p1_write_up

Chenran Ning (cn257)

1. Prevalence for each category

```
Prevalance in Race and ethnic background
White, Non-Hispanic : 0.11263072316783225
Black, Non-Hispanic : 0.16032157860058327
Asian, Non-Hispanic : 0.05818606756283354
American Indian/Alaskan Native, Non-Hispanic : 0.21731123388581952
Hispanic : 0.09116956555727374
Other race, Non-Hispanic : 0.10262628240029854

Prevalance in Gender:
Male:  0.1317797475863744
Female:  0.10058772436186747

Prevalance in BRFSS categorical age
Age 18 to 24  :  0.009802946995984593
Age 25 to 29  :  0.01486187322439348
Age 30 to 34  :  0.02685343737463315
Age 35 to 39  :  0.034828360518639
Age 40 to 44  :  0.04297458525195933
Age 45 to 49  :  0.07793353540397412
Age 50 to 54  :  0.09218509553545576
Age 55 to 59  :  0.11989310847825907
Age 60 to 64  :  0.1343341206366263
Age 65 to 69  :  0.15359074912591827
Age 70 to 74  :  0.18554337284178948
Age 75 to 79  :  0.1787952237368863
Age 80 or older  :  0.1361165723351632
```

These data are the output of my `calculate_statistics(joined_df)` function to calculate each prevalance for each categories.

The function is defined below:

```
def calculate_statistics(joined_df):
    """
    Calculate prevalence statistics

    :param joined_df: the joined df

    :return: None
    """
```

```
    #add your code here
    # Race and ethnic background
    # Gender
    # BRFSS categorical age
    # SEX,_LLCPWT,_AGEG5YR,_IMPRACE, DIBEV1
    df = joined_df

    with open('output.txt', 'w') as f:
        # Race
        races = {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"}
        print("Prevalance in Race and ethnic background", file = f)
        for key, value in races.items():
            prevalance = df.filter((df._IMPRACE == key) & (df.DIBEV1 == 1)).count() /
df.filter(df._IMPRACE == key).count()
            print(value + " : " + str(prevalance), file = f)

        # Gender
        male = df.filter((df.SEX == 1) & (df.DIBEV1 == 1)).count() / df.filter(df.SEX
== 1).count()
        female = df.filter((df.SEX == 2) & (df.DIBEV1 == 1)).count() / df.filter(df.SEX
== 2).count()
        print("\nPrevalance in Gender:", file = f)
        print("Male: ", male , file = f)
        print("Female: ", female, file = f)

        # BRFSS categorical age
        ages = {1:"Age 18 to 24",2:"Age 25 to 29",3:"Age 30 to 34",4:"Age 35 to
39",5:"Age 40 to 44",
                6:"Age 45 to 49",7:"Age 50 to 54",8:"Age 55 to 59",9:"Age 60 to
64",10:"Age 65 to 69",
                11:"Age 70 to 74",12:"Age 75 to 79",13:"Age 80 or older"}
        print("\nPrevalance in BRFSS categorical age", file = f)
        for category, title in ages.items():
            prevalance = df.filter((df._AGEG5YR == category) & (df.DIBEV1 ==
1)).count() / df.filter(df._AGEG5YR == category).count()
            print(title, " : ", prevalance, file = f)

    return
```

2. Research

- Gender

```
males was 14.0% and 12.8% among females

# calculated by me
Prevalance in Gender:
Male:  0.1317797475863744
Female:  0.10058772436186747
```

- Race

```
14.5% of American Indians/Alaskan Natives
12.1% of non-Hispanic blacks
11.8% of Hispanics
9.5% of Asian Americans
7.4% of non-Hispanic whites

# Calculated by me
White, Non-Hispanic : 0.11263072316783225
Black, Non-Hispanic : 0.16032157860058327
Asian, Non-Hispanic : 0.05818606756283354
American Indian/Alaskan Native, Non-Hispanic : 0.21731123388581952
Hispanic : 0.09116956555727374
Other race, Non-Hispanic : 0.10262628240029854
```

- Ages

```
18-44 3.3% 45-64 11.7% >=65 11.5%
# Calculated by me
Prevalance in BRFSS categorical age
Age 18 to 24  :  0.009802946995984593
Age 25 to 29  :  0.01486187322439348
Age 30 to 34  :  0.02685343737463315
Age 35 to 39  :  0.034828360518639
Age 40 to 44  :  0.04297458525195933
Age 45 to 49  :  0.07793353540397412
Age 50 to 54  :  0.09218509553545576
Age 55 to 59  :  0.11989310847825907
Age 60 to 64  :  0.1343341206366263
Age 65 to 69  :  0.15359074912591827
Age 70 to 74  :  0.18554337284178948
Age 75 to 79  :  0.1787952237368863
Age 80 or older  :  0.1361165723351632
```

3. Comparison

- For gender categories, my data for Male: 0.1317797475863744 and Female: 0.10058772436186747 are pretty close to the actual data which was 14.0% among males and 12.8% among females. Prevalence in males are a little bit higher than in females.

- For ages, the prevalence for each age group is about two times of the actual dataset. I think this may be caused by redundant data in the origin dataset which contributed to multiple duplicates.
- For races, the prevalences are also greater than the searched result. I think it's the same problem above.

4. Access

There may be some duplicates because of the collecting method of BRFSS. As mentioned in the documentation:

```
users should understand that the data set
they need is based on the location of the questions either in the core or in optional
modules. Users
should keep in mind that there are 4 possible data sets from which they will need to
include data
```

Thus, these datasets may have duplicates because one person's data may exist in multiple different datasets.