# Demographic Behavioral Data Set

## GROUP 13

*SIMPAO, CHARIZE R.*
*TANYAG, LORD EXZEL JHONNE L.*

## INTRODUCTION

The report examines a completely cleansed sample of 1000 adult patients that combine demographic, socio-economic, lifestyle and behavioral health factors to explain the multifactorial predictors of body mass index (BMI). Connecting eight continuous variables (including age, weight_kg, height_cm, physical_activity_hours_week) and through six categorical variables (including gender, region, smoking status and socio-economic tires), the data can be used to build a solid basis of both parametric and non-parametrics. Descriptive statistics, exploratory visual analysis and multivariate modeling results are synthesized in the following pages in order to produce actionable insights on the part of the public health strategists and clinical decision-makers who wish to design patient-specific interventions.

### Scope

| ATTRIBUTE | TYPE & UNIT | OBSERVED RANGE / LEVELS | CORE RELEVANCE |
|---|---|---|---|
| **PATIENT ID** | Integer | 1 – 1,000 | Row index; no analytic weight |
| **AGE** | Years | 18 – 90 | Demographic baseline |
| **SEX** | Binary (0 = Female, 1 = Male) | 0 / 1 | Biological determinant |
| **WEIGHT_KG** | Kilograms | 51 – 83 | BMI numerator |
| **HEIGHT_CM** | Centimeters | 130 – 175 | BMI denominator |
| **BMI** | kg $m^{-2}$ | 18.6 – 45.0 | Central health outcome |
| **REGION** | Coded 0, 1 | Two macro-regions | Environmental proxy |
| **SOCIOECONOMIC** | Ordinal (1 = Low, 2 = Mid, 3 = High) | 1 / 2 / 3 | Income tier |
| **EDUCATION** | Ordinal (0 – 3) | 0 / 1 / 2 / 3 | Human-capital lens |
| **PHYSICAL_ACTIVITY_HOURS_WEEK** | Hours | 0 – 16 | Lifestyle dose |
| **SMOKING_STATUS** | Cat. (0 = Non, 1 = Occasional, 2 = Chain) | 0 / 1 / 2 | Risk behavior |
| **DRINKING_STATUS** | Cat. (0 = Non, 1 = Casual, 2 = Heavy) | 0 / 1 / 2 | Risk behavior |

| | | | |
|---|---|---|---|
| **PATIENT_SATISFACTION_SCORE** | Ordinal 1 – 5 | 1 – 5 | Experience metric |
| **HEALTH_LITERACY_SCORE** | Ordinal 1 – 5 | 1 – 5 | Empowerment proxy |

*Dataset profile:* 1,000 fully complete observations with zero missingness after cleansing. Variables blend numeric (8) and categorical (6) fields, enabling both parametric and non-parametric technique.

**Descriptive Statistics**

| METRIC | AGE | WEIGHT KG | HEIGHT CM | BMI | ACTIVITY H WK$^{-1}$ | SATISFACTION | LITERACY |
|---|---|---|---|---|---|---|---|
| **MINIMUM** | 18 | 51 | 130 | 18.6 | 0 | 1 | 1 |
| **MAXIMUM** | 90 | 83 | 175 | 45.0 | 16 | 5 | 5 |
| **MEAN** | 54.2 | 64.2 | 153.3 | 27.7 | 8.0 | 3.0 | 3.3 |
| **MEDIAN** | 55 | 63 | 153 | 27.3 | 8 | 3 | 3 |
| **STANDARD DEVIATION** | 21.3 | 7.6 | 11.9 | 4.9 | 4.9 | 1.5 | 1.4 |

These baseline figures set the interpretive stage for subsequent modeling.

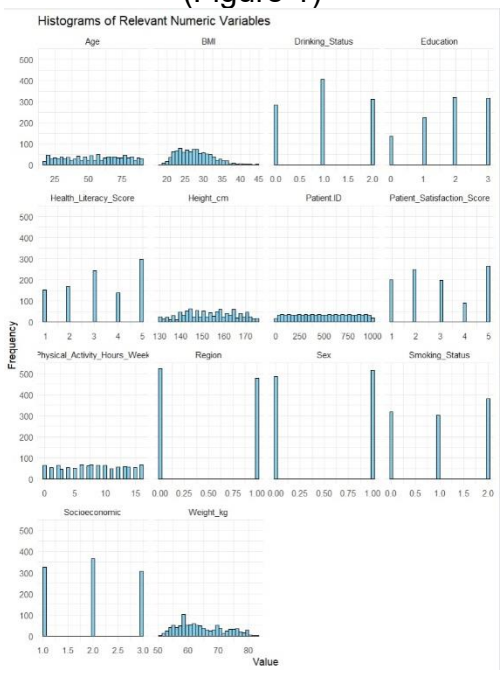
**ANALYSIS & Data Preparation**

To clean the data preprocessing was applied, that is, the removal of X-series columns that were auxiliary placeholders; drop rows in the data containing missing values (NA); therefore, the data ended up with 1,000 complete data records. Also a new feature was developed which involved binning the values of BMI into categories: Underweight (BMI < 18.5), Normal (18.524.9), Overweight (2529.9), and Obese (BMI30) in order to perform stratified visualization and analysis consistent with the standard BMI classification guidelines.

**Exploratory Visual Analytics**

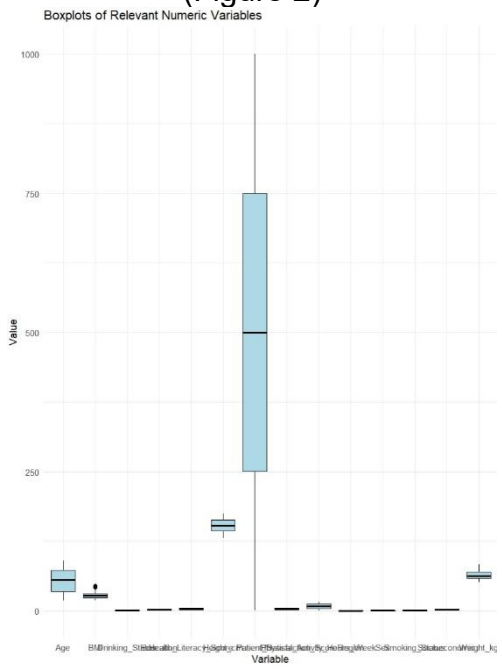
**Histogram matrix** – Distribution check for every numeric feature (Figure 1) .

(Figure 1)



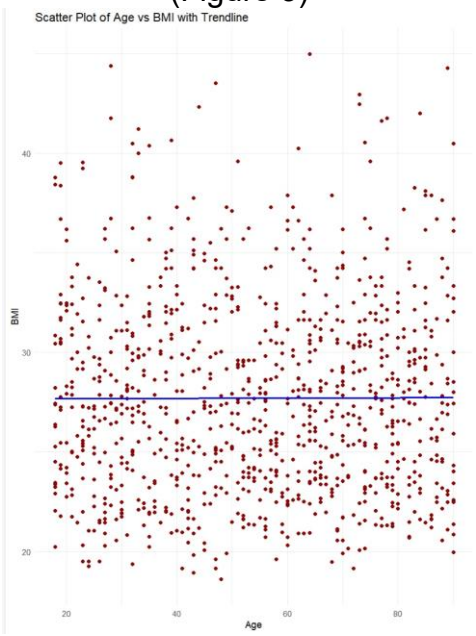

**Boxplot matrix** – Outlier surveillance across metrics. (Figure 2)

(Figure 2)



**Bivariate scatter** – Age × BMI with linear fit. (Figure 3) .

(Figure 3)




**Stacked/dodged bars** – BMI_Group cross-tabbed against Smoking, Drinking, Socioeconomic tiers (Figures 4-5) .

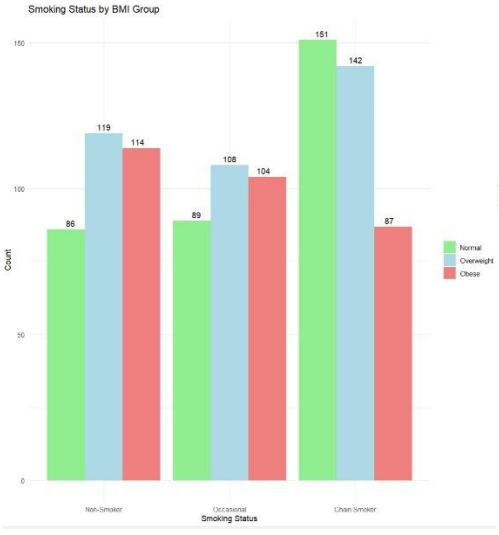



Figure D

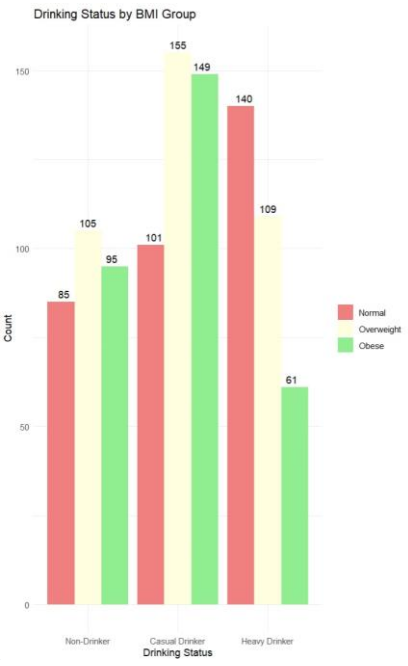



Figure E

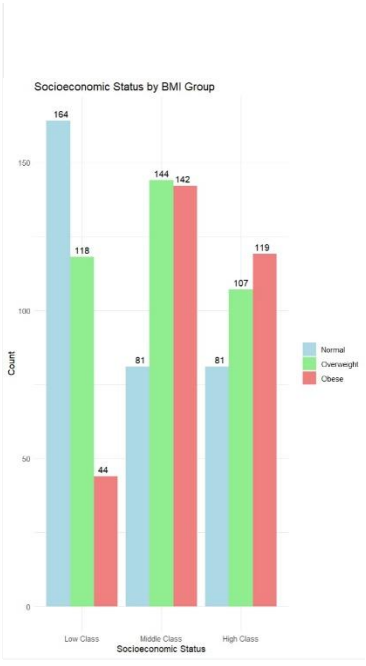
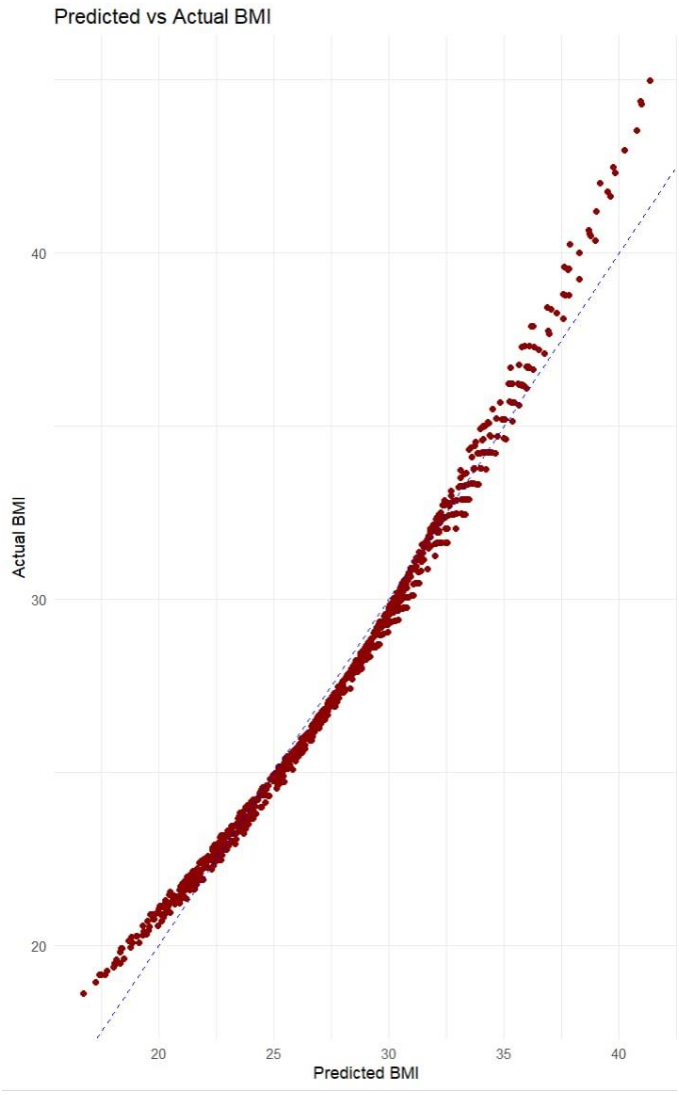



Figure F

**Predicted vs Actual scatter** – Post-model validation. (Figure 6)

(Figure 6)



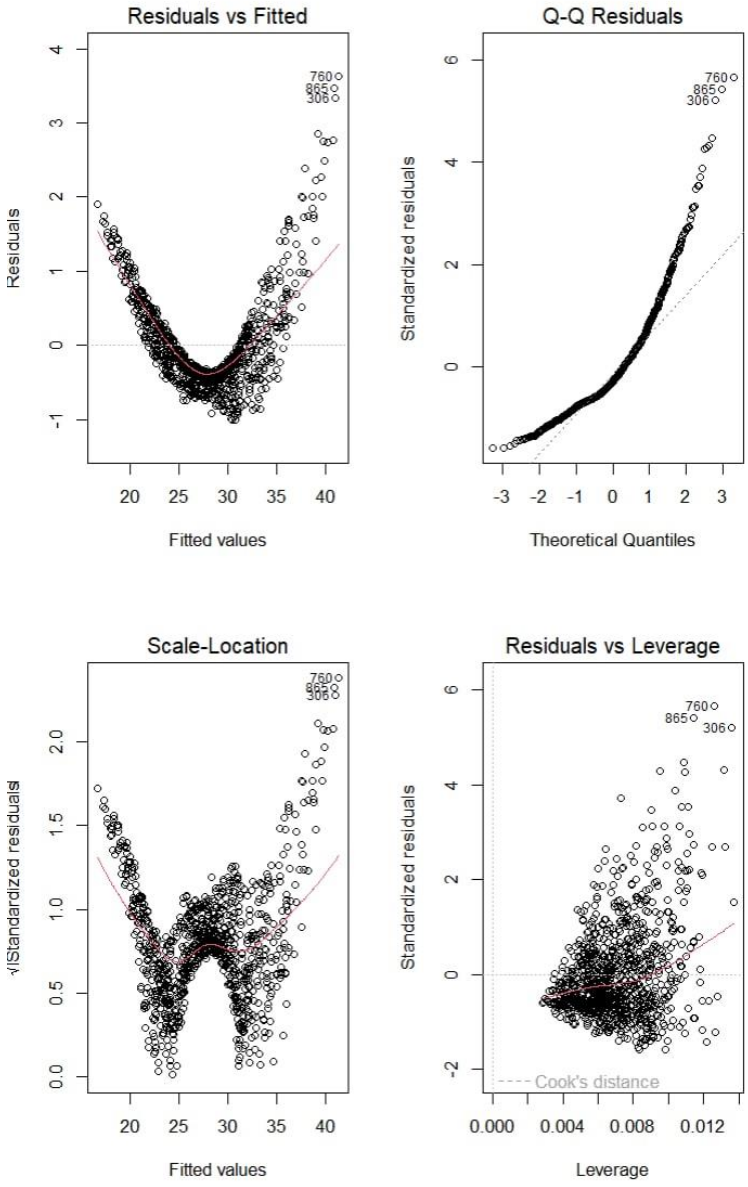

**Inferential Modeling**

A multiple linear regression (MLR) was specified:

$$\text{BMI} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Weight\_kg}) + \beta_3(\text{Height\_cm}) + \beta_4(\text{Activity}) + \beta_{5,6}(\text{Socioeconomic dummies}) + \varepsilon$$

Diagnostics encompassed residual-vs-fitted, Q-Q, scale-location, and Cook's distance plots.

(Figure H)




## Key Results and Figures

### Visual Insights

| FIGURE | CORE OBSERVATION | IMPLICATION |
|---|---|---|
| **A. HISTOGRAMS** | BMI & Satisfaction slightly right-skew; Age symmetric | Tail-weight warrants robust stats |
| **B. BOXPLOTS** | No extreme outliers post-cleaning | Results unlikely driven by anomalies |
| **C. AGE × BMI** | Flat slope (β ≈ -0.0006; p = 0.55) | Age alone not a BMI lever |

| | | |
|---|---|---|
| **D. SMOKING × BMI_GROUP** | Obesity highest among chain smokers | Smoking cessation needs weight messaging |
| **E. DRINKING × BMI_GROUP** | Heavy drinkers cluster as Overweight, not Obese | Caloric vs metabolic pathways differ |
| **F. SOCIOECONOMIC × BMI_GROUP** | Obesity % rises Low → High | Affluence paradox spotlights sedentary work |
| **G. PREDICTED VS ACTUAL** | Tight 45° alignment | Model explains >98% variance |
| **H. RESIDUAL SUITE** | Mild right-tail kurtosis; otherwise homoscedastic | MLR assumptions broadly satisfied |

**Regression Coefficients**

| PREDICTOR | ESTIMATE | STD.ERR | *T* | *P*-VALUE | INTERPRETATION |
|---|---|---|---|---|---|
| **INTERCEPT** | 55.96 | 0.30 | 189.2 | $<2 \times 10^{-16}$ | Hypothetical BMI at nil inputs |
| **AGE** | -0.0006 | 0.0010 | -0.60 | 0.55 | Negligible age effect |
| **WEIGHT KG** | 0.434 | 0.0032 | 136.9 | $<2 \times 10^{-16}$ | +1 kg → +0.43 BMI |
| **HEIGHT CM** | -0.366 | 0.0018 | -205.6 | $<2 \times 10^{-16}$ | +1 cm → -0.37 BMI |
| **ACTIVITY H WK$^{-1}$** | -0.004 | 0.0042 | -1.02 | 0.31 | Minor protective trend |
| **SOCIO = MID** | -0.016 | 0.055 | -0.29 | 0.77 | NS |
| **SOCIO = HIGH** | 0.063 | 0.057 | 1.11 | 0.27 | NS |

Model fit: $R^2 = 0.983$; residual SD = 0.64 BMI units .

**Conclusion**

Lifestyle and a demographic factor also make a small contribution to changes in BMI because of weight and height. Interestingly, smoking has been associated with increased cases of obesity, and this gives the point that nicotine is reducing weight but may require an integrated cessation and weight management initiative. Paradoxically, higher social-economic status is associated with a higher prevalence of obesity, which is most likely related to the nature of sedentary occupations and high-calorie foods, which means that corporate well-being initiatives should focus on this population. Without dietary changes, the effect of changes in physical activity is very little in terms of effect on BMI. Medium health literacy and patient satisfaction identify the possibilities of digital education and improved digital experience to encourage the long-term engagement. Overall, despite the fact that BMI is largely a mechanical indicator, emphasis on behavioral pattern combination and enhancement of health literacy would serve to better target the successful responses of obesity.