**Introduction:**

**The first dataset (‘*NhanesDemoAdapted.csv*’) provides the following information:**

1. SEQN – Integer: ID of person

2. Gender – String: ['Female', 'Male']

3. Age – Integer: age of person (0-80, 80 given for anyone aged 80 or over)

4. Ethnicity – String: ['Black', 'White', 'Asian', 'Mexican-American', 'Other Hispanic', 'Others']

5. US born – Integer: {1: Us born, 2: Other}

6. Education – Integer: Highest education level achieved for persons aged 20+ {1:<9th Grade,

2: 9th-11th grade, 3:HighSchool graduate, 4:Some college, 5: College graduate or above}

7. Marital Status – Integer: Persons aged 20+ {1:Married, 2:Widowed, 3:Divorced, 4:Separate, 5: Single, 6: Living with Partner}

8. HouseholdSize – Integer: Number of people in house {1-7, 7 given for houses with 7 or more

people}

9. AgeUnder6 – Number of household members under the age of 6

10. Age6to18– Number of household members between the ages of 6 and 18

11. AgeOver60 – Number of household members over the age of 60

12. HouseholdIncome – Total household income in 1000s of dollars

13. IncomePovertyRatio – Float: Ratio of family income to poverty {0.0-5.0, 5 given for ratios of 5 or higher}

There are 9254 observations and 13 predictors (variables).

**The second dataset (*‘NhanesFoodAdapted.csv’*) provides the following information:**

1. SEQN – Integer: ID of person

2. dGRMS – Float: Gram weight of meal

3. dKCAL – Integer: Energy (kcal) of meal

4. dPROT – Float: Protein content (gm) of meal

5. dCARB – Float: Carbohydrate content (gm) of meal

6. dSUGR – Float: Total sugars content (gm)t of meal

7. dFIBE – Float: Dietary fibre content (gm) of meal

8. dTFAT – Float: Total fat content (gm) of meal

9. dSFAT – Float: Total saturated fatty acids (gm) of meal

10. dCHOL – Float: Cholestorol (mg) of meal

11. dVITC – Float: Vitamin C content (mg) of meal

12. dVITD – Float: Vitamin D content (mcg) of meal

13. dCALC – Float: Calcium content (mg) of meal

14. dCAFF – Float: Caffeine content (mg) of meal

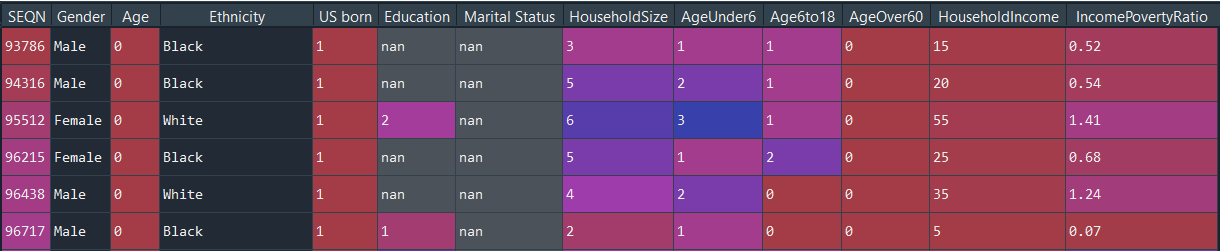
15. dALCO – Float: Alcohol content (gm) of meal

There are 9891 observations and 15 predictors (variables).

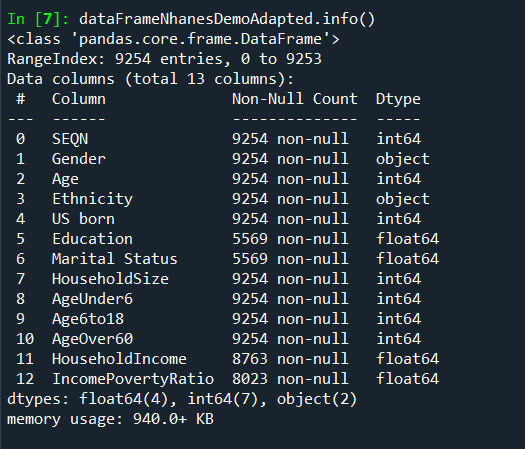
**Data Cleaning:**

Before performing any operation in the datasets, the data cleaning must be done. First, we need to identify the data that has categorical information.

**Sample of Demo dataset:**

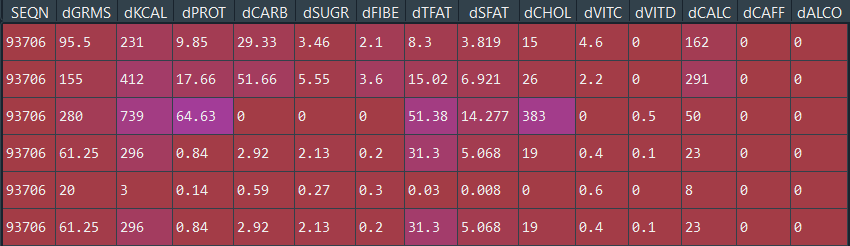


**Structure of Demo dataset:**

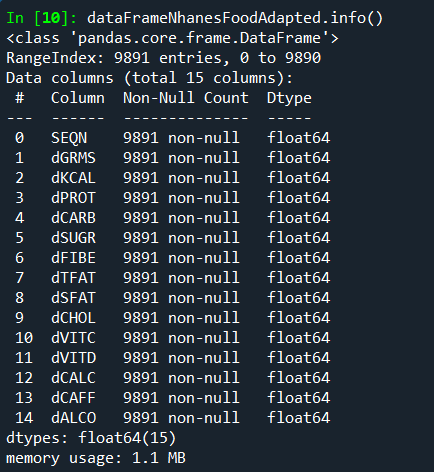


In the Demo dataset above, the variables "US born", "Education", "Marital Status", "HouseholdSize" are categorical data, despite holding numeric values. These variables/predictors must be converted to string type. This does not apply to the Diet dataset displayed below, since there is no categorical information.

**Sample of Food/Diet dataset:**

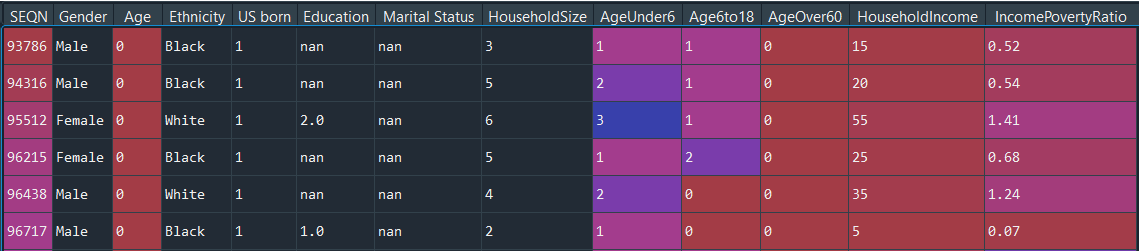


**Structure of Food/Diet dataset:**

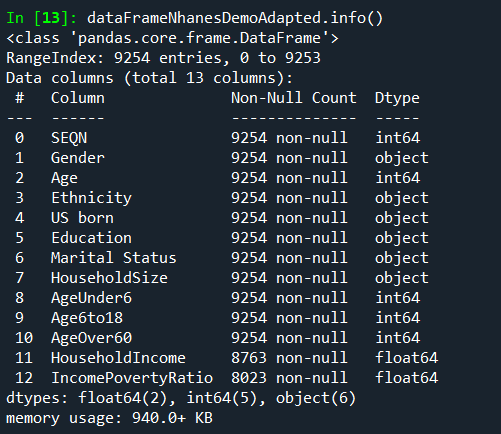


The converted Demo data set is displayed below. If there is any missing data, it would also be converted as string. For example, nan or na would become "nan" or "na" respectively.

**Sample of Demo dataset after conversion:**

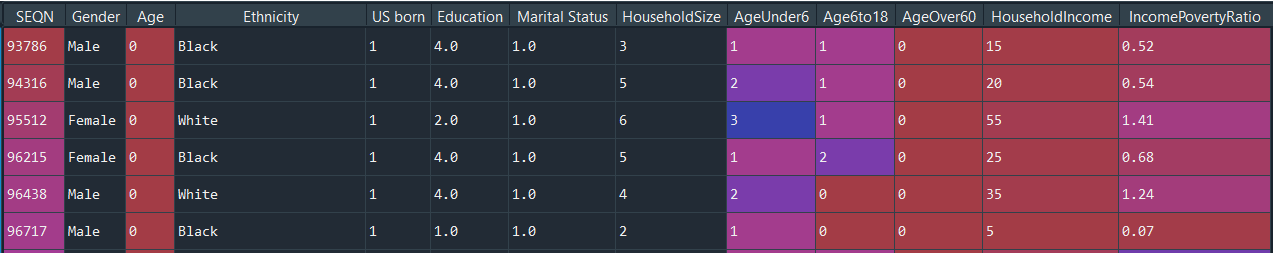


**Structure of Demo dataset after conversion:**

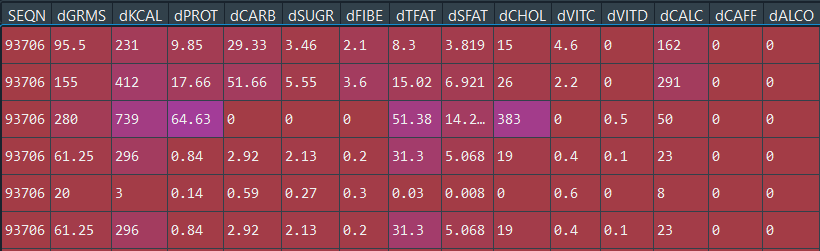


If there are missing values in the numeric/float data in a column, then they can be replaced with its mean. If there are missing values in the categorical data in a column, then they can be replaced with its mode.

**Sample of Demo dataset after data cleaning:**

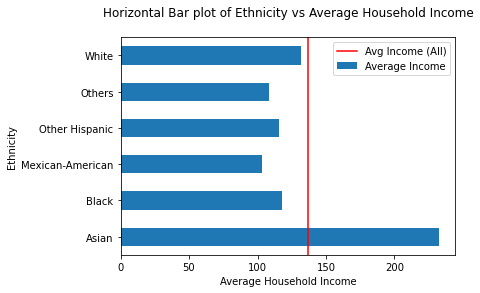


**Sample of Food/Diet dataset after data cleaning:**

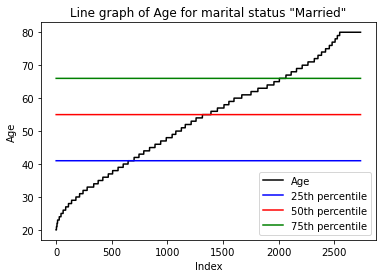


Apparently, there are no missing values in the Diet Dataset.

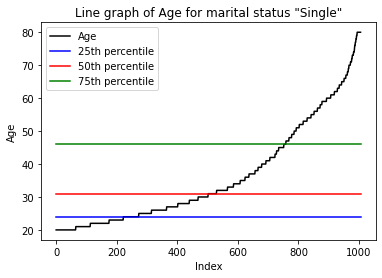
From the Horizontal bar plot below, we can see that among the adults, the Asians hold the highest average household income and the Mexican-Americans hold the least average household income. The average income of Blacks is higher than the Hispanics'. The Others have the second least average income. When compared to the average income of all the adults, the average of the Asians is higher. The rest of the averages fall below the overall average. This means that there are social disadvantages as the Asians are the only people with the highest average income, while the rest are below average.



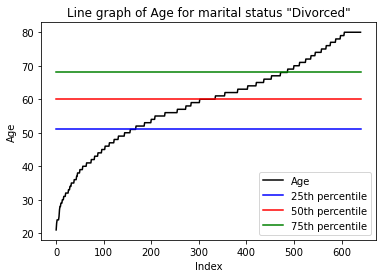
The red vertical line denotes the average household income of all the adults which is equal to 136.8 (in 1000s of dollars). From the line graph below, we can see that out of all the adults, the highest percentage belongs to the Married people. Among them, 25% of peoples' age is lesser than 41, 50% of peoples' age is less than 55 and 75% of peoples' age is more than or equal to 66.



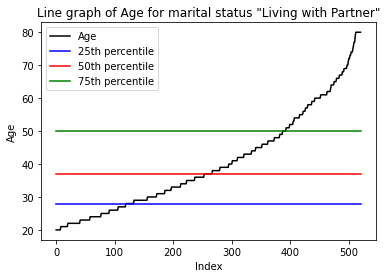
From the line graph below, we can see that the next highest percentage is from the category of singles. Among them, 25% of peoples' age is lesser than 24, 50% of peoples' age is less than 31 and 75% of peoples' age is more than or equal to 46.



From the line graph below, we can see that the 3rd highest percentage is from the category of divorced people. Among them, 25% of peoples' age is lesser than 51, 50% of peoples' age is less than 60 and 75% of peoples' age is more than or equal to 68.



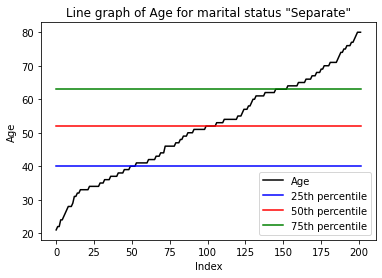
From the line graph below, we can see that the 4th highest percentage is from the category of people living with their partners. Among them, 25% of peoples' age is lesser than 28, 50% of peoples' age is less than 37 and 75% of peoples' age is more than or equal to 50.



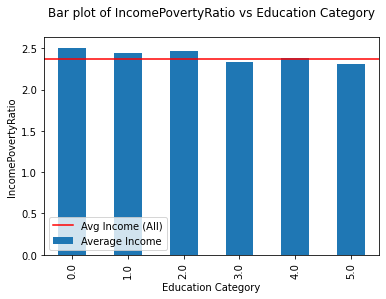
From the line graph below, we can see that the 2nd least percentage is from the category of people who are widowed. Among them, 25% of peoples' age is lesser than 65, 50% of peoples' age is less than 75 and 75% of peoples' age is more than or equal to 80.



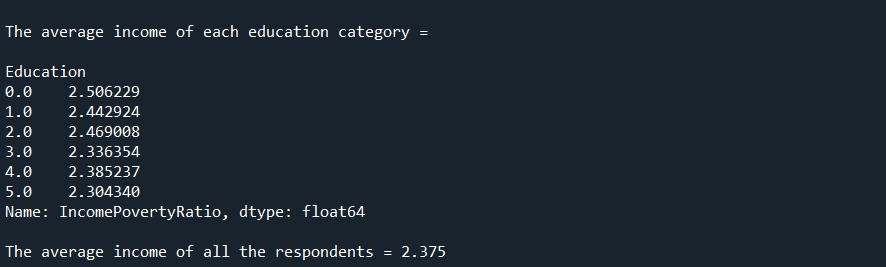
From the line graph below, we can see that the least percentage is from the category of people who are separated from their partners. Among them, 25% of peoples' age is lesser than 40, 50% of peoples' age is less than 52 and 75% of peoples' age is more than or equal to 63.



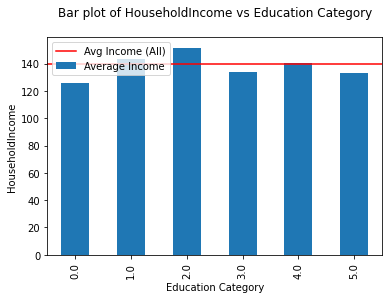
From the Simple bar plot below, we can see that the IncomePovertyRatio is 99% uniform across all education categories. The average IncomePovertyRatio of all respondents = 2.375. Category 0 has the highest ratio, whereas the people with 9th-11th grade as the highest education level hold the 2nd highest average income ratio. The income of respondents who at max have completed 9th grade have an income greater than people who have graduated in some college. Respondents who are either High School graduates or college graduates hold the similar income. Overall, 67% of the categories fall slightly below the average line and 33% of the categories fall slightly above the average line. This is almost a uniform distribution.



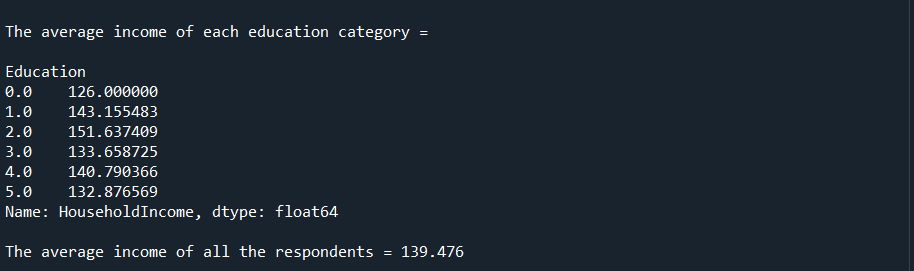
The below image displays the average income of each category and the average income of all the respondents.



From the Simple bar plot below, we can see that the HouseholdIncome is almost uniform across all education categories. The average HouseholdIncome of all respondents = 139.476. Category 0 has the least income, whereas the people with 9th-11th grade as the highest education level hold the highest average income. The income of respondents who at max have completed 9th grade have an income greater than people who have graduated in some college. Both categories are just above the overall average mark. Respondents who are either High School graduates or college graduates hold the similar income. Overall, 50% of the categories fall slightly below the average line and 50% of the categories fall slightly above the average line. This is akin to a uniform distribution.



The below image displays the average household income of each category and the average income of all the respondents.



From this we can say that Higher level of education does not mean that one can get a higher income.

**Bibliography:**

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.replace.html>

<https://pandas.pydata.org/pandas-docs/stable/development/index.html#development>

<https://matplotlib.org/api/pyplot_api.html>

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.api.types.is_integer_dtype.html>