and demographics.

2. Are you doing supervised or unsupervised learning? Classification or regression?

- We will use **supervised learning** with **regression** to predict continuous outcomes like health status or mental health measures (e.g., depression or anxiety levels). You might also perform **classification** tasks, such as predicting whether someone will report a health condition based on demographic and socioeconomic factors.

3. What models or algorithms do you plan to use in your analysis? How?

- For regression, we will start with **linear regression** to model relationships between socioeconomic factors (like income, education) and self-reported health outcomes.

- For classification, we will use **logistic regression** or **decision trees** (whichever one performs better) to predict binary outcomes like the presence of chronic conditions or mental health disorders.

- We might use **PCA** (Principal Component Analysis) for dimensionality reduction if the dataset has highly correlated features.

- **Random forests** and **LASSO** can be useful for improving predictive accuracy and handling multicollinearity if previous methods do not give the accuracy we need.

4. How will you know if your approach "works"? What does success mean?

- Success will be measured by **R^2**, **RMSE**, or **classification accuracy** (for categorical outcomes). For regression models, low RMSE and high R^2 are signs of good model performance. For classification, accuracy, **F1 score**, **sensitivity**, and **specificity** can indicate success.

- **Cross-validation** will be used to help assess how well our model generalizes to unseen data.

5. What are weaknesses that you anticipate being an issue? How will you deal with them?

- **Data Missingness:** There may be missing data in some years. Plan to handle this through imputation or by excluding certain rows/columns.

- **Bias in Data:** Early years (1993) may have data skewed toward more affluent populations due to limited phone access. You could consider weighting the data or adjusting for demographic variables.

- **Multicollinearity:** If predictor variables are highly correlated, it could affect regression models. You can use **LASSO** for regularization or **PCA** for dimensionality reduction.

- **Model Overfitting:** This can occur if the model becomes too complex. Use **cross-validation** and **regularization** techniques to mitigate this.

6. Feature Engineering:

- Prepare the data by **one-hot encoding** categorical variables like region or age group.

- If necessary, create **interaction terms** for socio-economic variables that might jointly affect health outcomes.

- Consider **scaling** continuous variables for algorithms sensitive to scale, like logistic regression and decision trees.

7. Results:

- Communicate results through **tables of coefficients** (for regression models) and **confusion matrices** (for classification models).