What Really Influences BMI? — Insights from NHANES (2021–2023)

Introduction to Data Science - University of Helsinki

Team: (**Saba, Abdullah**, **Imaan)**

Abstract

We study associations between lifestyle variables and Body Mass Index (BMI) using recent NHANES data. Our pipeline covers data selection, cleaning, feature engineering, exploratory analysis, and light-weight machine learning. We communicate insights via a public Streamlit dashboard designed for a non-technical audience.

1. Introduction

BMI is a widely used proxy for weight categories that may lead to health problems. Our goal was to find which daily habits most clearly connect to BMI in the NHANES data. We focus on sleep, daily sugar and calorie intake, plus basic demographics (age, gender, and race). The aim is not causal inference but clear, data-driven patterns communicated in plain language. The outcome is both a written analysis and a web app that the general public can explore.

2. Data Source & Variables

Dataset: **NHANES 2021–2023** (Centers for Disease Control). We selected participants with complete measurements for the variables below. Table 1 lists the core fields and derived features used throughout the analysis.

Variable	Description / Notes
BMXBMI	Body Mass Index (target). Continuous.
DR1TSUGR, DR2TSUGR	Day 1 & Day 2 sugar (g). Used to compute sugar_avg.
DR1TKCAL, DR2TKCAL	Day 1 & Day 2 calories (kcal). Used to compute kcal_avg.
SLD012, SLD013	Self-reported sleep hours. Used to compute sleep_avg and sleep_
RIAGENDR	Gender (1=Male, 2=Female) → gender_label.
RIDRETH1	Race/Ethnicity → race_label.
BMXWT, BMXHT	Weight (kg) and height (cm); help contextualize BMI and correlatio

3. Pre-processing & Feature Engineering

Cleaning: We removed impossible or missing values (e.g., BMI ≤ 10 or > 80). For diet variables with two daily recalls we computed averages: sugar_avg = mean (DR1TSUGR, DR2TSUGR) and kcal_avg = mean (DR1TKCAL, DR2TKCAL). Sleep hours were averaged between SLD012 and SLD013 to produce sleep_avg. We then created sleep_group categories: <6h, 6–7h, 7–9h, >9h. Numeric codes for gender and race were converted into readable labels. All steps were implemented in Python (pandas/numpy) and reproduced inside the Streamlit app so the online version computes the same derived fields from the uploaded CSV.

4. Tools & Technologies Used

We used Python for all data processing, visualization, and modeling. Key libraries include pandas, numpy, matplotlib, seaborn, plotly, and scikit-learn. The interactive dashboard was built and deployed using Streamlit and Streamlit Cloud, with GitHub for version control.

5. Exploratory Data Analysis (EDA)

We began with univariate distributions followed by bivariate relationships. The figures below summaries the most informative patterns.

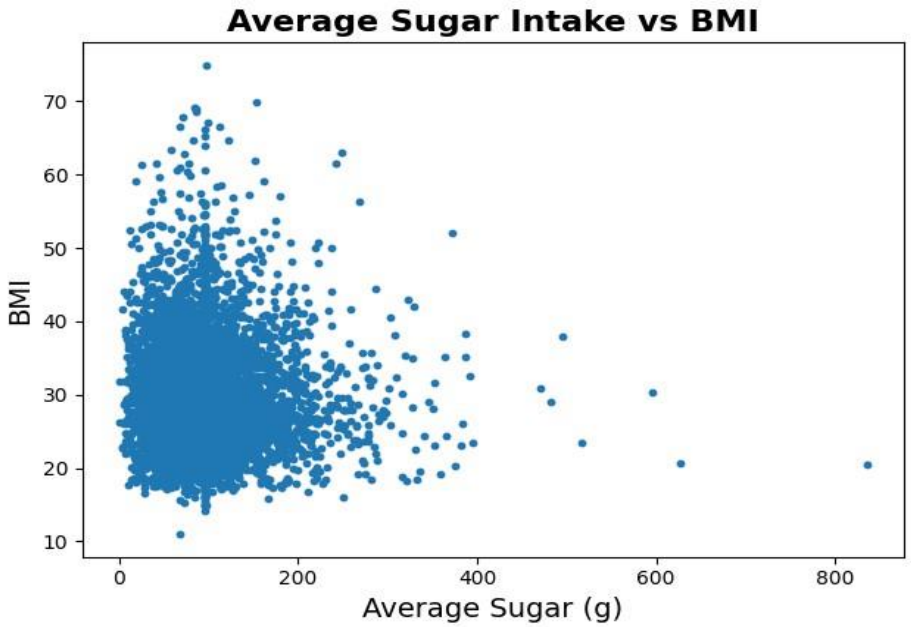

Figure 1 — Average sugar intake vs BMI. The cloud is diffuse; no strong linear trend is visible.

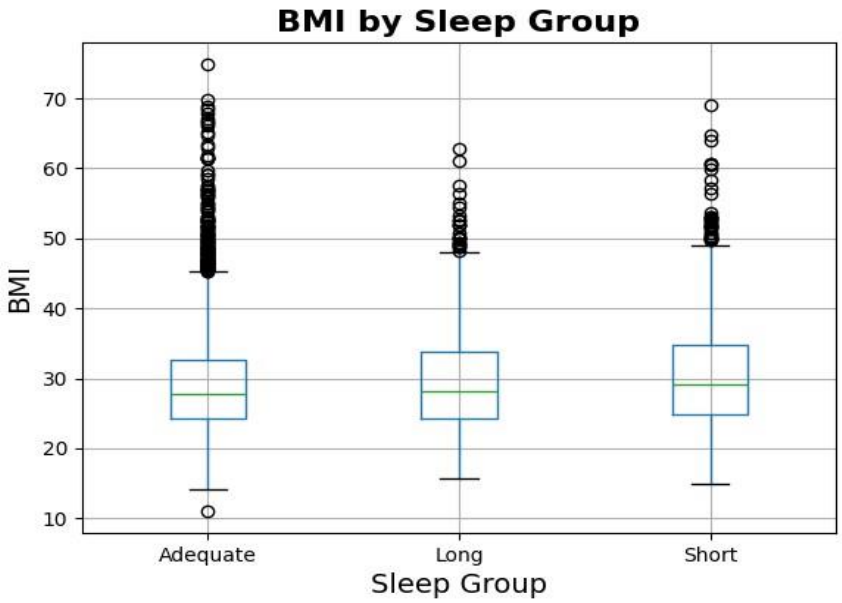

Figure 2 — BMI by sleep group. The 7–9h group tends to have lower median BMI and fewer extreme values.

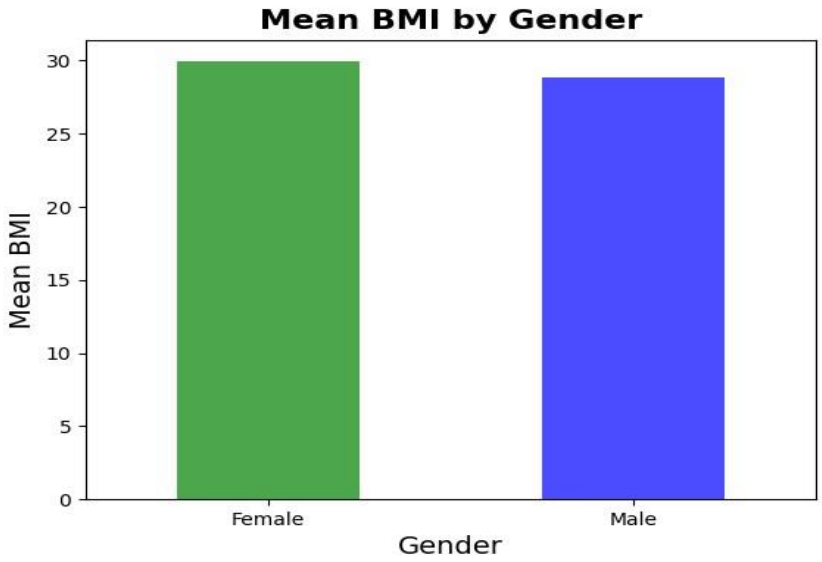

Figure 3 — Mean BMI by gender. Females show a slightly higher average BMI than males.

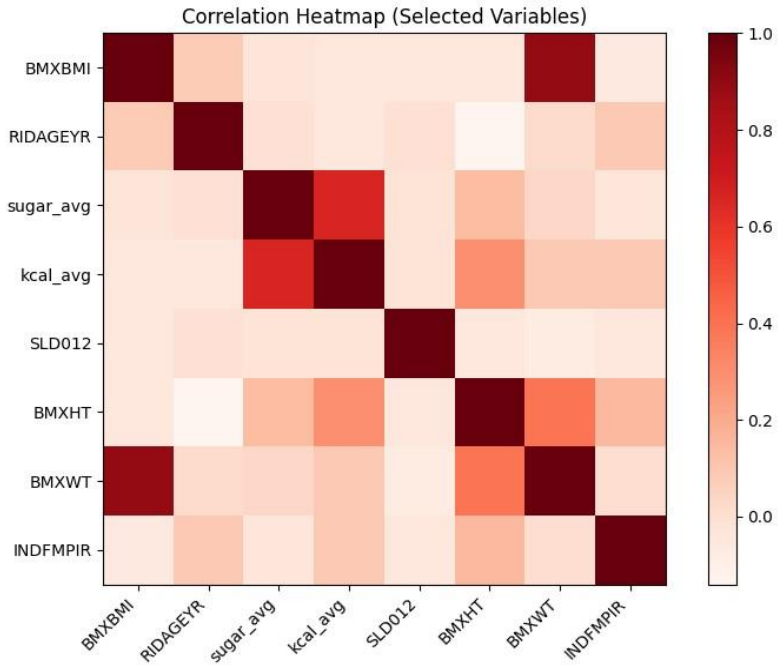

Figure 4 — Correlation heatmap. Weight (BMXWT) is strongly associated with BMI, as expected; other variables are weak to moderate.

From our observations: (1) sleep quality/quantity shows clearer separation in BMI than sugar alone; (2) sugar and calories correlate moderately, suggesting dietary patterns move together; (3) demographic differences exist but are smaller than the BMI–weight relationship.

6. Machine Learning

Problem formulation: we framed BMI as a 3-class target (Normal / Overweight / Obese) using standard BMI cutoffs. We split the data 80/20 into train/test with a fixed random seed for reproducibility. A simple preprocessing pipeline scaled continuous features and one-hot encoded categorical ones. We trained Logistic Regression, Random Forest, and XGBoost using Scikit-learn defaults, plus light hyperparameter tuning via grid search (n_estimators and max_depth for Random Forest; learning_rate and n_estimators for XGBoost). Metrics reported: Accuracy and F1-score (macro) to account for class imbalance.

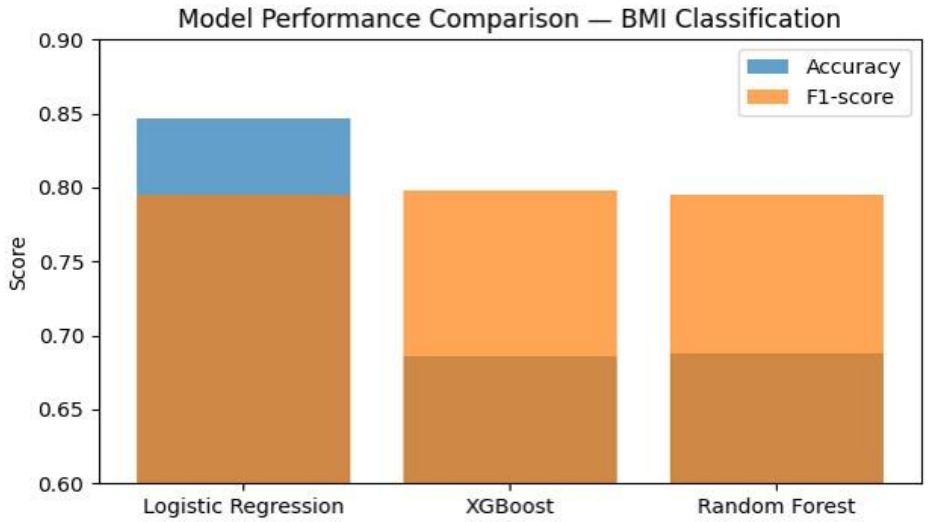

Figure 5 — Model comparison. Random Forest was the most stable performer (≈0.80 accuracy, ≈0.79 F1). Feature inspections indicated age and sleep variability as meaningful predictors alongside calories.

Interpretation: while predictive performance is moderate, models support the EDA story sleep-related features carry signal even when dietary quantities (sugar_avg) are noisy. We avoided over-fitting by keeping models simple and reporting test-set metrics only.

7. Communicating Results — Streamlit Dashboard

To reach a non-technical audience we built a Streamlit app with a left sidebar for filters (gender, sugar range, sleep hours). Tabs present the five views used in this report. The app re-computes derived features so any compatible NHANES subset can be uploaded. The live deployment is available at: https://nhanes-bmi-app-teambinary.streamlit.app/.

8. Limitations & Ethical Notes

NHANES is observational; we do not claim causality. Self-reported sleep and 24-hour dietary recalls are noisy and may include bias. BMI is an imperfect proxy for health across different body types and ethnic groups. All results are for educational purposes only and should not be used as medical guidance.

9. Reflection & Learning Outcomes

Team roles: Saba led data collection/cleaning and feature engineering; Abdullah ran EDA and built the Streamlit dashboard and prepared the technical report; Imaan implemented and compared ML models.

What worked: clear task split, reproducible code, and a simple public interface.

Challenges: handling missing values across NHANES files and aligning column names; choosing a small but meaningful set of features; keeping visuals consistent across the app and slides.

Future work: include physical-activity measures, dietary quality indices, and uncertainty bands; expand the model with calibrated probabilities and better class balancing.

References

NHANES (CDC).

Scikit-learn, pandas, numpy, matplotlib/plotly, Streamlit documentation. University of Helsinki course materials.