# Question 1: Extract, Transform Load (ETL)

## How to run the code

```
python3 p1.py data/nhis_input.csv data/brfss_input.json -o output
```

## Write-up for step 5

Based on the calculated prevalence of diabetes from the joined dataset and the actual prevalence data from the CDC, we can draw several comparisons and conclusions. Here's how the found prevalence in each category stacks up against the actual prevalence reported by the CDC:

### Gender

- **Men**: The calculated prevalence was 13.15%, while the CDC reports a prevalence of 12.6%.
- **Women**: The calculated prevalence was 10.05%, compared to the CDC's reported 10.2%.

The prevalence rates for gender are relatively close to the CDC's statistics, with a slight overestimation for men and an underestimation for women in the calculated data.

### Race/Ethnic Background

- **White, Non-Hispanic**: Calculated prevalence is 11.26% versus the CDC's 8.5%.
- **Black, Non-Hispanic**: Calculated prevalence is 16.03%, compared to the CDC's 12.5%.
- **Asian, Non-Hispanic**: Calculated prevalence is 5.82%, while the CDC reports 9.2%.
- **American Indian/Alaskan Native, Non-Hispanic**: Calculated prevalence is 23.60%, significantly higher than the CDC's 16.0%.
- **Hispanic**: Calculated prevalence is 8.92%, slightly lower than the CDC's 10.3%.

The discrepancies in race/ethnic background suggest a possible overestimation or underestimation of diabetes prevalence in certain groups, notably overestimating in American Indian/Alaskan Native and underestimating in Asian and Hispanic populations.

### Age

- **18-44**: The calculated prevalence of 2.575% is slightly lower than the CDC's 3.0% for a younger demographic.
- **45-64**: The calculated prevalence of 11.352% is lower than the CDC's 14.5%.
- **65+**: The calculated prevalence of 16.091% is significantly lower than the CDC's 24.4%.

The age group analysis indicates an underestimation of diabetes prevalence, especially in the older age groups, highlighting the greatest disparity in the 65+ category.

### Assessment and Improvements

The differences between the calculated and actual prevalence rates suggest a need for adjustment in the methodology. Several factors could account for these discrepancies:

- **Sample Representation**: The BRFSS dataset might not fully represent the U.S. population's diversity or its health status accurately.
- **No Duplication Assumption**: The assumption that there is no duplication between the NHIS and BRFSS datasets during the join operation may not be correct.

To improve the calculated prevalence rates:

- Ensure datasets accurately represent the demographic diversity of the U.S. population.
- Reassess and validate the assumption of no duplication in the datasets.