📊 Hackathon Data Overview

You'll work with **longitudinal datasets** (data collected from the same people over many years).

🎯 **Goal:**
Detect early *"weak signals"* of health risks using self-reported health and lifestyle data — before clinical diagnosis.

You may use:

- ✅ Any of the 3 recommended datasets below
- ✅ Or any publicly available & properly licensed dataset with self-reported health/lifestyle data

National Longitudinal Survey of Youth 1997 (NLSY97)

🧠 What It Is

Tracks 8,984 Americans born between 1980–1984 from adolescence into adulthood.

👥 Population

Youth aged 12–16 in 1997

📅 Time Span

1997–2022
21 rounds (annual → then every 2 years)

📦 Sample Size

8,984 individuals

🔍 What Data You Get

Health

- Health conditions
- Substance use
- Risk behaviors

Education

- Highest grade completed
- Degrees earned

Economic

- Employment status
- Income
- Poverty ratio
- Net worth

Social

- Family background
- Attitudes & expectations

🔑 Important Variables

Variable	Meaning
PUBID	Unique person ID
CV_HGC_EVER	Highest grade completed
CV_HIGHEST_DEGREE_EVER	Highest degree
CV_ESR	Employment status
CV_INCOME_FAMILY	Family income

CV_HH_POV_RATIO	Poverty ratio

🔓 Access

- Free via NLS Investigator
- Documentation: https://www.bls.gov/nls/nlsy97.htm

Dataset	Best For	Age Group	Years Covered	Sample Size
NLSY97	Youth → adulthood life trajectory	12–16 (start)	1997–2022	8,984

-
- 🎓 Education, employment & health transitions → **NLSY97**