PREVALENCE:

DIBEV1 Prevalence by Race (_IMPRACE):

- White, Non-Hispanic: 11.26% (29450159 out of 261475361)
- Black, Non-Hispanic: 16.03% (683486 out of 4263219)
- Asian, Non-Hispanic: 5.82% (22140 out of 380268)
- American Indian/Alaskan Native, Non-Hispanic: 23.6% (22928 out of 97158)
- Hispanic: 8.92% (474787 out of 5320518)
- Other race, Non-Hispanic: 9.18% (28446 out of 309723)

DIBEV1 Prevalence by Gender (SEX):

- Male: 13.15% (14266074 out of 108477351)
- Female: 10.05% (16415872 out of 163368896)

DIBEV1 Prevalence by Age Group (_AGEG5YR):

- Age 18–24: 0.98% (126534 out of 12965168)
- Age 25–29: 1.46% (147302 out of 10061594)
- Age 30–34: 2.68% (324417 out of 12086977)
- Age 35–39: 3.51% (431783 out of 12314645)
- Age 40–44: 4.33% (471310 out of 10875373)
- Age 45–49: 7.81% (1207538 out of 15460714)
- Age 50–54: 9.24% (2096156 out of 22690546)
- Age 55–59: 12.0% (3853245 out of 32106924)
- Age 60–64: 13.45% (5255536 out of 39079332)
- Age 65–69: 15.37% (5912198 out of 38454540)
- Age 70–74: 18.57% (5007216 out of 26970563)
- Age 75–79: 17.88% (2381834 out of 13317955)
- Age 80-99: 13.62% (3466877 out of 25461916)

Total prevalence in population:
(14266074 + 16415872) / (108477351 + 163368896) = 0.11286507111 = 11.29% of the population

----

The prevalence by race seems off by this source, but proportionally, the differences between the races seem very similar:
-- "The rates of diagnosed diabetes by race/ethnic background are: 7.6 percent of non-Hispanic whites; 9 percent of Asian Americans; 12.8 percent of Hispanics; 13.2 percent of non-Hispanic blacks; 15.9 percent of American Indians/Alaskan Natives."
(https://diabetescaucus-degette.house.gov/facts-and-figures#:~:text=The%20rates%20of%20diagnosed%20diabetes,of%20American%20Indians%2FAlaskan%20Natives)

--"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#:~:text=13.6%25%20of%20American%20Indians/Alaskan,of%20non%2DHispanic%20White%20adults)

The prevalence in the total population seems roughly accurate, but isn't truly:

-- "Prevalence: In 2021, 38.4 million Americans, or 11.6% of the population, had diabetes." This is close to our 11.29% of the population calculated. At a second glance, our figure from NHIS is for diagnosed, but this website claims a similar percentage for diagnosed and undiagnosed. (https://diabetes.org/about-diabetes/statistics/about-diabetes#:~:text=13.6%25%20of%20American%20Indians/Alaskan,of%20non%2DHispanic%20White%20adults)

It is unclear if our prevalence calculated for ages 65+ is accurate

Our stats for 65 and up:

(5912198 + 5007216 + 2381834 + 3466877) / (38454540 + 26970563 + 13317955 + 25461916) = 0.16091482359 = 16.09% of the population over 65.

Note that this is explicitly "diagnosed"

-- The percentage of Americans age 65 and older remains high, at 29.2%, or 16.5 million seniors (diagnosed and undiagnosed)." (https://diabetes.org/about-diabetes/statistics/about-diabetes#:~:text=13.6%25%20of%20American%20Indians/Alaskan,of%20non%2DHispanic%20White%20adults)

Conclusion here is unclear with current stats.

For the prevalence difference between sexes:

--"Total and diagnosed diabetes prevalence was higher in men (18.0% and 12.9%, respectively) than in women (13.7% and 9.7%, respectively)." (https://www.cdc.gov/nchs/products/databriefs/db516.htm#:~:text=Total%20and%20diagnosed%20diabetes%20prevalence,increased%20with%20increasing%20weight%20status.)

We have the following (diagnosed) prevalence:

- Male: 13.15% (14266074 out of 108477351)

- Female: 10.05% (16415872 out of 163368896)

Our diagnosed prevalences line up pretty well with the CDC. Differences may be due to our NA value exclusion (the way we handled missing data), the dates of the reports,
or the sampling techniques used to derive the data (if sampling techniques were used for either)

In order to improve the prevalence statistics that we found,

While our calculated prevalence is close to reputable statistics found online, especially for gender and overall rates, there are some discrepancies.
This may stem from a few things, including imperfect sampling techniques, our method of nullifying entire rows with NA values (as opposed to imputing),
and our generally imperfect data (due to missing values). Also, we could improve our analysis accounting properly for both diagnosed and undiagnosed diabetes.
Our analysis doesn't considered undiagnosed prevalence, which is an important assumption that will lead to inaccuracies in our final reporting (as there
are clearly large percentages of the population that go undiagnosed, and we completely discount that in our reporting).

Running the code (documentation):

Simply run:

spark-submit p1.py ./data/brfss_input.json ./data/nhis_input.csv -o ./data/joined_output

- The prevalence report will show up in a file “summary_stats.txt” within the same directory that the script is run.
- Data is assumed to exist in a data folder, and the output is written to a subdirectory called joined_output within the data directory. This incudes the coalesced joined output in a csv.