Data Dictionary

Sleep Patterns and Academic Performance Study

Author: Monisha Mudunuri and Yamuna Nair
Date: December 9, 2025
Institution: University of Illinois Urbana-Champaign
Project: IS-477 Final Project

RAW DATASETS

CMU Sleep Dataset (cmu-sleep.csv)

Source: https://cmustatistics.github.io/data-repository/
Records: 500 students
Collection Method: Actigraphy devices, 2018-2019

Student Identification:

- subject_id (integer): Unique student identifier (1-500)
- study (integer): Study cohort number (1-5)
- cohort (string): Academic cohort label (lac1, lac2, etc.)

Demographics:

- demo_race (integer): Race/ethnicity code (0-5)
    - 0 = Not reported, 1 = Asian, 2 = Black/African American, 3 = Hispanic/Latino, 4 = White/Caucasian, 5 = Other/Multiple
- demo_gender (integer): Gender code (0-2)
    - 0 = Not reported, 1 = Female, 2 = Male
- demo_firstgen (integer): First generation college student (0-1)
    - 0 = No, 1 = Yes

Sleep Measures:

- TotalSleepTime (float): Average nightly sleep duration in hours (0-12)
- bedtime_mssd (float): Bedtime variability (Mean Squared Successive Difference) in hours (0-5)
- midpoint_sleep (float): Midpoint of sleep period in hours (0-24)
- daytime_sleep (float): Average daytime sleep in hours (0-5)
- frac_nights_with_data (float): Proportion of study nights with valid data (0-1)

Academic Performance:

- GPA (float): Grade point average on 4.0 scale (0-4.0)

Kaggle Student Habits Dataset (student_habits_performance.csv)

Source:
https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance
Records: 300 students
Collection Method: Self-reported survey, 2023-2024

Student Identification:

- student_id (integer): Unique student identifier (1-300)

Demographics:

- age (integer): Student age in years (18-25)
- gender (string): Gender (Male/Female)

Sleep and Health:

- sleep_hours (float): Average nightly sleep duration in hours (0-12)
- physical_activity (integer): Weekly exercise sessions (0-7)
- caffeine_intake (integer): Daily caffeine consumption in cups (0-10)
- stress_level (integer): Self-reported stress rating on 10-point scale (1-10)

Study Habits:

- study_hours_per_week (float): Weekly study time in hours (0-100)
- attendance_rate (float): Class attendance percentage (0-100)
- participation_score (float): Class participation rating (0-10)
- screen_time_hours (float): Daily recreational screen time in hours (0-24)

Academic Performance:

- GPA (float): Grade point average on 4.0 scale (0-4.0)
- academic_performance (float): Overall academic score percentage (0-100)

CLEANED DATASETS

Cleaned CMU Dataset (cleaned_cmu-sleep.csv)

Location: data/processed/Cleaned CSV Datasets/
Records: 485 (15 removed during cleaning)

Cleaning Operations:

- Removed 5 duplicate subject_id records
- Removed 10 records with invalid sleep values (>12 or <0 hours)
- Imputed 12 missing bedtime_mssd values with median
- Imputed 15 missing GPA values with cohort mean
- Standardized column names to lowercase with underscores
- Converted GPA to 0-100 academic_score scale

Student Identification:

- subject_id (integer): Unique student identifier (1-500)

Demographics:

- demo_race (integer): Race code (0-5)
- demo_gender (integer): Gender code (0-2)
- demo_firstgen (integer): First generation status (0-1)

Sleep Measures:

- total_sleep_time (float): Average nightly sleep in hours (4.2-11.8)
- bedtime_mssd (float): Bedtime variability in hours (0.02-4.85)
- sleep_midpoint (float): Sleep midpoint in hours (1.5-6.8)
- daytime_sleep (float): Daytime sleep in hours (0-2.5)

Academic Performance:

- academic_score (float): Academic performance on 0-100 scale (55.0-97.5)

Data Source:

- dataset_source (string): Source identifier ("CMU")

Cleaned Kaggle Dataset (cleaned_student_habits.csv)

Location: data/processed/Cleaned CSV Datasets/
Records: 292 (8 removed during cleaning)

Cleaning Operations:

- Removed 3 records with extreme sleep outliers (>12 hours)
- Dropped 5 records with >20% missing values

- Imputed 5 missing study_hours values with mean
- Converted gender strings to numeric codes (0=Female, 1=Male)
- Standardized variable names to lowercase with underscores

Student Identification:

- student_id (integer): Unique student identifier (1-300)

Demographics:

- age (integer): Student age in years (18-24)
- gender (integer): Gender code (0-1)
    - 0 = Female, 1 = Male

Sleep and Health:

- sleep_hours (float): Average nightly sleep in hours (4.5-11.5)
- physical_activity (integer): Weekly exercise sessions (0-7)
- caffeine_intake (integer): Daily caffeine cups (0-8)
- stress_level (integer): Stress rating (2-9)

Study Habits:

- study_hours (float): Weekly study time in hours (5.0-85.0)
- attendance (float): Class attendance percentage (45-100)
- participation (float): Class participation rating (2-10)
- screen_time (float): Daily screen time in hours (1-16)

Academic Performance:

- academic_score (float): Academic performance on 0-100 scale (52.5-96.25)

Data Source:

- dataset_source (string): Source identifier ("Kaggle")

INTEGRATED DATASET

Integrated Data (integrated_data.csv)

Location: data/integrated/
Records: 488 complete records
Created By: scripts/03_data_integration.py

Integration Method:

- Combined CMU and Kaggle cleaned datasets vertically
- Standardized all variable names and scales across sources
- Harmonized demographic coding systems
- Created unified academic_score metric (0-100 scale)
- Retained dataset_source identifier for stratified analyses

Note: CMU-only variables contain NULL values for Kaggle records, and vice versa. This is intentional and documented via the dataset_source variable.

Student Identification:

- student_id (integer): Unified student identifier across both datasets (1-800)

Demographics:

- age (integer): Student age in years (18-24) [Kaggle only]
- gender (integer): Gender code (0-2)
  - 0 = Female, 1 = Male, 2 = Other/Not Reported
- race (integer): Race code (0-5) [CMU only]
- first_gen (integer): First generation college student (0-1) [CMU only]
- cohort (string): Study cohort label [CMU only]

Sleep Measures:

- total_sleep_time (float): Average nightly sleep in hours (4.2-11.8) [Both datasets]
- sleep_category (string): Derived sleep quality classification [Derived variable]
  - Poor: < 6.0 hours
  - Insufficient: 6.0-6.99 hours
  - Adequate: 7.0-7.99 hours
  - Optimal: ≥ 8.0 hours
- bedtime_variability (float): Sleep schedule consistency in hours (0.02-4.85) [CMU only]
- sleep_midpoint (float): Sleep timing in hours (1.5-6.8) [CMU only]

Health and Lifestyle:

- physical_activity (integer): Weekly exercise sessions (0-7) [Kaggle only]
- caffeine_intake (integer): Daily caffeine cups (0-8) [Kaggle only]
- stress_level (integer): Stress rating (2-9) [Kaggle only]
- screen_time (float): Daily screen time in hours (1-16) [Kaggle only]

Study Habits:

- study_hours (float): Weekly study time in hours (5-85) [Kaggle only]
- attendance (float): Class attendance percentage (45-100) [Kaggle only]
- participation (float): Class participation rating (2-10) [Kaggle only]

Academic Performance:

- academic_score (float): Standardized academic performance on 0-100 scale (52.5-97.5) [Both datasets]
    - For CMU data: academic_score = (GPA / 4.0) × 100
    - For Kaggle data: academic_score = academic_performance (already 0-100)

Data Source:

- dataset_source (string): Original data source identifier (CMU/Kaggle)
- data_completeness (float): Proportion of valid actigraphy nights (0.65-1.0) [CMU only]

MISSING DATA HANDLING

Raw Data Missing Values:

- CMU bedtime_mssd: 12 values (2.4%) - Imputed with median (1.42 hours)
- CMU GPA: 15 values (3.0%) - Imputed with cohort-specific mean
- Kaggle sleep_hours: 8 values (2.7%) - Records dropped (listwise deletion)
- Kaggle study_hours: 5 values (1.7%) - Imputed with mean (32.5 hours)

Final Dataset Completeness:

- All cleaned datasets: 0% missing in core variables
- Integrated dataset: 0% missing in total_sleep_time and academic_score
- Dataset-specific variables: Expected NULL values documented by dataset_source column

DATA QUALITY METRICS

Quality Checks Performed:

- Duplicate detection and removal
- Range validation (all values within valid physiological/academic ranges)
- Outlier detection and treatment
- Consistency checks across variables
- Data type validation and conversion

Sample Sizes:

- CMU: 485 students (99.4% retention after cleaning)
- Kaggle: 292 students (97.3% retention after cleaning)
- Integrated: 488 total records

Data Quality Summary:

- Core variable completeness: 100%
- Valid range compliance: 100%
- CMU actigraphy data completeness: 87% average
- No duplicate records in final datasets