Problem Identification:

Problem statement formation:

To what extent does perceived vulnerability to health risks — inferred from patterns of preventive health behavior — predict an individual's likelihood to smoke, and how does this relationship change across different age groups?

Context :

Smoking remains one of the leading causes of preventable death and disease. Despite widespread awareness of its risks, smoking persists — and varies substantially across age groups. To better understand why, this project explores how **perceived vulnerability to health issues**, approximated by **engagement in preventive behaviors** (e.g., annual checkups, screenings, insurance coverage), relates to smoking behavior. This relationship is framed using the **Health Belief Model**, which posits that individuals who feel more susceptible to illness are more likely to adopt protective behaviors — such as avoiding smoking.

Using the **Behavioral Risk Factor Surveillance System (BRFSS)**, a rich dataset of self-reported health behaviors among U.S. adults, this study examines whether smoking prevalence is associated with health-risk perception proxies and how this association shifts across age groups.

Criteria for success:

Develop a **logistic regression model** that predicts smoking status (`current smoker` vs `non-smoker`) using behavioral proxies for perceived vulnerability (e.g., routine checkups, chronic illness, health coverage).

Demonstrate **statistically significant relationships** between perceived vulnerability features and smoking likelihood.

Identify whether **age moderates** this relationship — i.e., whether the effect of perceived vulnerability differs for younger vs older adults.

Provide **interpretable marginal effects or interaction plots** to communicate risk profiles clearly.

Use results to inform **behaviorally-informed, age-specific health messaging strategies**.

Scope of solution space:

Focused on survey years **2011–2015** of BRFSS data.

Binary outcome: **current smoker vs. not**.

Methods will include:

- Descriptive EDA
- Logistic regression with interaction terms
- Marginal effect visualization by age group

Constraints:

The data is **self-reported**, which can introduce bias (e.g., underreporting of smoking, overreporting of health behaviors).

Measures of **perceived vulnerability** are **indirect proxies** (e.g., using checkups to imply perceived risk).

Cross-sectional data limits **causal inference** — results will reflect **association, not causation**.

BRFSS uses complex survey design; any generalization must consider appropriate weighting and design corrections.

Stakeholders :

Public health agencies and behavior change policymakers (e.g., CDC, state health departments)

Healthcare providers and insurers interested in behavioral segmentation for smoking prevention

Behavioral scientists and psychologists exploring models of risk perception and health behavior

Public education and communication teams crafting targeted, age-specific anti-smoking campaigns

Data sources:

Dataset: **Behavioral Risk Factor Surveillance System (BRFSS)**

Source: Kaggle BRFSS Repository or CDC BRFSS official site

Sample: U.S. adults aged 18+, one randomly selected individual per household

Years: 2011–2015