Social Survey Data Wrangling

T Hill

2022-09-23

```
library(sjPlot)
```

```
## Learn more about sjPlot with 'browseVignettes("sjPlot")'.
```

```
library(ggplot2)
library(forcats)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
health <- read.csv("C:/Users/thill/Desktop/math130/data/AddHealth_Wave_IV.csv", header=TRUE)
```

DATA INFORMATION

Add Health

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States during the 1994-95 school year. The Add Health cohort has been followed into young adulthood with four in-home interviews, the most recent in 2008, when the sample was aged 24-32. Add Health is re-interviewing cohort members in a Wave V follow-up from 2016-2018 to collect social, environmental, behavioral, and biological data with which to track the emergence of chronic disease as the cohort moves through their fourth decade of life ("https://www.norcalbiostat.com/data/"). The survey was conducted in 2008 and 2009.

INTRODUCTION

This data set presents an opportunity to briefly look at some correlations between several variables that outline individuals income, education level, and married life. The data is first trimmed down from its numbered inputs and refined into the corresponding values they represent. Each variable has been grouped and relabeled to be better formatted in R.

```
datastat <- select(health, H4ED2, H4EC1, H4EC7, H4RD7A)
datastat$H4EC1[datastat$H4EC1==98] <- NA
datastat$H4EC1[datastat$H4EC1==96] <- NA
datastat$H4EC7[datastat$H4EC7==98] <- NA
datastat$H4EC7[datastat$H4EC7==96] <- NA
datastat$H4RD7A[datastat$H4RD7A==98] <- NA
datastat$H4RD7A[datastat$H4RD7A==96] <- NA
datastat$H4ED2[datastat$H4ED2==12] <- NA
datastat$H4ED2[datastat$H4ED2==13] <- NA
datastat$H4EC1[datastat$H4EC1==1] <- NA
datastats <- na.omit(datastat)

table(datastats$H4EC1,useNA="always")
```

```
##
##    2    3    4    5    6    7    8    9   10   11   12 <NA>
##  106  156  154  221  245  469  523 1045  626  425  209    0
```

```
table(datastats$H4ED2,useNA="always")
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11 <NA>
##    8  264  648  148  274 1458  888  175  231   56   29    0
```

```
table(datastats$H4RD7A,useNA="always")
```

```
##
##    1    2    3    4    5 <NA>
## 2316 1262  367  139   95    0
```

```
table(datastats$H4EC7,useNA="always")
```

```
##
##    1    2    3    4    5    6    7    8    9 <NA>
##  680  506  789  749  646  484  194   86   45    0
```

```
datastats$Education_Level <- factor(datastats$H4ED2, labels=c("HS Graduate or Less", "HS Graduate or Le
datastats$Marriage_Happiness <- factor(datastats$H4RD7A, labels=c("Happy", "Happy", "Neutral", "Unhappy
datastats$Household_Assets_Thousands <- factor(datastats$H4EC7, labels=c("< 5.0", "10.0", "25.0", "$50.
datastats$Household_Income <- factor(datastats$H4EC1, labels=c("Low", "Low", "Low", "Low", "Med", "Med"

table(datastats$Household_Assets_Thousands, useNA="always")
```

```
##
##      < 5.0      10.0      25.0     $50.0    $100.0    $250.0    $500.0   $1000.0
##        680       506       789       749       646       484       194        86
## > $1000.0      <NA>
##         45         0
```

```
table(datastats$Household_Income, useNA="always")
```

```
##
##  Low  Med High <NA>
##  637 1237 2305    0
```

```
table(datastats$Education_Level, useNA="always")
```

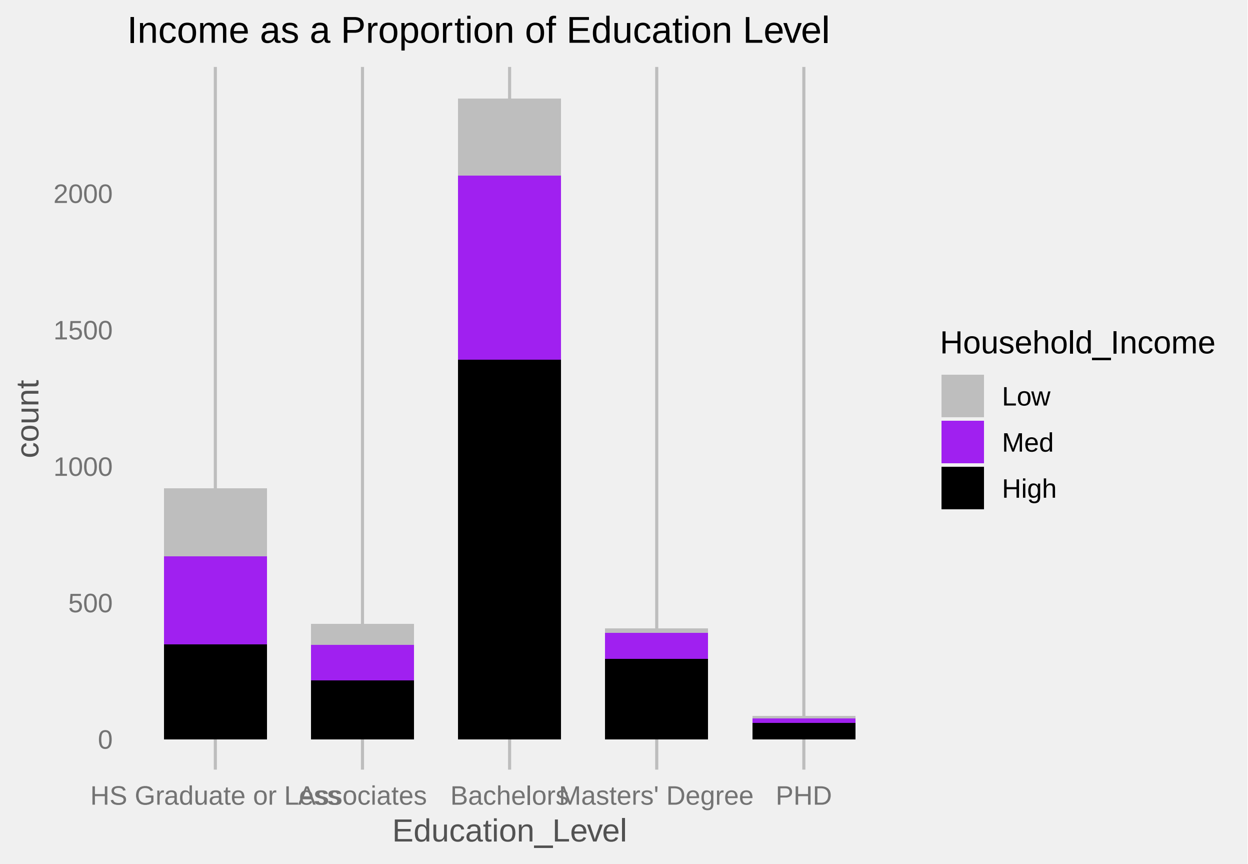
```
##
## HS Graduate or Less          Associates           Bachelors     Masters' Degree
##                 920                 422                2346                 406
##                 PHD                <NA>
##                  85                   0
```

```
table(datastats$Marriage_Happiness, useNA="always")
```

```
##
##   Happy Neutral Unhappy    <NA>
##    3578     367     234       0
```

Below is a bar plot of Education_Level made out of its component Household_Income levels.

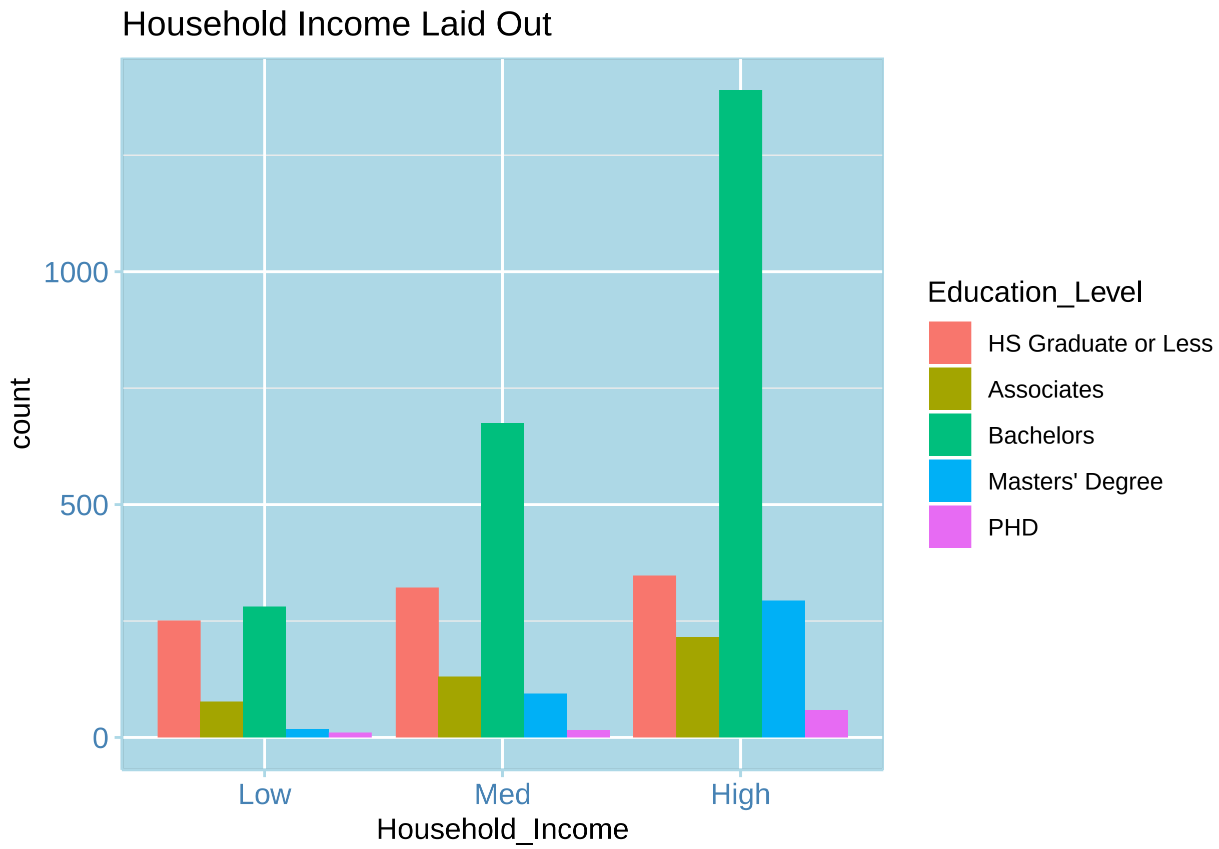
```
ggplot(datastats, aes(x=Education_Level, fill=Household_Income)) + geom_bar(width = .7) +
  scale_fill_manual(values=c("gray", "purple", "black")) + theme_538() +
ggtitle("Income as a Proportion of Education Level")
```



This plot has a count on the y-axis. The count of Bachelors is clearly pretty high so at first glance we can conclude the majority of participants have earned Bachelors' degrees. It also appears that low income earners are mostly contained to the domain of bachelors and below.

```
theme_bluewhite <- function (base_size = 11, base_family = "") {
    theme_bw() %+replace%
    theme(
      panel.grid.major  = element_line(color = "white"),
      panel.background = element_rect(fill = "lightblue"),
      panel.border = element_rect(color = "lightblue", fill = NA),
      axis.line = element_line(color = "lightblue"),
      axis.ticks = element_line(color = "lightblue"),
      axis.text = element_text(color = "steelblue"))}

ggplot(datastats, aes(x=Household_Income, fill=Education_Level)) + geom_bar(position = "dodge") + theme_
```

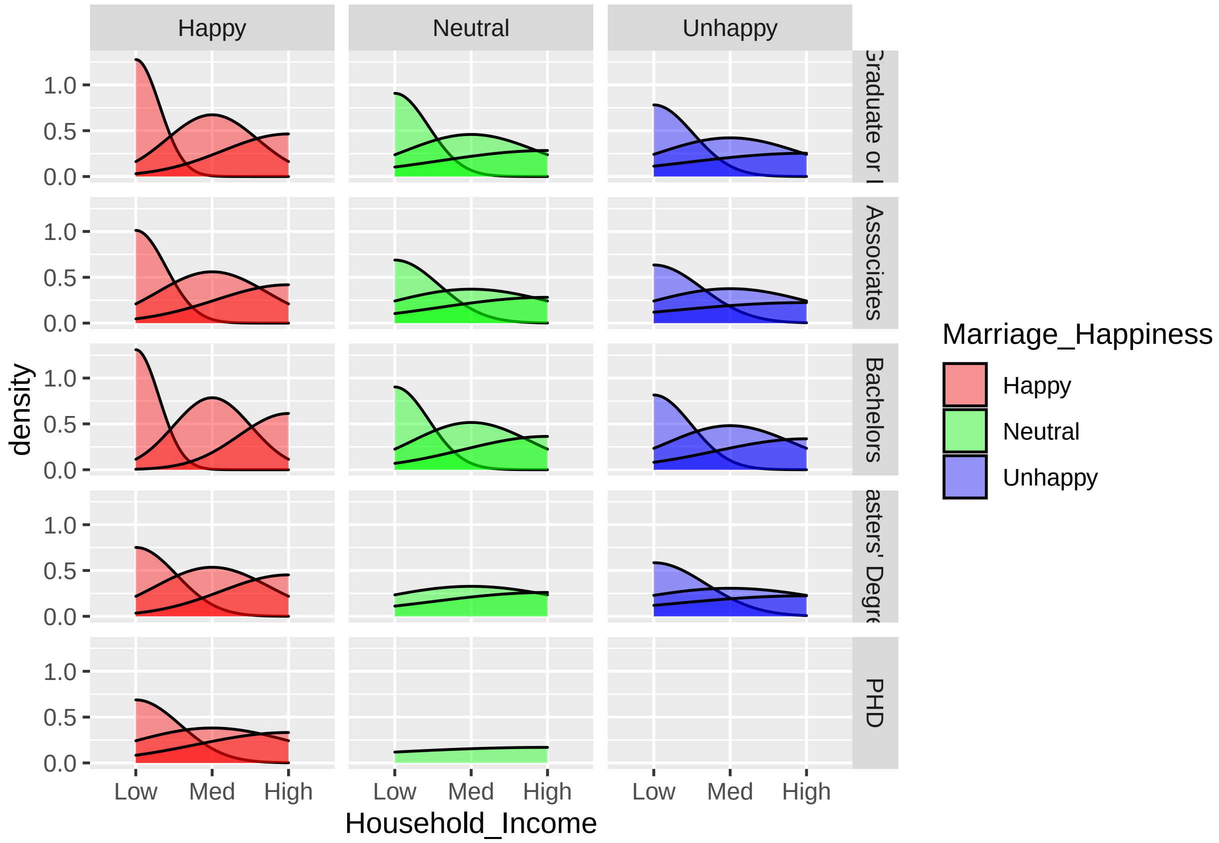


Here we see Education_Level as a function of Household_Income. Each subset of each variable can be clearly seen and compared within its own grouping. The y-axis is just a count of the totals within each category of income. What should be compared and is very interesting are the spread of each Education_Level across all income categories, here we can see that the lowest education level is fairly well represented across all income categories. The data suggests fairly even chances of being in any income level with a low amount of education.

```
datastats %>% select(Education_Level, Marriage_Happiness,Household_Income) %>% na.omit() %>%
  ggplot(aes(x=Household_Income, fill=Marriage_Happiness)) +
  geom_density(alpha=.4) + facet_grid(Education_Level~Marriage_Happiness) +  scale_fill_manual(values=c
## Warning: Groups with fewer than two data points have been dropped.
## Groups with fewer than two data points have been dropped.
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```

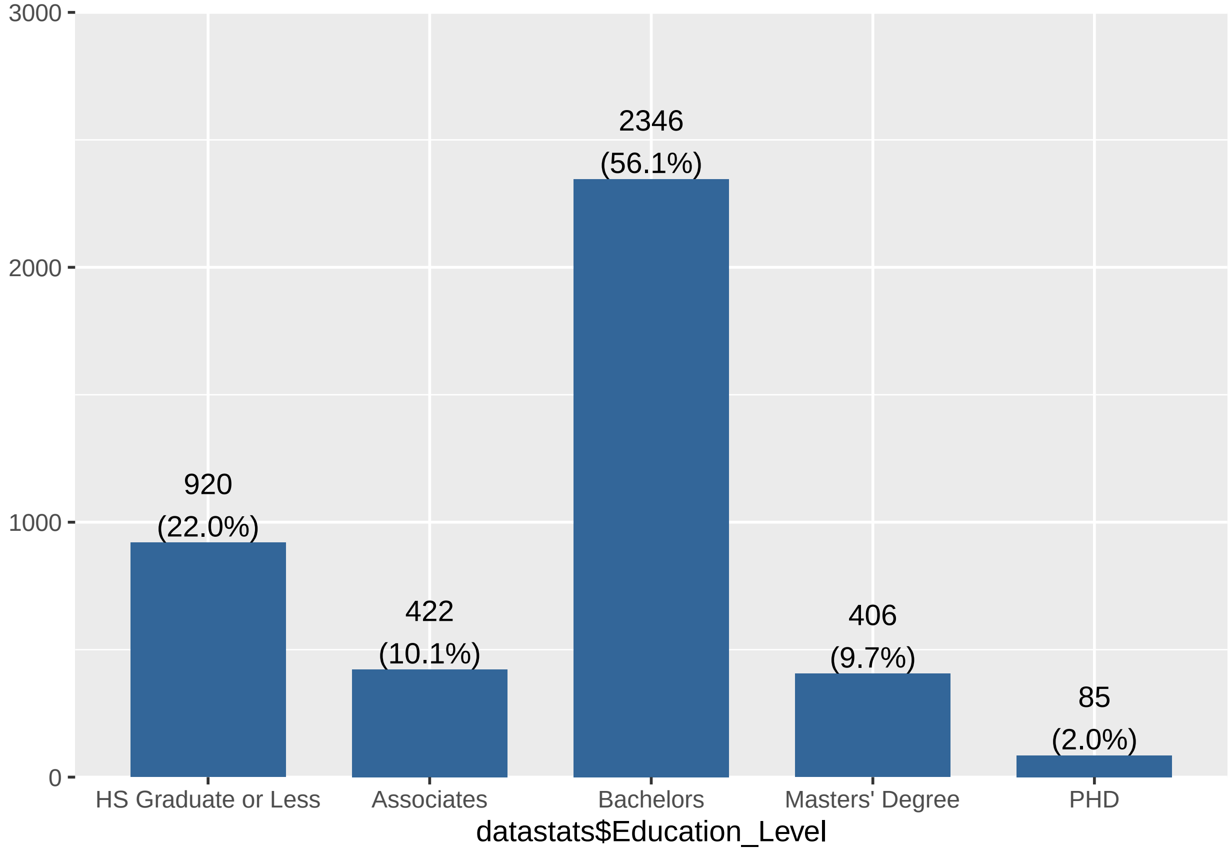


```
table(datastats$Education_Level)
```

```
## 
## HS Graduate or Less          Associates           Bachelors     Masters' Degree 
##                 920                 422                2346                 406 
##                 PHD 
##                  85
```

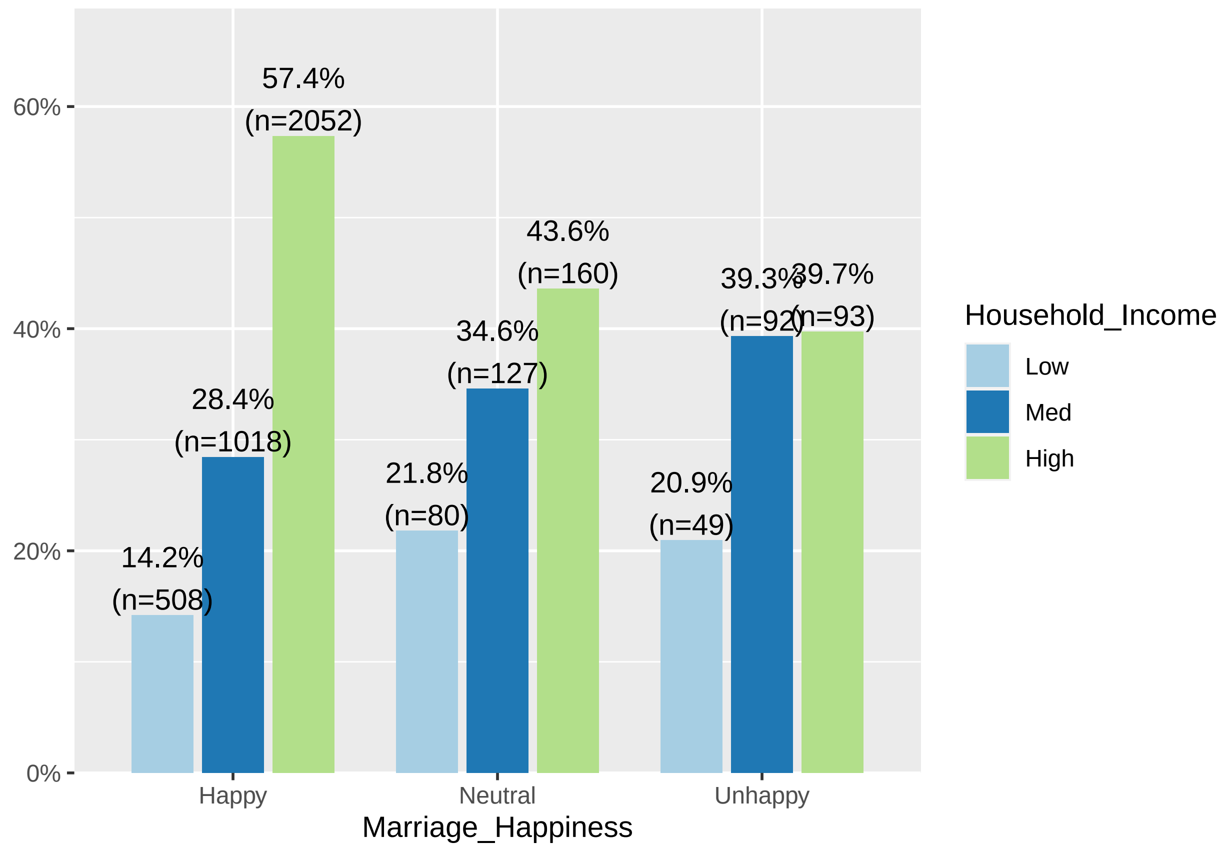
This chart looks pretty confusing and it is pretty confusing. All 5 possible values of the Education_level are arranged on the right side, forming their own rows across which the Houshold_Income levels are represented. The chart is further broken down into 3 columns, each of which represent a value within the grouped Marriage_Happiness category. This set of charts suggests earning a PHD is associated with a happier marriage.

```
plot_frq(datastats$Education_Level)
```



This is a simple breakdown of the survey participants Education_Level as a percent of the whole. Bachelors degrees are more common than all other education levels combined.

```
plot_xtab(datastats$Marriage_Happiness, datastats$Household_Income, margin='row', show.total = FALSE)
```



And finally here's a chart of Marriage_Happiness broken down by Household_Income. The data suggests a majority of people are happy and earn a high income from within this sample.