Data_Read

February 6, 2021

1 NHANES Dataset

[1]:

```
import pandas as pd

nhanes = pd.read_csv('../Data/nhanes.csv')
nhanes.head()
nhanes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 823012 entries, 0 to 823011
Data columns (total 58 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Unnamed: 0          823012 non-null  int64
 1   SEQN                823012 non-null  int64
 2   WTDRD1              823012 non-null  float64
 3   DR1ILINE            823012 non-null  int64
 4   DR1FS               773582 non-null  float64
 5   DR1IFDCD            823012 non-null  float64
 6   DR1IGRMS            817914 non-null  float64
 7   DR1.020             823012 non-null  int64
 8   DR1.030Z            823012 non-null  int64
 9   DR1.040Z            817610 non-null  float64
 10  DR1IKCAL            817914 non-null  float64
 11  DR1IPROT            817914 non-null  float64
 12  DR1IPFAT            817914 non-null  float64
 13  DR1IVARA            817914 non-null  float64
 14  DR1IVB12            817914 non-null  float64
 15  DR1ICALC            817914 non-null  float64
 16  DR1IIRON            817914 non-null  float64
 17  DR1IZINC            817914 non-null  float64
 18  DR1ISELE            817914 non-null  float64
 19  DR1IP205            817914 non-null  float64
 20  DR1IP226            817914 non-null  float64
 21  RIAGENDR            823012 non-null  int64
 22  RIDAGEYR            823012 non-null  int64
 23  RIDRETH1            823012 non-null  int64
 24  DMDEDUC3          203442 non-null  float64
 25  DMDEDUC2          491353 non-null  float64
 26  DMDHHSIZ          823012 non-null  int64
 27  DMDFMSIZ          823012 non-null  int64
 28  INDHHIN2          814606 non-null  float64
 29  INDFMIN2          815421 non-null  float64
 30  INDFMPIR          762405 non-null  float64
 31  SDMVPSU           823012 non-null  int64
 32  SDMVSTRA          823012 non-null  int64
 33  RIDRETH3          379378 non-null  float64
 34  DESCRIPTION       823012 non-null  object
 35  DR1I_PF_SEAFD_HI  817914 non-null  float64
 36  DR1I_PF_SEAFD_LOW 817914 non-null  float64
 37  DR1I_PF_MEAT      817914 non-null  float64
 38  DR1I_PF_CUREDMEAT 817914 non-null  float64
 39  DR1I_PF_ORGAN     817914 non-null  float64
 40  DR1I_PF_POULT     817914 non-null  float64
 41  DR1I_PF_MPS_TOTAL 817914 non-null  float64
 42  DR1I_PF_EGGS      817914 non-null  float64
 43  DR1I_PF_SOY       817914 non-null  float64
 44  DR1I_PF_NUTSDS    817914 non-null  float64
 45  DR1I_PF_LEGUMES   817914 non-null  float64
 46  DR1I_PF_TOTAL     817914 non-null  float64
 47  DR1I_D_TOTAL      817914 non-null  float64
 48  DR1I_D_MILK       817914 non-null  float64
 49  DR1I_D_YOGURT     817914 non-null  float64
 50  DR1I_D_CHEESE     817914 non-null  float64
 51  SDDSRVYR          823012 non-null  int64
 52  WTDRD1_6YR        823012 non-null  float64
 53  DR1I_PF_SEAFD_TOT 817914 non-null  float64
 54  DR1I_PF_MEAT_TOT  817914 non-null  float64
 55  species           9978 non-null    object
 56  species_code      9978 non-null    float64
 57  DR1.030Z_2        823012 non-null  int64
dtypes: float64(42), int64(14), object(2)
memory usage: 364.2+ MB
```

Data Description Results

Source - Dietary interview, Day 1:

SEQN: Respondent sequence number. This variable is required for grouping the analysis by individual.

WTDRD1: Dietary day one sample weight. Not a variable of interest for this case. This variable can be dropped.

DR1ILINE: One of the key variables for the file. The primary key variables are SEQN and DR1ILINE. Can potentially be used for individual food item analysis.

DR1FS: Source of food. This variable can potentially be used for sourcing analysis. For categorical value mapping, see https://wwwn.cdc.gov/nchs/nhanes/search/default.aspx. Mapping may be different for each survey.

DR1IFDCD: USDA food code. Maybe useful for categorizing food items. Reference the USDA for a mapping.

DR1IGRMS: Gram weight of the food/individual component. May be useful for understanding seafood consmuption by weight.

DR1.020: Time of eating occasion (HH:MM). May be useful for temporal analysis. May also be useful to understand relationship between side dishes consumed with seafood. I.e. confirming that it is indeed a seafood side dish, using time as a determinant.

DR1.030Z: Name of eating occasion. May be useful for grouping by meal. A code mapping is available in https://wwwn.cdc.gov/nchs/nhanes/search/default.aspx, but may be different for each survey.

DR1.040Z: Did you eat this meal at home? May be useful for food sourcing. Needs mapping from documentation.

DR1IKCAL: Energy (kcal). May be useful to undestand proportional seafood consumption by calories, similar to using food weight in grams.

DR1IPROT: Protein (gm). May be useful to understand consumption based on a protein/carb/fat relationship.

DR1IPFAT: Total fat (gm). May be useful to understand consumption based on a protein/carb/fat relationship.

DR1IVARA: Vitamin A, RAE (mcg). Not required for this analysis.

DR1IVB12: Vitamin B12 (mcg). Not required for this analysis.

DR1ICALC: Calcium (mg)

DR1IIRON: Iron (mg)

DR1IZINC: Zinc (mg)

DR1ISELE: Selenium (mcg)

DR1IP205: PFA 20:5 (Eicosapentaenoic) (gm)

DR1IP226: PFA 22:6 (Docosahexaenoic) (gm)

Source - Demographics:

RIAGENDR: Gender of the sample person

RIDAGEYR: Best age in years of the sample person at time of HH screening. Individuals 80 and over are topcoded at 80 years of age.

RIDRETH1: Recode of reported race and ethnicity information.

DMDEDUC3: Education Level - Children/Youth 6-19, (SP Interview Version) What is the highest grade or level of school {you have/SP has} completed or the highest degree {you have/s/he has} received?

DMDEDUC2: Education Level - Adults 20+, (SP Interview Version) What is the highest grade or level of school {you have/SP has} completed or the highest degree {you have/s/he has} received?

DMDHHSIZ: Total number of people in the Household

DMDFMSIZ: Total number of people in the Family

INDHHIN2: Annual Household Income

INDFMIN2: Annual Family Income

INDFMPIR: Ratio of family income to poverty

SDMVPSU: Masked Variance Unit Pseudo-PSU variable for variance estimation

SDMVSTRA: Masked Variance Unit Pseudo-Stratum variable for variance estimation

SDDSRVYR: Data Release Number.

Source - Unknown:

DESCRIPTION: Main variable of interest. Contains the text that will analyzed to assess the dietary complements for seafood consumption.

RIDRETH3

DR1I_PF_SEAFD_HI

DR1I_PF_SEAFD_LOW

DR1I_PF_CUREDMEAT

DR1I_PF_ORGAN

DR1I_PF_POULT

DR1I_PF_MPS_TOTAL

DR1I_PF_EGGS

DR1I_PF_NUTSDS

DR1I_PF_LEGUMES

DR1I_PF_TOTAL

DR1I_D_TOTAL

DR1I_D_MILK

DR1I_D_YOGURT

DR1I_D_CHEESE

WTDRD1_6YR

DR1I_PF_SEAFD_TOT: Another main variable of interest. Seems to indetify if food item is a seafood item.

DR1I_PF_MEAT_TOT

species

species_code

DR1.030Z_2

```
[2]: #Obtain survey years included in the data
     nhanes['SDDSRVYR'].describe()
```

```
[2]: count    823012.000000
     mean          6.381709
     std           1.693001
     min           4.000000
     25%           5.000000
     50%           6.000000
     75%           8.000000
     max           9.000000
     Name: SDDSRVYR, dtype: float64
```

SDDSRVYR min/max is 4/9. So this dataset contains survey years 2005-2006 (4), 2007-2008 (5), 2009-2010 (6), 2011-2012 (7), 2013-2014 (8), 2015-2016 (9).

Data Read Summary/Questions

The provided dataset contains 58 variable columns. A search for the description of all these variables was performed at the NHANES survey website from the CDC (https://wwwn.cdc.gov/nchs/nhanes/search/default.aspx). The following conclusions are made about the dataset, with questions for the client:

1. The NHANES survey is performed every two years, and this dataset contains the interview and demographic data for the following survey years:

2005-2006, 2007-2008, 2009-2010, 2011-2012, 2013-2014, 2015-2016

Is there a particular reason why only these surveys are included? Does this temporal scope satisfy the analysis needs of the client? Should the cope be more broad or more narrow?

2. The NHANES open source data is available in a SAS format. The dataset obtained from the client has been converted to a more inclusive .csv format. How was the .csv file assembled and what was the thought behind what to include?

3. The provided dataset has been compiled from different data modules provided by NHANES. A subset of the variables are from the dietary interview data, another subset is from the demographic data. There is another subset of variables that was not found during the variable search. This subset includes some variables of interest such as:

DESCRIPTION: Main variable of interest. Contains the text that will analyzed to assess the dietary complements for seafood consumption.

DR1I_PF_SEAFD_TOT: Another main variable of interest. Seems to indetify if food item is a seafood item.

RIDRETH3

DR1I_PF_SEAFD_HI

DR1I_PF_SEAFD_LOW

DR1I_PF_CUREDMEAT

DR1I_PF_ORGAN

DR1I_PF_POULT

DR1I_PF_MPS_TOTAL

DR1I_PF_EGGS

DR1I_PF_NUTSDS

DR1I_PF_LEGUMES

DR1I_PF_TOTAL

DR1I_D_TOTAL

DR1I_D_MILK

DR1I_D_YOGURT

DR1I_D_CHEESE

WTDRD1_6YR

DR1I_PF_MEAT_TOT

species

species_code

DR1.030Z_2

Where were these variables obtained from? Is DR1I_PF_SEAFD_TOT indeed the filter key for identifying seafood items? Are there any other useful variables here, like species or species_code?

4. The demographic data does not contain any information on the location of the participant. Such a variable could be found on the NHANES demographic data description (Example for 2007-2008 survey: https://wwwn.cdc.gov/Nchs/Nhanes/2007-2008/DEMO_E.htm#SDDSRVYR). The client requested an analysis of diet based on location. Did the client have a variable under consideration for this?