

# Final\_Report

2025-11-23

## 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 `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.

## Day 1

```
library(readr)#loading package
library(tidyverse)#loading package
```

```
## — Attaching core tidyverse packages — tidyverse 2.0.0 —
## ✓ dplyr      1.1.4      ✓ purrr      1.0.4
## ✓ forcats    1.0.0      ✓ stringr    1.5.1
## ✓ ggplot2    4.0.0      ✓ tibble     3.2.1
## ✓ lubridate  1.9.4      ✓ tidyr      1.3.1
## — Conflicts — tidyverse_conflicts() —
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
setwd("/Users/lesleythompson/Desktop/Pubh 422/NHIS2021_Group3/Data")#set working directory within
files on my computer
NHIS_2021 <- read.csv("NHIS_Data_2021.csv", header = TRUE) #load hints dataset
View(NHIS_2021)#look at whole data set
str(NHIS_2021)#gives data structure
```

```
## 'data.frame':    29482 obs. of  18 variables:
## $ DEMENEV_A : int  2 2 2 2 2 2 2 2 2 2 ...
## $ COPDEV_A  : int  2 2 2 2 2 2 2 2 2 2 ...
## $ HYPEV_A   : int  1 1 2 1 2 1 2 2 2 2 ...
## $ DEPEV_A   : int  2 2 2 2 2 2 2 2 2 1 ...
## $ CANEV_A   : int  2 2 2 2 2 2 2 2 2 2 ...
## $ DIBEV_A   : int  2 1 2 2 2 2 2 2 2 2 ...
## $ AGEV_A    : int  50 53 56 57 25 55 45 41 26 71 ...
## $ SEX_A     : int  1 1 1 2 1 1 1 1 2 2 ...
## $ HISPALLP_A : int  2 3 2 2 3 3 2 3 3 2 ...
## $ MARSTAT_A : int  5 6 5 7 9 9 1 1 7 1 ...
## $ EDUCP_A   : int  1 7 8 5 4 5 9 5 4 9 ...
## $ PHSTAT_A  : int  2 2 2 4 3 3 1 1 2 1 ...
## $ LSATIS4R_A : int  2 1 3 2 8 8 1 1 1 1 ...
## $ SMKIGST_A : int  3 4 3 3 9 9 4 4 4 4 ...
## $ RATCAT_A  : int  7 12 14 11 6 6 14 14 7 14 ...
## $ BMICAT_A  : int  3 3 3 4 4 3 9 3 4 2 ...
## $ WEIGHTLBTC_A : int  199 205 160 190 250 200 997 206 996 127 ...
## $ HEIGHTTC_A : int  69 75 67 63 72 69 67 72 96 63 ...
```

```
head(NHIS_2021)#print first 6 rows of data
```

```
## DEMENEV_A COPDEV_A HYPEV_A DEPEV_A CANEV_A DIBEV_A AGEV_A SEX_A HISPALLP_A
## 1          2          2          1          2          2          2          50          1          2
## 2          2          2          1          2          2          1          53          1          3
## 3          2          2          2          2          2          2          56          1          2
## 4          2          2          1          2          2          2          57          2          2
## 5          2          2          2          2          2          2          25          1          3
## 6          2          2          1          2          2          2          55          1          3
## MARSTAT_A EDUCP_A PHSTAT_A LSATIS4R_A SMKIGST_A RATCAT_A BMICAT_A
## 1          5          1          2          2          3          7          3
## 2          6          7          2          1          4          12         3
## 3          5          8          2          3          3          14         3
## 4          7          5          4          2          3          11         4
## 5          9          4          3          8          9          6         4
## 6          9          5          3          8          9          6         3
## WEIGHTLBTC_A HEIGHTTC_A
## 1          199          69
## 2          205          75
## 3          160          67
## 4          190          63
## 5          250          72
## 6          200          69
```

```
summary(NHIS_2021)# gives summary of data
```

```
##      DEMENEV_A      COPDEV_A      HYPEV_A      DEPEV_A
## Min.    :1.000    Min.    :1.000    Min.    :1.000    Min.    :1.000
## 1st Qu.:2.000    1st Qu.:2.000    1st Qu.:1.000    1st Qu.:2.000
## Median :2.000    Median :2.000    Median :2.000    Median :2.000
## Mean   :1.993    Mean   :1.951    Mean   :1.648    Mean   :1.829
## 3rd Qu.:2.000    3rd Qu.:2.000    3rd Qu.:2.000    3rd Qu.:2.000
## Max.   :9.000    Max.   :9.000    Max.   :9.000    Max.   :9.000
##      CANEV_A      DIBEV_A      AGEV_A      SEX_A      HISPALLP_A
## Min.    :1.000    Min.    :1.0    Min.    :18.00    Min.    :1.000    Min.    :1.000
## 1st Qu.:2.000    1st Qu.:2.0    1st Qu.:37.00    1st Qu.:1.000    1st Qu.:2.000
## Median :2.000    Median :2.0    Median :53.00    Median :2.000    Median :2.000
## Mean   :1.882    Mean   :1.9    Mean   :52.63    Mean   :1.547    Mean   :2.203
## 3rd Qu.:2.000    3rd Qu.:2.0    3rd Qu.:68.00    3rd Qu.:2.000    3rd Qu.:2.000
## Max.   :9.000    Max.   :9.0    Max.   :99.00    Max.   :9.000    Max.   :7.000
##      MARSTAT_A      EDUCP_A      PHSTAT_A      LSATIS4R_A
## Min.    :1.000    Min.    : 1.000    Min.    :1.00    Min.    :1.000
## 1st Qu.:1.000    1st Qu.: 4.000    1st Qu.:2.00    1st Qu.:1.000
## Median :4.000    Median : 6.000    Median :2.00    Median :2.000
## Mean   :3.858    Mean   : 6.447    Mean   :2.39    Mean   :1.748
## 3rd Qu.:7.000    3rd Qu.: 8.000    3rd Qu.:3.00    3rd Qu.:2.000
## Max.   :9.000    Max.   :99.000    Max.   :9.00    Max.   :9.000
##      SMKCIgst_A      RATCAT_A      BMICAT_A      WEIGHTLBTC_A
## Min.    :1.000    Min.    : 1.000    Min.    :1.000    Min.    :100.0
## 1st Qu.:3.000    1st Qu.: 7.000    1st Qu.:2.000    1st Qu.:150.0
## Median :4.000    Median :11.000    Median :3.000    Median :180.0
## Mean   :3.579    Mean   : 9.848    Mean   :3.121    Mean   :248.8
## 3rd Qu.:4.000    3rd Qu.:14.000    3rd Qu.:4.000    3rd Qu.:215.0
## Max.   :9.000    Max.   :14.000    Max.   :9.000    Max.   :999.0
##      HEIGHTTC_A
## Min.    :59.00
## 1st Qu.:64.00
## Median :67.00
## Mean   :68.72
## 3rd Qu.:70.00
## Max.   :99.00
```

## Day 2

```
attach(NHIS_2021)
subNHIS <- NHIS_2021 %>%
  select(AGEV_A, WEIGHTLBTC_A, HEIGHTTC_A, SEX_A, HISPALLP_A, EDUCP_A, PHSTAT_A, LSATIS4R_A)
#allows us to select only the variables required for analysis, helps prevent the removal of more p
#articipants than necessary
View(subNHIS)#making sure only the selected variables show up
sum(is.na(subNHIS))
```

```
## [1] 0
```

```
#missing values in the code book are 97-99,7, 9, and 996-999 depending on the variable, excluded b
#elow by only including values outside of missing
NHIS_omit <- subNHIS[c(AGEV_A <97 & WEIGHTLBTC_A <996 & HEIGHTTC_A <96 & SEX_A <3 & HISPALLP_A <8
& EDUCP_A <11 & PHSTAT_A <6 & LSATIS4R_A <5),]
#Checking to make sure missing values have been ommitted
summary(NHIS_omit)
```

```
##      AGEP_A      WEIGHTLBTC_A      HEIGHTTC_A      SEX_A      HISPALLP_A
## Min.   :18.00   Min.   :100.0   Min.   :59.0   Min.   :1.00   Min.   :1.000
## 1st Qu.:37.00   1st Qu.:147.0   1st Qu.:64.0   1st Qu.:1.00   1st Qu.:2.000
## Median :54.00   Median :173.0   Median :66.0   Median :2.00   Median :2.000
## Mean   :52.57   Mean   :176.8   Mean   :66.7   Mean   :1.54   Mean   :2.197
## 3rd Qu.:67.00   3rd Qu.:200.0   3rd Qu.:70.0   3rd Qu.:2.00   3rd Qu.:2.000
## Max.   :85.00   Max.   :299.0   Max.   :76.0   Max.   :2.00   Max.   :7.000
##      EDUCP_A      PHSTAT_A      LSATIS4R_A
## Min.   : 1.000   Min.   :1.000   Min.   :1.000
## 1st Qu.: 4.000   1st Qu.:2.000   1st Qu.:1.000
## Median : 6.000   Median :2.000   Median :2.000
## Mean   : 6.032   Mean   :2.346   Mean   :1.583
## 3rd Qu.: 8.000   3rd Qu.:3.000   3rd Qu.:2.000
## Max.   :10.000   Max.   :5.000   Max.   :4.000
```

```
#seeing how much data was removed from the dataset
str(NHIS_omit)
```

```
## 'data.frame': 26037 obs. of 8 variables:
## $ AGEP_A : int 50 53 56 57 41 71 69 44 69 59 ...
## $ WEIGHTLBTC_A: int 199 205 160 190 206 127 100 208 165 225 ...
## $ HEIGHTTC_A : int 69 75 67 63 72 63 63 69 71 70 ...
## $ SEX_A : int 1 1 1 2 1 2 2 1 1 1 ...
## $ HISPALLP_A : int 2 3 2 2 3 2 2 2 2 2 ...
## $ EDUCP_A : int 1 7 8 5 5 9 9 8 4 8 ...
## $ PHSTAT_A : int 2 2 2 4 1 1 1 2 2 3 ...
## $ LSATIS4R_A : int 2 1 3 2 1 1 1 2 1 1 ...
```

```
#creating new levels from 1-4 based on the codebook by excluding all values not in that section
NHIS_omit$EDUCP_A[NHIS_omit$EDUCP_A >=0 & NHIS_omit$EDUCP_A <=3] = 1
NHIS_omit$EDUCP_A[NHIS_omit$EDUCP_A == 4] = 2
NHIS_omit$EDUCP_A[NHIS_omit$EDUCP_A >=5 & NHIS_omit$EDUCP_A <=7] = 3
NHIS_omit$EDUCP_A[NHIS_omit$EDUCP_A >=8 & NHIS_omit$EDUCP_A <=10] = 4
#making sure the new levels are working
summary(NHIS_omit$EDUCP_A)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000  2.000   3.000   2.971  4.000   4.000
```

```
#adding the labels to the new levels for education using the NHIS 2021 codebook
NHIS_2021_clean <- NHIS_omit %>%
  mutate(EDUCP_A = factor(EDUCP_A,
                          levels = c(1,2,3,4),
                          labels = c("less than High School", "High School Graduate", "Some College
Education", "College Graduate or better")))
#Making sure data labels are showing up properly
summary(NHIS_2021_clean$EDUCP_A)
```

```
##      less than High School      High School Graduate
##              2676              5765
##      Some College Education College Graduate or better
##              7238              10358
```

```
#making sure that education is now a factor variable with meaningful labels
str(NHIS_2021_clean$EDUCP_A)
```

```
## Factor w/ 4 levels "less than High School",...: 1 3 4 3 3 4 4 4 2 4 ...
```

```
View(NHIS_2021_clean)
#creating the cleaned .csv file for submission, with the help function
?write.csv
write.csv(NHIS_2021_clean, "nhis_clean.csv")
```

## Day 3

```
# Load required library for ggplot2
library(ggplot2)

# Check first few rows and column names
head(NHIS_2021_clean)
```

```
##      AGEP_A WEIGHTLBT A HEIGHTTC_A SEX_A HISPALLP_A      EDUCP_A
## 1      50      199      69      1      2      less than High School
## 2      53      205      75      1      3      Some College Education
## 3      56      160      67      1      2 College Graduate or better
## 4      57      190      63      2      2      Some College Education
## 8      41      206      72      1      3      Some College Education
## 10     71      127      63      2      2 College Graduate or better
##      PHSTAT_A LSATIS4R_A
## 1          2          2
## 2          2          1
## 3          2          3
## 4          4          2
## 8          1          1
## 10         1          1
```

```
names(NHIS_2021_clean)
```

```
## [1] "AGEP_A"      "WEIGHTLBT A" "HEIGHTTC_A"  "SEX_A"      "HISPALLP_A"
## [6] "EDUCP_A"      "PHSTAT_A"    "LSATIS4R_A"
```

```
#####DAY 3
#### Task 1
###Summary Statistics
```

```
#Quantitative Variables
```

```
# 1. Age (AGEP_A)
cat("=== AGE (AGEP_A) ===\n")
```

```
## === AGE (AGEP_A) ===
```

```
summary(NHIS_2021_clean$AGEP_A)      # Min, 1st Qu., Median, Mean, 3rd Qu., Max
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.00   37.00   54.00   52.57   67.00   85.00
```

```
cat("Mean:", mean(NHIS_2021_clean$AGEP_A, na.rm=TRUE), "\n")
```

```
## Mean: 52.57107
```

```
cat("Median:", median(NHIS_2021_clean$AGEP_A, na.rm=TRUE), "\n")
```

```
## Median: 54
```

```
cat("Standard Deviation:", sd(NHIS_2021_clean$AGEP_A, na.rm=TRUE), "\n\n")
```

```
## Standard Deviation: 18.33484
```

```
# 2. Weight (WEIGHTLBTC_A)
cat("=== WEIGHT (WEIGHTLBTC_A) ===\n")
```

```
## === WEIGHT (WEIGHTLBTC_A) ===
```

```
summary(NHIS_2021_clean$WEIGHTLBTC_A)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    100.0   147.0   173.0   176.8   200.0   299.0
```

```
cat("Mean:", mean(NHIS_2021_clean$WEIGHTLBTC_A, na.rm=TRUE), "\n")
```

```
## Mean: 176.8261
```

```
cat("Median:", median(NHIS_2021_clean$WEIGHTLBTC_A, na.rm=TRUE), "\n")
```

```
## Median: 173
```

```
cat("Standard Deviation:", sd(NHIS_2021_clean$WEIGHTLBTC_A, na.rm=TRUE), "\n\n")
```

```
## Standard Deviation: 39.59538
```

```
# 3. Height (HEIGHTTC_A)
cat("=== HEIGHT (HEIGHTTC_A) ===\n")
```

```
## === HEIGHT (HEIGHTTC_A) ===
```

```
summary(NHIS_2021_clean$HEIGHTTC_A)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	59.0	64.0	66.0	66.7	70.0	76.0

```
cat("Mean:", mean(NHIS_2021_clean$HEIGHTTC_A, na.rm=TRUE), "\n")
```

```
## Mean: 66.70108
```

```
cat("Median:", median(NHIS_2021_clean$HEIGHTTC_A, na.rm=TRUE), "\n")
```

```
## Median: 66
```

```
cat("Standard Deviation:", sd(NHIS_2021_clean$HEIGHTTC_A, na.rm=TRUE), "\n")
```

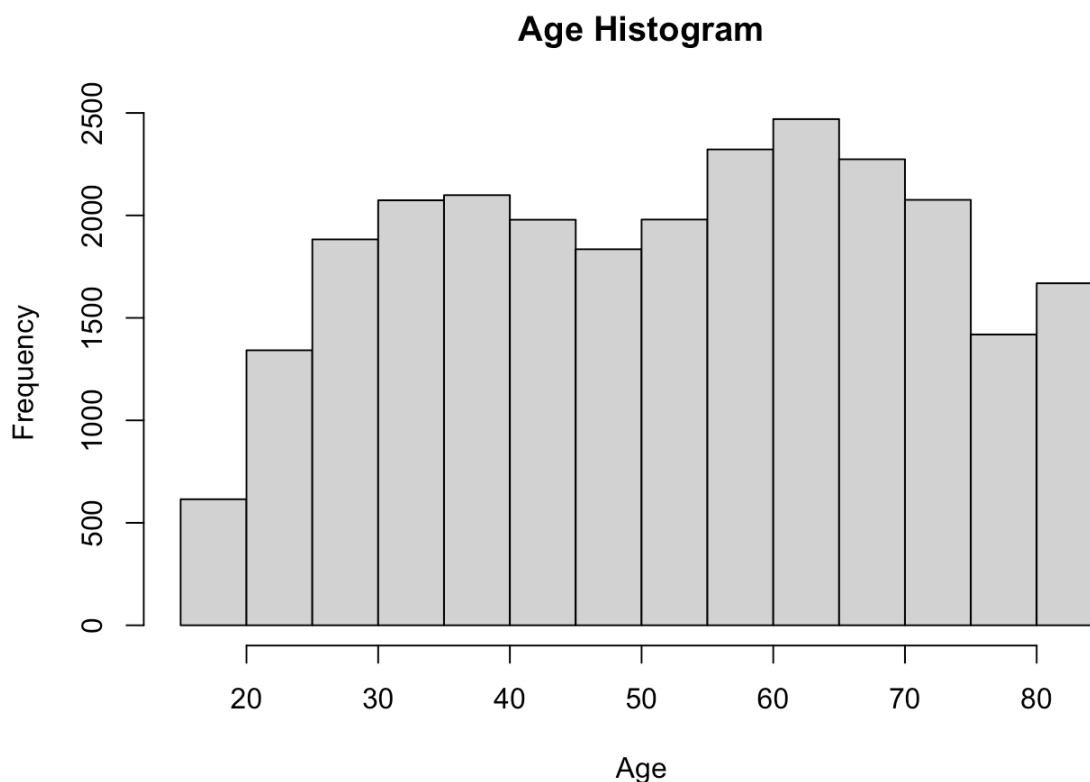
```
## Standard Deviation: 3.898793
```

### Interpretation:

The NHIS\_2021\_clean data set had a mean age of 52.57 (SD = 18.33), the minimum age was 18, and the maximum age was 85. The mean weight was 176.8 (SD = 39.59), with a minimum of 100, and a maximum of 299. The mean height was 66.7 inches (3.89) with a minimum of 59, and a maximum of 79. Looking at these descriptive statistics give a preliminary understanding of who this data is representing, and is important when drawing conclusions for future statistical analyses.

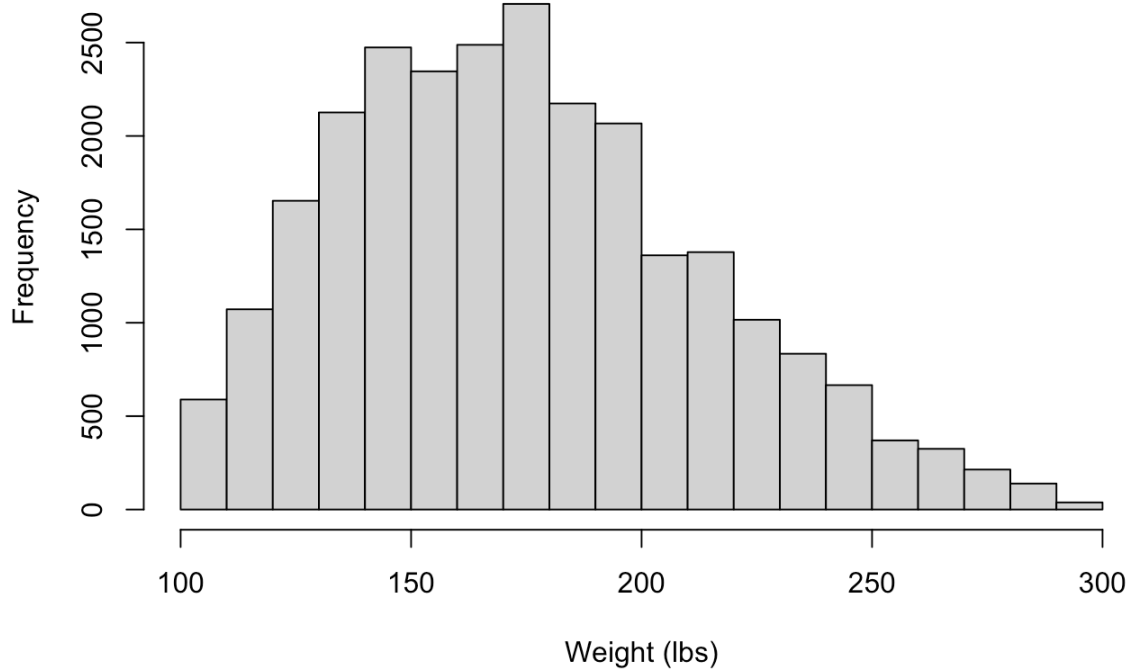
```
### Base R Histograms
```

```
hist(NHIS_2021_clean$AGEP_A, main="Age Histogram", xlab="Age")
```



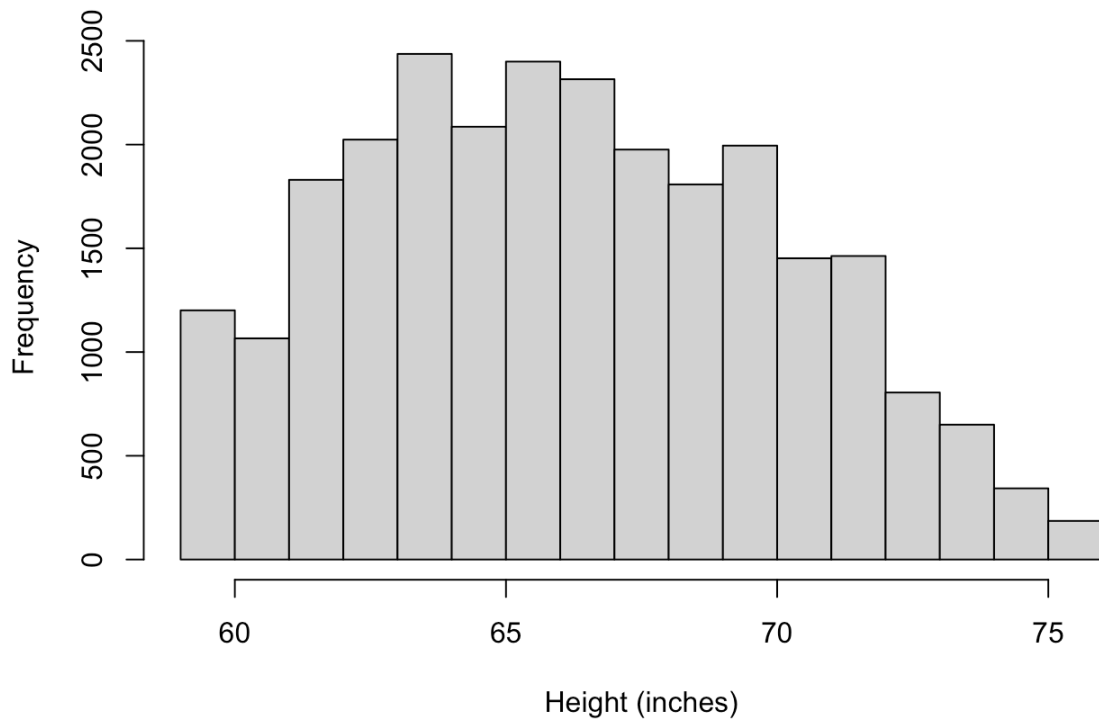
```
hist(NHIS_2021_clean$WEIGHTLBT_A, main="Weight Histogram", xlab="Weight (lbs)")
```

## Weight Histogram



```
hist(NHIS_2021_clean$HEIGHTTC_A, main="Height Histogram", xlab="Height (inches)")
```

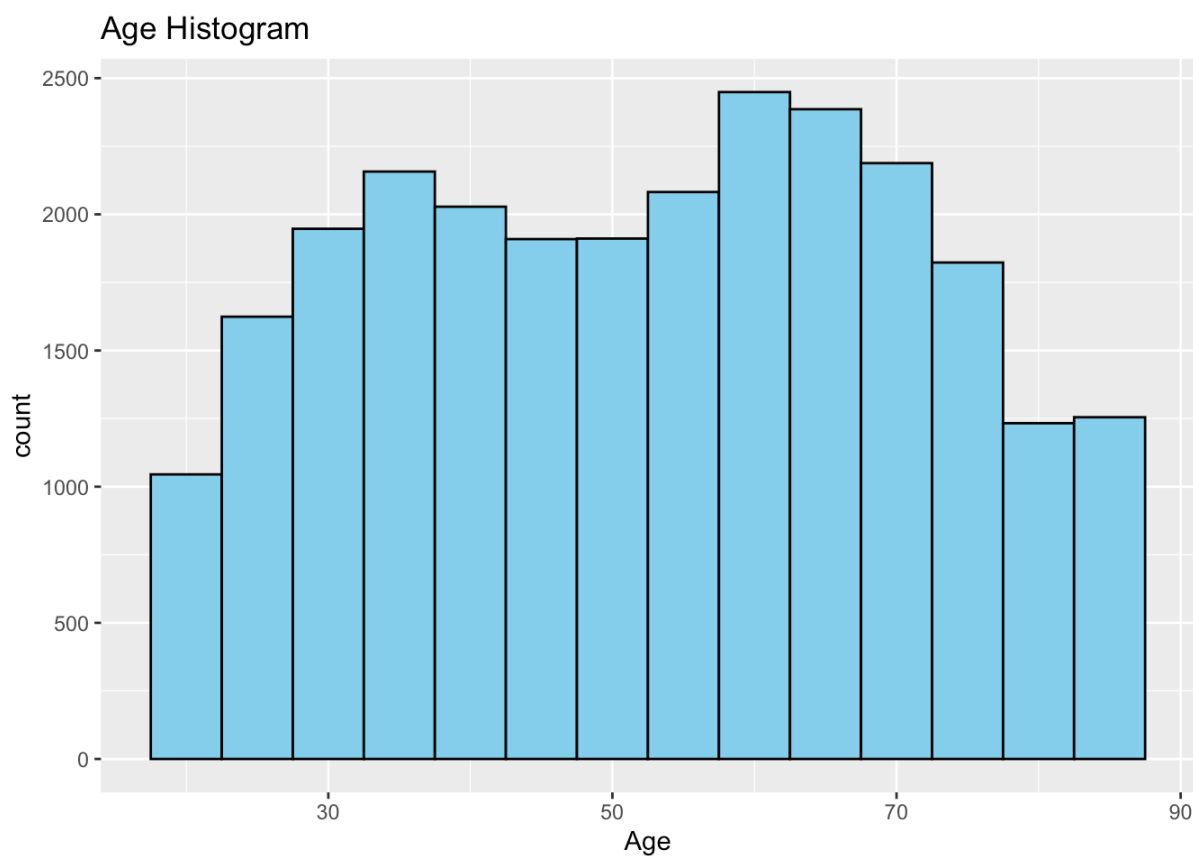
## Height Histogram





```
### ggplot2 Histograms
```

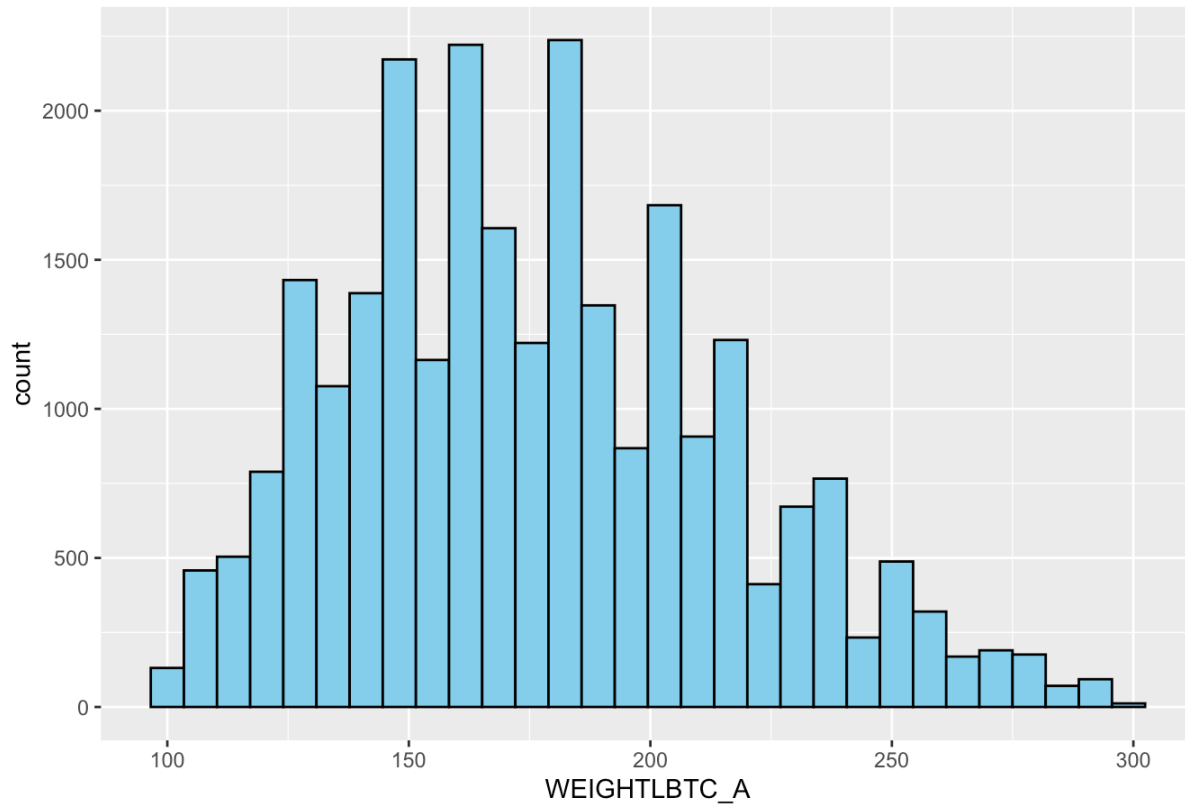
```
ggplot(NHIS_2021_clean, aes(x=AGEP_A)) + geom_histogram(binwidth=5, fill="skyblue", color="black")  
+ labs(title="Age Histogram", x="Age")
```



```
ggplot(NHIS_2021_clean, aes(x=WEIGHTLBTC_A)) + geom_histogram(fill="skyblue", color="black") + lab  
s(title="Weight Histogram")
```

```
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```

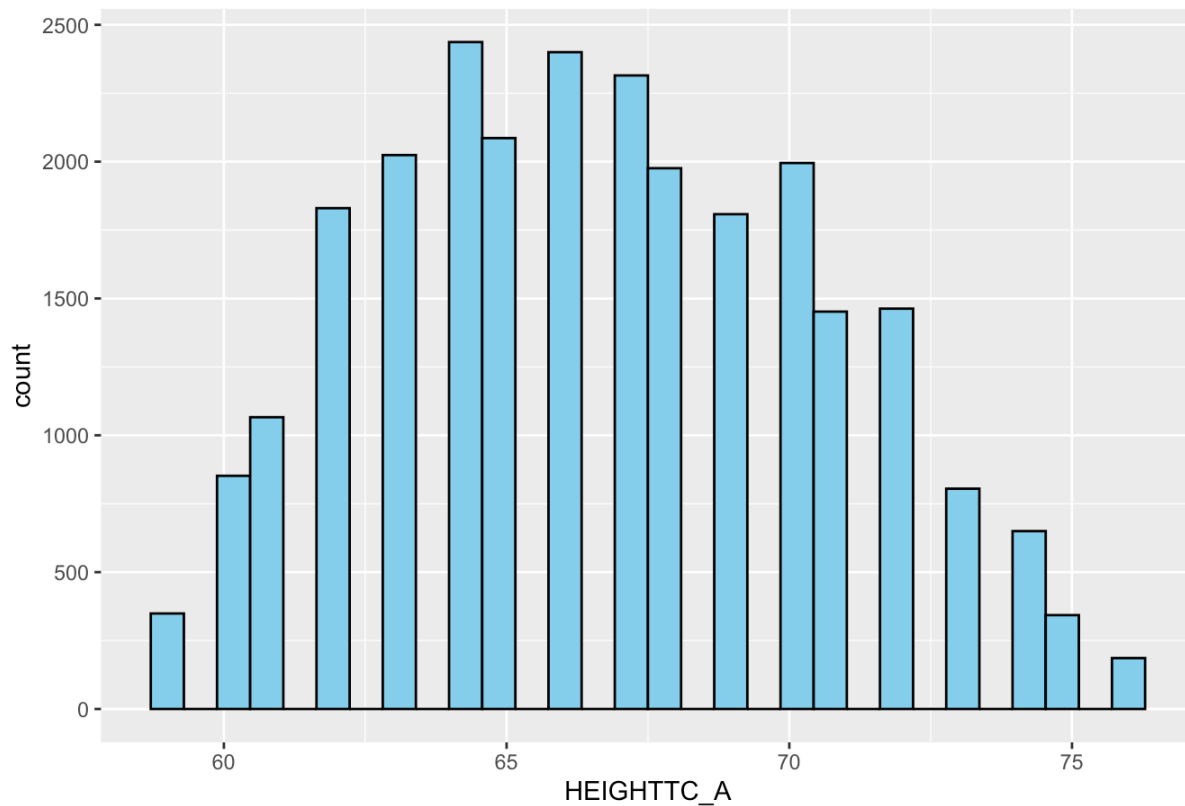
Weight Histogram



```
ggplot(NHIS_2021_clean, aes(x=HEIGHTTC_A)) + geom_histogram(fill="skyblue", color="black") + labs
(title="Height Histogram")
```

```
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```

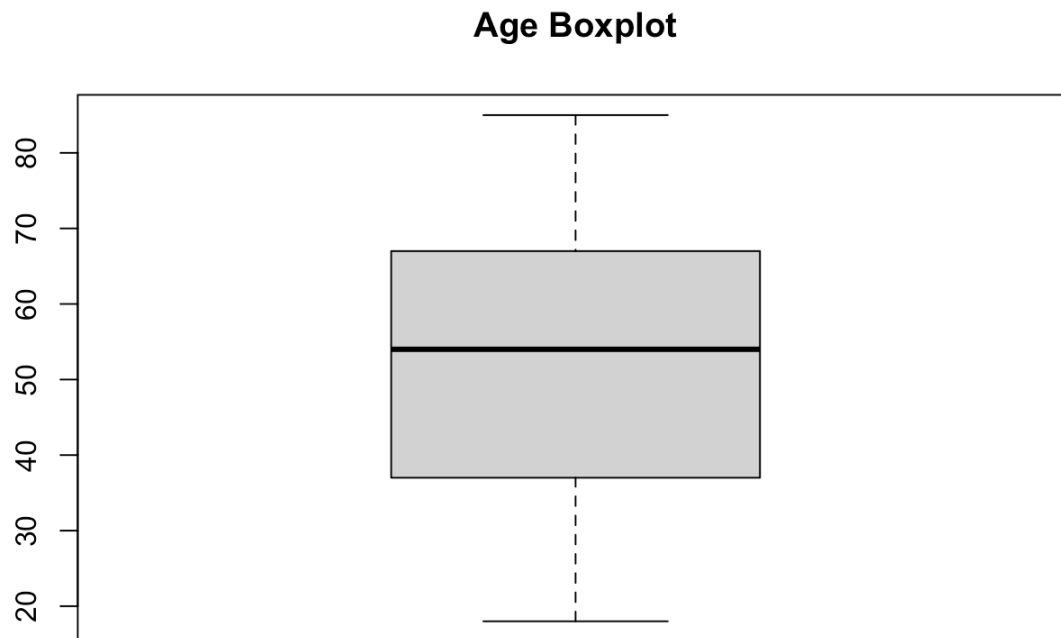
Height Histogram



**Interpretation:**

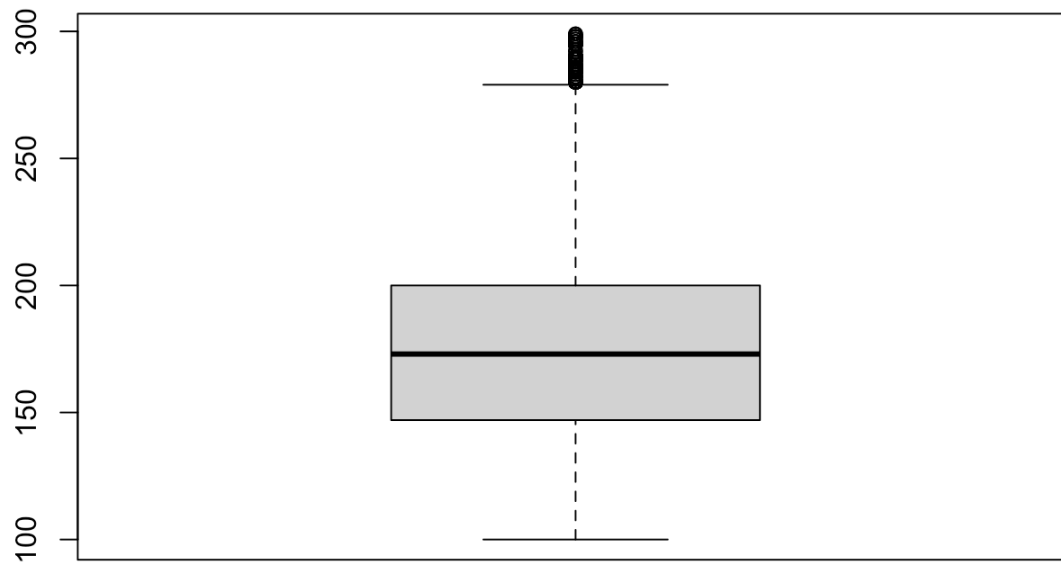
The sample contains adults across a wide age range that seems normally distributed, with a concentration in middle-aged groups.

```
#Boxplots Base R  
boxplot(NHIS_2021_clean$AGEP_A, main="Age Boxplot")
```



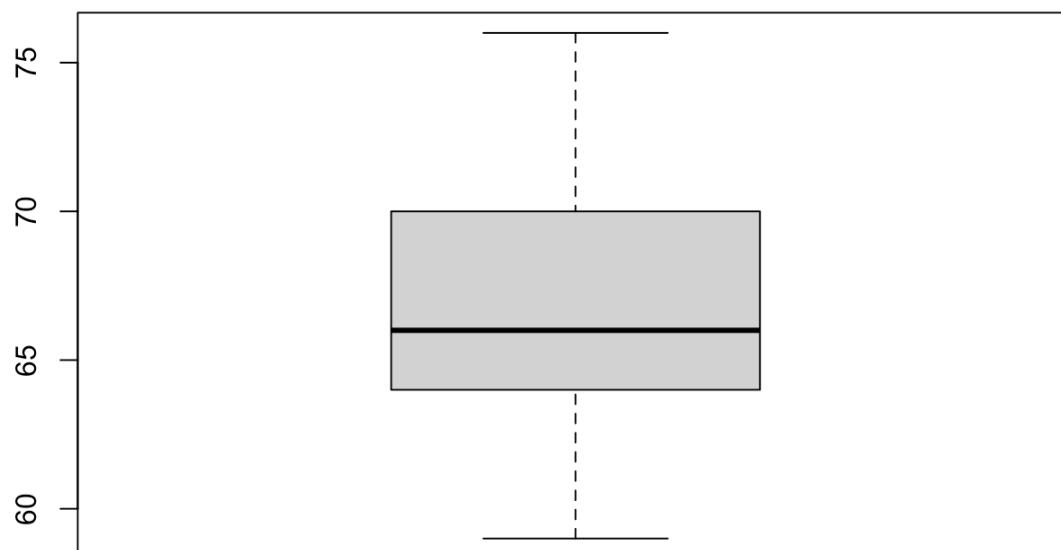
```
boxplot(NHIS_2021_clean$WEIGHTLBTC_A, main="Weight Boxplot")
```

## Weight Boxplot



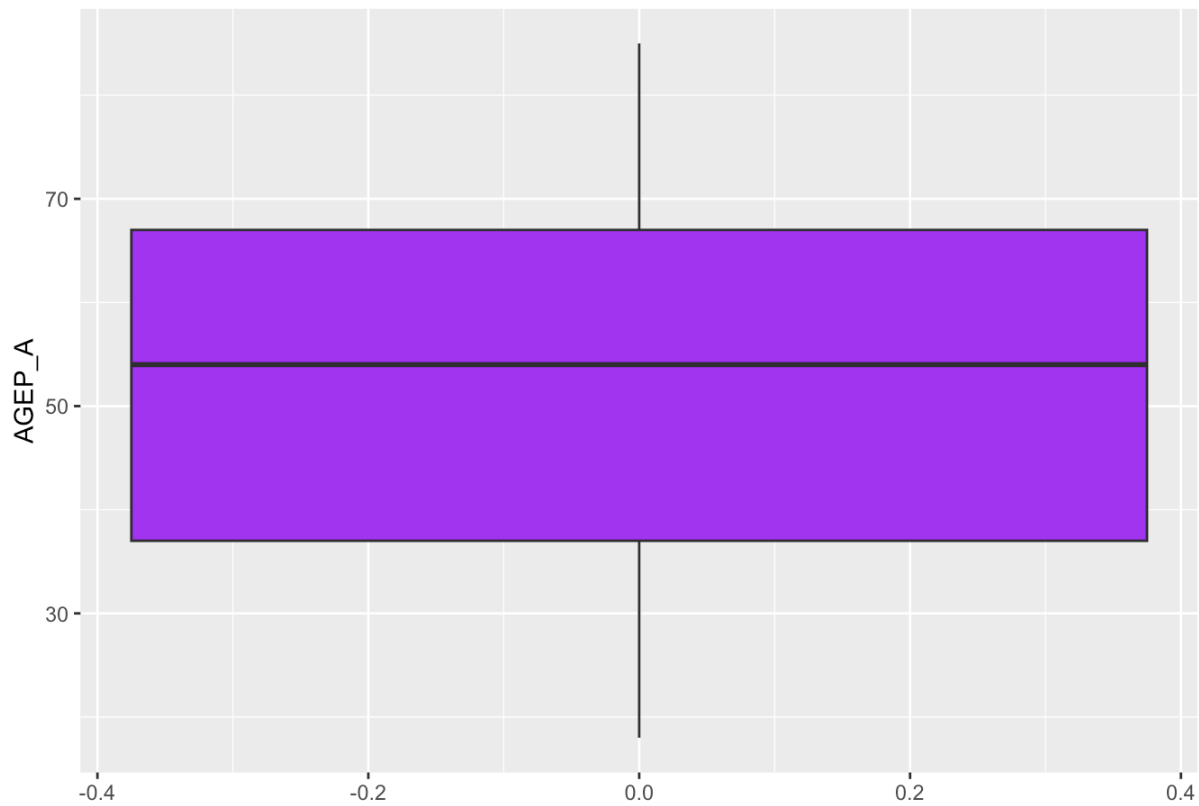
```
boxplot(NHIS_2021_clean$HEIGHTTC_A, main="Height Boxplot")
```

## Height Boxplot



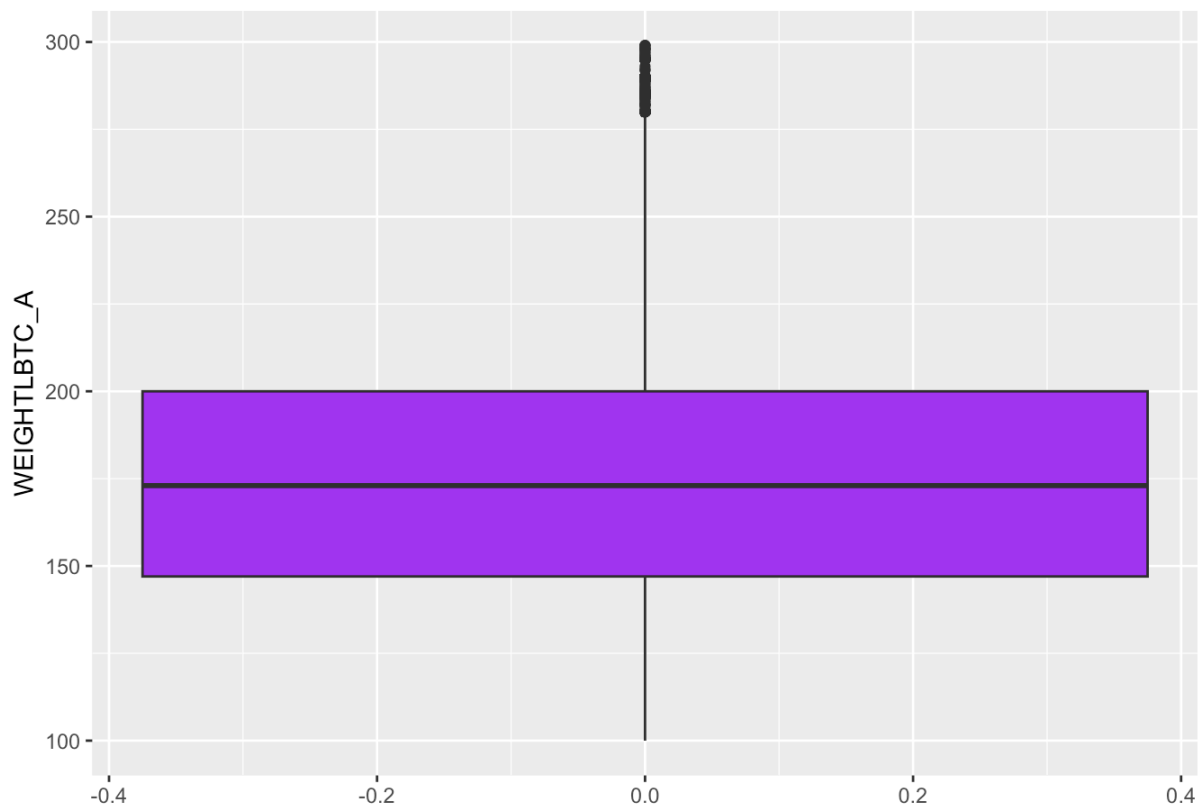
```
#Boxplot ggplot2  
ggplot(NHIS_2021_clean, aes(y=AGEP_A)) + geom_boxplot(fill="purple") + labs(title="Age Boxplot")
```

Age Boxplot

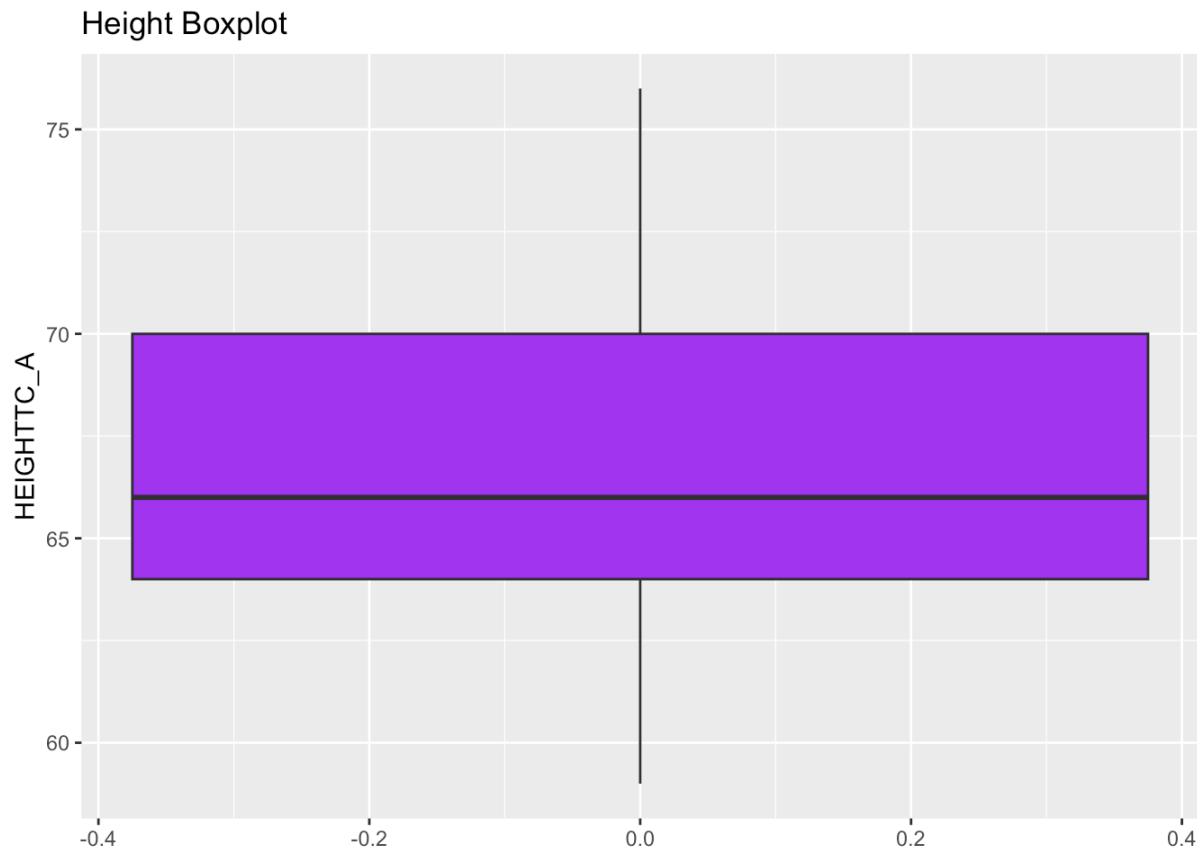


```
ggplot(NHIS_2021_clean, aes(y=WEIGHTLBT_C_A)) + geom_boxplot(fill="purple") + labs(title="Weight Bo  
xplot")
```

Weight Boxplot



```
ggplot(NHIS_2021_clean, aes(y=HEIGHTTC_A)) + geom_boxplot(fill="purple") + labs(title="Height Boxplot")
```



#### Interpretation:

When looking at the height variable, M=most participants fall between 65 and 70 inches.

###Qualitative Variables

##Frequency Table

```
table(NHIS_2021_clean$SEX_A)
```

```
##
##      1      2
## 11967 14070
```

```
table(NHIS_2021_clean$HISPALLP_A)
```

```
##
##      1      2      3      4      5      6      7
## 3533 17617 2645 1564 153 196 329
```

```
table(NHIS_2021_clean$EDUCP_A)
```

```
##
##      less than High School      High School Graduate
##                2676                5765
##      Some College Education College Graduate or better
##                7238                10358
```

```
table(NHIS_2021_clean$PHSTAT_A)
```

```
##
##      1      2      3      4      5
## 6065 9185 7287 2717  783
```

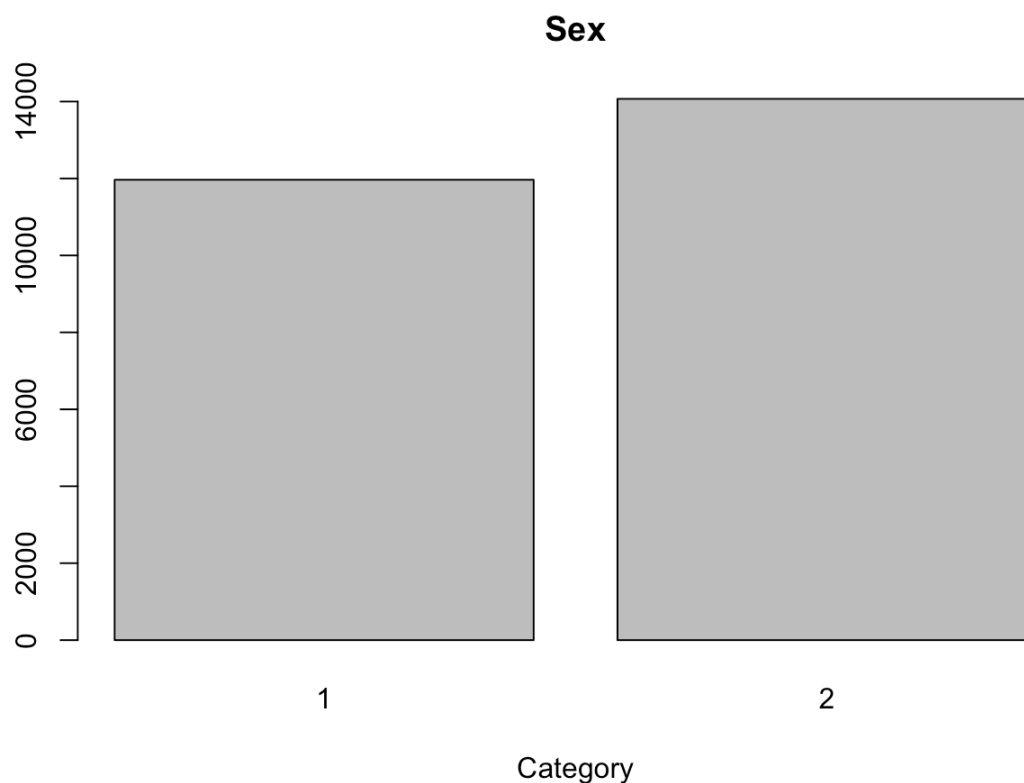
```
table(NHIS_2021_clean$LSATIS4R_A)
```

```
##
##      1      2      3      4
## 12458 12266  1025   288
```

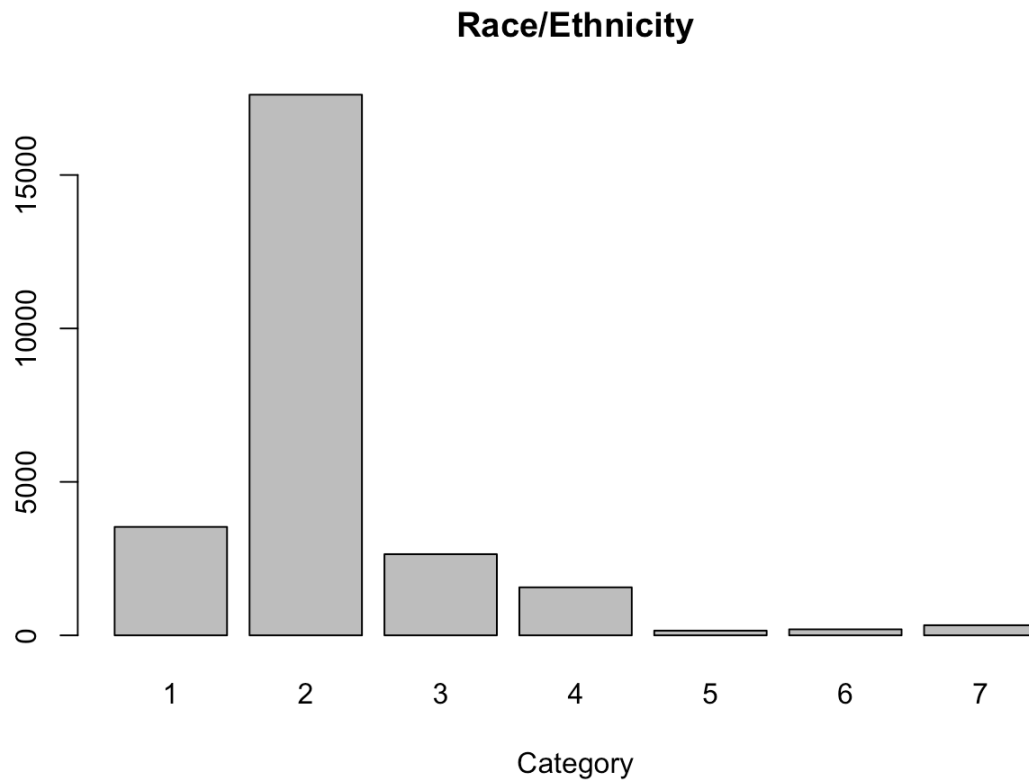
### Interpretation:

Many respondents have at least a high school diploma, with most reporting some college or a college degree at 10,358 participants. There were more females than males within the data, most participants were non hispanic white, had very good health status, and were very satisfied or satisfied with their lives. From a public health standpoint, results indicate that this dataset has higher numbers of participants that could be considered mentally and physically content, and further research could be done to see if any to variables with what could be considered positive results are related to one another.

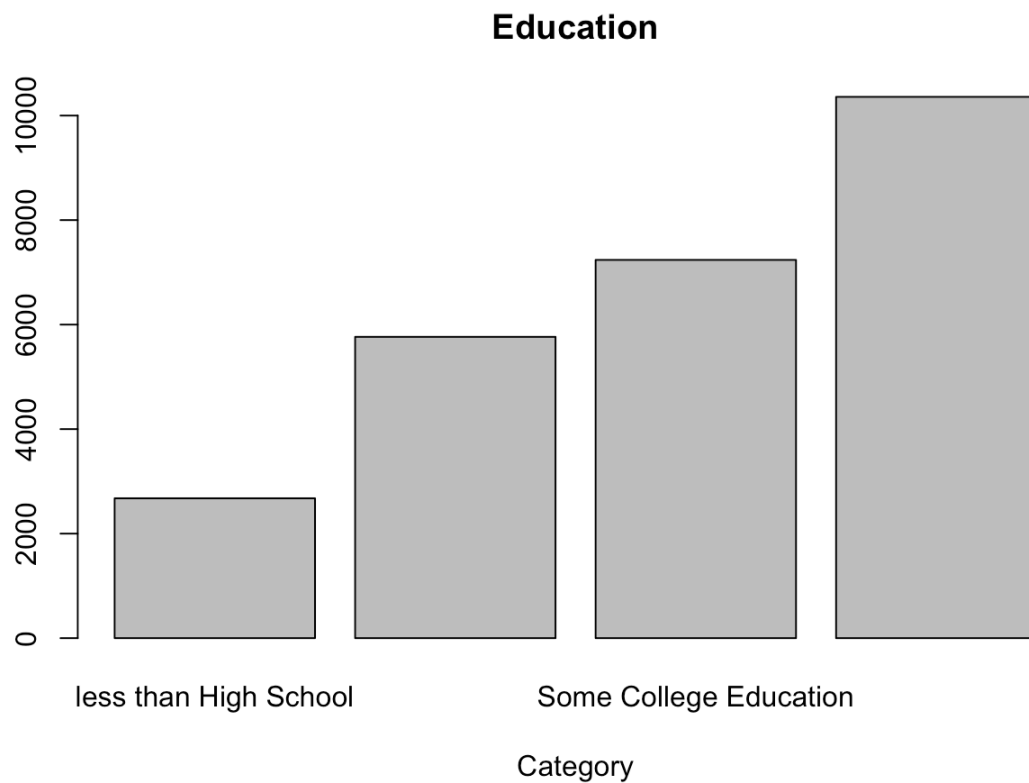
```
## Base R Bar plots for qualitative variables
barplot(table(NHIS_2021_clean$SEX_A), main="Sex", xlab="Category")
```



```
barplot(table(NHIS_2021_clean$HISPALLP_A), main="Race/Ethnicity", xlab="Category")
```

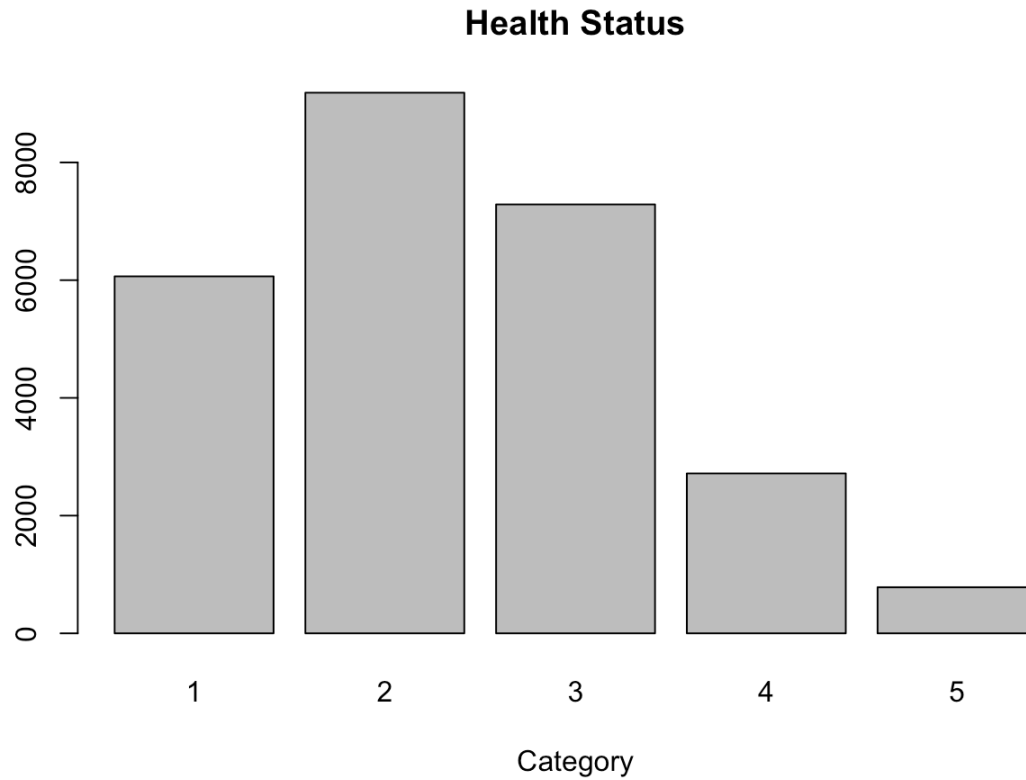


```
barplot(table(NHIS_2021_clean$EDUCP_A), main="Education", xlab="Category")
```

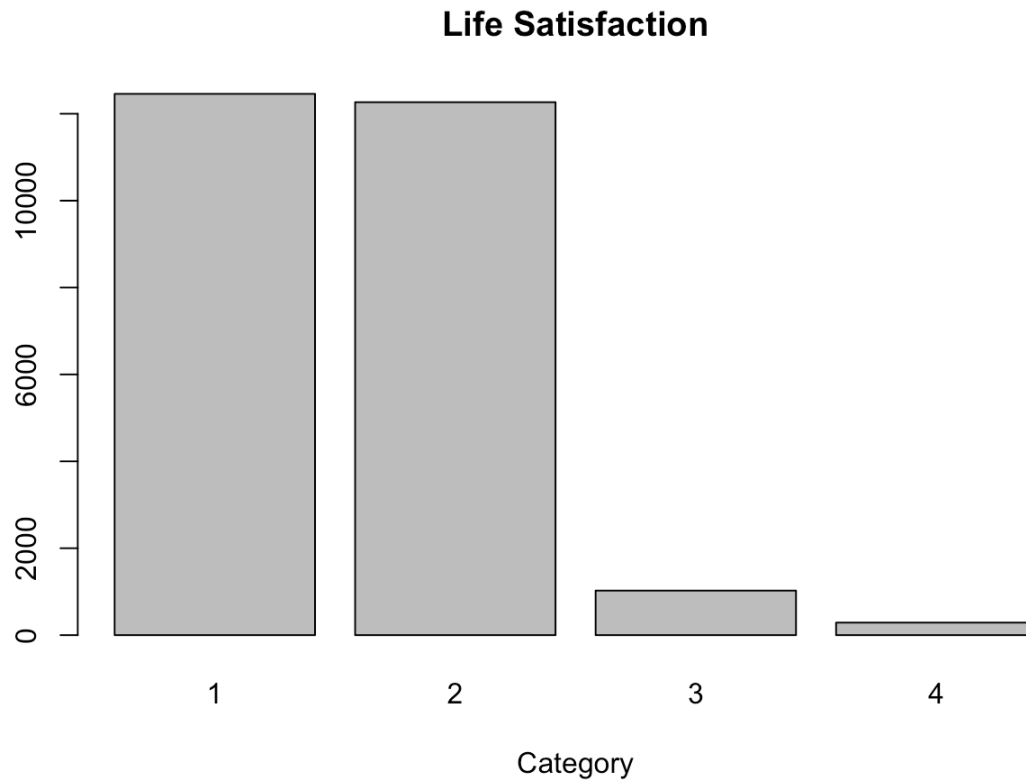




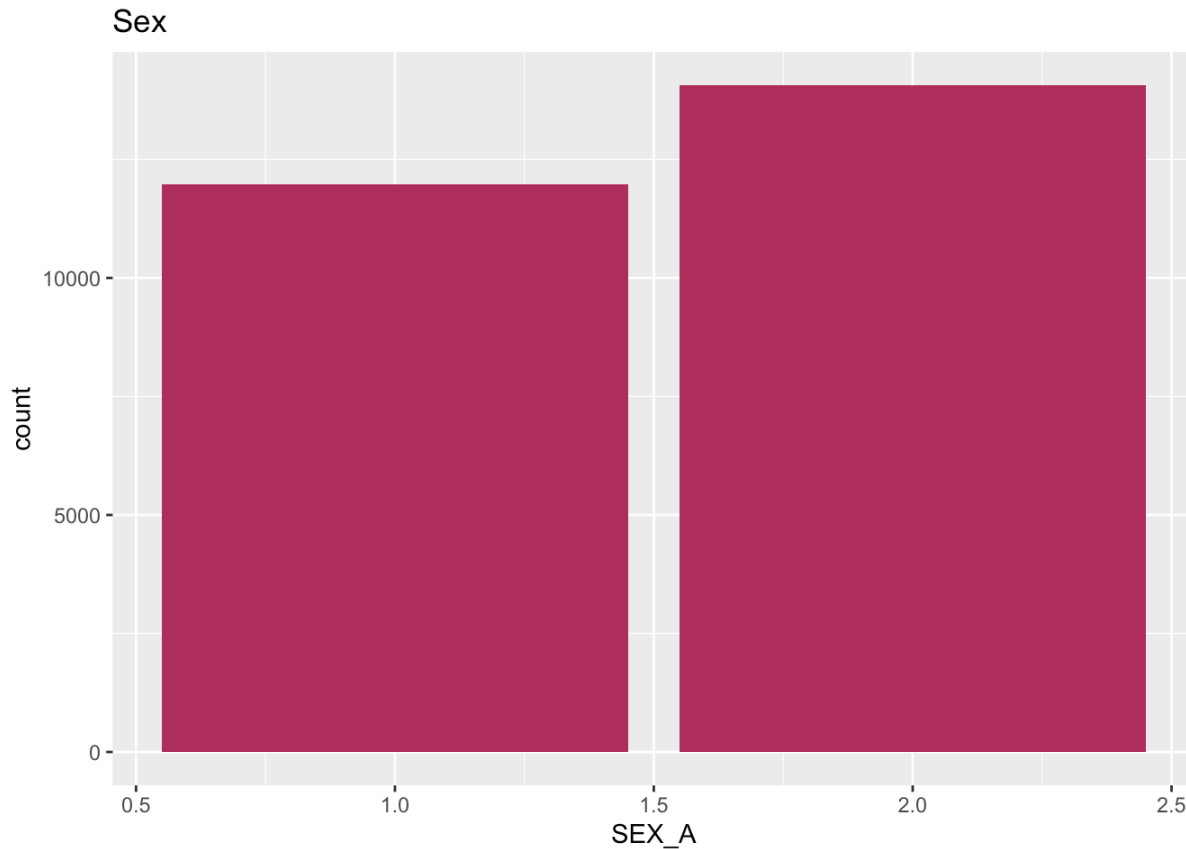
```
barplot(table(NHIS_2021_clean$PHSTAT_A), main="Health Status", xlab="Category")
```



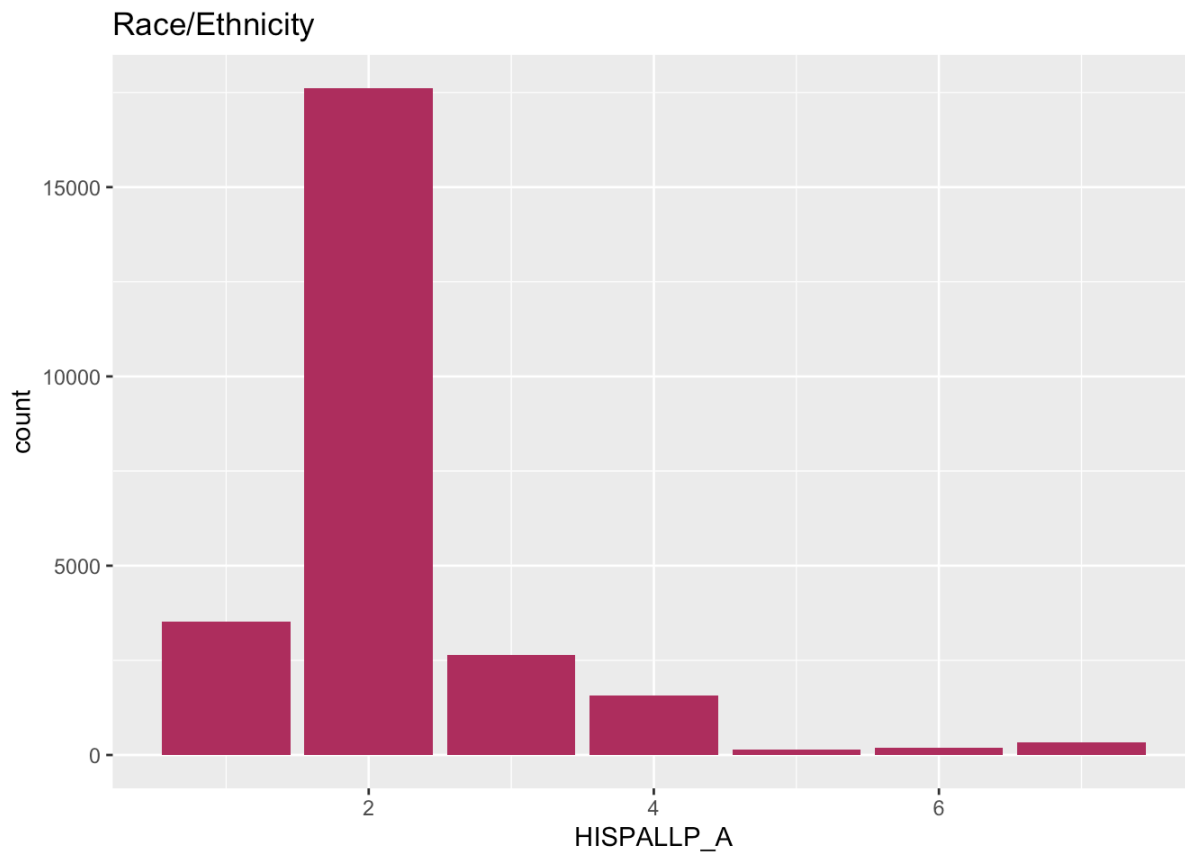
```
barplot(table(NHIS_2021_clean$LSATIS4R_A), main="Life Satisfaction", xlab="Category")
```



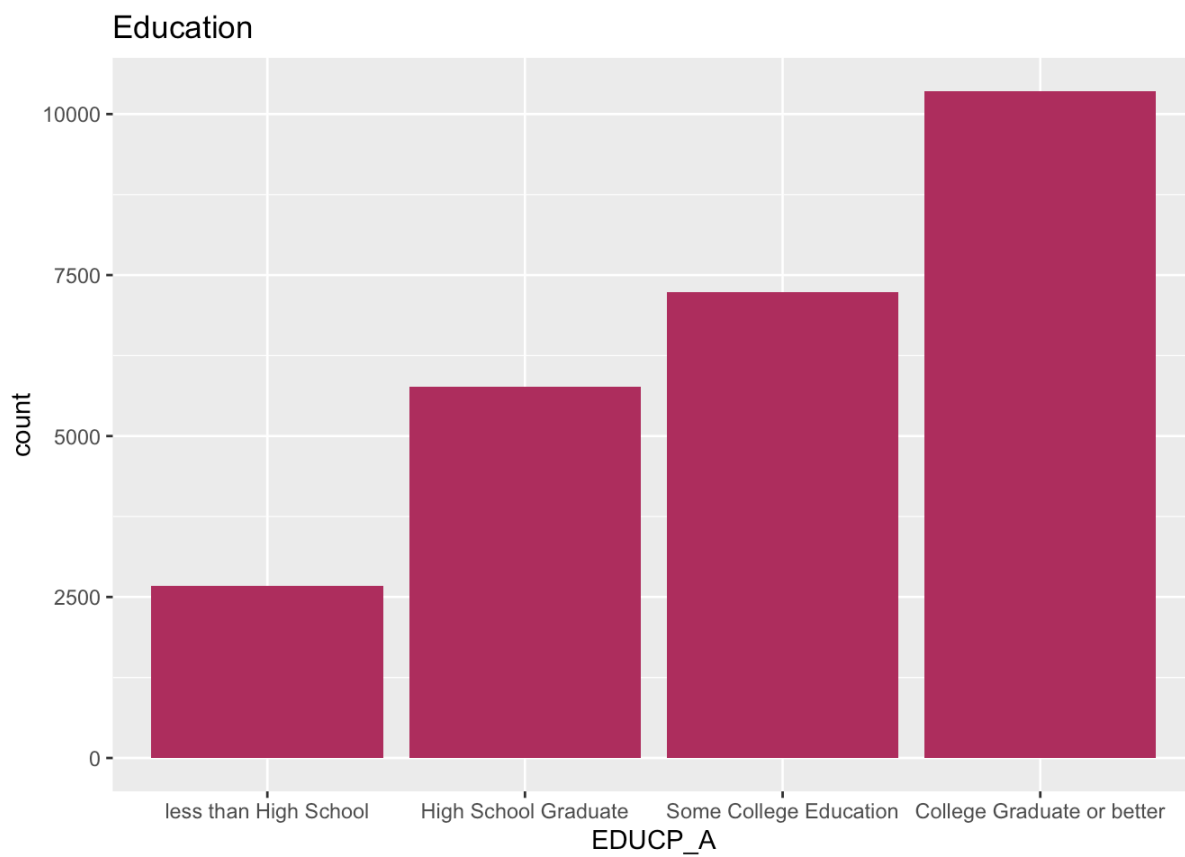
```
##ggplot2 bar plots  
ggplot(NHIS_2021_clean, aes(x=SEX_A)) + geom_bar(fill="maroon") + labs(title="Sex")
```



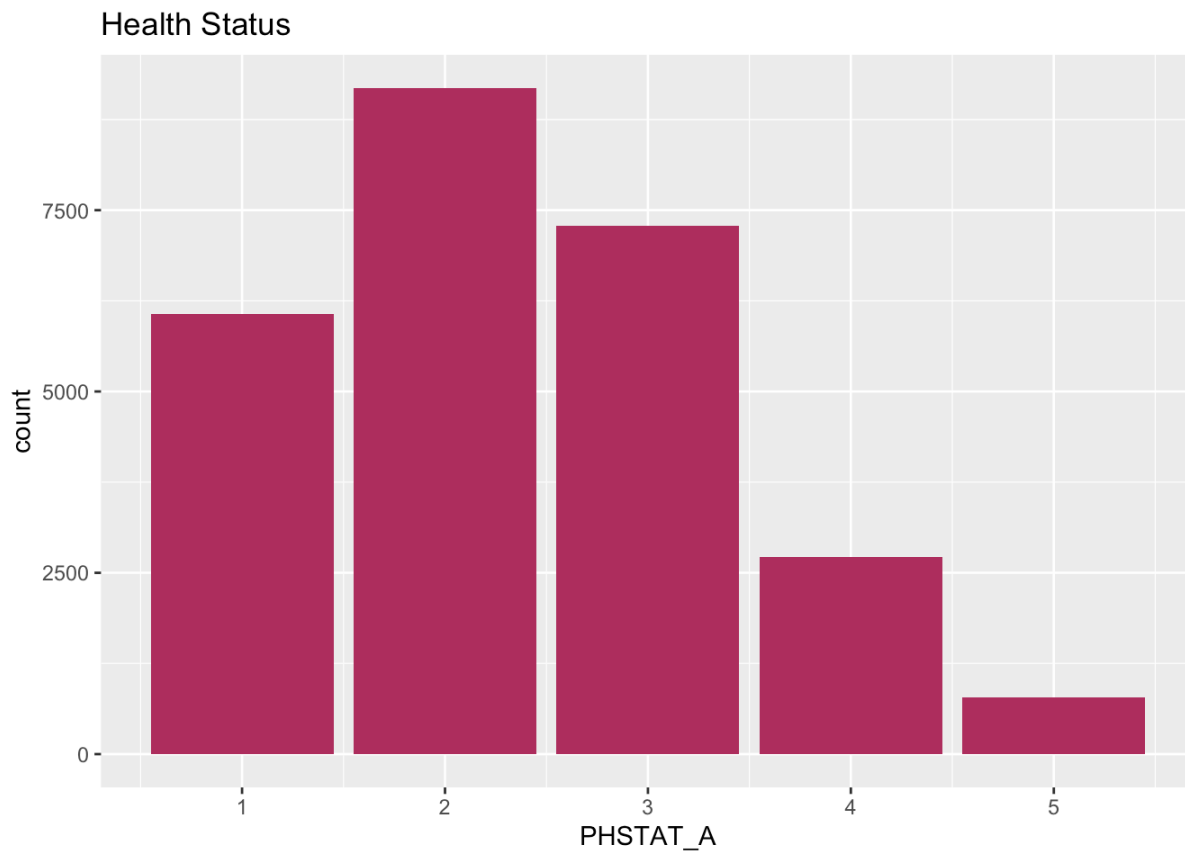
```
ggplot(NHIS_2021_clean, aes(x=HISPALLP_A)) + geom_bar(fill="maroon") + labs(title="Race/Ethnicit  
y")
```



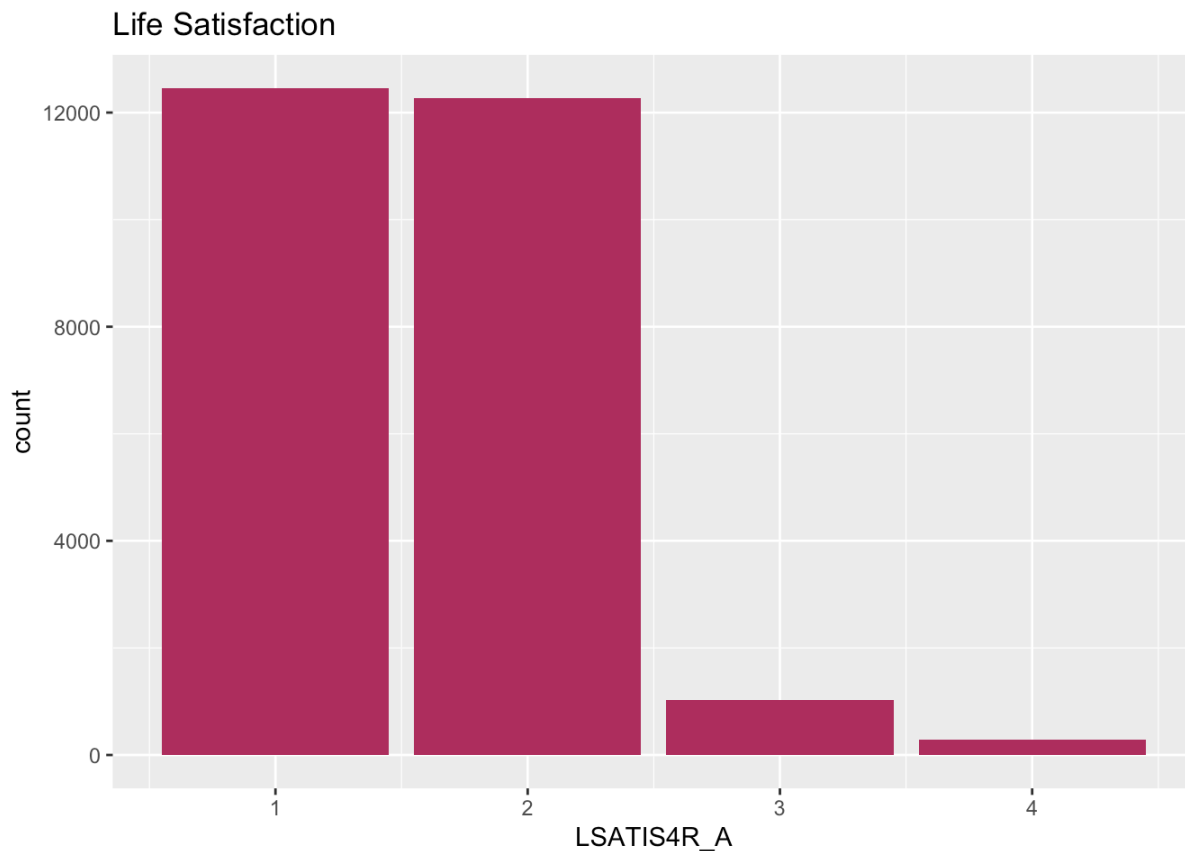
```
ggplot(NHIS_2021_clean, aes(x=EDUCP_A)) + geom_bar(fill="maroon") + labs(title="Education")
```



```
ggplot(NHIS_2021_clean, aes(x=PHSTAT_A)) + geom_bar(fill="maroon") + labs(title="Health Status")
```



```
ggplot(NHIS_2021_clean, aes(x=LSATIS4R_A)) + geom_bar(fill="maroon") + labs(title="Life Satisfacti  
on")
```

**Interpretation:**

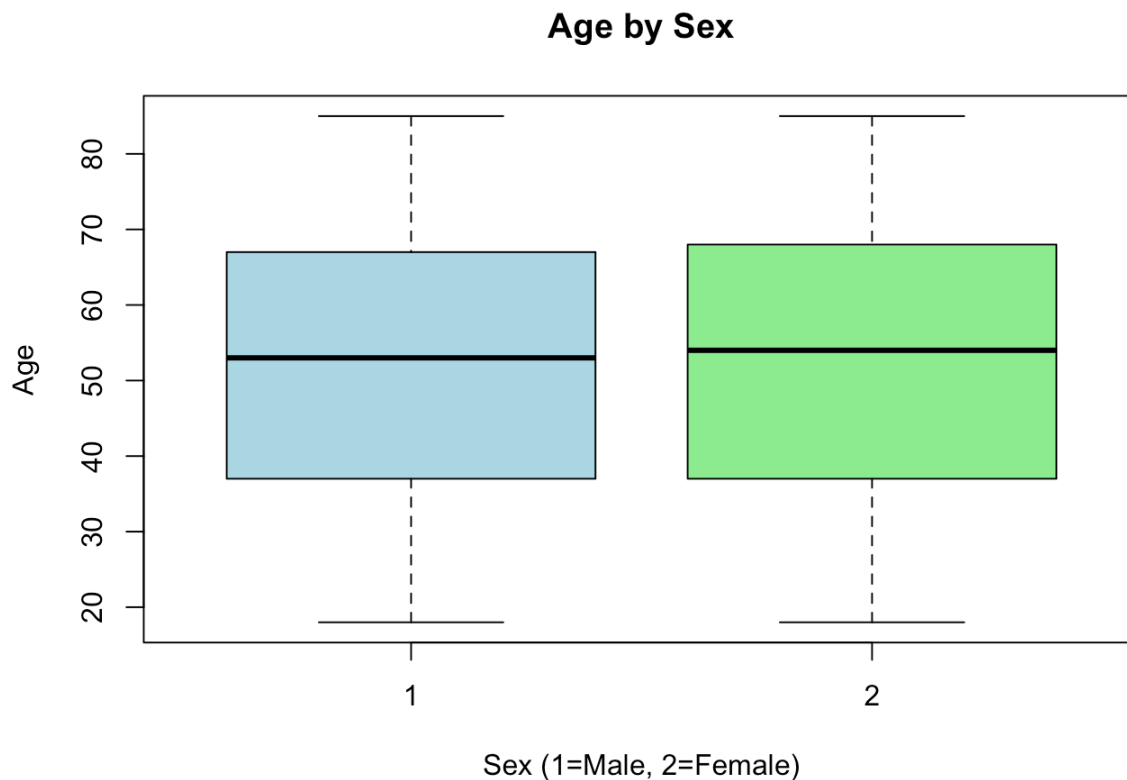
Most respondents have at least a high school diploma, with many reporting some college or a college degree.

## #Day 3 Task 2: Bivariate Analysis

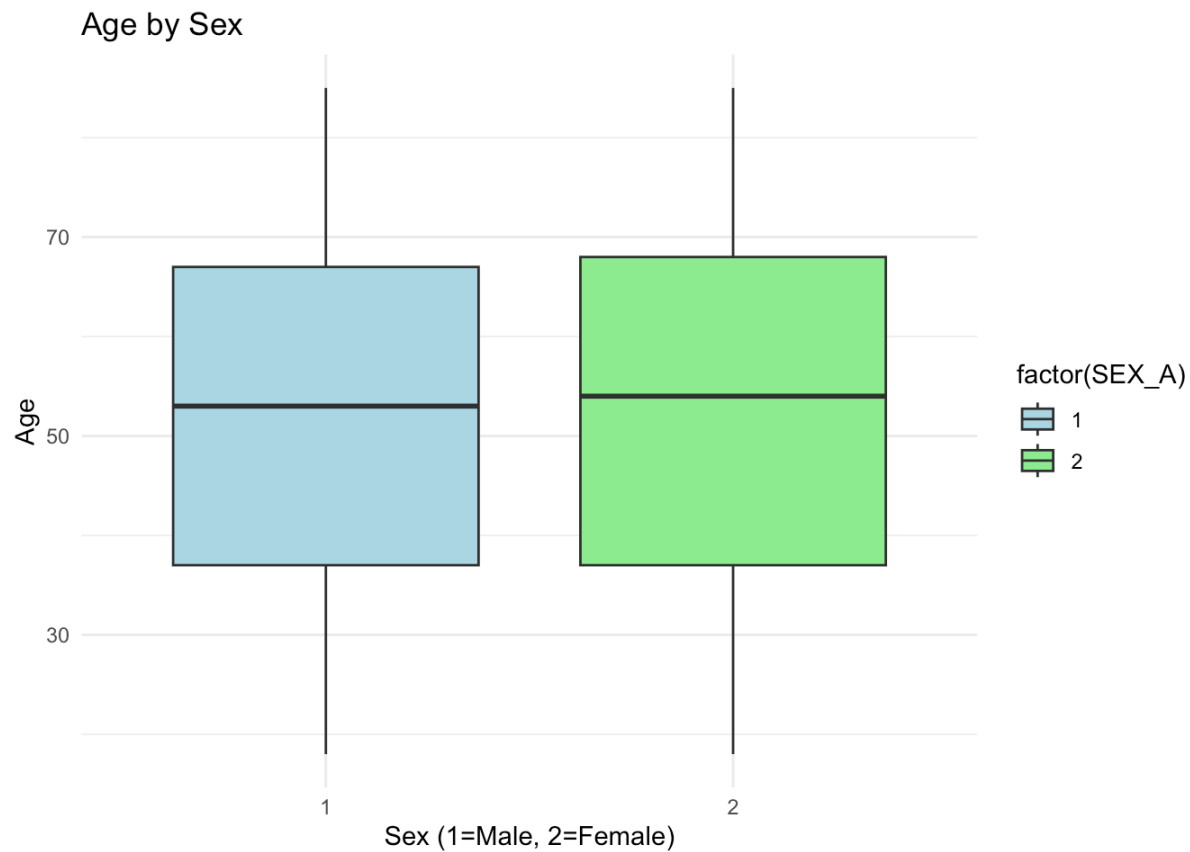
## ## 1. Quantitative vs Qualitative

```
# Load libraries
library(ggplot2)

## 1A: AGE by SEX_A - Base R
boxplot(AGEP_A ~ SEX_A, data=NHIS_2021_clean,
        main="Age by Sex",
        xlab="Sex (1=Male, 2=Female)",
        ylab="Age",
        col=c("lightblue", "lightgreen"))
```



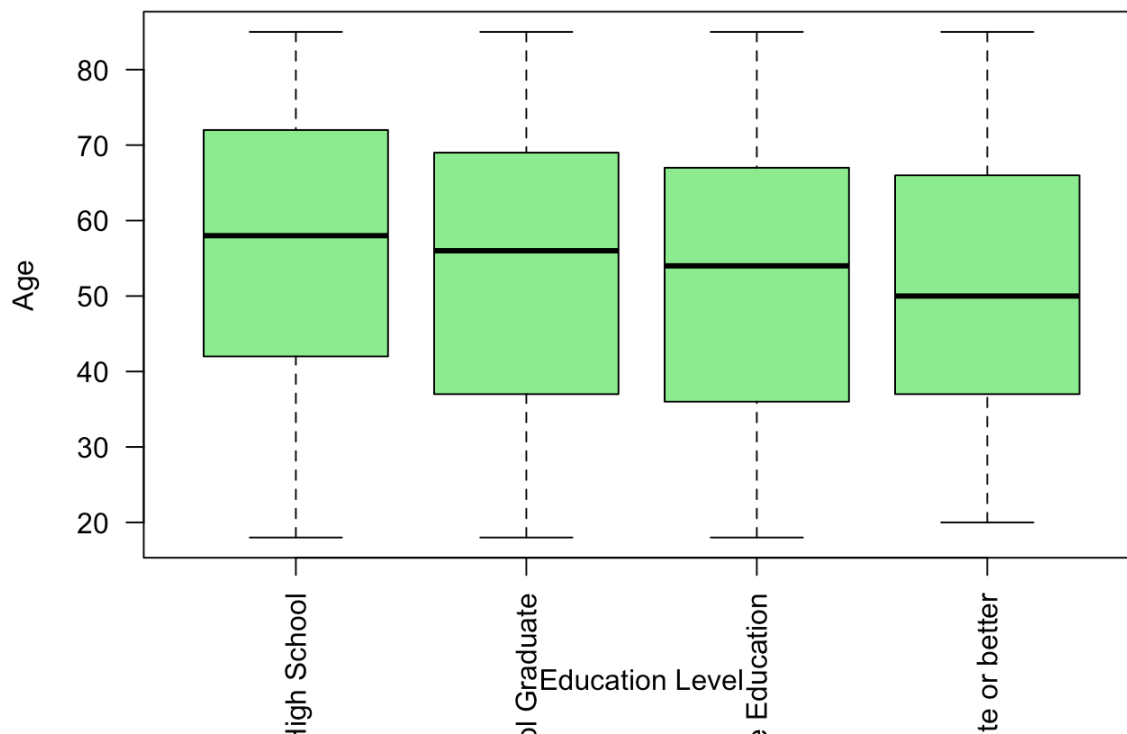
```
# 1B: AGE by SEX_A ggplot2
ggplot(NHIS_2021_clean, aes(x=factor(SEX_A), y=AGEP_A, fill=factor(SEX_A))) +
  geom_boxplot() +
  labs(title="Age by Sex", x="Sex (1=Male, 2=Female)", y="Age") +
  scale_fill_manual(values=c("lightblue", "lightgreen")) +
  theme_minimal()
```

**Interpretation:**

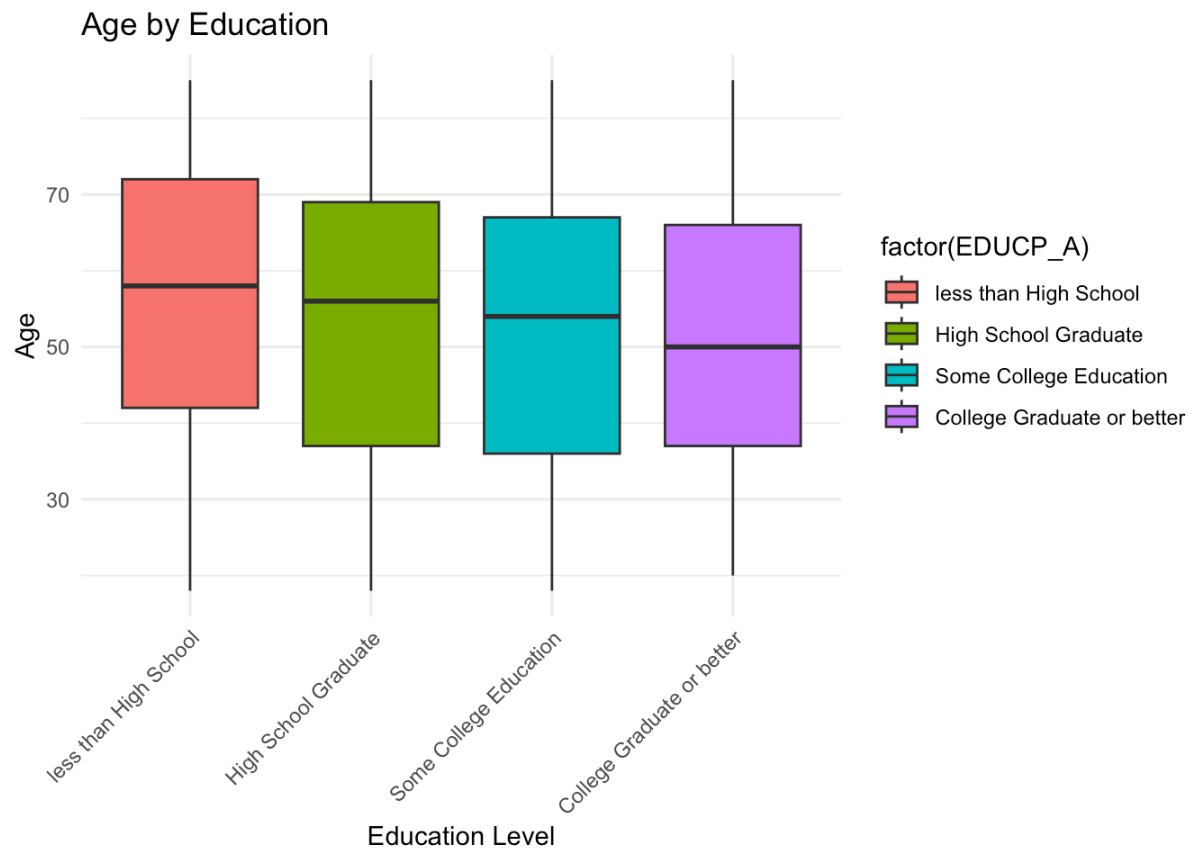
Within the sample, men and women have a similar distribution looking at age.

```
# 1C Age by EDUCP_A Base R
boxplot(AGEP_A ~ EDUCP_A, data=NHIS_2021_clean,
        main="Age by Education",
        xlab="Education Level",
        ylab="Age",
        col="lightgreen",
        las=2)
```

## Age by Education



```
# 1D Age by EDUCP_A ggplot2
ggplot(NHIS_2021_clean, aes(x=factor(EDUCP_A), y=AGEP_A, fill=factor(EDUCP_A))) +
  geom_boxplot() +
  labs(title="Age by Education", x="Education Level", y="Age") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle=45, hjust=1))
```



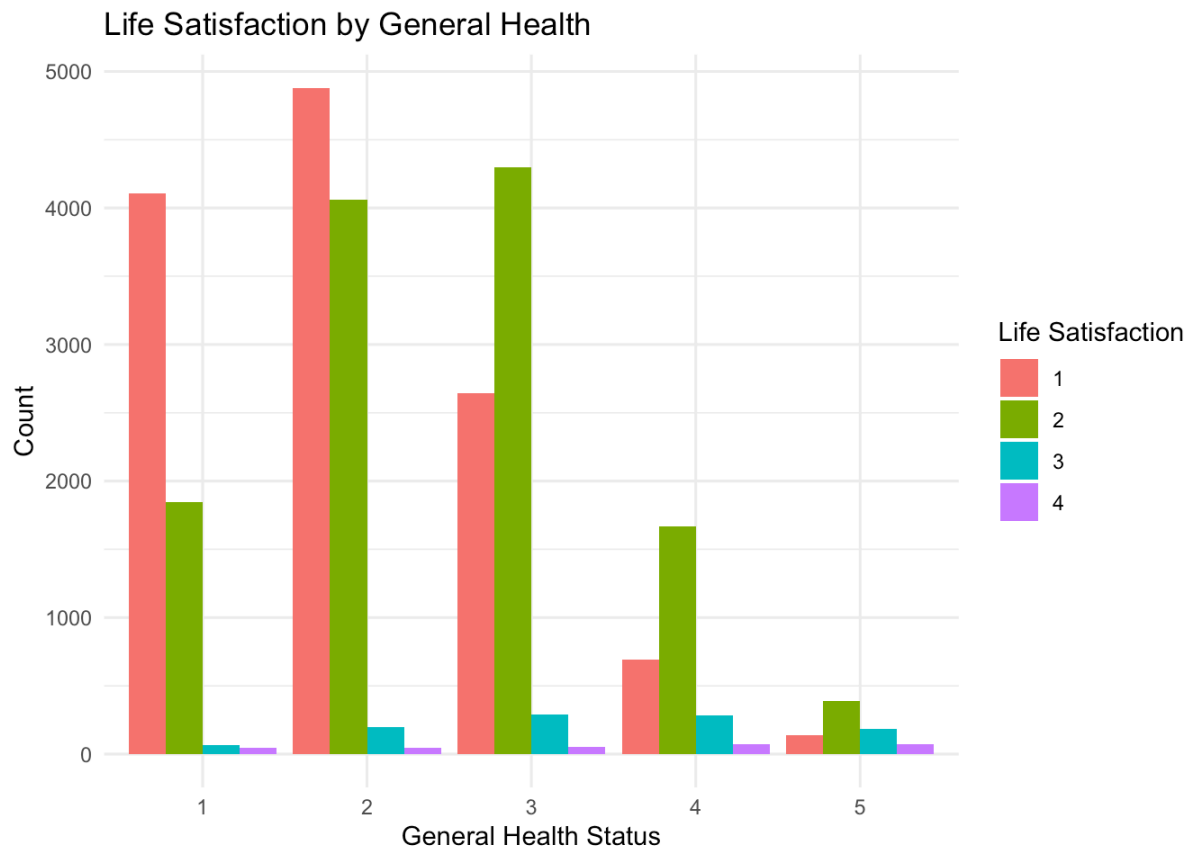
### Interpretation:

When looking at education, age seems to decrease with higher levels of education.

## 2. Qualitative vs Qualitative

```
# Clustered bar chart ggplot2
ggplot(NHIS_2021_clean, aes(x=factor(PHSTAT_A), fill=factor(LSATIS4R_A))) +
  geom_bar(position="dodge") +
  labs(title="Life Satisfaction by General Health",
       x="General Health Status",
       y="Count",
       fill="Life Satisfaction") +
  theme_minimal()
```



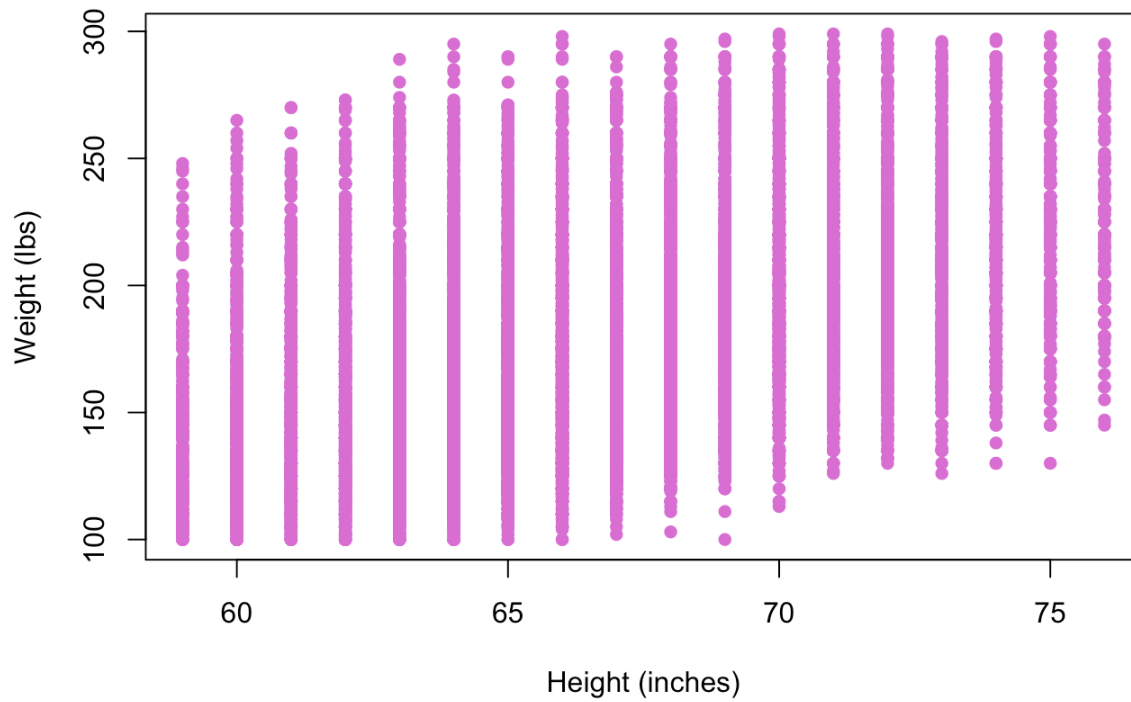
**Interpretation:**

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

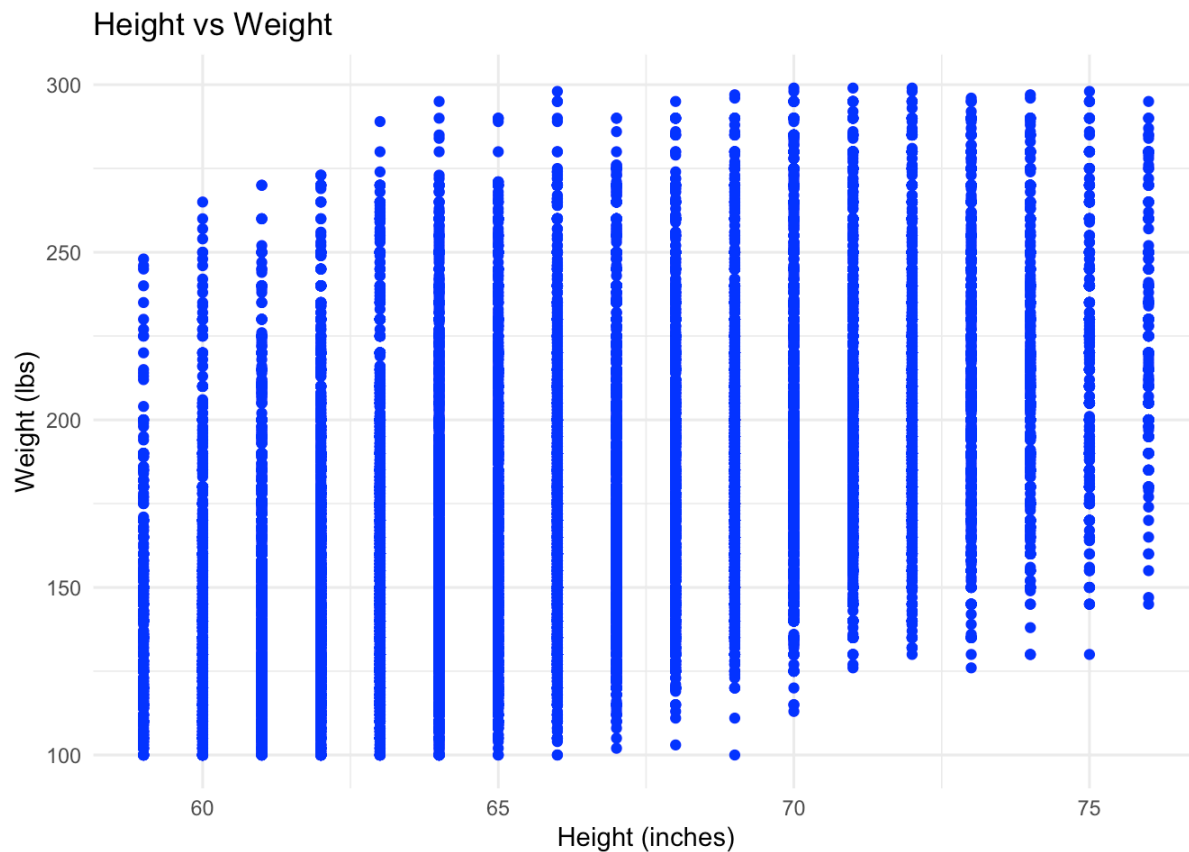
### 3. Quantitative vs Quantitative

```
# 3A: Height vs weight Base R version
plot(NHIS_2021_clean$HEIGHTTC_A, NHIS_2021_clean$WEIGHTLBTC_A,
     main="Height vs Weight",
     xlab="Height (inches)",
     ylab="Weight (lbs)",
     col="orchid", pch=16)
```

## Height vs Weight



```
# 3B Height vs weight ggplot2
ggplot(NHIS_2021_clean, aes(x=HEIGHTTC_A, y=WEIGHTLBTC_A)) +
  geom_point(color="blue") +
  labs(title="Height vs Weight", x="Height (inches)", y="Weight (lbs)") +
  theme_minimal()
```

**Interpretation:**

This graph shows that low height is related to lower weight, while larger height is related to higher weight.

```
# 3C Correlation coefficient
```

```
cor_value <- cor(NHIS_2021_clean$HEIGHTTC_A, NHIS_2021_clean$WEIGHTLBTC_A, use="complete.obs")
cat("Correlation coefficient (Height vs Weight):", cor_value, "\n")
```

```
## Correlation coefficient (Height vs Weight): 0.5023037
```

**Interpretation:**

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

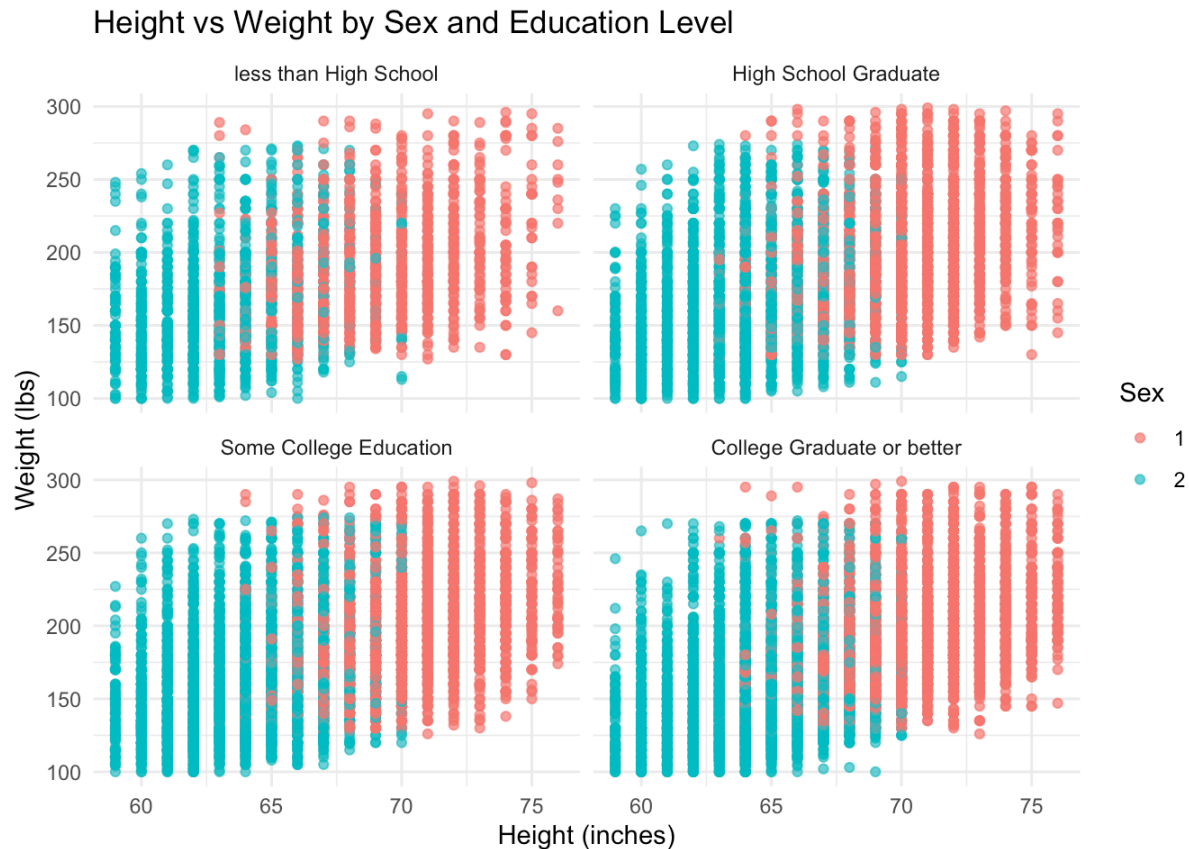
## Multivariate Visualization

```
# Load packages
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
## %+%, alpha
```

```
##### Task 1: Part 1 Enhancing Scatter Plot (Height vs Weight)
# Colored by SEX_A and faceted by EDUCP_A
ggplot(NHIS_2021_clean, aes(x = HEIGHTTC_A, y = WEIGHTLBTC_A, color = factor(SEX_A))) +
  geom_point(alpha = 0.7) +
  facet_wrap(~ EDUCP_A) +
  labs(
    title = "Height vs Weight by Sex and Education Level",
    x = "Height (inches)",
    y = "Weight (lbs)",
    color = "Sex"
  ) +
  theme_minimal()
```



#### Interpretation:

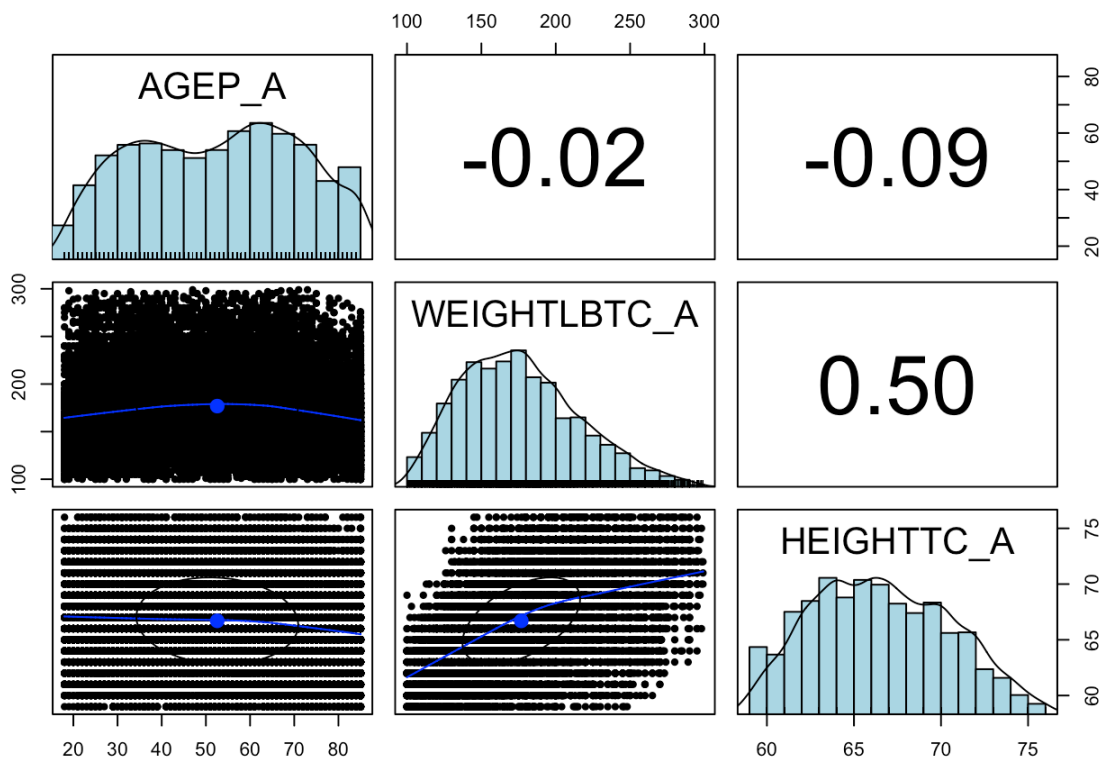
The height-weight trend is consistent across sexes and education groups, though men tend to weigh more at comparable heights.

## Part 2 Correlation Plot Matrix

```
# Variables: Age, Weight, Height

vars_for_corr <- NHIS_2021_clean[, c("AGEP_A", "WEIGHTLBTC_A", "HEIGHTTC_A")]

# Psych package scatter matrix
pairs.panels(
  vars_for_corr,
  method = "pearson",      # method used
  hist.col = "lightblue",  # color of histograms
  density = TRUE,
  ellipses = TRUE
)
```



### 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.