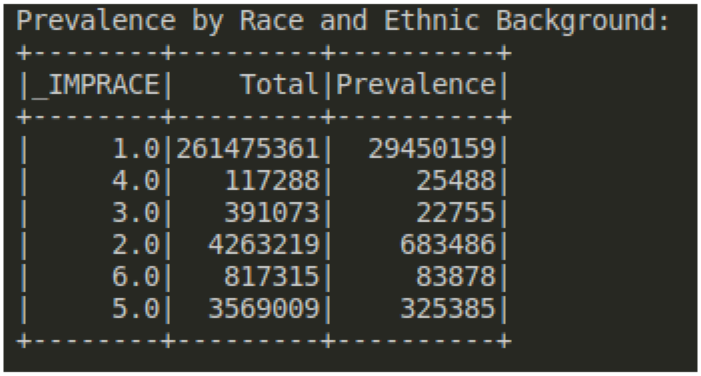
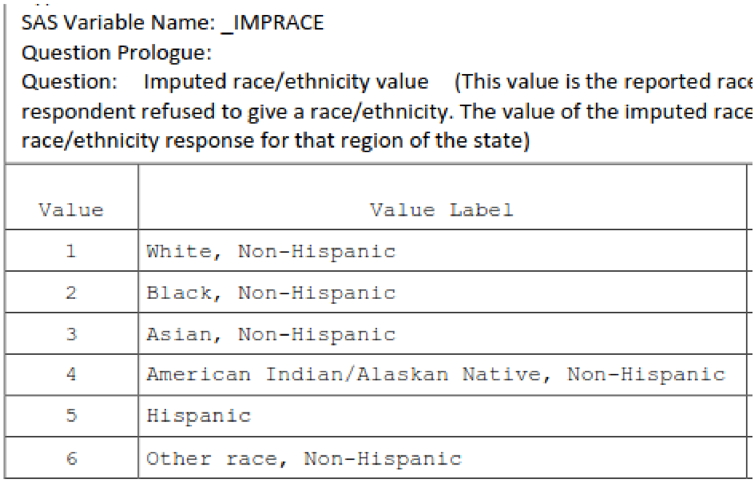
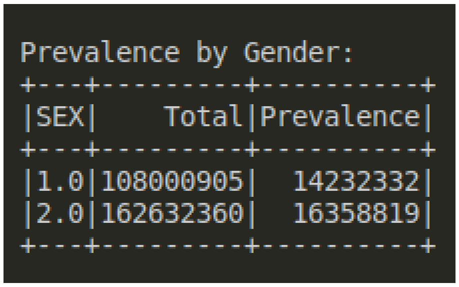


```
Prevalence by Race and Ethnic Background:
+--------+---------+----------+
|_IMPRACE|    Total|Prevalence|
+--------+---------+----------+
|     1.0|261475361|  29450159|
|     4.0|   117288|     25488|
|     3.0|   391073|     22755|
|     2.0|  4263219|    683486|
|     6.0|   817315|     83878|
|     5.0|  3569009|    325385|
+--------+---------+----------+
```

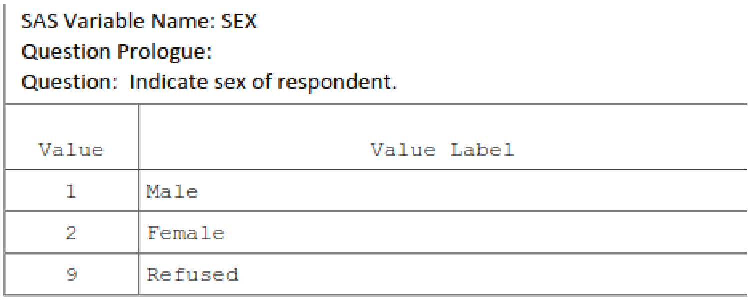


SAS Variable Name: _IMPRACE
Question Prologue:
Question: Imputed race/ethnicity value (This value is the reported race
respondent refused to give a race/ethnicity. The value of the imputed race
race/ethnicity response for that region of the state)

Value	Value Label
1	White, Non-Hispanic
2	Black, Non-Hispanic
3	Asian, Non-Hispanic
4	American Indian/Alaskan Native, Non-Hispanic
5	Hispanic
6	Other race, Non-Hispanic

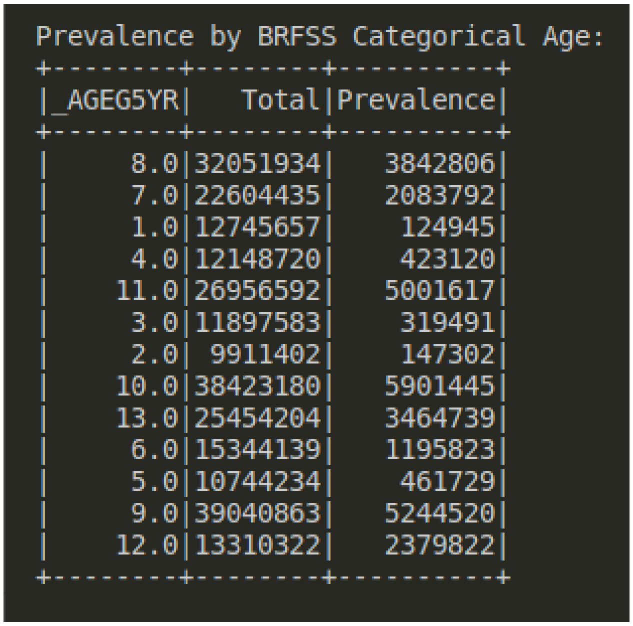


```
Prevalence by Gender:
+---+---------+----------+
|SEX|    Total|Prevalence|
+---+---------+----------+
|1.0|108000905|  14232332|
|2.0|162632360|  16358819|
+---+---------+----------+
```

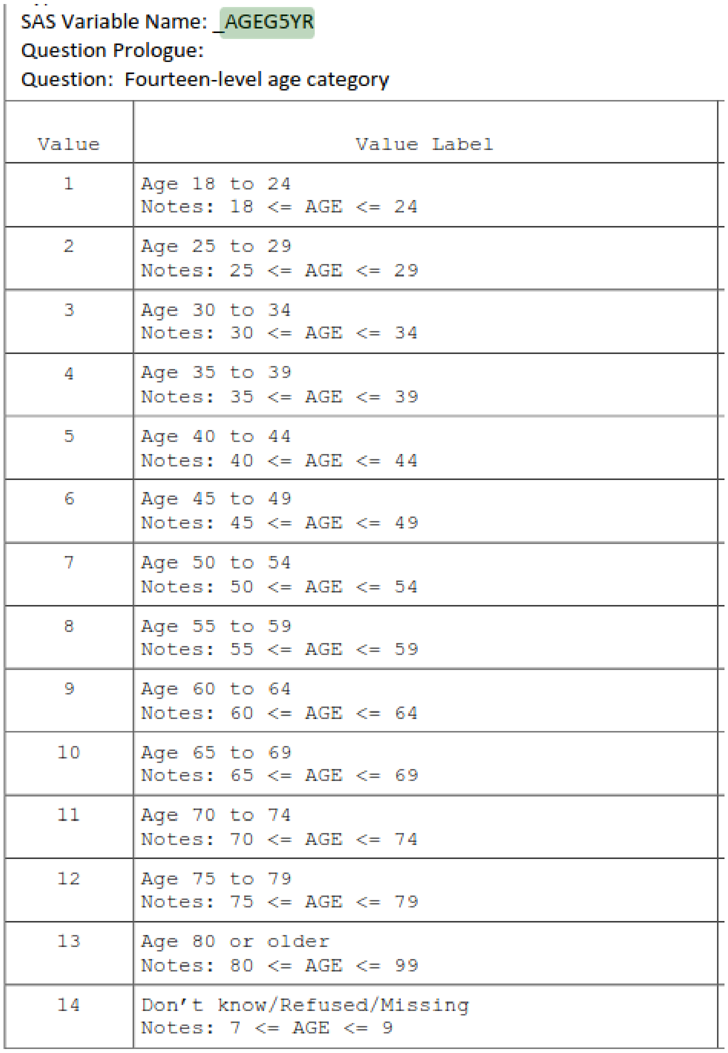


SAS Variable Name: SEX
Question Prologue:
Question: Indicate sex of respondent.

Value	Value Label
1	Male
2	Female
9	Refused



```
Prevalence by BRFSS Categorical Age:
+--------+--------+----------+
|_AGEG5YR|   Total|Prevalence|
+--------+--------+----------+
|     8.0|32051934|   3842806|
|     7.0|22604435|   2083792|
|     1.0|12745657|    124945|
|     4.0|12148720|    423120|
|    11.0|26956592|   5001617|
|     3.0|11897583|    319491|
|     2.0| 9911402|    147302|
|    10.0|38423180|   5901445|
|    13.0|25454204|   3464739|
|     6.0|15344139|   1195823|
|     5.0|10744234|    461729|
|     9.0|39040863|   5244520|
|    12.0|13310322|   2379822|
+--------+--------+----------+
```



SAS Variable Name: _AGEG5YR
Question Prologue:
Question: Fourteen-level age category

Value	Value Label
1	Age 18 to 24 Notes: 18 <= AGE <= 24
2	Age 25 to 29 Notes: 25 <= AGE <= 29
3	Age 30 to 34 Notes: 30 <= AGE <= 34
4	Age 35 to 39 Notes: 35 <= AGE <= 39
5	Age 40 to 44 Notes: 40 <= AGE <= 44
6	Age 45 to 49 Notes: 45 <= AGE <= 49
7	Age 50 to 54 Notes: 50 <= AGE <= 54
8	Age 55 to 59 Notes: 55 <= AGE <= 59
9	Age 60 to 64 Notes: 60 <= AGE <= 64
10	Age 65 to 69 Notes: 65 <= AGE <= 69
11	Age 70 to 74 Notes: 70 <= AGE <= 74
12	Age 75 to 79 Notes: 75 <= AGE <= 79
13	Age 80 or older Notes: 80 <= AGE <= 99
14	Don't know/Refused/Missing Notes: 7 <= AGE <= 9

Research:

The rates of diagnosed diabetes in adults by race/ethnic background are:

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

Prevalence in seniors: The percentage of Americans age 65 and older remains high, at 29.2%, or 16.5 million seniors (diagnosed and undiagnosed).

[Sources:
https://diabetes.org/about-diabetes/statistics/about-diabetes#:~:text=Diabetes%20by%20race%2Fethnicity&text=12.1%25%20of%20non%2DHispanic%20black,of%20non%2DHispanic%20white%20adults]

The study findings showed that out of 590 patients with diabetes, 310 (52.5%) were males and 280 (47.5%) were females.

[Sources: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10071047/]

Comparison and observation:

Age: The result shows that categories _AGEG5YR 9,10,11 have noticeable higher Prevalence, meaning that the chances of finding a diabetes patient to be greater than or equal to 60 years old is higher.
The dataset's age-related prevalence follows a similar trend to the actual prevalence, with noticeably higher prevalence in older age groups. However, the dataset's prevalence rates might be higher, and the specific age groups may need adjustment for better alignment.

Gender: Prevalence by Gender shows that the number of female patients is slightly greater than male patients. The overall trend seems that gender is not a major factor that increases the likelihood to get diabetes.
The dataset's gender prevalence does not align with the research findings. Females in the dataset show slightly higher prevalence, whereas the actual prevalence is slightly higher in the male group.

Race: The found prevalence generally aligns with actual prevalence trends. White individuals in the dataset show higher prevalence, consistent with the actual prevalence. Black and Hispanic populations also exhibit notable prevalence, resembling real-world trends. The dataset's Asian population's prevalence is relatively higher compared to the actual prevalence, suggesting potential discrepancies or variations.

How to run the code:

```
python3 p1.py -o path/to/output brfss_input.json nhis_input.csv
spark-submit p1.py /path/to/brfss.json /path/to/nhis.csv -o path_to_output
```