Scientific Programming in Python Project

by McNamee_R00207204

Introduction

The objective of this project is to produce an application that allows the user to explore some of the most interesting aspects of two adapted datasets. It is required to use Pandas as a means of analysing the data and to incorporate visualisation as a method of illustrating results.

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.

The first dataset provided ('NhanesDemoAdapted.csv') has been adapted and changed from the demographic set. It includes the following variables :

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:<9 th Grade, 2: 9 th -11 th 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:
8. Single, 6: Living with Partner}
9. HouseholdSize – Integer: Number of people in house {1-7, 7 given for houses with 7 or more
10. people}
11. AgeUnder6 – Number of household members under the age of 6
12. Age6to18– Number of household members between the ages of 6 and 18
13. AgeOver60 – Number of household members over the age of 60
14. HouseholdIncome – Total household income in 1000s of dollars
15. IncomePovertyRatio – Float: Ratio of family income to poverty {0.0-5.0, 5 given for ratios of 5 or higher}

The second dataset ('NhanesFoodAdapted.csv') involves the diet of individuals. Again this has been adapted and changed from the original Nhanes dataset. The reduced dataset has multiple entries for each individual (as identified by the same SEQN as in the first dataset). Each line involves a different recorded meal for an individual and contains the nutritional information of the meal such as grams, calories, protein, etc. It includes the following variables :

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

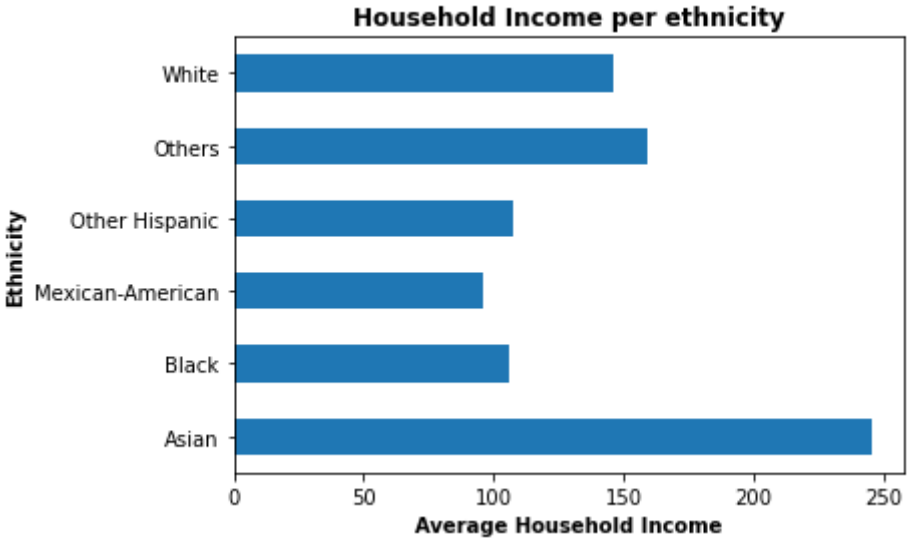
10. dCHOL – Float: Cholestorol (mg) of meal
11. dVITC – Float: Vitamin C content (mg) of meal
12. dVITD – Float: Vitamin D content (mcg) of meal
13. dCALC – Float: Calcium content (mg) of meal
14. dCAFF – Float: Caffeine content (mg) of meal
15. dALCO – Float: Alcohol content (gm) of meal

The application reads the two datasets as pandas dataframes and manipulated to present various plot representations of the data via a Python console menu for each of the following options:

1. Household Income of ethnicities
2. Marital status
3. Education and income
    1. Income Poverty Ratio
    2. Household Income
4. Diet Analysis
5. Exit

Household Income of ethnicities

This function generates a series from the main demo dataframe as a count of each category of ethnicity and printed to the console as a table. Then another series is generated by first grouping the data by ethnicity, then finding the mean for each category, and finally sorting by category to generate a plot of Average Household Income by Ethnicity.

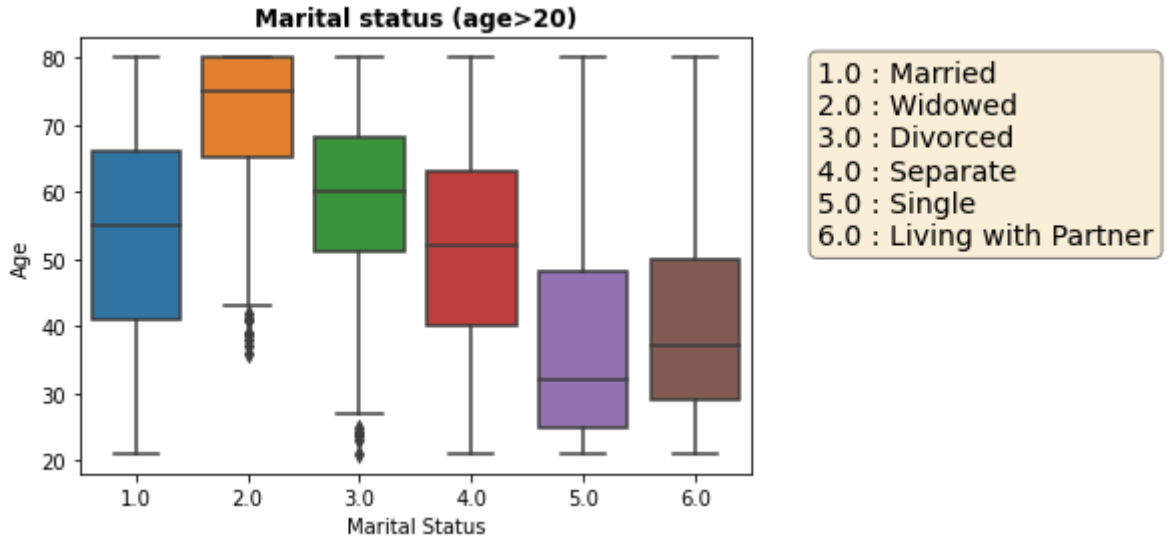


Observations

What stands out is the disparity of income by ethnicity, most notably the Asian mean income is an outlier at twice the overall average, and then the Black, Mexican-American and Other Hispanic which are significantly lower then the White, Others and of course the Asians.

Marital status

This function generates a series from the main demo dataframe as a count of each category of marital status for all respondents with age > 20, sorted by category and printed to the console as a table, and then generates a plot.

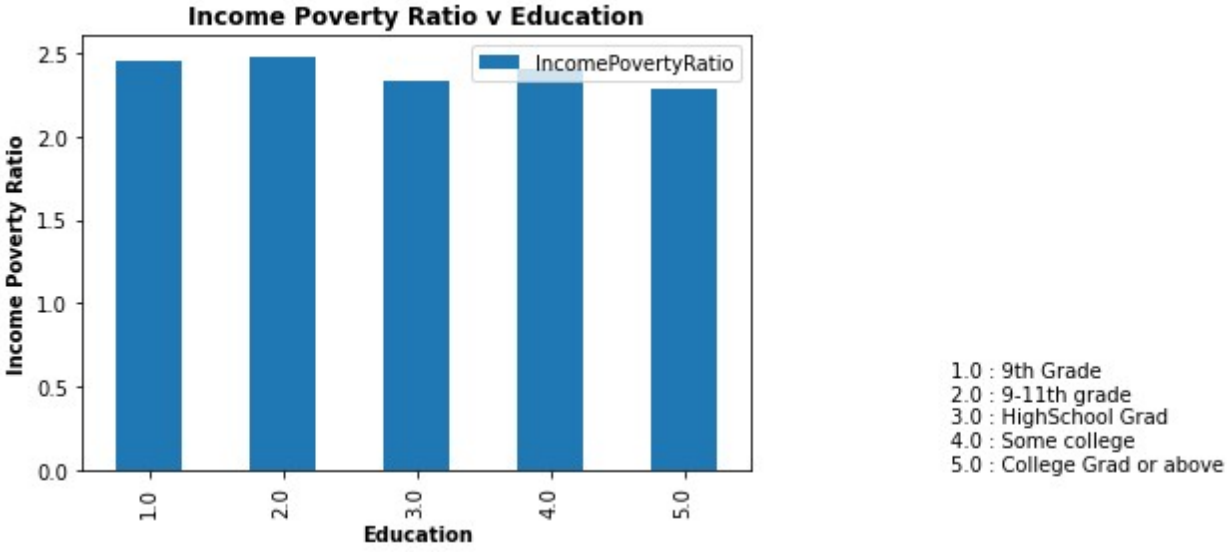


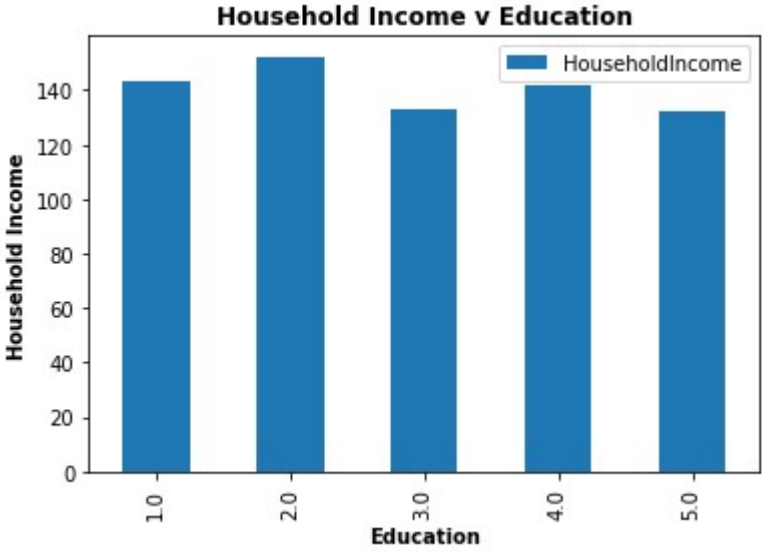
Observations

Each box plot shows the various percentiles, median and range for each marital status by age. It provides an immediate visual understanding of the data distribution. It clearly follows the cycle of life with many young respondents, either single or living with partners, then the next older group are the married and separated respondents, followed last by the respondents who are divorced or widowed. There are many inferences that could be made if time allowed (college workload is intense) but what surprises me most is the separated category – it resembles the married category but not the divorced category, whereas, I would have expected the IQ2 to begin at a greater age and more closely resemble the divorced category.

Education and income

This function generates a two new subset dataframes from the main demo dataframe, the average Income Poverty Ratio and the average Household Income, both grouped by education level, and then generates a plot.





1.0 : 9th Grade
2.0 : 9-11th grade
3.0 : HighSchool Grad
4.0 : Some college
5.0 : College Grad or above

Observations
There is no smoking gun here in terms of a strong linear relationship, both plots are quite uniform. For Income Poverty Ratio, there appears to be a weak positive linear relationship with Education. For Household Income, there appears to be also a weak relationship as shown by a feint bell curve with second level education at the peak and low and high levels of education at the sides.

I was quite unsatisfied with the results of this analysis, I expected a strong positive linear relationship for both. Not trusting my newly acquired pandas dataframe manipulation skills, I looked to more solid ground in Excel and checked my results using more familiar ground. I discovered 40% of the dataset had no education level recorded – ok, not so bad that leaves 5556 datapoints. I reproduced the two tables generated by the application, and when allowing for null values, found they match.

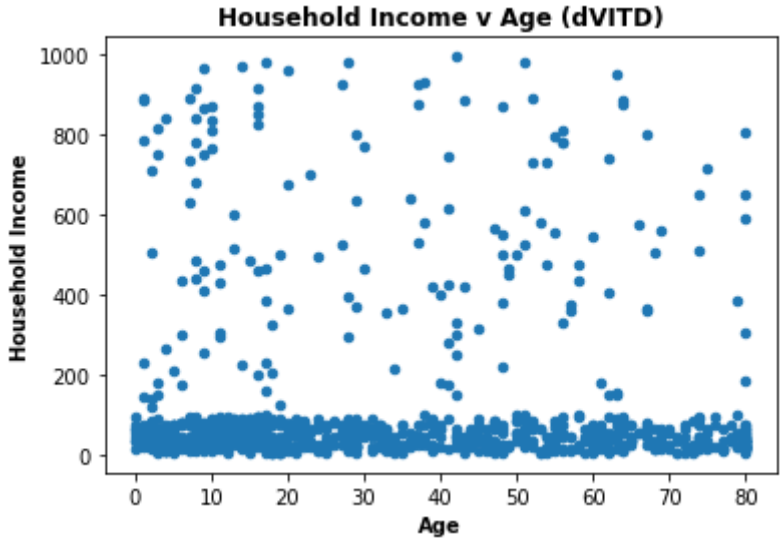
My curiosity has me second guessing the results. My liberal world view says that more education equals more income so I expect a one to one relationship yet that is not what I see in the plots. I want to know more about the dataset – what year? is it representative of the general US population? has white America fallen so far? has Asian America risen so far? is this why Trump was elected (disenfranchised White voters and newly conservative Aisian voters)?

Diet Analysis
This function takes two dataframes as input (demographic data and dietary data), reduces the food dataset (using pandas groupby functionality) to the average food intake per meal per individual, merges both dataframes via pandas, and then writes this merged dataframe to a csv file called ‘McNamee_R00207204_Merged_merged.csv’.

The application then prompts the user to select from a list of nutrients and then two plots are generated, one for the categories of Household Income and Age, and the other for Gender, Ethnicity and Education.

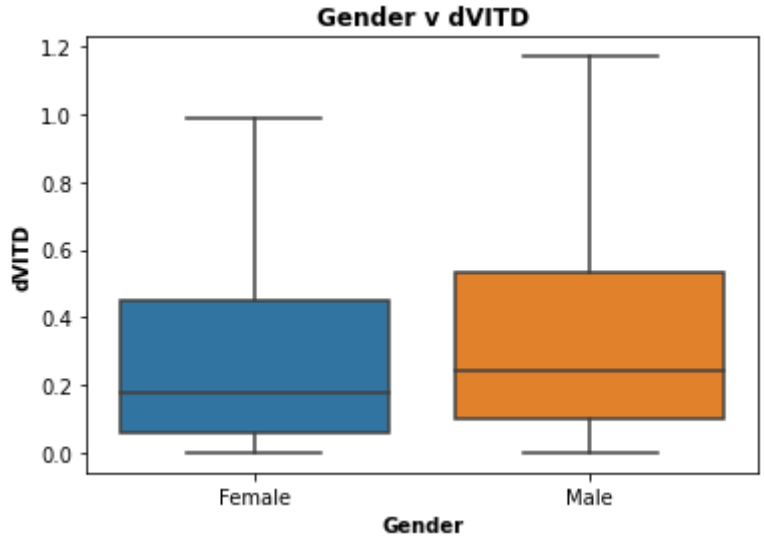
Vitamin D example - There are 11 food options so in the context of this project, there are too many to report so one has been selected at random, namely, Vitamin D.

Household Income v Age.
Here is an example with Vitamin D selected through the menu to produce a scatter plot of Household Income v Age.

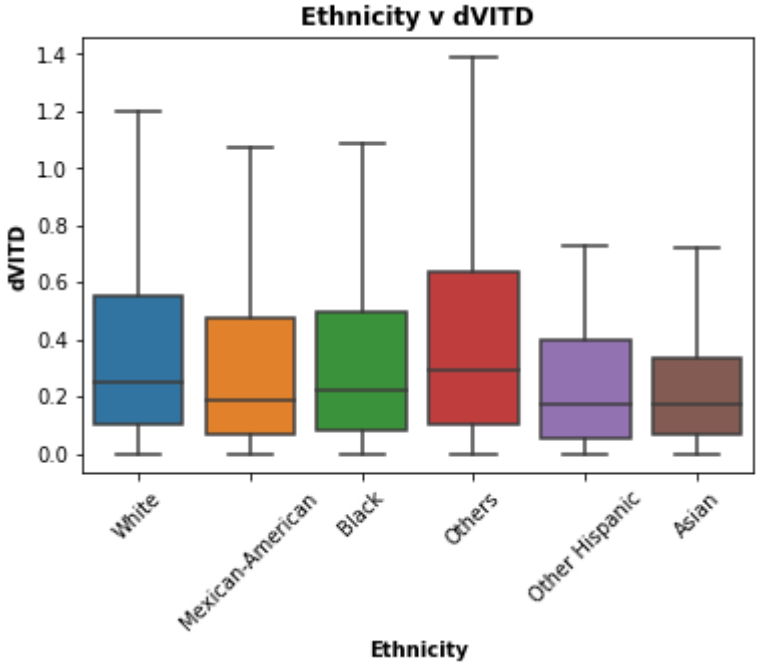


Observation
The plot shows most respondents have Vitamin D in food below 100mg for all ages - shown by a concentration of observations at the bottom. There are increased levels of Vitamin D then for all ages for higher income which suggests these respondents can afford a better diet.
There is a concentration of observations in the bottom left of the plot showing those respondents under 20 showing a large number of < 20 respondents fare no better than their parents in sharing the same low Vitamin D diet of their low income parents.
However, the sample data is only a small subset of the population (9254 observations), only 908 suitable observations (not empty values ignored by dataframe slicing), and a repeating pattern can be seen in the upper (ages 20 – 80) part of plot which may be evidence the data was manufactured, eg a few hundred rows were copied and pasted several times.

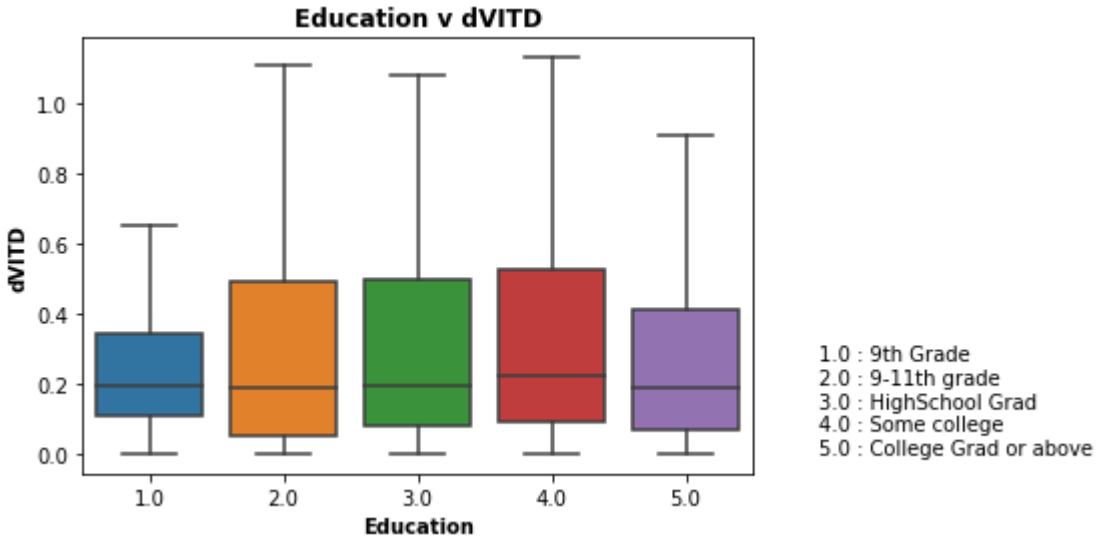
Gender, Ethnicity and Education
Here is an example with Vitamin D selected through the menu to produce a box plot for each of the Gender, Ethnicity and Education variables. Outliers were removed from the plotted data for each of the boxplots below to avoid skewing the data so much that each plot appeared squashed and unreadable.



Observations
Men appear to have slightly higher levels of Vitamin D in their diet overall suggesting men eat better than women.



Observations
There appears to be quite a variety of ranges, medians and IQs in this plot of Vitamin D diet for all ethnicities. Asians appear to have a diet with relatively low Vitamin D levels. Others the opposite. Note, Vitamin D is mostly manufactured by the body from other nutrients through exposure to sunlight so it seems the Asians may be at risk especially as they may be more likely to be in employment indoors (evidenced by high househld income earlier), and similarly, dark skinned respondents may suffer as a result of a reduced ability to manufacture Vitamin D combined with the lower diet levels, eg Mexican-American, Hispanic and Black ethonicities.



Observations

There appears to be a relationship between education and Vitamin D levels in respondents diet. It appears that more education leads to better diet (if Vitamin D representative of diet here). It is particularly pronounced at the lowest level of education, 9th grade, with low Vitamin D levels. However, unexpectedly, the trend does not continue to the highest level of education, College Grad or above, with Vitamin D levels falling off (IQ2 is well below the same quantile for education levels 2, 3 and 4). This may be confirmed by the similar tailing off of household income by education level seen earlier, ie the respondents cannot afford as good a diet as their less educated but better paid neighbours.