# Demographic Behavioral Data

**Baniqued, Dwayne Timothy[1], Loyola, Audrey Faith, R.[1]**

[1]Students, BI120L, CON29, School of Health Sciences, Mapúa University

## INTRODUCTION

This report provides a comprehensive analysis of a demographic and behavioral dataset comprising information from 1000 individual patients. The dataset encompasses a variety of variables, including patient identification, age, sex, body measurements such as weight and height, Body Mass Index (BMI), geographical region, socioeconomic status, educational attainment, weekly physical activity hours, smoking habits, alcohol consumption patterns, patient satisfaction scores, and health literacy levels. The primary goal of this analysis is to thoroughly investigate the interrelationships among these variables, understand the distribution characteristics of key demographic and behavioral factors, and identify any statistically significant differences or correlations that may exist within the data.

## METHODS

The entire analytical process was executed using R, leveraging its powerful capabilities for data manipulation, descriptive statistical computation, and advanced data visualization.

The initial step involved loading the raw dataset, which was provided in a CSV file format, directly into an R data frame for processing. Following this, column names were systematically standardized using the janitor::clean_names() function to ensure consistency and ease of access. A crucial part of the data preparation involved handling missing values; specifically, any columns exhibiting more than 80% missing data were entirely removed from the dataset. Furthermore, rows containing missing values in variables deemed critical for the core analysis—namely age, sex, BMI, physical activity hours, and smoking status—were excluded to maintain the integrity and reliability of subsequent analyses. Finally, several key variables transformed: the sex variable was converted into a factor with clearly defined levels of "Female" (represented by 0) and "Male" (represented by 1), and similarly, smoking_status was transformed into a factor with descriptive labels "Non-Smoker" (0), "Occasional" (1), and "Chainsmoker" (2).

To gain an initial understanding of the dataset's characteristics, comprehensive descriptive statistics were generated. The summary() function provided a concise overview of all variables, presenting key metrics such as quartiles, means, and counts for categorical (factor) variables. For a more in-depth statistical description of the numeric variables, the psych::describe() function was employed. This yielded detailed insights, including standard deviation, median, trimmed mean, skewness, and kurtosis, offering a robust statistical profile of the quantitative data.

ggplot2, a highly versatile visualization package in R, was utilized to create a series of informative plots. These visualizations were instrumental in exploring data distributions and relationships visually. A histogram was generated to illustrate the distribution of health literacy scores, segmented by sex, allowing for a direct comparison of patterns between genders. A bar plot was created to display the frequency of different smoking statuses, providing a clear count for each category. A box plot was used to visualize the distribution of BMI across females and males, effectively highlighting their respective medians, interquartile ranges, and any existing outliers. Lastly, a scatter plot was constructed to examine the relationship between BMI and weekly physical activity hours, with individual data points color-coded by sex and augmented with a linear model smooth line to indicate overall trends.

Beyond descriptive statistics and visualizations, formal statistical tests were conducted to infer relationships and differences within the data. A Pearson correlation test (cor.test()) was performed to quantitatively assess the linear association between BMI and physical activity hours per week. To investigate potential differences in mean BMI between females and males, a Welch Two-Sample t-test (t.test()) was carried out. This particular t-test was chosen for its robustness in situations where the variances of the two groups might not be equal. Finally, a comprehensive correlation matrix was computed for all numeric variables in the dataset, providing a tabular summary of pairwise linear relationships across the quantitative measures.

## RESULTS AND FIGURES

The cleaned dataset, denoted as df_clean, consists of 1000 complete observations, ready for analysis. The age of participants spans a wide range, from 18 to 90 years, with an average age of approximately 55.28 years. The dataset exhibits a nearly balanced distribution between sexes, comprising 490 females and 510 males. Body Mass Index (BMI) values in the dataset range from a minimum of 19.05 to a maximum of 44.29, with the average BMI calculated at 27.27. In terms of physical activity, participants reported engaging in activities for 0 to 16 hours per week, with the mean weekly activity standing at 7.85 hours. Regarding smoking status, the dataset shows that 265 individuals identify as Non-Smokers, 287 as Occasional smokers, and a notable proportion of 448 individuals are classified as Chainsmokers.

The histogram illustrating the Health Literacy Score Distribution by Sex (Figure 1) reveals that both female and male participants exhibit similar patterns in their health literacy scores. Across both groups, there is a tendency for scores to be higher, clustering towards the maximum score of 5.

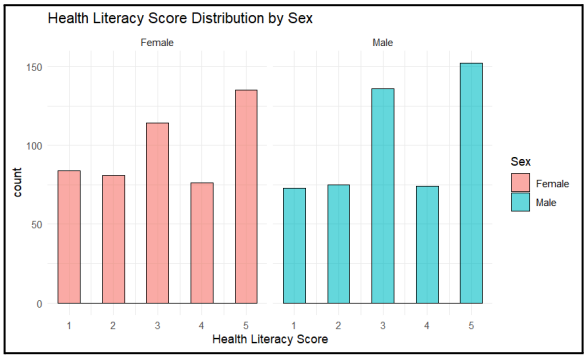


Figure 1: Health Literacy Score Distribution by Sex

The bar plot depicting Smoking Status Frequency (Figure 2) clearly shows the distribution of smoking habits within the dataset. "Chainsmokers" represent the largest group, followed by "Occasional" smokers, and then "Non-Smokers" as the smallest category.

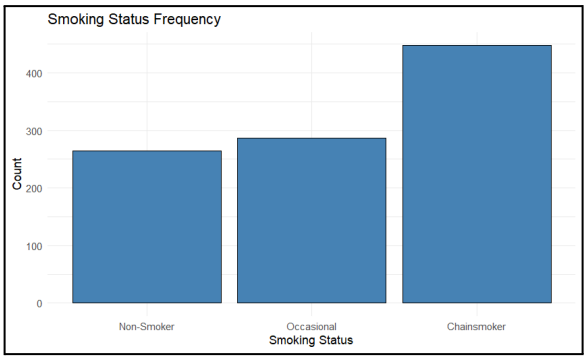


Figure 2: Smoking Status Frequency

The box plot of BMI by Sex (Figure 3) indicates a noticeable difference in BMI distributions between females and males. Females generally exhibit a higher median BMI and a wider interquartile range compared to males, suggesting a higher average BMI and greater variability among female participants.

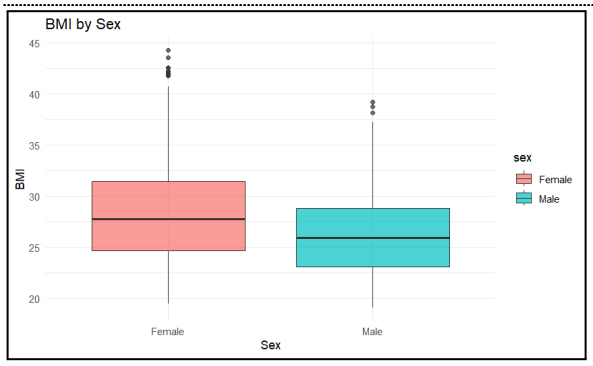


Figure 3: BMI by Sex

The scatter plot illustrating BMI vs Physical Activity (Figure 4) explores the relationship between these two variables, with points differentiated by sex. The plot includes a dashed black line representing a linear regression fit, which appears almost flat. This visual suggests a very weak, if any, linear relationship between physical activity hours per week and BMI for both sexes combined.

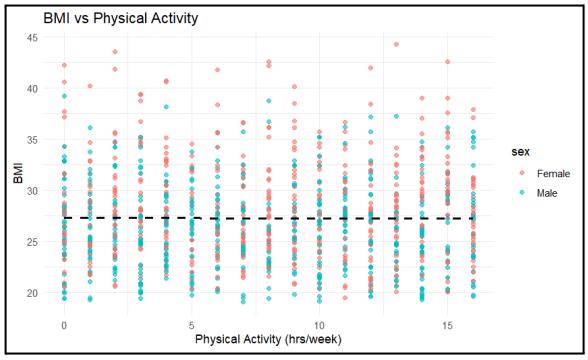


Figure 4: BMI vs Physical Activity

Table 1. Correlation: BMI and Physical Activity

```
Pearson's product-moment correlation

data:  df_clean$bmi and
df_clean$physical_activity_hours_week
t = -0.011874, df = 998, p-value = 0.9905
alternative hypothesis: true correlation is not equal
to 0
95 percent confidence interval:
 -0.06236748  0.06161866
sample estimates:
         cor
-0.0003758547
```

A Pearson's product-moment correlation test was conducted between BMI and physical activity hours per week. The results showed a correlation coefficient (r) of -0.0003758547, with a p-value of 0.9905. This indicates an extremely weak and statistically non-significant linear relationship between BMI and physical activity. The 95% confidence interval for the correlation coefficient ranges from -0.06236748 to 0.06161866, further confirming the lack of a substantial linear association.

Table 2. t-test: BMI by Sex

```
Welch Two Sample t-test

data:  bmi by sex
t = 6.8713, df = 963.93, p-value = 1.141e-11
alternative hypothesis: true difference in means
between group Female and group Male is not
equal to 0
95 percent confidence interval:
 1.435791 2.583777
sample estimates:
mean in group Female   mean in group Male
           28.29114            26.28135
```

A Welch Two Sample t-test was performed to compare the mean BMI between females and males. The test yielded a t-statistic of 6.8713 with 963.93 degrees of freedom, and a highly significant p-value of . The sample estimates show that the mean BMI for females is 28.29114, while for males it is 26.28135. The 95% confidence interval for the true difference in means between females and males is [1.435791, 2.583777]. This result strongly suggests that there is a statistically significant difference in mean BMI between females and males, with females having a higher average BMI.

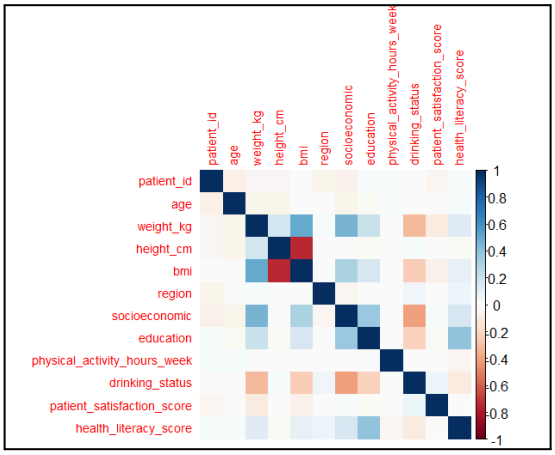


Figure 5. Correlation Matrix

The correlation matrix reveals several notable relationships among the variables. There is a strong negative correlation between Body Mass Index (BMI) and height (r = -0.75), indicating that as an individual's height increases, their BMI tends to decrease—an expected outcome given the formula used to calculate BMI. Additionally, BMI shows a moderate positive correlation with weight (r = 0.51), which is also consistent with expectations. Socioeconomic status is moderately positively correlated with weight (r = 0.46), suggesting that individuals with higher socioeconomic standing tend to weigh more. Education level demonstrates a moderate positive correlation with health literacy score (HLS) (r = 0.39), implying that higher levels of education are associated with better health literacy. Lastly, there is a moderate negative correlation between drinking status and socioeconomic status (r = -0.41), indicating that more frequent alcohol consumption tends to be associated with lower socioeconomic status.

## INTERPRETATION AND CONCLUSION

This analysis of the demographic and behavioral dataset provides several key insights. The data cleaning process successfully prepared the dataset for robust analysis by handling missing values and transforming variables appropriately. Descriptive statistics offered a clear overview of the patient population, highlighting the age distribution, near-equal sex representation, and the prevalence of different smoking statuses, with "Chainsmokers" being the most common group.

Visualizations further elucidated these patterns. The health literacy scores show similar distributions across both sexes, generally leaning towards higher scores. The box plot for BMI by sex revealed a statistically significant difference, with females having a higher average BMI compared to males, as confirmed by the Welch Two-Sample t-test (p\<0.001). This difference is a notable finding that warrants further investigation into potential contributing factors.

Conversely, the scatter plot and Pearson correlation test between BMI and physical activity hours per week indicated virtually no linear relationship. The correlation coefficient was extremely close to zero, and the p-value was very high (0.9905), leading to the conclusion that physical activity, as measured in this dataset, does not have a linear association with BMI. This unexpected finding suggests that other factors or more nuanced relationships might be at play, or that the measurement of physical activity might not fully capture its impact on BMI.

The correlation matrix provided a broader view of relationships among numeric variables. Strong expected correlations were observed between BMI, Wt. (kg), and Ht. (cm). Additionally, positive correlations between Educ. and HLS, and Socioeco. status and Wt. (kg) were identified. A negative correlation between Drink. Status and Socioeco. status also emerged.

In conclusion, while the dataset revealed a significant difference in BMI between sexes and interesting correlations among socioeconomic, education, and health literacy variables, the anticipated linear relationship between BMI and physical activity was not observed. These findings highlight the complexity of health-related behaviors and demographic factors, suggesting avenues for more in-depth research, potentially involving non-linear models or additional confounding variables, to fully understand these interactions.