Health Data Dictionary

Complete description of all variables in the Healthcare Risk Assessment dataset (`health_data.csv`).

Dataset Overview

- **Filename**: `health_data.csv`
- **Observations**: 1,000 patients
- **Variables**: 15 total (3 demographic, 5 lifestyle, 6 physiological, 2 outcome)
- **Missing Values**: None
- **Data Type**: Cross-sectional observational study

Variable Definitions

Identifiers

Variable	Type	Range	Unit	Description
`patient_id`	Integer	1-1000	-	Unique patient identifier

Demographics

Variable	Type	Range	Unit	Description
`age`	Numeric	25-80	years	Patient age at assessment
`bmi`	Numeric	18.0-45.0	kg/m²	Body Mass Index (weight/height²)

BMI Categories (for reference):

- Underweight: < 18.5
- Normal weight: 18.5-24.9
- Overweight: 25.0-29.9
- Obese: ≥ 30.0

Lifestyle Factors

These variables represent modifiable behavioral risk factors for cardiovascular disease.

Variable	Type	Range	Unit	Description	Notes
`exercise_hours_week`	Numeric	0.0-12.0	hours	Weekly exercise/physical activity	Includes all moderate-to-vigorous activity
`smoking_years`	Numeric	0.0-40.0	years	Cumulative years of smoking	0 indicates never-smoker or <1 year

Variable	Type	Range	Unit	Description	Notes
`alcohol_units_week`	Numeric	0.0-25.0	units	Weekly alcohol consumption	1 unit = 10mL pure alcohol (standard drink)
`stress_score`	Numeric	1.0-10.0	scale	Self-reported stress level	1=minimal stress, 10=extreme stress
`sleep_hours`	Numeric	4.0-10.0	hours	Average nightly sleep duration	Self-reported typical night

Lifestyle Risk Thresholds (clinical guidelines):

- **Exercise**: < 2.5 hours/week considered insufficient
- **Smoking**: Any smoking history increases CVD risk
- **Alcohol**: > 14 units/week (women) or > 21 units/week (men) considered high risk
- **Stress**: > 7 considered high stress
- **Sleep**: < 6 or > 9 hours associated with increased CVD risk

Physiological Measurements

These variables represent objective clinical measurements of cardiovascular and metabolic health.

Variable	Type	Range	Unit	Description	Clinical Interpretation
`systolic_bp`	Integer	90-180	mmHg	Systolic blood pressure	< 120: Normal 120-139: Elevated 140-159: Stage 1 Hypertension ≥ 160: Stage 2 Hypertension
`diastolic_bp`	Integer	60-110	mmHg	Diastolic blood pressure	< 80: Normal 80-89: Elevated 90-99: Stage 1 Hypertension ≥ 100: Stage 2 Hypertension
`cholesterol`	Integer	120-300	mg/dL	Total cholesterol	< 200: Desirable 200-239: Borderline high ≥ 240: High
`glucose`	Integer	70-200	mg/dL	Fasting blood glucose	70-99: Normal 100-125: Prediabetes ≥ 126: Diabetes
`triglycerides`	Integer	50-300	mg/dL	Triglyceride levels	< 150: Normal 150-199: Borderline high 200-499: High ≥ 500: Very high

Variable	Type	Range	Unit	Description	Clinical Interpretation
`hdl`	Integer	25-80	mg/dL	HDL "good" cholesterol	< 40 (men) or < 50 (women): Low (risk factor) ≥ 60: High (protective)

Notes:

- Blood pressure measurements assume resting state
- Cholesterol and glucose values assume fasting state (8-12 hours)
- HDL inversely related to CVD risk (higher is better)
- All other markers directly related to CVD risk (lower is better)

Outcome Variables

Variable	Type	Range	Values	Description
`cvd_risk_high`	Binary	0-1	0 = Low Risk 1 = High Risk	Cardiovascular disease risk classification
`treatment_group`	Categorical	-	Control Intervention	Lifestyle intervention program assignment

CVD Risk Classification:

- Based on composite risk score incorporating all lifestyle and physiological variables
- High risk: Elevated probability of cardiovascular event within 10 years
- Used as binary outcome for logistic regression

Treatment Groups:

- **Control**: Standard care (health education materials only)
- **Intervention**: 12-week lifestyle program (diet counseling, exercise plan, stress management)
- Random assignment (approximately 50/50 split)
- Physiological measurements reflect post-intervention values

Variable Relationships

Multicollinearity Among Predictors

Lifestyle Factors:

- Exercise and BMI: Moderately negatively correlated (r ≈ -0.35)
- Smoking and alcohol: Weakly positively correlated (r ≈ 0.22)
- Stress and sleep: Moderately negatively correlated (r ≈ -0.28)

Physiological Measurements:

- Systolic BP and diastolic BP: Strongly correlated (r ≈ 0.75)
- Cholesterol and triglycerides: Moderately correlated (r ≈ 0.48)
- HDL and triglycerides: Moderately negatively correlated (r ≈ -0.42)

Between Sets (Lifestyle → Physiological):

- Exercise → BP, cholesterol: Negative correlations (protective)
- Smoking → BP, cholesterol: Positive correlations (risk)
- Stress → BP: Positive correlation
- Sleep → Glucose: Negative correlation

Canonical Structure

Lifestyle Canonical Variate 1 (Unhealthy Pattern):

- High smoking, high alcohol, high stress, low exercise, poor sleep

Physiological Canonical Variate 1 (CVD Risk Profile):

- Elevated BP, high cholesterol, high glucose, high triglycerides, low HDL

Strong canonical correlation (r ≈ 0.71) indicates lifestyle patterns strongly predict physiological health status.

Data Generation Notes

Synthetic Data Properties

This is a **synthetic dataset** generated for educational purposes with the following characteristics:

1. **Realistic distributions**: Variables follow distributions typical of real cardiovascular health data
2. **Known relationships**: Correlations reflect established medical literature
3. **Controlled structure**: Designed to demonstrate specific multivariate methods
4. **No real patients**: All data is simulated; no actual patient information

Generation Process

1. **Lifestyle factors**: Generated from appropriate distributions (normal, exponential)
2. **Physiological measurements**: Linear combinations of lifestyle + age + BMI + random noise
3. **Intervention effect**: Modest improvements (3-8 points) in physiological measures for intervention group
4. **CVD risk**: Composite risk score based on all predictors, dichotomized at median

Limitations for Teaching

- **Simplified relationships**: Real biology is more complex with nonlinear effects and interactions
- **No missing data**: Real healthcare data has substantial missingness
- **Cross-sectional**: No temporal dynamics or repeated measurements
- **No confounding**: Simplified causal structure for clarity

- **Balanced groups**: Real studies often have unequal sample sizes

Statistical Analysis Suitability

Logistic Regression

- **Binary outcome**: `cvd_risk_high` (0/1)
- **Multiple predictors**: All lifestyle and physiological variables
- **Sample size**: n=1000 adequate for ~15 predictors
- **Separation**: Reasonable overlap between risk groups (not perfectly separable)

Hotelling's T-squared

- **Two groups**: cvd_risk_high (0 vs. 1) or treatment_group (Control vs. Intervention)
- **Multiple outcomes**: 6 physiological measurements
- **Assumptions**: Approximately multivariate normal, equal covariance matrices

MANOVA

- **Independent variable**: treatment_group (2 levels)
- **Dependent variables**: systolic_bp, diastolic_bp, cholesterol, glucose (4 outcomes)
- **Sample size**: n=1000 with balanced groups (adequate power)
- **Assumptions**: Multivariate normality, homogeneity of covariance (testable with Box's M)

Canonical Correlation

- **Set 1 (Lifestyle)**: 5 variables (exercise_hours_week, smoking_years, alcohol_units_week, stress_score, sleep_hours)
- **Set 2 (Physiological)**: 6 variables (systolic_bp, diastolic_bp, cholesterol, glucose, triglycerides, hdl)
- **Number of pairs**: min(5, 6) = 5 canonical correlations
- **Sample size**: n=1000 adequate for 11 total variables

Box's M Test

- **Purpose**: Test equality of covariance matrices between treatment groups
- **Variables**: 6 physiological measurements
- **Groups**: 2 (Control vs. Intervention)
- **Sensitivity**: Test is sensitive to non-normality; interpret cautiously

Data Quality

Completeness

- No missing values (complete case analysis)
- All measurements within physiologically plausible ranges
- No data entry errors or outliers beyond clinical possibility

Validity

- Distributions match population norms for cardiovascular health studies
- Correlations consistent with epidemiological literature
- Treatment effects reflect realistic intervention impact

Reliability

- Measurements assumed to be taken with standard clinical protocols
- Random seed (42) ensures reproducibility of synthetic data
- Consistent units and scales across all observations

Usage Recommendations

Data Preprocessing

1. **Standardization**: Recommended for analyses sensitive to scale (canonical correlation, factor analysis)
2. **Outlier detection**: Check for extreme values, though rare in this synthetic dataset
3. **Assumption checking**: Test multivariate normality (Mardia's test, Q-Q plots) before parametric methods

Variable Selection

- Use all lifestyle factors for comprehensive risk prediction
- Consider removing one of systolic_bp/diastolic_bp if multicollinearity problematic
- HDL may need sign reversal for interpretation (higher is better)

Interpretation

- Remember synthetic nature when discussing clinical implications
- Focus on methodological demonstration rather than clinical discovery
- Compare statistical findings to known cardiovascular risk literature

References

Clinical Guidelines:

- American Heart Association (AHA) guidelines for blood pressure
- National Cholesterol Education Program (NCEP) ATP III guidelines

- American Diabetes Association (ADA) glucose thresholds
- World Health Organization (WHO) BMI categories

Statistical Methods:

- Logistic regression for binary outcomes
- Hotelling's T-squared for multivariate mean comparison
- MANOVA for multiple dependent variables
- Canonical correlation for relating variable sets