Write-up Part 1

0. Run Code:

Structure your data so that the brfss_input.json file and the nhis_input.csv files are in the folder: ./data/p1/

Then run command: '" spark-submit p1.py brfss_input.json nhis_input.csv -o output/ "'

1. Record the found prevalence you calculated for each category

Sex	Description	Percentage
1	Male	13.15%
2	Female	10.05%

Race	Description	Percentage
1	White	11.26%
2	Black	16.03%
3	Asian	5.82%
4	American Indian/ Alaskan	23.60%
5	Hispanic	8.92%
6	Other Race	9.18%

Age-Range	Description	Percentage
1	Age 18 to 24	12.00%
2	Age 25 to 29	9.24%
3	Age 30 to 34	0.98%
4	Age 35 to 39	3.51%
5	Age 40 to 44	18.57%
6	Age 45 to 49	2.68%
7	Age 50 to 54	1.46%
8	Age 55 to 59	15.37%
9	Age 60 to 64	13.62%
10	Age 65 to 69	7.81%
11	Age 70 to 74	4.33%
12	Age 75 to 79	13.44%
13	Age 80 or older	17.88%

2. Research what the actual prevalence is

As of 2021, the Centers for Disease Control and Prevention (CDC) reported these statistics:

Sex	Description	Percentage
1	Male	9.00%
2	Female	9.70%

Race	Description	Percentage
1	White	7.50%
2	Black	12.30%
3	Asian	8.40%
4	American Indian/ Alaskan	14.30%
5	Hispanic	13.50%
6	Other Race	Not Found

Age-Range	Description	Percentage
1	Age 18 to 44	3.10%
2	Age 45 to 64	11.50%
3	Age 65 to 74	22.90%
13	Age 75 or older	22.00%

3. Write a short paragraph comparing your found prevalence within the joined dataset to the actual prevalence, by gender, race/ethnic background, and age

Comparing the actual deta reported by the CDC and the computed prevalence data of diabetes, we find some significant class imbalances. Gender-wise, males with positive diabetes cases are significantly over-represented, while females just slightly. This might be a result of the self-reporting nature of the survey, that might result in males generically being more open about their medical condition.

Race-wise, while positive diabetes cases in the white, black, asian and american indian population are consistently over-represented in BRFSS by 3-4%, they are significantly under-represented in the Latino population (by roughly 5%). Over-representations are normal in BRFSS since the survey is deliberately over-representing older sub-strata of the population to catch more data on medical conditions, so the under-representation in the Hispanic community is interesting.

Finally Age-wise, CDC does not provide data for the exact same age-ranges as BRFSS so it is a bit more complex to carry an analysis. However, we can assume that the diabetes prevalences are roughly linear within the age-range. By doing so, we notice that diabetes cases in the younger age-ranges (18-44) are severly over-represented, while the survey is more accurate for medium ranges and even under-represents the older age-ranges (65 and older).

4. Assess how you might improve the prevalence you calculated.

The BFRSS data is strongly biased towards certain categories, in all three demographics. Firstly regarding age-range, the study deliberately oversamples older age-ranges as we can see from the tables above for 50+ year old ranges. That is because the survey captures healthcare information of the respondents, hence oversampling elders can lead to a better presence of certain health conditions that are more rare in the younger population. Secondly regarding gender, it seems to overrepresent females, probably mainly because BFRSS deliberately oversamples older age ranges to gather more accurate data points on certain healthcare conditions. Due to females' higher life's expectancy, this oversampling might have introduced some inbalance in gender towards females. Finally regarding race, white is extremely overrepresented, taking 96% of the whole dataset. This is probably the result of a mixture of components of the survey. Firstly, the survey is carried through landline phones, which statistically are more present in the white population. Furthermore, the study oversamples older age ranges, that again have an inbalance towards white people.

Because of the significant unbalance in the datasets, the computed prevalence data is intrinsically biased too. To improve it, we could try to weight datapoints differently based on their demographics combination. An easy way to skew the prevalences more towards the actual distributions of the US population would be to apply weightings to each demographic combination based on the actual distribution of genders, ethnicities and age ranges in the US.