Sharif Univerity of Technology
Industrial Engineering Department

Final Project

Dataset Documentation

Advanced Programming Course

Instructed by:

Dr. Habibi

Dr. Lashkari

Document created by:

Mohammad Hossein Mahmoudi

1402 - Fall

Data Source

About the National Health and Nutrition Examination Survey (NHANES)

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 health statistics for the Nation.

The NHANES program began in the early 1960s and has been conducted as a series of surveys focusing on different population groups or health topics. In 1999, the survey became a continuous program that has a changing focus on a variety of health and nutrition measurements to meet emerging needs. The survey examines a nationally representative sample of about 5,000 persons each year. These persons are located in counties across the country, 15 of which are visited each year.

The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.

Findings from this survey will be used to determine the prevalence of major diseases and risk factors for diseases. Information will be used to assess nutritional status and its association with health promotion and disease prevention. NHANES findings are also the basis for national standards for such measurements as height, weight, and blood pressure. Data from this survey will be used in epidemiological studies and health sciences research, which help develop sound public health policy, direct and design health programs and services, and expand the health knowledge for the Nation. [1]

Which Datasets Have Been Used for This Project?

The meticulous process of data collection embarked with a methodical selection of subjects integral to the project's focus. This discerning task was executed in collaboration with a domain expert, specializing in the medical field. The chosen subjects are pivotal in shaping the foundation of the research endeavor. They encompass:

1. Demographic Variables and Sample Weights

1.1. Demographic Variables and Sample Weights

This dataset serves as the cornerstone, containing essential demographic variables such as age, gender, and ethnicity, along with sample weight information. It provides a comprehensive overview of the study's participants, ensuring representativeness in the analyses.

From this dataset, the features named "Gender", "Age in years at screening", and "Race/Hispanic origin" have been chosen.

2. Examination Data

2.1. Body Measures

Focusing on physical health, this dataset encompasses measurements like weight and body mass index (BMI). These metrics are pivotal for understanding health trends and their correlations.

From this dataset, the features named "Weight (kg)" and "Body Mass Index (kg/m**2)" have been chosen.

3. Questionnaire Data

3.1. Alcohol Use

Detailing participants' alcohol consumption patterns, this dataset aids in assessing the prevalence of alcohol use and its potential health implications.

From this dataset, the features named "Ever had a drink of any kind of alcohol" and "Past 12 mo how often drink alcoholic bev" have been chosen.

3.2. Cardiovascular Health

Providing insights into various aspects of cardiovascular health, this dataset captures essential data for understanding heart conditions and related factors.

From this dataset, the features named "SP ever had pain or discomfort in the chest" and "Shortness of breath on stairs/inclines" have been chosen.

3.3. Diabetes

This dataset is invaluable for analyzing diabetes prevalence and factors contributing to this chronic condition.

From the dataset, the features named "Respondent sequence number" and "Doctor told you have diabetes" have been chosen.

3.4. Income

Income data is pivotal for assessing the socio-economic aspects of participants, which can influence health outcomes.

The "Family monthly poverty level category" has been chosen from the dataset.

3.5. Physical Activity

Information on physical activity levels is essential for understanding the role of exercise in health.

From the dataset, the feature named "Vigorous work activity" has been chosen.

3.6. Smoking - Cigarette Use

This dataset captures smoking habits, which are critical for analyzing the impact of smoking on health.

From the dataset, the feature named "Smoked at least 100 cigarettes in life" has been chosen.

Final Features

As you can see, each subject comprises a spectrum of datasets, each offering a unique perspective. To curate the most pertinent dataset attributes, the guidance and expertise of a medical professional were sought. It is imperative to make reasonable selections, ensuring that the chosen dataset features align closely with the research objectives.

Through a sequence of meticulous steps and a dedicated approach, this amalgamation of datasets culminates in the creation of the final dataset. This dataset serves as the bedrock for the comprehensive analysis and insights to be derived in the subsequent phases of this project.

The final features of these datasets, which have been described, are:

Demographic Variables and Sample Weights

- Respondent sequence number (SEQN)
- Gender (RIAGENDR)
- Age in years at screening (RIDAGEYR)
- Race/Hispanic origin (RIDRETH1)

BodyMeasures (Examination Data)

- Weight (kg) (BMXWT)
- Body Mass Index (kg/m**2) (BMXBMI)

Diabetes (QuestionnaireData)

- Doctor told you have diabetes (DIQ010)

PhysicalActivity (QuestionnaireData)

- Vigorous work activity (PAQ605)

Smoking - Cigarette Use (Questionnaire Data)

- Smoked at least 100 cigarettes in life (SMQ020)

Income (Questionnaire Data)

- Family monthly poverty level category (INDFMMPC)

Alcohol Use (Questionnaire Data)

- Ever had a drink of any kind of alcohol (ALQ111)
- Past 12 mo how often drink alcoholic bev (ALQ121)

Cardiovascular Health (Questionnaire Data)

- SP ever had pain or discomfort in chest (CDQ001)
- Shortness of breath on stairs/inclines (CDQ010)

You can access more information about each feature in the supplementary file.

Data Cleanliness Assessment

In this section, the cleanliness of the data used for this project has been assessed. Clean data is essential for accurate analysis and reliable results. To obtain the highest data quality from the main dataset, records with more than 20% of features missing have been removed. In the final dataset, there are some missing values that need to be addressed. To resolve this issue, working on filling or imputing these missing values should be done. The table below shows the proportion of non-missing values for the features in the final dataset, which containes 8111 records and 14 features.

Features	**Proportion of Non-Missing Values**
SEQN	100.00
DIQ010	100.00
PAQ605	100.00
SMQ020	100.00
RIAGENDR	100.00
RIDAGEYR	100.00
RIDRETH1	100.00
BMXWT	99.30
BMXBMI	99.09
ALQ111	96.39
INDFMMPC	94.77
ALQ121	89.95
CDQ001	71.95
CDQ010	71.95

Table 1. Proportion of Missing Values in the Final Dataset

References

[1] Centers for Disease Control and Prevention (CDC) (2023) National Health and Nutrition Examination Survey (NHANES). Retrieved from https://www.cdc.gov/nchs/nhanes/about_nhanes.htm