REPORT AND ANALYSIS

Why Health Literacy Doesn't Predict Patient Satisfaction

Demographic Behavioral data

Prepared By

Brylle Matthew A. Lupac

Naomi Christienne Tiama

1. Introduction

This dataset provides detailed demographic, anthropometric, lifestyle, and subjective health information for 1000 patients, designed to support health profiling and population-level analyses. Each record includes patient-specific variables such as age, sex, weight (kg), height (cm), and calculated Body Mass Index (BMI), allowing for the assessment of physical health status across individuals. In addition, the dataset captures contextual factors like geographic region (urban or rural), socioeconomic class, and level of educational attainment, which are crucial for understanding social determinants of health. Behavioral data, including weekly physical activity hours, smoking status, and drinking status, offer insight into each individual's health behaviors and risk factors. Subjective measures such as patient satisfaction scores and health literacy ratings are also included, enabling a broader exploration of healthcare engagement and knowledge. The dataset uses both coded numerical values and descriptive labels (e.g., "extremely satisfied," "non-smoker") to support flexible analysis. Its multidimensional nature makes it suitable for studies on health disparities, behavior-health outcome associations, and the influence of sociodemographic variables on perceived and actual well-being.

2. Methods

2.1. Data Cleaning and Preparation

- The dataset titled *2_Demographic_Behavioral_data_Group_005.csv* was imported using the read.csv() function from base R. The tidyverse package was loaded to enable streamlined data manipulation and visualization. The colnames() function was used to inspect column names and identify relevant variables for analysis. The group then selected the most relevant columns using the select() function from the dplyr package.
- After selection, data types were cleaned and converted using the mutate() function. Specifically, several categorical variables were recoded into factor format with descriptive labels for clarity:
    - **Sex** was labeled as "Female" and "Male."
    - **Education** was mapped to "Uneducated," "Primary," "Secondary," and "Tertiary."
    - **Smoking_Status** was relabeled as "Non-Smoker," "Occasional Smoker," and "Chainsmoker."
    - **Drinking_Status** was relabeled as "Non-drinker," "Casual drinker," and "Heavy drinker."
    - Other variables such as **Region** and **Socioeconomic** status were also converted into factor types.
- To ensure clean data, all incomplete rows were removed using the drop_na() function, which eliminated records with missing values across any variable.

2.2. Descriptive Statistical Analysis

- Summary statistics for all variables were generated using the summary() function to view minimum, maximum, and quartile distributions. Additionally, a more detailed summary of selected continuous variables was computed using the summarise_all() function, which calculated:
    - **Mean** – the arithmetic average
    - **Median** – the midpoint of ordered values
    - **Standard Deviation (SD)** – the spread or variability around the mean
- Variables included in this summary were:
    - **Age**

- **Weight (kg)**, **Height (cm)**, and **BMI**
- **Physical Activity Hours per Week**

2.3. Data Visualization

- **Histogram:** A histogram was plotted to show the distribution of physical activity (in hours per week). The histogram used 12 bins and was styled with blue coloring and a minimal theme.
- **Boxplot:** A boxplot was generated to compare **BMI** across different **Drinking Status** categories. Each box represented one drinking level: Non-drinker, Casual drinker, or Heavy drinker.
- **Scatter Plot:** A scatter plot was created to assess the relationship between **Health Literacy Score** and **Patient Satisfaction Score**. A linear regression line (geom_smooth(method = "lm")) was added to visualize the trend.
- **Bar Plot:** A bar graph was plotted to display the frequency distribution of **Education levels**. Each education level was labeled and color-coded using a pastel palette for clear interpretation.

2.4. Advanced Statistical Test

- To examine the potential association between **Health Literacy Score** and **Patient Satisfaction Score**, the group conducted a **Pearson correlation test** using the cor.test() function in R. This test returned a **correlation coefficient (r)** and a **p-value** to indicate:
  - The strength and direction of the relationship
  - Whether the association was statistically significant

3. Results and Discussion

3.1. Summary Statistics

Table 1. Descriptive Statistics for Selected Variables

Variable	Mean	Median	Standard Deviation
Age (years)	54.8	55	21.34
Weight (kg)	63.88	63	7.42
Height (cm)	153.61	154	11.79
Body Mass Index (BMI)	27.44	26.78	4.74

Physical Activity (hours/week)	7.87	8	4.79

- The average age of patients was 54.8 years, with a median of 55, indicating a relatively balanced age distribution. The standard deviation of 21.34 suggests a wide age range.
- Patients had an average weight of 63.88 kg and height of 153.61 cm, with moderate variability (SD = 7.42 and 11.79, respectively). The average BMI was 27.44, slightly above the normal range, indicating that many patients may be overweight.
- On average, patients engaged in 7.87 hours of physical activity per week, with a median of 8 hours, showing generally consistent physical activity levels across the group (SD = 4.79).

3.1. Data Visualizations

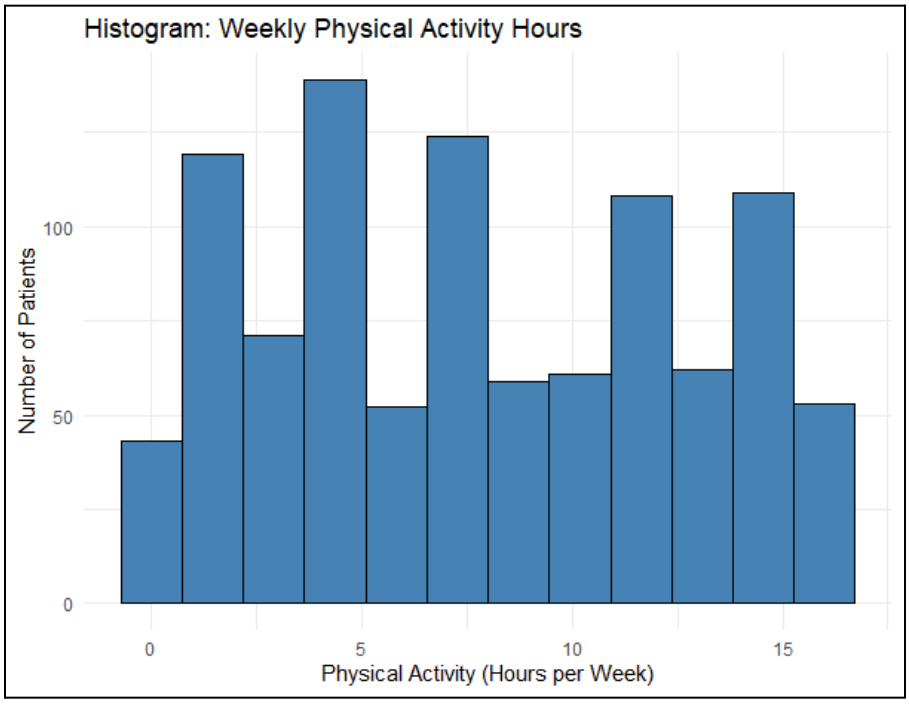

Figure 1. Histogram: Weekly Physical Activity Hours

- The number of patients engaging in physical activity ranged from 0 to 16 hours per week.
- The most common activity levels were around 4 to 8 hours per week, with frequencies ranging from 63 to 76 patients, indicating this as the typical range.
- A small number of patients (around 43) reported 0 hours of activity, showing a sedentary group.
- The distribution appears fairly balanced, with no extreme peaks or gaps, suggesting that most patients engage in some form of physical activity each week.
- Overall, the histogram shows a moderate to active population, with very few individuals reporting no activity at all.

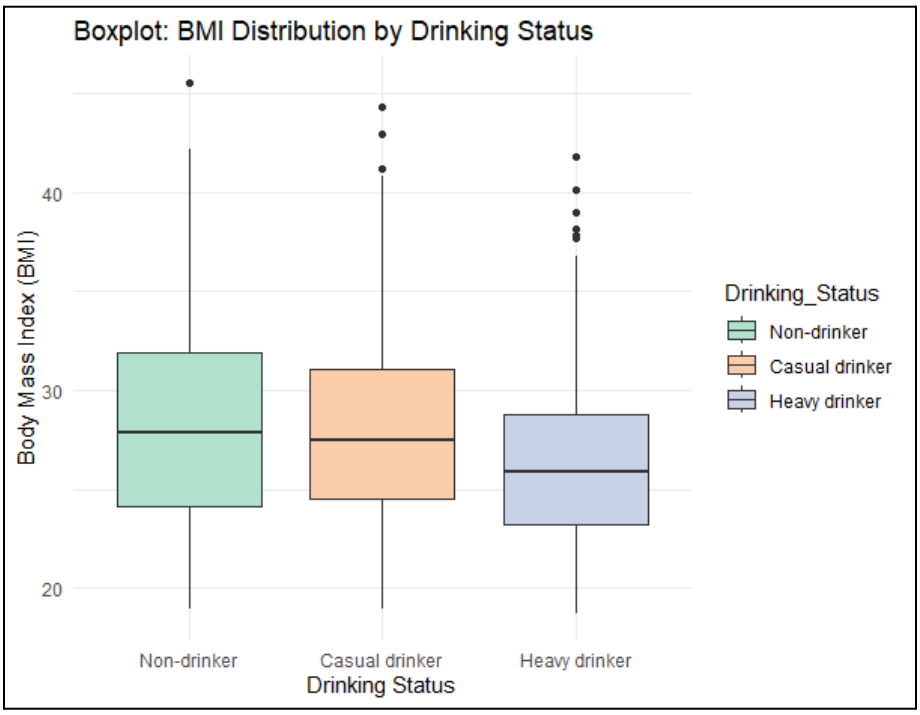

Figure 2. Boxplot: BMI Distribution by Drinking Status

- Most patients were casual drinkers (388), followed by heavy drinkers (364), and the fewest were non-drinkers (248).
- The median BMI was similar across all drinking groups, suggesting minimal differences in central tendency.
- Heavy drinkers had the widest BMI range, indicating more variability in body weight.
- Outliers were present in each group, showing individual differences not directly linked to drinking status.
- Overall, drinking status did not show a strong impact on BMI distribution.

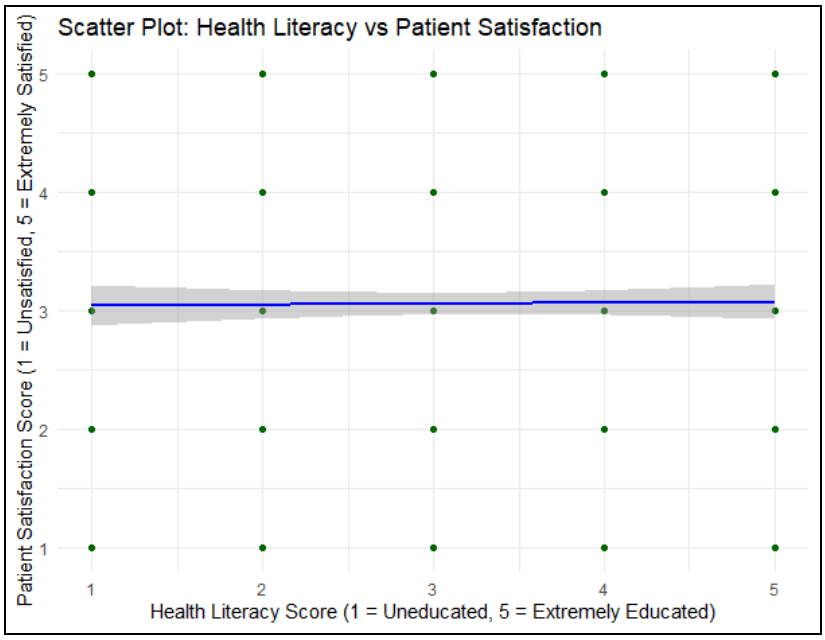

Figure 3. Scatter Plot: Health Literacy vs Patient Satisfaction

- The scatter plot shows a weak to no visible correlation between health literacy and patient satisfaction.
- The trend line is nearly flat, indicating no meaningful upward or downward relationship between the two variables.
- Patients with higher health literacy scores did not consistently report higher or lower satisfaction levels.

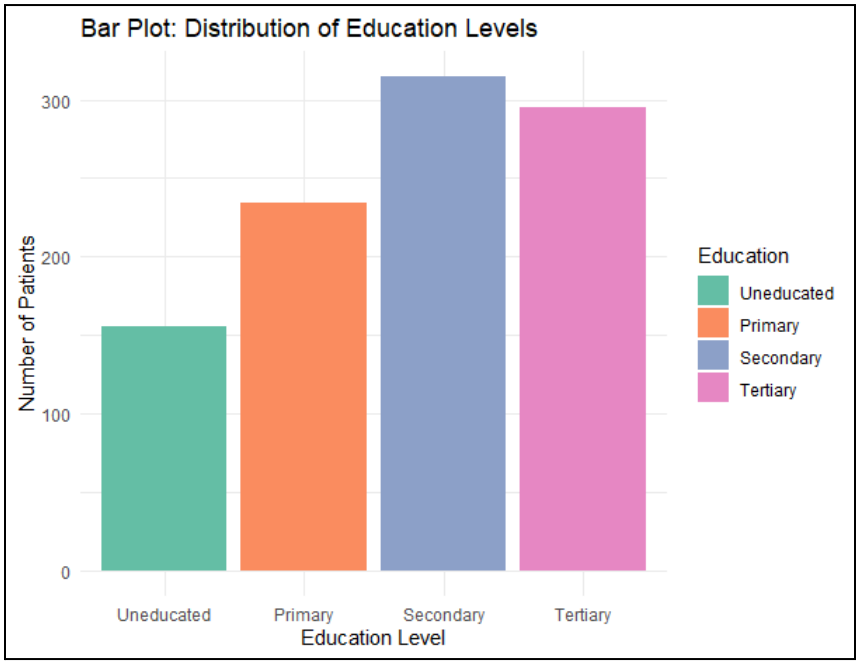

Figure 4. Bar Plot: Distribution of Education Levels

- The majority of patients had either a secondary (315) or tertiary education (295).
- Primary education was reported by 234 patients, making it the third most common level.

- The fewest patients (156) were uneducated, indicating limited representation of this group.
- Overall, the bar plot shows that most patients had at least some formal education, with higher education levels being more common.

2.3 Advanced Statistical Insight

Table 2. Pearson Correlation between Health Literacy and Patient Satisfaction

Statistic	Value
Correlation Coefficient (r)	0.0076
t-value	0.2392
Degrees of Freedom (df)	998
p-value	0.811
95% Confidence Interval (CI)	-0.0544 to 0.0695
Interpretation	No significant correlation

- The Pearson correlation coefficient (r = 0.0076) indicates almost no linear relationship between health literacy and patient satisfaction.
- The p-value of 0.811 is much higher than 0.05, meaning the correlation is not statistically significant.
- The 95% confidence interval ranges from -0.0544 to 0.0695, which includes zero—further confirming no meaningful correlation.
- This suggests that in this dataset, higher health literacy does not predict or influence patient satisfaction.

5. Conclusion

This analysis of a diverse patient dataset revealed key insights into demographics, health behaviors, and subjective health perceptions. Descriptive statistics showed that the average patient was middle-aged, slightly overweight, and moderately active. Visual analyses highlighted that most individuals engaged in 4–8 hours of weekly physical activity, and BMI levels were relatively consistent across drinking status groups, with heavy drinkers showing slightly more variability. Education was skewed toward higher levels, with most patients having secondary or tertiary education. Importantly, the correlation analysis between health literacy and patient satisfaction revealed no significant association ($r = 0.0076$, $p = 0.811$), indicating that in this population, increased health literacy did not translate into higher satisfaction with healthcare. These findings suggest that while demographic and behavioral patterns are observable, subjective perceptions like satisfaction may be influenced by other, unmeasured factors beyond health knowledge.