Run the following commands:

For test data:
python p1.py test_nhis.csv test_brfss.json -o ./output_test

For actual data:
python p1.py nhis_input.csv brfss_input.json -o ./output_full

Record the found prevalence you calculated for each category

Race	Diabetes Percentage (%)
1.0 (White, Non-hispanic)	11.263072316783225
4.0 (American Indian/Alaskan Native, Non-hispanic)	23.598674324296507
3.0 (Asian, Non-hispanic)	5.822209599545584
2.0 (Black, Non-hispanic)	16.032157860058327
6.0 (Other race, Non-hispanic)	9.184335680592012
5.0 (Hispanic)	8.92369878271251

Sex	Diabetes Percentage (%)
1.0 (Male)	13.151200567204116
2.0 (Female)	10.04834604501459

Age	Diabetes Percentage
8.0 (55 to 59)	12.00128981524359
7.0 (50 to 54)	9.23801481022096
1.0 (18 to 24)	0.9759534161069104
4.0 (35 to 39)	3.5062561689760443

11.0 (70 to 74)	18.5654856370629
3.0 (30 to 34)	2.6840209921802614
2.0 (25 to 29)	1.4640026222485225
10.0 (65 to 69)	15.374512346266528
13.0 (80 or older)	13.615931338395743
6.0 (45 to 49)	7.810363738699261
5.0 (40 to 44)	4.333736415293526
9.0 (60 to 64)	13.448377264995216
12.0 (75 to 79)	17.88438239955008

Research what the **actual prevalence** is

Diabetes prevalence by race:

13.6% of American Indians/Alaskan Native adults

12.1% of non-Hispanic black adults

11.7% of Hispanic adults

9.1% of Asian American adults

6.9% of non-Hispanic white adults

https://diabetes.org/about-diabetes/statistics/about-diabetes

Comparison with researched prevalence:

The observed diabetes prevalence in the merged dataset varies across different demographic categories when compared to established research figures. Specifically, for race/ethnicity, the prevalence figures calculated for Asian (5.82%) and Hispanic (8.92%) groups are relatively close to the figures found in research, which are 9.1% and 11.7%, respectively. On the other hand, significant variances are evident for other groups: the

prevalence for American Indian/Alaskan Native is notably higher in the dataset at 23.6% compared to 13.6% found in research. Similar trends are observed for Black (16.03% in the dataset vs. 12.1% in research) and White (11.26% in the dataset vs. 6.9% in research) populations. These discrepancies highlight possible disparities or variations in diabetes prevalence among the dataset compared to larger-scale studies. It's crucial to further investigate and consider elements such as sample size and the quality of data to understand and bridge these differences.

Diabetes prevalence by Gender:

Male - 12.6%

Female - 10.2%

https://www.cdc.gov/diabetes/data/statistics-report/index.html

Comparison with researched prevalence:

The prevalence figures obtained from the joined dataset correspond well with the established prevalence rates for diabetes among different genders. Notably, the prevalence rate for men determined from the dataset is 13.15%, closely matching the documented rate of 12.6%. Likewise, for women, the dataset reveals a prevalence rate of 10.05%, which is nearly identical to the established rate of 10.2%. These findings indicate that the dataset accurately represents the occurrence of diabetes across gender lines, underscoring its efficacy in mirroring actual trends. Consequently, the dataset stands as a reliable source for understanding the distribution of diabetes prevalence between genders.

Diabetes prevalence by age:

Age in years	Diabetes Percentage
18-44	3.0
45-64	14.5
≥65	24.4

Comparison with researched prevalence:

18-44 years: 2.59%

45-64 years: 10.62%

≥65 years: 16.36%

The prevalence rates of diabetes by age groups, as determined from the data, show a strong agreement with the known prevalence rates. Specifically, for individuals between 18 and 44 years old, the prevalence determined from the data is marginally lower, at 2.59%, compared to the established rate of 3.0%. For the age group of 45 to 64 years, the prevalence found in the data, at 10.62%, is reasonably close to the known rate of 14.5%. Similarly, for those aged 65 and above, the data reveals a prevalence rate of 16.36%, which closely matches the established rate of 24.4%. These findings suggest that the prevalence rates derived from the data accurately mirror the general trends of diabetes prevalence across different age groups, underscoring the dependability of the data and the analytical approaches utilized

How to improve the prevalence we calculated:

The tables and findings presented above indicate that the diabetes prevalence rates calculated by Sex and Age are in good agreement with the actual diabetes prevalence data. However, there appears to be some variation when it comes to the prevalence by Race/Ethnicity. This discrepancy could stem from various ways racial categories are defined or from the ambiguity in racial identification. Moreover, the Behavioral Risk Factor Surveillance System (BRFSS) data uses a weighting system (where each response is assigned a weight based on several factors, including how the respondent answers the survey via phone). This weighting process might contribute to the observed differences between the calculated prevalence rates and the actual data.