Group Project 2 – Project 2 Group 2

PUBH 422

Priscilla Almestar, Alyssa Bautista, and Isabella Velazquez

November 23, 2025

Here we present an overview of the study using simple, transparent summaries of both continuous and categorical variables. Boxplots and histograms describe the distributions of age, height, and weight, which highlight typical values, spread, and any outliers. Bar charts display frequencies of gender, race/ethnicity, education, general health status, and quality of life. Our goal is to provide a quick, reliable picture of who is in the sample and how key characteristics are distributed.

On day 2, the dataset was cleaned by removing missing values and using only the necessary variables for the analysis. Education was recoded into different categories for better understanding, and the analysis used RStudio, tidyverse, ggplot2, and psych. For day 3, univariate statistics were used to examine differences in variables such as age, height, and weight, and to calculate the mean, median, standard deviation, and range within the dataset. ggplot2 was used to create histograms and boxplots for the quantitative and qualitative variables. On day 4, the associations among the variables were extended. ggplot2 was used to create a scatterplot comparing height and weight by sex and education level. A correlation matrix was also used to compare age, height, and weight, and to assess the association in the scatterplot.

This report provides a clear, big-picture view of our cohort by summarizing how age, weight, height, gender, race/ethnicity, education, general health status, and quality of life are distributed. Heights are tightly centered at mid-60 inches, with a median of 66. Weights are more variable and right-skewed. Most participants fall between 150 and 200 pounds. Age spans early adulthood through older age, with participants densest between 50 and 70 years. For the coded categorical measures, bar charts reveal a distribution that is highly concentrated at two points for education: the high-school graduate code (4) and the college-or-better range (8-10). Overall, most participants have at least some college. Additionally, most participants rate their general health status as good to very good, with few reporting fair or poor health. This pattern suggests a generally healthy sample. Race/ethnicity is highly concentrated in the group NH white. Gender is well represented as the distribution is nearly balanced between both male and female. The female bar is slightly taller. Lastly, the sample reports a high quality of life overall, with only a small fraction indicating dissatisfaction. Together, these distributions give a clear baseline for the cohort and will help guide future analyses.

The bivariate analysis shows important variables within the dataset. Boxplots are used to compare and make associations with sex and education levels, as well as to see the comparisons in different sexes where the data varies among education levels. Using clustered bar charts in the dataset helps to show health status to examine life satisfaction. The scatterplots also showed a positive association between height and weight, where more important variable relations exist than quantitative variables.

A scatterplot was run for day 4 task 1 to show the relationships between height (HEIGHTTC_A) and weight (WEIGHTLBTC_A), by sex, and faceted by EDUCP_A_recoded. The plot shows that there is a relationship between height and weight, as height increases, so does weight. Clusters showed sex differences, with males heavier and taller than females. In the education group data, heights and weights differ across groups. There is an association between height and weight, and between education and both, regardless of sex. A correlation matrix was generated to examine the associations among age, AGEP_A, height, HEIGHTTC_A, and weight, WEIGHTLBTC_A. A link between height and weight is shown in the scatterplot. With a factor like age, the comparison is weaker, since age isn't linearly related to height and weight. The matrix shows that height and weight have the strongest relationship in this dataset.

The analyses show how patterns show within adult health data. Height and weight have a strong association, whereas age does not. Participants report good health and life satisfaction, as indicated by variables such as education levels in the dataset. The bivariate and multivariate graphs show the relationships and support this finding. This project shows how datasets can be used to understand data across groups to make sense of them using different techniques.

This project demonstrates how NHIS data can be analyzed using descriptive statistics, visualization, and other methods to examine health data. Strong associations can be observed between height and weight, whereas those with different variables may be weaker. The project demonstrates data cleaning and visualization, as well as the use of interpretation skills for future public health work.