Project

Tina Qian, Yanlin Li

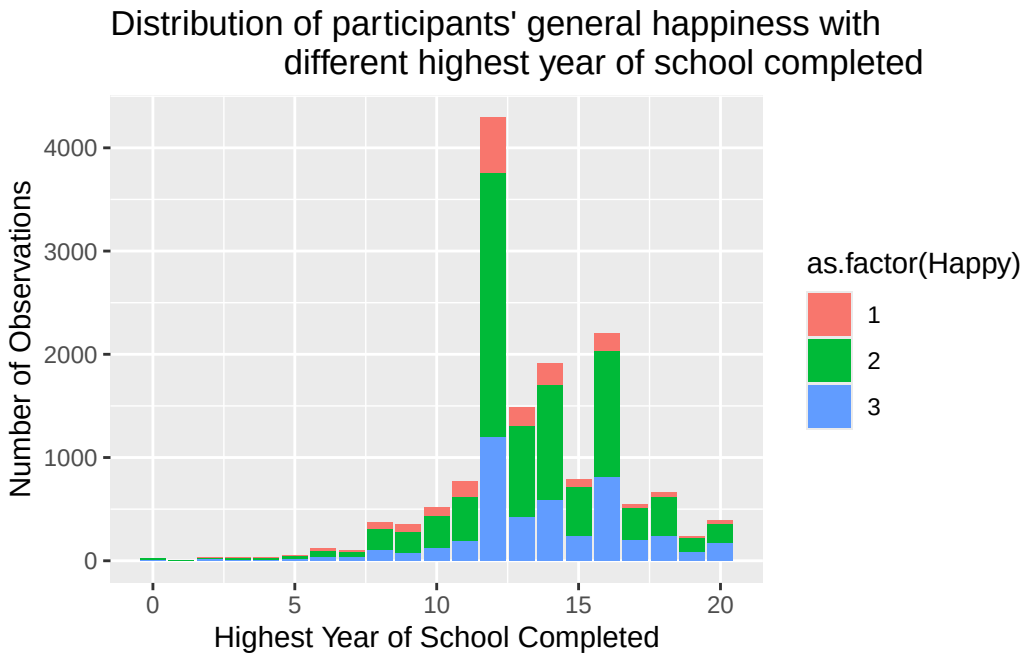
```
library(tidyverse)
library(tidymodels)
library(readr)
library(tidyr)
library(ggplot2)
happiness <- read_csv("happiness.csv")
```

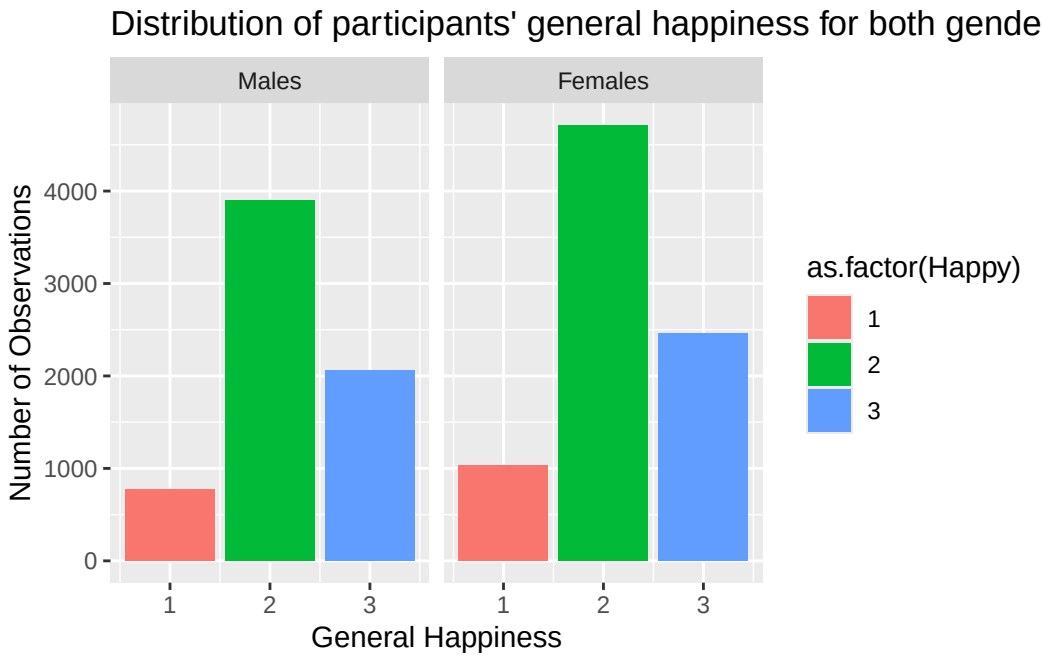
Introduction: Motivation and Exploratory Data Analysis

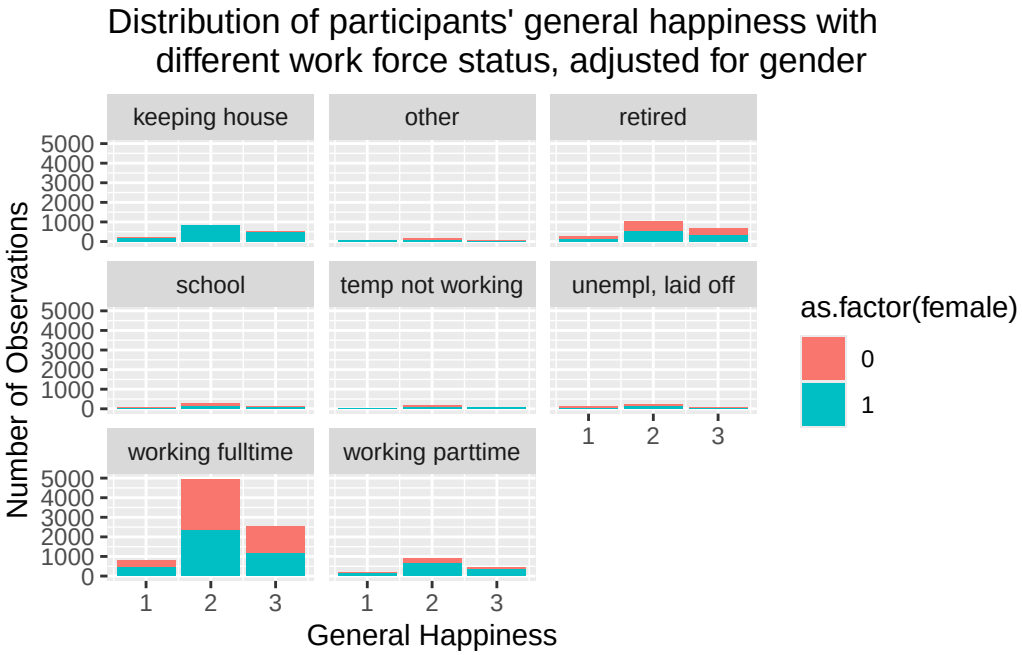
Happiness has long been one of people's most important aspires, and the relationship between socioeconomic status (SES) and happiness has become a focal point of multidisciplinary research on individual well-being and societal cohesion. Socioeconomic status, encompassing elements such as income, education, and occupation stands as a pivotal determinant in shaping the quality of life for individuals. Understanding the complexities of this relationship holds significant implications for policymakers, educators, and social scientists alike. While conventional wisdom suggests that higher SES often translates to greater happiness, we are curious to confirm this assumption with data-backed evidence.

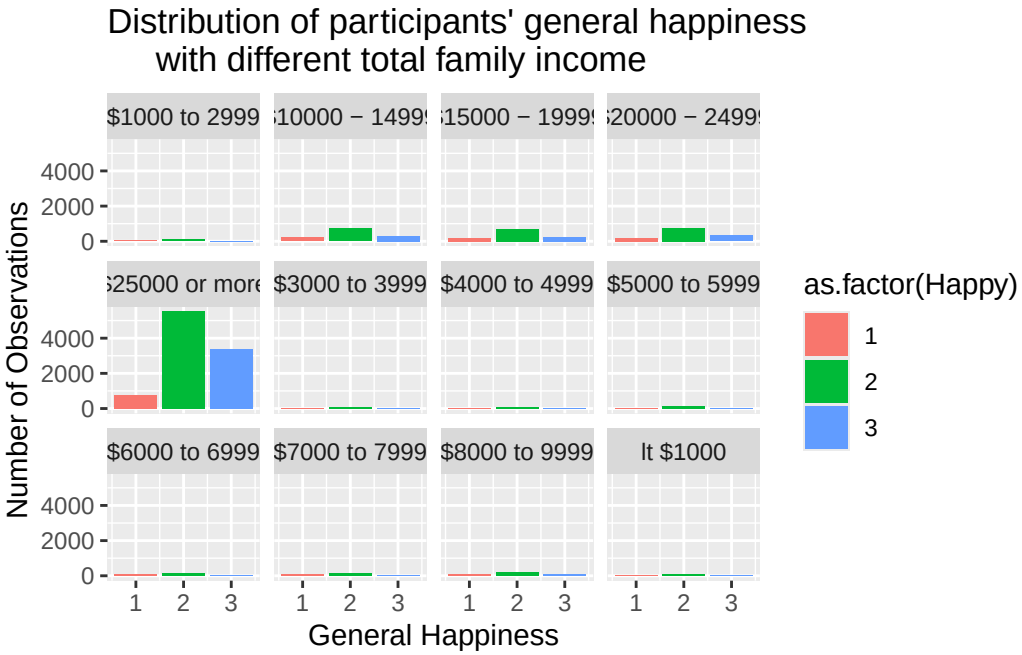
We have come across the dataset labeled "happiness," provided by Wooldridge Data, which enables us to explore the correlation between various aspects of socioeconomic status and an individual's self-assessed happiness. This dataset, comprising 33 variables, contains demographic details of 17,137 individuals. Key variables include "happy" (indicating an individual's overall happiness), "workstat" (employment status), "income" (total family income), "educ" (highest level of education attained), "female" (1 if female), "babies" (household members under 6 years old), "preteens" (household members aged 6 to 12), and "teens" (household members aged 13 to 17).

Our research question in this instance is whether individuals with higher socioeconomic standing tend to report greater levels of happiness. Our hypothesis is a positive correlation; more specifically, an employed male with high household income, high educational level and fewer young members at home is anticipated to be happier.


Distribution of participants' general happiness with
different highest year of school completed
Number of Observations
4000
3000
2000
1000
0
0
5
10
15
20
Highest Year of School Completed
as.factor(Happy)
1
2
3


Distribution of participants' general happiness for both gender
Males
Females
Number of Observations
4000
3000
2000
1000
0
1
2
3
1
2
3
General Happiness
as.factor(Happy)
1
2
3


Distribution of participants' general happiness with different work force status, adjusted for gender
keeping house
other
retired
school
temp not working
unempl, laid off
working fulltime
working parttime
Number of Observations
0
1000
2000
3000
4000
5000
1
2
3
General Happiness
as.factor(female)
0
1


Distribution of participants' general happiness with different total family income
$1000 to 2999
$10000 − 14999
$15000 − 19999
$20000 − 24999
$25000 or more
$3000 to 3999
$4000 to 4999
$5000 to 5999
$6000 to 6999
$7000 to 7999
$8000 to 9999
lt $1000
Number of Observations
0
2000
4000
1
2
3
General Happiness
as.factor(Happy)
1
2
3

Concerning our response variable "happy," we assigned numerical rankings of "1," "2," and "3" to its three categories—"not too happy," "pretty happy," and "very happy," respectively—to facilitate analysis and model fitting.

To create the predictor variable that records the number of household members under 17, which was not originally included in our dataset, we aggregated data from three existing variables: the number of babies, preteens, and teenagers in the household. We believe it's more logical to consider these age groups collectively since each requires financial support and additional care from older members of the household, given that they are not yet fully independent.

Given the N/As in the "income" variable and the "young_members" variable, we used dropped those data with N/As for regression.

We generated visualizations to explore the relationships between the predictor variables and the response variable. Since most of our variables, including the response variable, are categorical, we opted for bar plots for all our exploratory data analyses.

The first bar plot examines the distribution of observations across different levels of highest year of school completed in the sample, accounting for their overall happiness. The majority of participants in the sample have completed 12-16 years of education, roughly equivalent to middle school and a high school diploma. Across almost every education level, more individuals selected "pretty happy" than "very happy" or "not too happy," possibly influenced by a tendency to prioritize these responses in general.
The second plot analyzes the distribution of observations for males and females separately, considering their overall happiness. Overall, participants tend to choose "pretty happy" more

frequently than "very happy" or "not too happy," and there are more female participants in the dataset than male.
The third plot examines the distribution of observations across different work force statuses, factoring in gender and overall happiness. Most participants in the dataset are employed full-time, with only a small proportion in school, temporarily unemployed, or not working. More female participants are homemakers or work part-time compared to male participants, while gender distribution is relatively even across other work force statuses. The overall pattern of happiness preference remains consistent across genders.
The fourth plot illustrates the distribution of observations across different household income levels, considering overall happiness. The majority of participants have a household income above $10,000, especially those with incomes exceeding $25,000, who strongly favor "pretty happy" and "very happy" over "not too happy." The general pattern of happiness preference persists across all income levels.
The fifth plot displays the distribution of observations for participants with different numbers of young household members, adjusting for overall happiness. Most participants do not have anyone under 17 in their household, while those who do typically have one or two such members. The overall pattern of happiness preference remains consistent across these groups.

Finally, we decided on predicting whether people with higher socioeconomic status are happier with these predictors: work force status ("workstat"), total family income ("income"), highest year of school completed ("educ"), gender ("female"), and the number of household members under 17 ("young_members").

Methodology: Interaction Assessment and Model Selection

Among our predictors, we examined two possible interactions: the interaction between work force status and gender, and that between household income and the count of household members under 17. Given the societal expectation for women to bear children and the stereotype of women primarily responsible for childcare, we anticipate a stronger correlation between being female and working part-time. Additionally, we anticipate that households with higher incomes will generally be able to support more members under 17, necessitating greater economic assistance from other family members.

F test: (all coef associated with the interaction term are equal to 0)

Hypothesis test: H0: there is no interaction between working status and whether one is woman.
Ha: there is an interaction between working status and whether one is woman.

```
Call:
lm(formula = Happy ~ as.factor(workstat) + educ + young_members +
    as.factor(income) + female, data = Happiness)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.4469 -0.2806 -0.1431  0.6833  1.4799

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           1.8338864  0.0525265  34.914  < 2e-16 ***
as.factor(workstat)other             -0.2208637  0.0371206  -5.950 2.74e-09 ***
as.factor(workstat)retired            0.0445325  0.0217085   2.051  0.04025 *
as.factor(workstat)school            -0.0351526  0.0338081  -1.040  0.29846
as.factor(workstat)temp not working -0.1269610  0.0389063  -3.263  0.00110 **
as.factor(workstat)unempl, laid off -0.3151793  0.0338523  -9.310  < 2e-16 ***
as.factor(workstat)working fulltime -0.0717166  0.0180047  -3.983 6.83e-05 ***
as.factor(workstat)working parttime -0.0687764  0.0220187  -3.124  0.00179 **
educ                                  0.0130338  0.0018223   7.153 8.92e-13 ***
young_members                         0.0076983  0.0047902   1.607  0.10805
as.factor(income)$10000 - 14999       0.1021878  0.0487596   2.096  0.03612 *
as.factor(income)$15000 - 19999       0.1083219  0.0492250   2.201  0.02778 *
as.factor(income)$20000 - 24999       0.1486357  0.0488262   3.044  0.00234 **
as.factor(income)$25000 or more       0.3069013  0.0463891   6.616 3.82e-11 ***
as.factor(income)$3000 to 3999       -0.0397120  0.0673747  -0.589  0.55559
as.factor(income)$4000 to 4999        0.0084535  0.0668954   0.126  0.89944
as.factor(income)$5000 to 5999        0.1109332  0.0622629   1.782  0.07482 .
as.factor(income)$6000 to 6999       -0.0929505  0.0627793  -1.481  0.13874
as.factor(income)$7000 to 7999       -0.0472827  0.0615892  -0.768  0.44267
as.factor(income)$8000 to 9999       -0.0042732  0.0548287  -0.078  0.93788
as.factor(income)lt $1000             0.0410543  0.0646611   0.635  0.52549
female                                0.0009382  0.0105805   0.089  0.92934
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6082 on 14916 degrees of freedom
Multiple R-squared:  0.05373,   Adjusted R-squared:  0.05239
F-statistic: 40.33 on 21 and 14916 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Model 1: Happy ~ as.factor(workstat) + educ + young_members + as.factor(income) +
    female
Model 2: Happy ~ as.factor(workstat) + educ + young_members + as.factor(income) +
    female + as.factor(workstat) * female
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
```

```
1  14916 5516.7
2  14909 5504.6  7    12.071 4.6705 3.059e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value for interaction term between full-time working status and female is 0.0001, which is lower than a significance level of 0.05 and statistically significant. We have sufficient evidence to reject the null hypothesis, and thus the relationship between happiness and whether one is working full-time is depended on whether the respondent is female.

```
Analysis of Variance Table

Model 1: Happy ~ as.factor(workstat) + educ + young_members + as.factor(income) +
    female + as.factor(workstat) * female
Model 2: Happy ~ as.factor(workstat) + educ + young_members + as.factor(income) +
    female + +as.factor(workstat) * female + as.factor(income) *
    young_members
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1  14909 5504.6
2  14898 5496.9 11    7.7621 1.9125 0.03307 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The f statistic value 0.033 for interaction term between household income (dollars per year) and number of children (young_members) across all levels are lower than the significance level 0.05. We have sufficient evidence to reject the null hypothesis, and thus the relationship between score of general happiness and household income (dollars per year) is depended on the number of children in the household.

In our study, we contemplated employing both multinomial regression and ordinal regression models for the categorical response variable. Ultimately, we opted for the latter. The decision was straightforward, as the response variable "happy" exhibits an ordinal nature, with discernible gradations from "not too happy" to "pretty happy" to "very happy."

After assessing and confirming all the above conditions, we went into fitting the ordinal regression model.

Results

```
Attaching package: 'MASS'
```

```
The following object is masked from 'package:dplyr':

    select
```

as.factor(workstat)other	as.factor(workstat)retired
0.4792577	1.1654807
as.factor(workstat)school	as.factor(workstat)temp not working
0.8878100	0.6694166
as.factor(workstat)unempl, laid off	as.factor(workstat)working fulltime
0.3521781	0.7812348
as.factor(workstat)working parttime	educ
0.7936719	1.0439826
young_members	as.factor(income)$10000 - 14999
1.0238657	1.4000486
as.factor(income)$15000 - 19999	as.factor(income)$20000 - 24999
1.4184065	1.6400305
as.factor(income)$25000 or more	as.factor(income)$3000 to 3999
2.7095532	0.8642704
as.factor(income)$4000 to 4999	as.factor(income)$5000 to 5999
1.0092246	1.4493184
as.factor(income)$6000 to 6999	as.factor(income)$7000 to 7999
0.7228762	0.8458438
as.factor(income)$8000 to 9999	as.factor(income)lt $1000
0.9733683	1.1070729
female	
1.0077983	

Based on the outcomes derived from the regression model, we have formulated interpretations corresponding to each of the predictor variables.

workstat:
According to the ordinal regression model, a person who retired is predicted to have 1.213 times the odds of being in the next higher score of happiness category compared to a person who keeps house, while adjusting for years of education, household income, the number of children they have, and whether being a woman. In the work force status categories, people with all other status are predicted to have lower possibilities of being in the next higher score of happiness category compared to those who keep houses, while adjusting for years of education, household income, the number of children in their household, and whether being a woman.

educ:
A person who has one more year of education is predicted to have 1.05 times the odds of being in the next higher score of happiness category compared to a person who has less years of

education, while adjusting for work force status, household income, the number of children in their household, and whether being a woman.

young_members:
A person who has one more number of children in their household is predicted to have 1.021 times the odds of being in the next higher score of happiness category compared to a person who has less numbers of children in their household, while adjusting for work force status, household income, years of education, and whether being a woman.

female:
A person who is a woman is predicted to have 1.007 times the odds of being in the next higher score of happiness category compared to a person who is a man, while adjusting for years of education, household income, and number of children in their household.

income:
According to the ordinal regression model, a person in a household with $25000+ total income is predicted to have 2.709 times the odds of being in the next higher score of happiness category compared to a person in a household with $1000-2999 total income, while adjusting for years of education, work force status, the number of children they have, and whether being a woman. In the total family income categories, people with all other status are predicted to have lower possibilities of being in the next higher score of happiness category compared to those with $1000-2999 total income, which is the lowest of all, while adjusting for other variables.

Discussion

In summary, the overarching conclusion is that employed females with higher household incomes, advanced educational attainment, and a greater number of children at home tend to report higher levels of happiness. However, two unexpected findings emerged: firstly, that females exhibit greater happiness compared to males, contrary to the assumption that males typically possess higher socioeconomic status and thus should be happier. This discrepancy may be attributed to societal and economic pressures placed on males, who are traditionally viewed as primarily responsible for household financial well-being. Secondly, the positive correlation between the presence of young household members and happiness may be explained by the emotional support and warmth provided by children to adults.

Some limitations in this research include: the potentially problematic strategy on dropping the N/A data (as some missing data may be due to lower SES status), the sampling bias caused by self-reported happiness, potential missing interactions, and potential missing components related to SES status.

In addition to addressing the limitations in our research, future studies in this field can concentrate on investigating the underlying reasons behind the two unexpected findings, evaluating the validity of the explanations we have proposed, and exploring additional factors that may influence the relationship between socioeconomic status and happiness.