2026 ASA South Florida Student Data Challenge

Dataset Description

The competition dataset consists of *n = 1,200* participants sampled from the 2024 National Health and Nutrition Examination Survey (NHANES), integrating information from the dietary recall interview, demographic questionnaire, laboratory assessments, and anthropometric examinations. The outcome of interest is **Direct HDL-Cholesterol (LBDHDD, mg/dL)**, a clinically relevant indicator of cardiometabolic health. The dataset includes **96 predictor variables** spanning daily nutrient intake (energy, macronutrients, sugars, fiber, fatty-acid subtypes, vitamins and minerals, caffeine, alcohol), dietary behavior (salt use, special diet status, water intake), sociodemographic factors (age, sex, race/ethnicity, income-to-poverty ratio, marital status), and body-measurement variables such as **BMI** and **waist circumference**. These features capture multiple dimensions of metabolic, nutritional, and socioeconomic variation, making HDL prediction both interpretable and scientifically grounded. The dataset will be provided as **1,000 observations for model training** and **200 observations for final testing**.

Download the training the test datasets at https://github.com/luminwin/ASASF
For co-organizers, the original dataset is available at GoogleDrive of asasfchapter@gmail.com

---

Competition Tasks

1. Prediction Task (Undergraduate and Graduate Divisions)

Participants will develop a statistical or machine-learning model to predict **HDL-Cholesterol (LBDHDD)** using the training dataset. Final rankings will be determined by the **Root Mean Squared Error (RMSE)** computed from predictions on the held-out test dataset. Any modeling approach is allowed as long as predictions follow the required submission format.

2. Data Visualization Task (High School Division Only, Optional --- have to reach out to high school teacher first to see possibility)

High-school participants will explore the full dataset to create compelling, clearly communicated visualizations illustrating how HDL levels vary across dietary, demographic, or body-measurement factors. Submissions should emphasize clarity, interpretability, and thoughtful storytelling rather than complex statistical modeling.

What to Submit

Prediction Track (Undergraduate & Graduate Divisions)
Participants must submit **two files**:

1. **Predictions File**
    - A single-column CSV containing the predicted HDL outcome (`LBDHDD_outcome`) for each record in the test dataset. The number of rows must match exactly the number of rows in the test dataset. The order of predictions must correspond exactly to the order of rows in the test dataset. No identifiers. No extra columns. No header variations.
    - Required format:

      ```
      pred
      52.31
      48.77
      ...
      ```

2. **Short Report (PDF, max 2 pages)**
   The report must briefly describe:
    - Modeling approach
    - Pre-processing or feature engineering
    - Model evaluation and selection
    - Any interpretative insights
    - Code summary (full code in supplement or link)

Data Visualization Track (High School Division)
Participants must submit:

1. **PDF report or poster (max 4 pages)** showing:
    - At least two well-constructed visualizations
    - Clear explanations of what the graphs show
    - A short discussion of patterns or insights
2. **Optional:** Code used to generate the visualizations.

How to Submit (still working on Google Form for submission instead of email)

Participants email their materials to:

ASA South Florida Data Challenge Committee
📧 **asasfchapter@gmail.com**

Subject line:
"2026 ASA Data Challenge Submission – [Team Name] – [Division]"

Attachments:

- Predictions CSV
- PDF report
- (Optional) Code + figures