NHANES Data for BMI714 Final Project
Fall 2024

The National Health and Nutrition Examination Survey (NHANES) is conducted by the National Center for Health and Statistics as provided by the National Health Survey Act of 1956 with the goal of "monitor[ing] the health of the United States population through the collection and analysis of data on a broad range of health topics" (CDC). The most recent publicly available version of NHANES contains data collected between 2017 and 2020 (pre pandemic).

We have provided you with an aggregated, lightly curated version of the NHANES 2017-2020 data as `BMI714_NHANES2020_Data.csv`. The BMI714 version of NHANES 2017-2020 contains 1595 features (variables) and 15560 observations. The first column, `SEQN`, provides a unique identifier for each participant. The remaining columns each contain a different type of data collected as part of NHANES. Note that data may be numeric, categorical, or text, and that the dataset has blanks indicating missing data, for example if someone did not participate in a particular questionnaire.

The column names in `BMI714_NHANES2020_Data.csv` are unfortunately not easily interpretable on their own. The features are briefly described in `BMI714_NHANES_VariableDictionary.csv`. In the variable dictionary each row provides more information about one column of the main dataset. `BMI_714_Variable_Name` corresponds to the column name for one dataset column. `Variable_Name` lists the original NHANES name for this feature. `Variable_Description` provides a brief text explanation of the feature. `Data_File_Name` lists the original NHANES filename in which the data was distributed. `Data_File_Description` provides a brief text explanation of the data file. `Component` indicates whether the data is categorized as Demographics, Dietary, Examination, Laboratory, or Questionnaire.

There is more detailed documentation available online. Starting from the NHANES 2017-2020 web page, click on a category under "Data, Documentation, Codebooks", and click on a link under the "Doc File" column. For example, here is documentation for NHANES 2017-2020 demographic data. Generally, you can substitute values from the `Data File Name` column of the data dictionary into `{DATA_FILE_NAME}` in the following URL to obtain information on that file.

`https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/{DATA FILE NAME}.htm`.

For example, youth weight history is distributed by NHANES in a file named `P_WHQMEC.xpt`, our data dictionary lists this data file name as `P_WHQMEC`, and a corresponding web reference is accessible at `https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_WHQMEC.htm` . Note that for categorical variables, these documentation web pages provide essential information about the meaning of each category.

We included many features to give you freedom to broaden the range of questions that can be addressed with this dataset. Before you begin your project, explore the variable dictionaries and read in more detail about the categories/variables contained in the .csv file. Take note of the variable types (categorical, continuous) and start to think about which relationships would be interesting to investigate.

Be aware of the nuances that surround different variables. Some variables may be proxies for other data, other variables may have subjective responses, and there are some variables with many missing values. Take these things into account when selecting which variables to consider and deciding how to incorporate them in your project.

More information about the NHANES 2017-2020 release is available at

- https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020
- https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewbrief.aspx?Cycle=2017-2020
- https://www.cdc.gov/nchs/data/nhsr/nhsr158-508.pdf

One quick way to browse for potentially interesting features would be to randomly sample the variable description table. For example:

```{r}
`%>%` <- magrittr::`%>%`
var_dict <- readr::read_csv("BMI714_NHANES_VariableDictionary.csv")
var_dict[sample(nrow(var_dict), 10,] %>% dplyr::select(Variable_Name,
      Variable_Description, Data_File_Name)
```

```
   Variable_Name Variable_Description                                                Data_File_Name
    <chr>         <chr>                                                              <chr>
 1 OSQ140Q       Please think about {your/SP's} use of prednisone or cortisone …    P_OSQ
 2 AUQ280        How much of a problem is this ringing, roaring, or buzzing in …    P_AUQ
 3 MCQ160B       Has a doctor or other health professional ever told {you/SP} t…    P_MCQ
 4 AUQ110        How often does {your/SP's} hearing cause {you/him/her}to feel …    P_AUQ
 5 LBDVBZLC      Blood Benzene Comment Code                                         P_VOCWB
 6 SMQ856        I will now ask you about tobacco smoke in other places. During…    P_SMQSHS
 7 LBDVOXLC      Blood o-Xylene Comment Code                                        P_VOCWB
 8 SMQ690D       Which of these products did {you/he/she} use?                      P_SMQRTU
 9 LBDLDL        LDL-Cholesterol, Friedewald equation (mg/dL). LBDLDL = (LBXTC-…    P_TRIGLY
10 DPQ070        [Over the last 2 weeks, how often have you been bothered by th…    P_DPQ
```