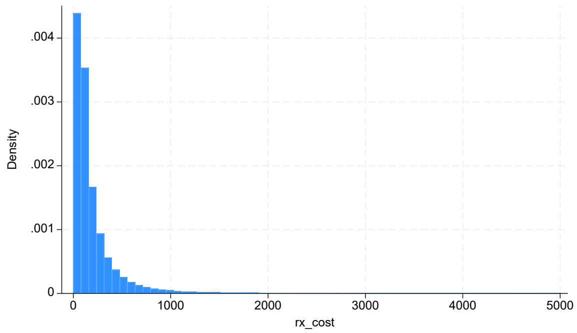
EXAMPLES FOR PART 1: CONCEPTS AND LOGIC (CLOSED-NOTE)

1. You are starting a research project on **whether having a baby leads to fewer hours of sleep**. There are two national data sets that ask questions about sleep habits and family size that you are considering using for this project:

> (1) National Longitudinal Survey of Youth 97: A longitudinal survey of a cohort of people who were teenagers in 1997, surveying those respondents annually from 1997 through 2025.
>
> (2) NHIS: A cross-sectional survey of a sample of Americans that has been conducted repeatedly from the 1950s to 2025.

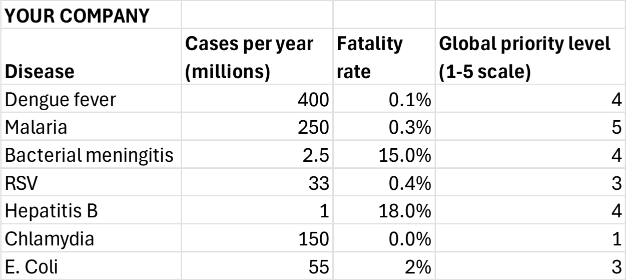
Just based on this information, which of these two datasets is likely to be better suited for your research question and why?

2. We are analyzing data from a nationally representative large-scale survey of 100,000 people that asks them about their out-of-pocket health care spending. You notice that the distribution of out-of-pocket prescription medication spending is very right-skewed (see histogram below). Your supervisor suggests that you could try analyzing "logged spending". **What does your supervisor mean by this, and what steps would you take create a variable of "logged spending"?**

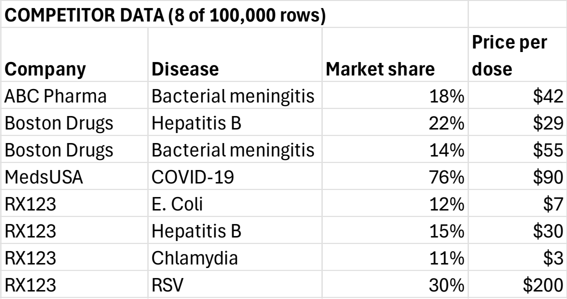


3. We want to use regression analyses to test whether there are differences in out-of-pocket prescription medication spending across different demographic groups. You decide to use survey weights in these analyses. **<u>In a few sentences, explain to your colleagues why you included survey weights in your analysis.</u>**

4. You work at a pharmaceutical company researching treatments for infectious diseases. You have a data set of 7 infectious diseases your company works on, with one row per disease. It lists key facts about each disease in the columns. You obtain a second data set with thousands of rows listing all of the other pharmaceutical companies who also make treatments for hundreds of infectious diseases. (Examples below.) You want to combine these two data sets so you can identify the competing products. Briefly explain the steps you'd take to combine these data sets and your rationale.



YOUR COMPANY			
Disease	**Cases per year (millions)**	**Fatality rate**	**Global priority level (1-5 scale)**
Dengue fever	400	0.1%	4
Malaria	250	0.3%	5
Bacterial meningitis	2.5	15.0%	4
RSV	33	0.4%	3
Hepatitis B	1	18.0%	4
Chlamydia	150	0.0%	1
E. Coli	55	2%	3



COMPETITOR DATA (8 of 100,000 rows)			
Company	**Disease**	**Market share**	**Price per dose**
ABC Pharma	Bacterial meningitis	18%	$42
Boston Drugs	Hepatitis B	22%	$29
Boston Drugs	Bacterial meningitis	14%	$55
MedsUSA	COVID-19	76%	$90
RX123	E. Coli	12%	$7
RX123	Hepatitis B	15%	$30
RX123	Chlamydia	11%	$3
RX123	RSV	30%	$200