NHIS 2021: Education, Health, and Well-Being

Group 3

2025-11-22

Introduction

The National Health Interview Survey (NHIS) is a large, nationally representative survey of the U.S. population. In this project, we examine how demographic characteristics, body measurements, general health, and life satisfaction relate to one another. Our analysis uses the 2021 NHIS Sample Adult File and focuses on age (AGEP_A), sex (SEX_A), education (EDUCP_A), height (HEIGHTTC_A), weight (WEIGHTLBTC_A), general health (PHSTAT_A), and life satisfaction (LSATIS4R_A).

Methods

We used the cleaned dataset created earlier in our project and saved as `data/nhis_clean.csv` . After importing the dataset, we converted several variables to factors with meaningful labels and produced descriptive statistics, univariate plots, bivariate comparisons, and multivariate visualizations. Height–weight trends and correlations among age, height, and weight were examined using both ggplot2 and psych's `pairs.panels()` function.

```
nhis <- readr::read_csv("data/nhis_clean.csv")

nhis <- nhis %>%
  mutate(
    SEX_A = factor(SEX_A, levels = c(1, 2),
                   labels = c("Male", "Female")),
    EDUCP_A = factor(EDUCP_A, levels = c(1, 2, 3, 4),
                     labels = c("Less than HS",
                                "HS graduate",
                                "Some college",
                                "College graduate+")),
    PHSTAT_A = factor(PHSTAT_A, levels = c(1, 2, 3, 4, 5),
                      labels = c("Excellent", "Very good", "Good", "Fair", "Poor")),
    LSATIS4R_A = factor(LSATIS4R_A, levels = c(1, 2, 3, 4),
                        labels = c("Very satisfied", "Satisfied", "Dissatisfied",
                                   "Very dissatisfied"))
  )

glimpse(nhis)
```

```
## Rows: 26,037
## Columns: 9
## $ ...1         <dbl> 1, 2, 3, 4, 8, 10, 11, 12, 15, 16, 19, 20, 21, 22, 23, 24…
## $ AGEP_A       <dbl> 50, 53, 56, 57, 41, 71, 69, 44, 69, 59, 41, 82, 74, 67, 7…
## $ WEIGHTLBTC_A <dbl> 199, 205, 160, 190, 206, 127, 100, 208, 165, 225, 150, 28…
## $ HEIGHTTC_A   <dbl> 69, 75, 67, 63, 72, 63, 63, 69, 71, 70, 66, 72, 66, 66, 6…
## $ SEX_A        <fct> Male, Male, Male, Female, Male, Female, Female, Male, Mal…
## $ HISPALLP_A   <dbl> 2, 3, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, …
## $ EDUCP_A      <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ PHSTAT_A     <fct> Very good, Very good, Very good, Fair, Excellent, Excelle…
## $ LSATIS4R_A   <fct> Satisfied, Very satisfied, Dissatisfied, Satisfied, Very …
```

Results

Descriptive statistics

```
nhis %>%
  select(AGEP_A, HEIGHTTC_A, WEIGHTLBTC_A) %>%
  summary()
```

```
##      AGEP_A        HEIGHTTC_A    WEIGHTLBTC_A
##  Min.   :18.00   Min.   :59.0   Min.   :100.0
##  1st Qu.:37.00   1st Qu.:64.0   1st Qu.:147.0
##  Median :54.00   Median :66.0   Median :173.0
##  Mean   :52.57   Mean   :66.7   Mean   :176.8
##  3rd Qu.:67.00   3rd Qu.:70.0   3rd Qu.:200.0
##  Max.   :85.00   Max.   :76.0   Max.   :299.0
```

Age distribution

```
ggplot(nhis, aes(x = AGEP_A)) +
  geom_histogram(bins = 30) +
  labs(x = "Age (years)", y = "Count")
```

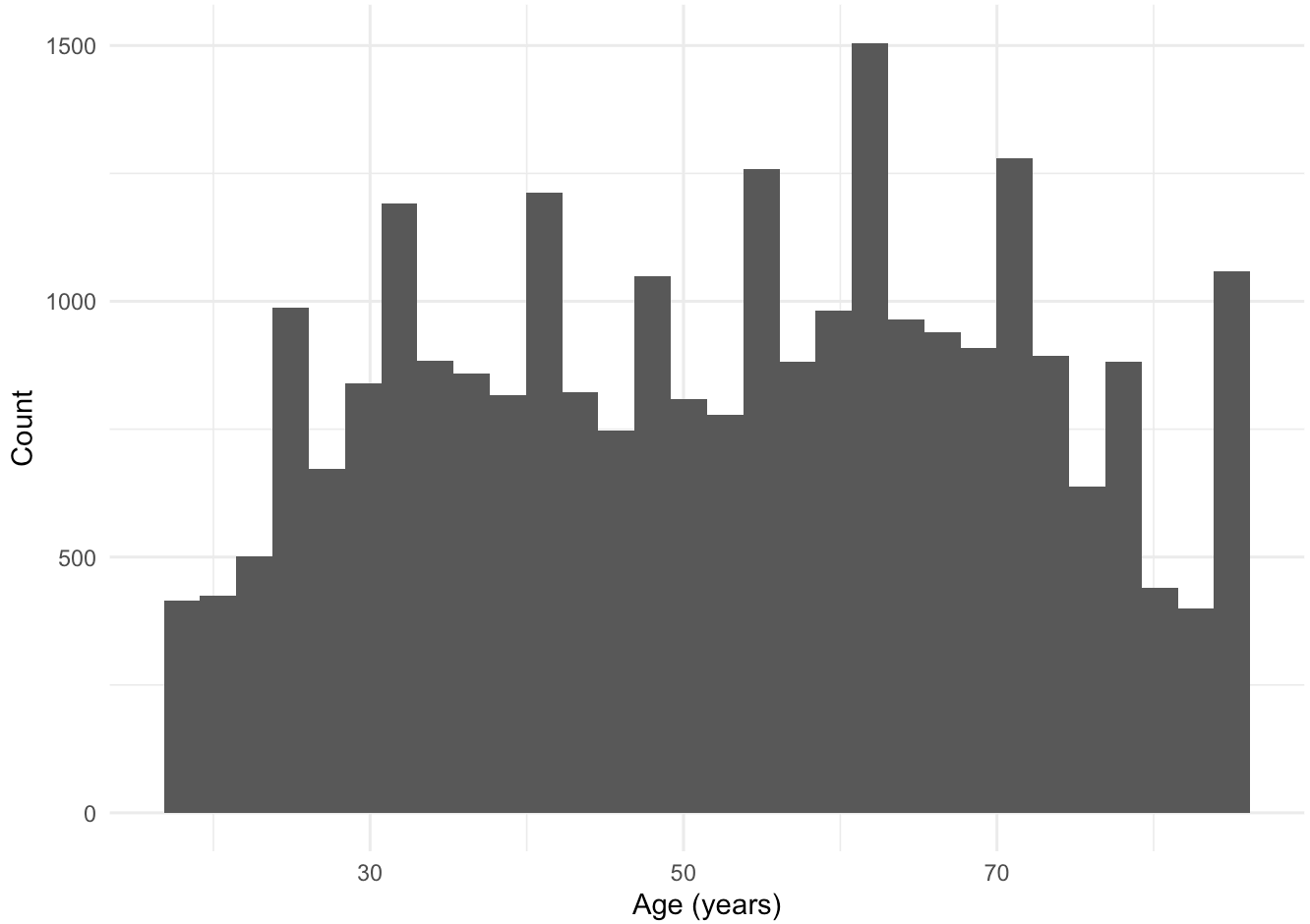

Figure 1. Distribution of age in the sample.

Interpretation:
The sample contains adults across a wide age range, with a concentration in middle-aged groups.

Education and health

```
nhis %>%
  count(EDUCP_A) %>%
  ggplot(aes(x = EDUCP_A, y = n)) +
  geom_col() +
  labs(x = "Education level", y = "Count")
```

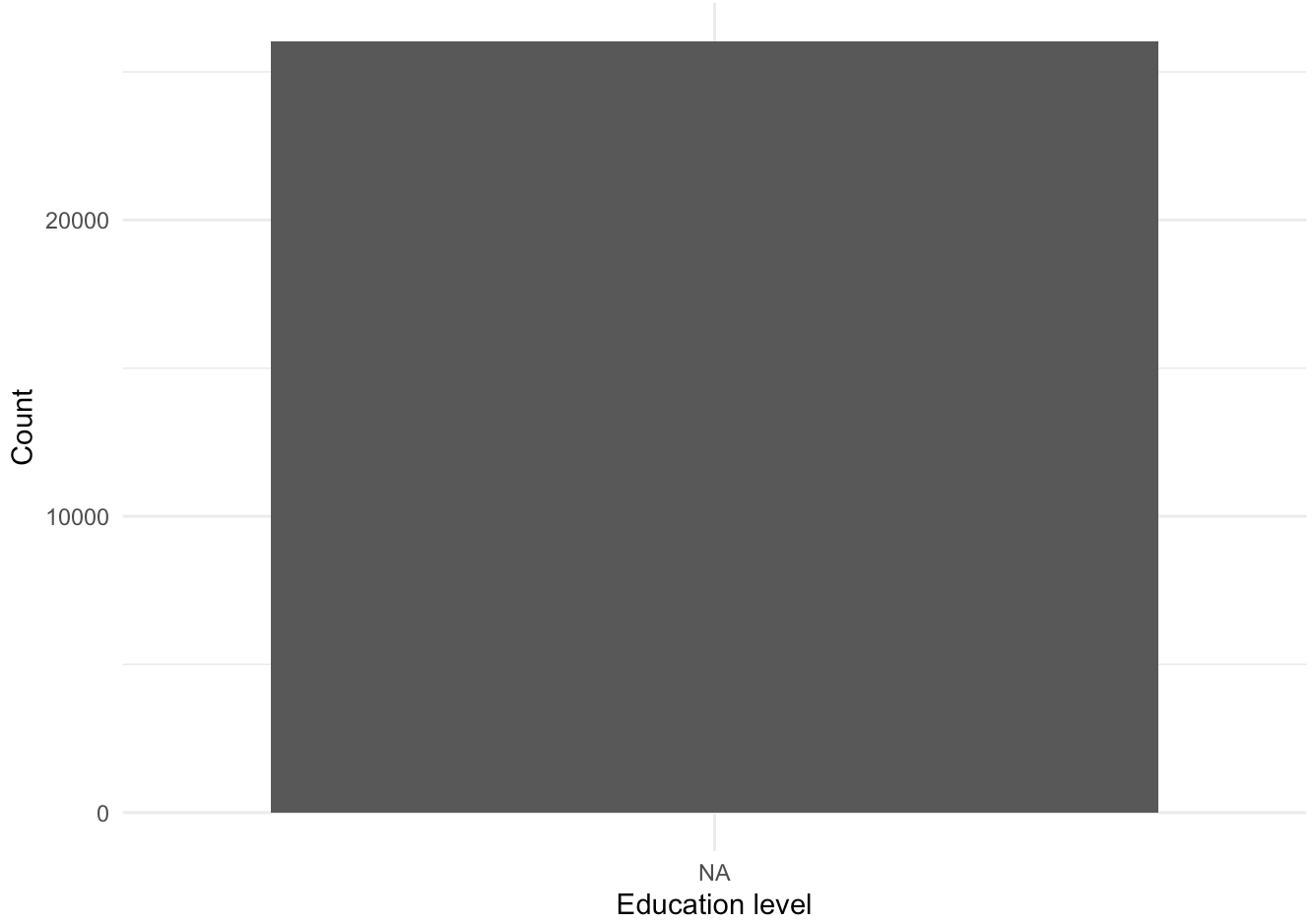

Figure 2. Education categories in the dataset.

Interpretation:
Most respondents have at least a high school diploma, with many reporting some college or a college degree.

General health × Life satisfaction

```
nhis %>%
  count(PHSTAT_A, LSATIS4R_A) %>%
  ggplot(aes(x = PHSTAT_A, y = n, fill = LSATIS4R_A)) +
  geom_col(position = "dodge") +
  labs(x = "General health", y = "Count", fill = "Life satisfaction")
```

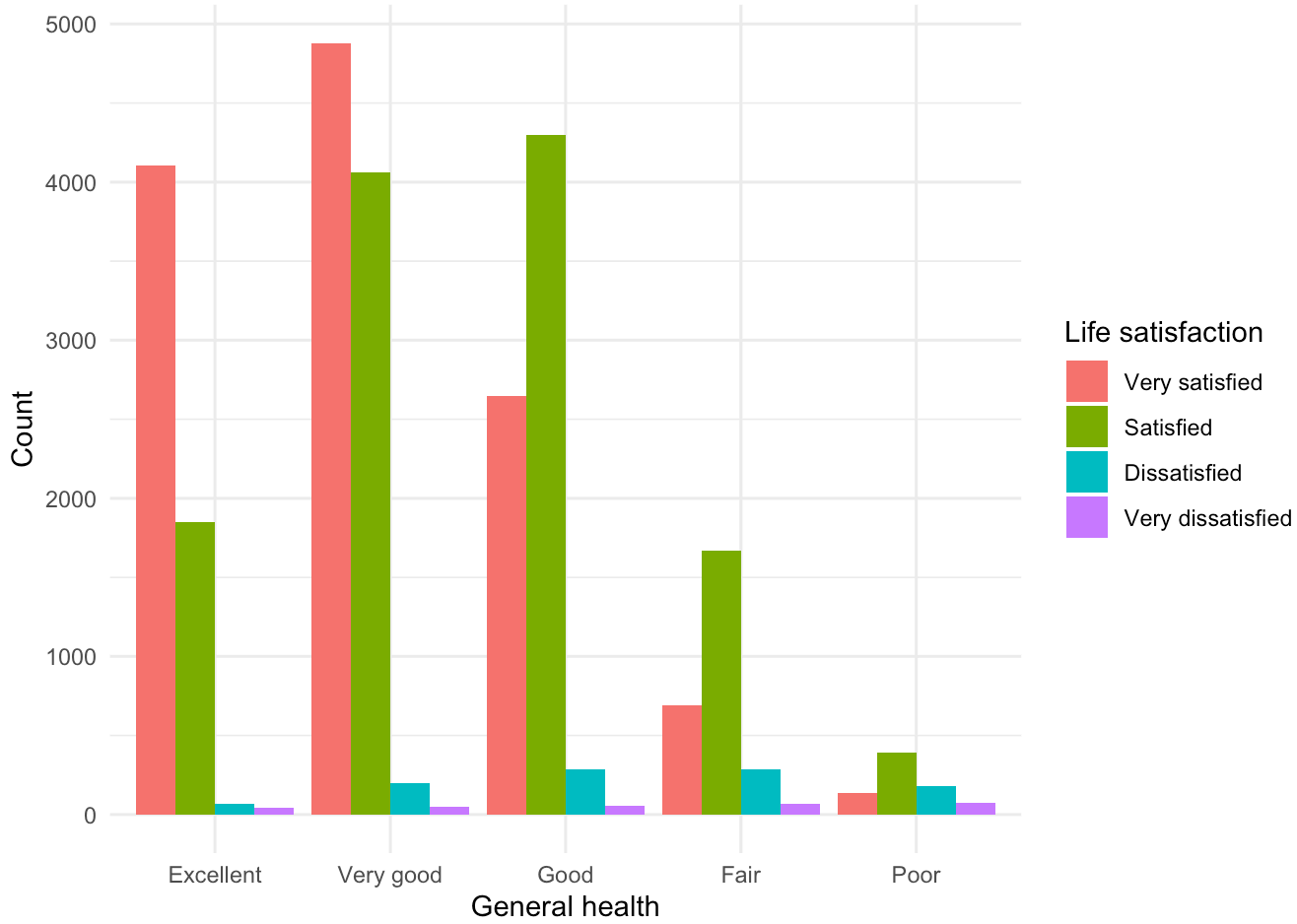

Figure 3. Health status by life satisfaction.

Interpretation:
Respondents with better general health tend to report higher life satisfaction.

Height & Weight Relationships

```
ggplot(nhis, aes(x = HEIGHTTC_A, y = WEIGHTLBTC_A)) +
  geom_point(alpha = .5) +
  labs(x = "Height (inches)", y = "Weight (lbs)")
```

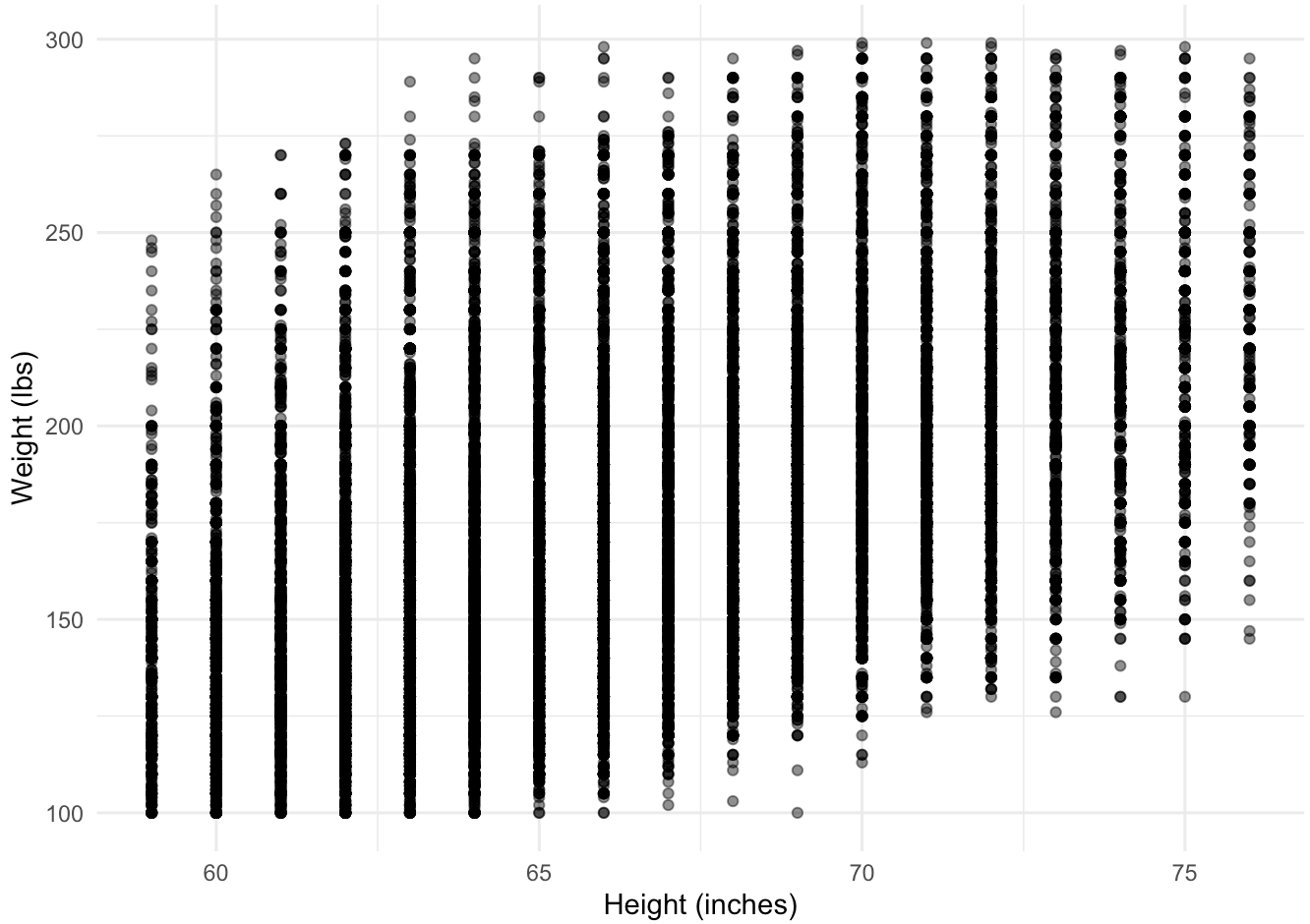

Figure 4. Scatterplot of height vs. weight.

```
cor(nhis$HEIGHTTC_A, nhis$WEIGHTLBTC_A, use = "complete.obs")
```

```
## [1] 0.5023037
```

Interpretation:
Height and weight show a strong positive correlation, consistent with expected body-size patterns.

Multivariate Analysis

Height, weight, sex, and education

```
ggplot(nhis, aes(x = HEIGHTTC_A, y = WEIGHTLBTC_A, color = SEX_A)) +
  geom_point(alpha = .5) +
  facet_wrap(~ EDUCP_A) +
  labs(x = "Height (inches)", y = "Weight (lbs)", color = "Sex")
```

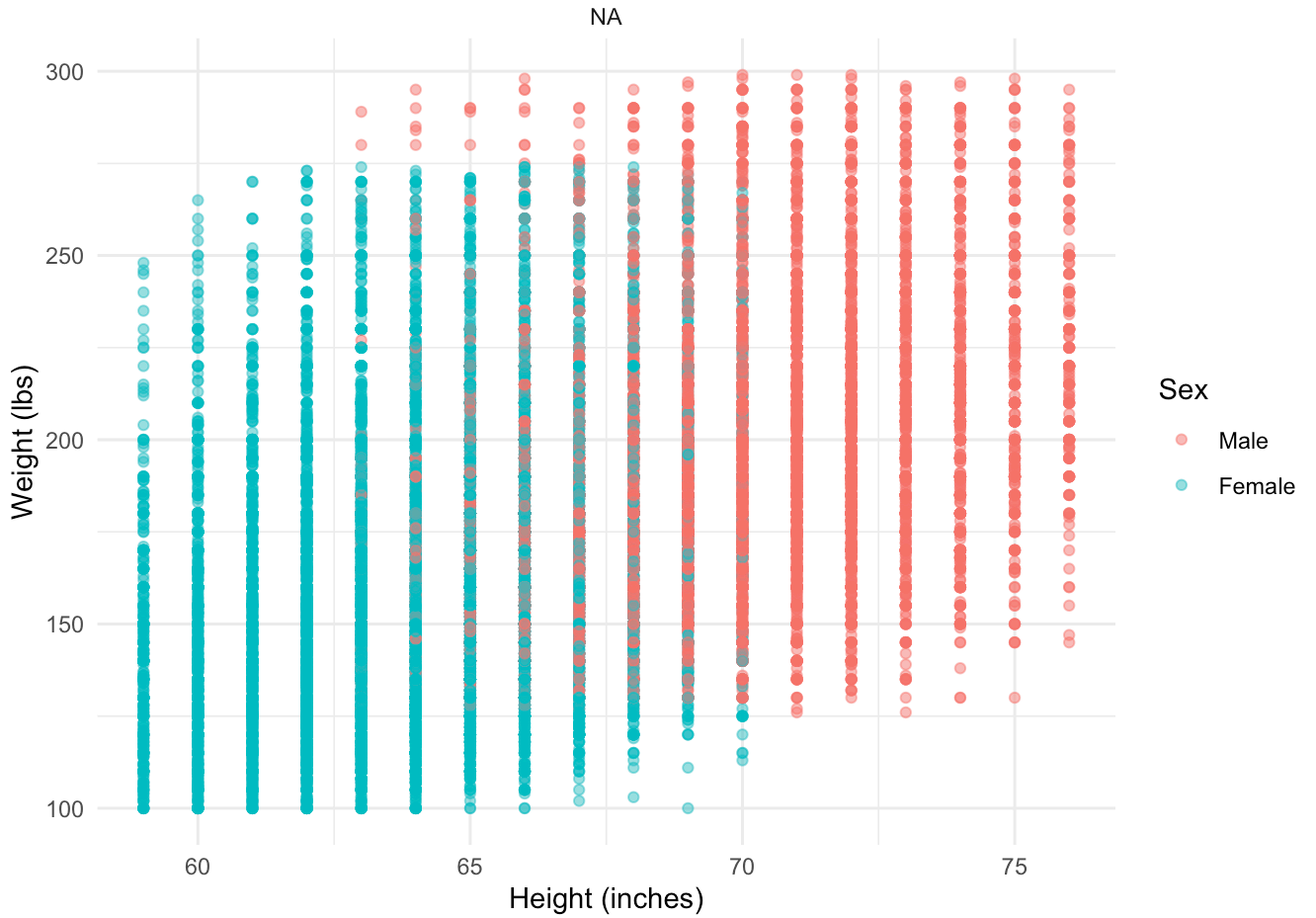

Figure 5. Weight vs. height by sex and education.

Interpretation:
The height–weight trend is consistent across sexes and education groups, though men tend to weigh more at comparable heights.

Correlation matrix

```
nhis %>%
  select(AGEP_A, HEIGHTTC_A, WEIGHTLBTC_A) %>%
  pairs.panels(method = "pearson",
               hist.col = "gray",
               density = TRUE,
               ellipses = FALSE)
```

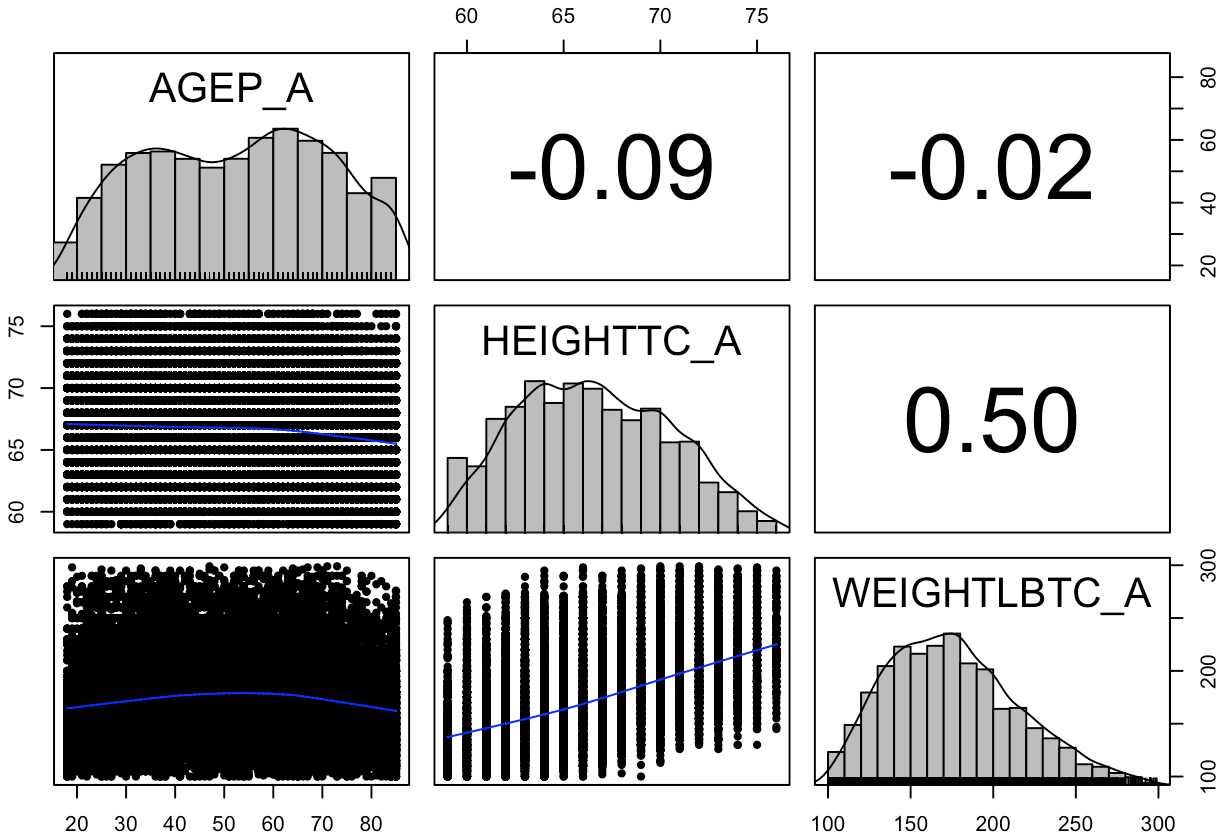

Figure 6. Correlation matrix.

Interpretation:
Height and weight have the strongest correlation; age has weaker associations but contributes to variation in weight.

Discussion

Our analysis of the 2021 NHIS dataset reveals several clear patterns. Self-rated health is strongly associated with life satisfaction, and the expected positive relationship between height and weight appears in all subgroups. Education shows meaningful differences in age distribution and may relate to health patterns indirectly. Because NHIS is cross-sectional, causal direction cannot be determined.

Conclusion

Overall, the results suggest that demographic factors, education, health status, and well-being are interconnected. Better general health aligns with higher life satisfaction, height and weight show predictable correlations, and education does not drastically change body-size relationships. These findings highlight the value of descriptive and multivariate approaches when analyzing public health survey data.