Background: This research paper intends to construct a model of alcohol beverage consumption and education, especially in terms of the number of high school graduates in the United States. We would like to **analyze the impact on alcohol beverage consumption by the change of the contemporary US education condition**, which can be reflected by a number of variables, such as the number of high school graduates, education investment per capita, and average school funding in the US.

Model Specification - briefly present a conceptual model that follows your problem statement. In this section, you should draw on **economics/finance theory if relevant.** What theory is appropriate? What are the variables in the model and how are they related? What are the "expected" signs for the parameters in the assumed model you have specified (Note: these are your hypotheses.)

Link: <https://www.bls.gov/data/>

<https://www.bls.gov/cex/2019/aggregate/educat.pdf>

<https://github.com/yiyichen1717/Healthcare-Data-Analytics/blob/master/Diabetes_rev3%20Zezhi%20Z.ipynb>

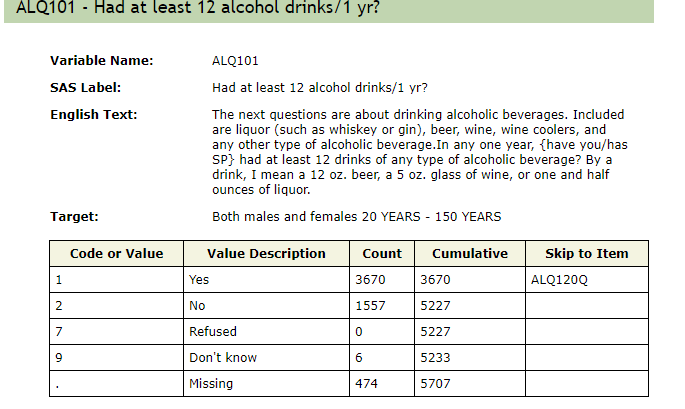
Data - describe your data set. List the variables that will be used in your econometric model. Identify all sources that you used to collect the data in proper bibliographic form. Also, discuss briefly the sample that you used. Was it a time-series? If so for what years were data collected? What is the frequency?

# of high school graduates, education expansion,

Database: NHANES

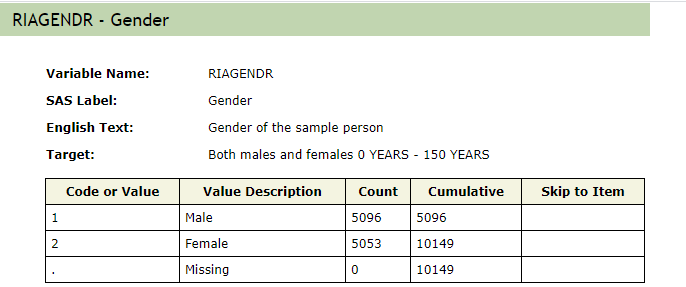
Dataset: from 2007-2016

Dependent variable: Alcohol consumption

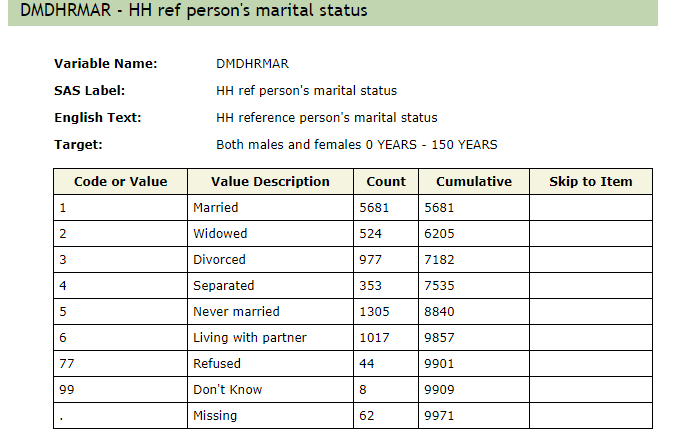


Features:

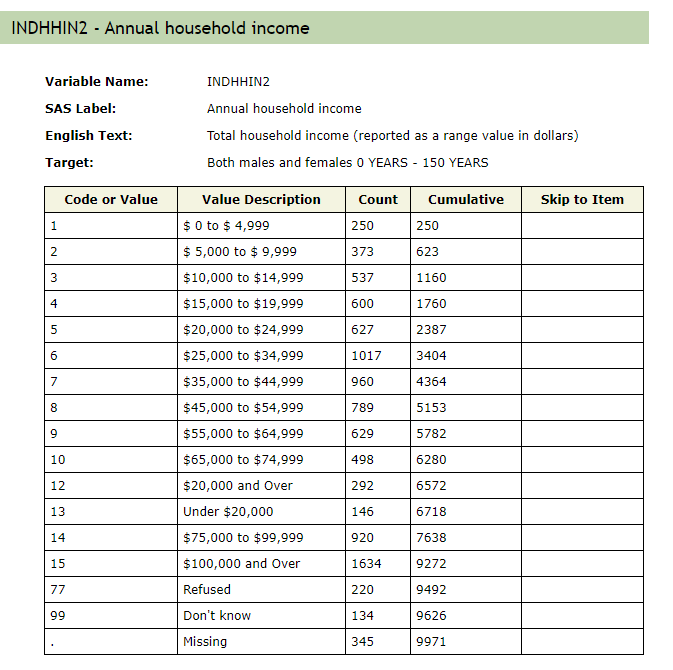
1. Demographic
   1. Gender



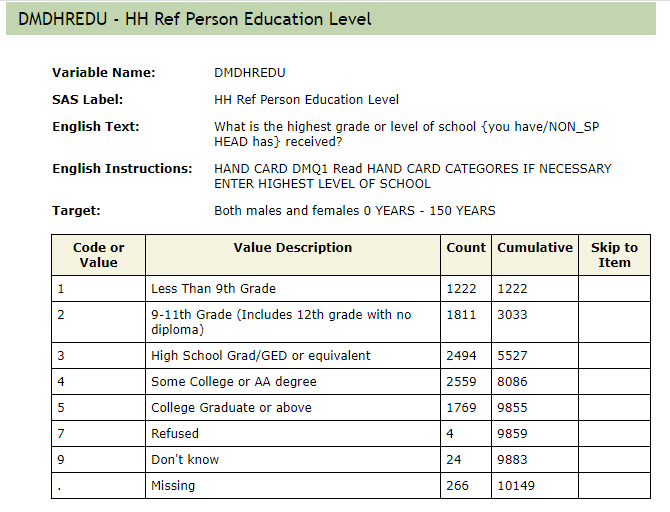
b. marital status,



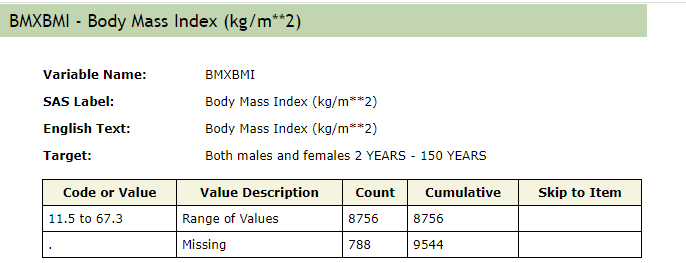
c. Annual household income



d. Education level



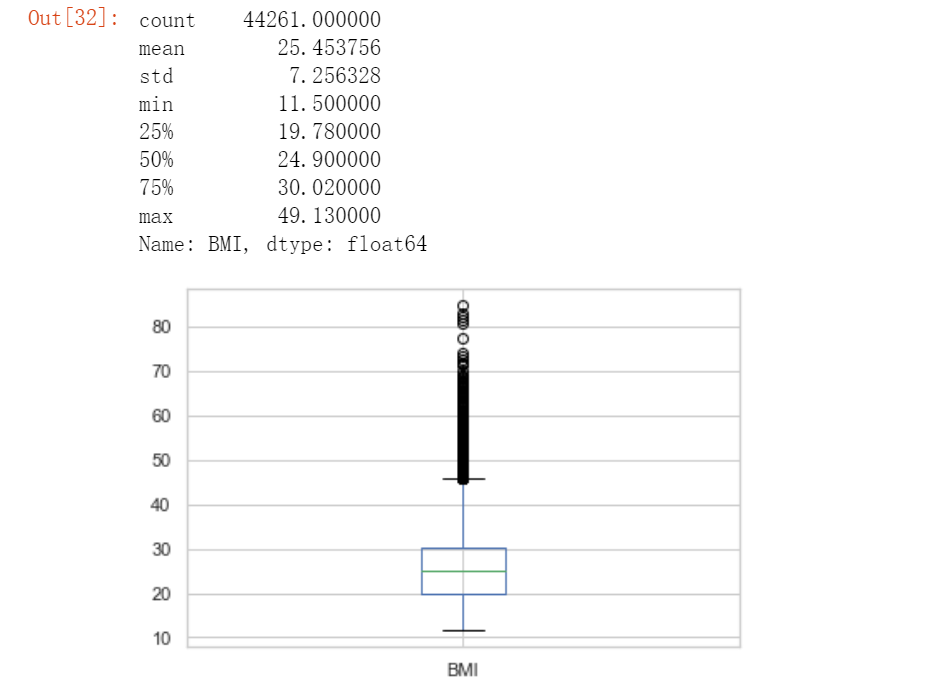
1. BMI

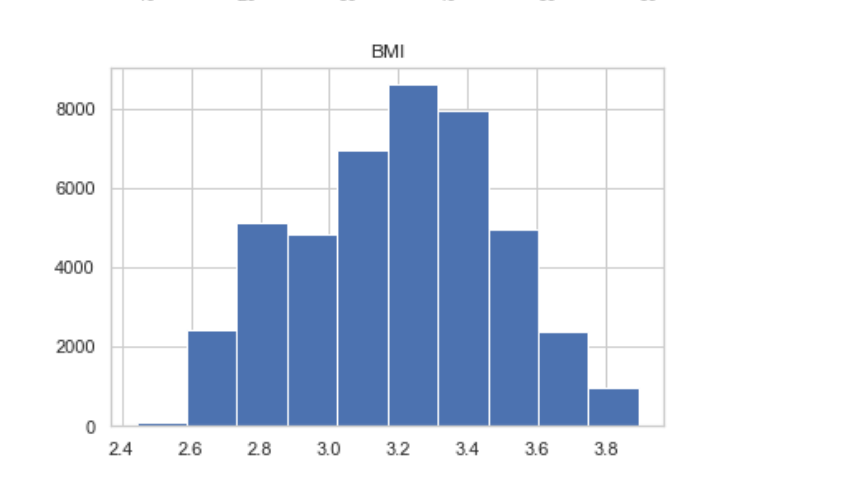


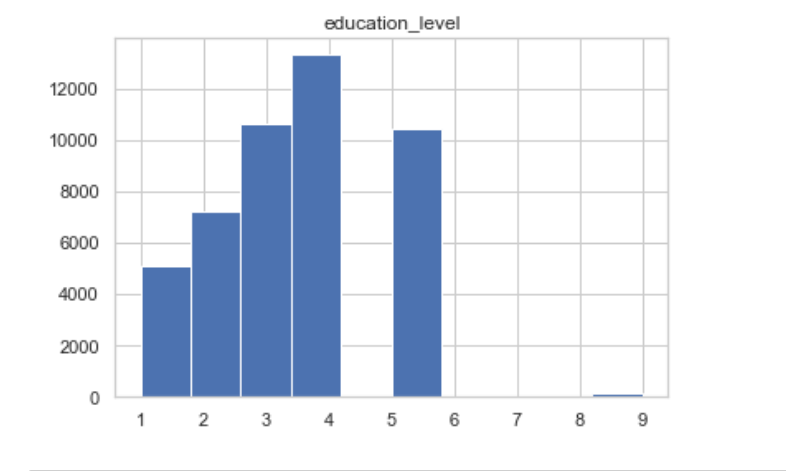
Data Preprocessing

The percentage of missing values is relatively low.

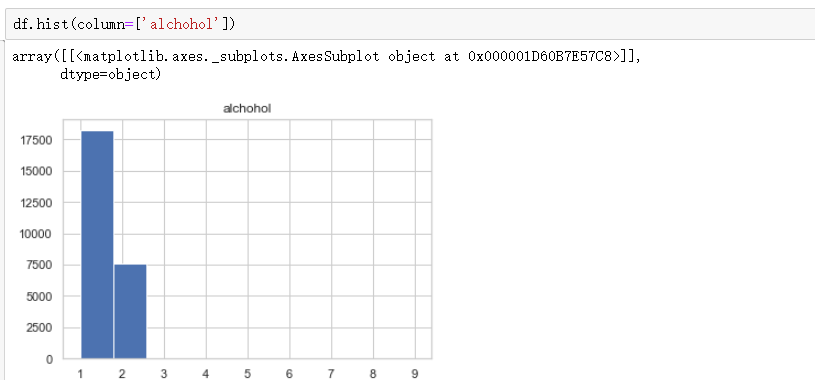
Outliers:







Dependent variable: Alcohol consumption at least 12 drinks last 1 year



After deleting not known, refuse to answer, and missing values, the number of data points is 23,420.

Meaning of categorical variable

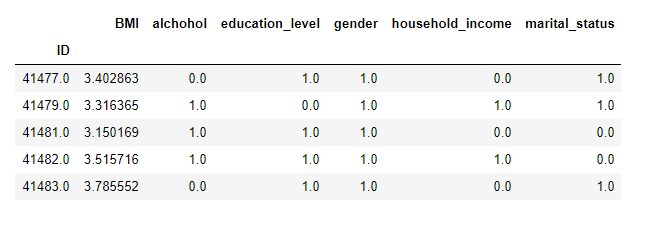
#Change 2(NHANES for Female) to 0(False - boolean) for features with YES/NO answers

# a household income <$45000: low income.change low income to 0, higher than that to 1

# change education lower than high school graduates to 0

# Alcohol: Change 2(NHANES for No) to 0(False - boolean) for features with YES/NO answers

#change married people and people living with a partner to 1, Otherwise 0: single.



Correlation, eda, importance, different models,

Alcohol use

Education level: <https://wwwn.cdc.gov/Nchs/Nhanes/2007-2008/DEMO_E.htm#DMDHREDU>

Alcohol: <https://wwwn.cdc.gov/Nchs/Nhanes/2007-2008/ALQ_E.htm>

不要写人种

Evaluations: t-test, accuracy, recall rate