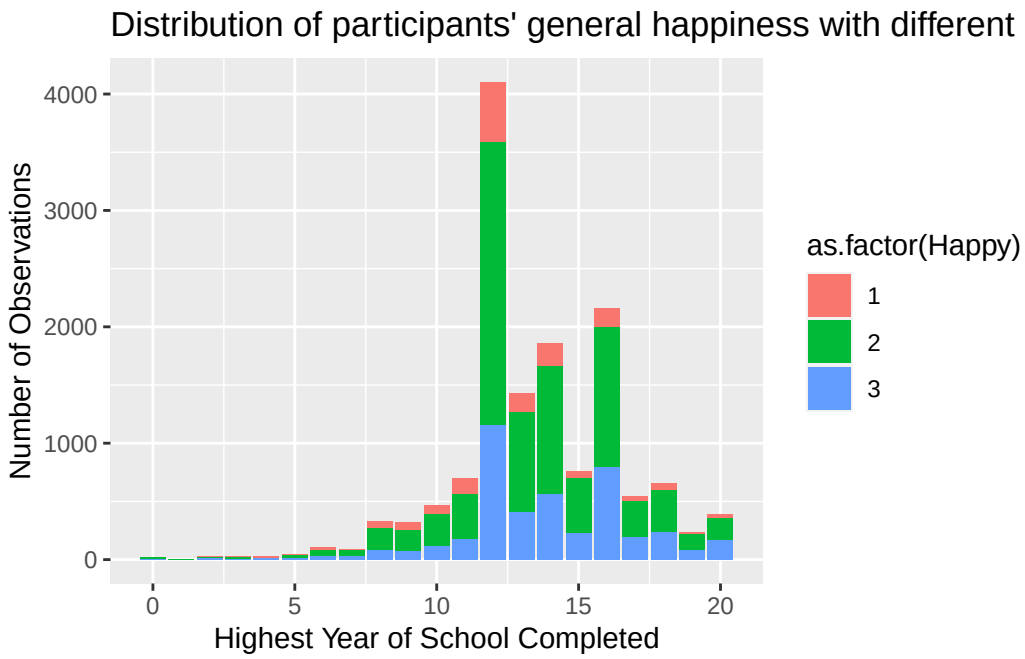
Project

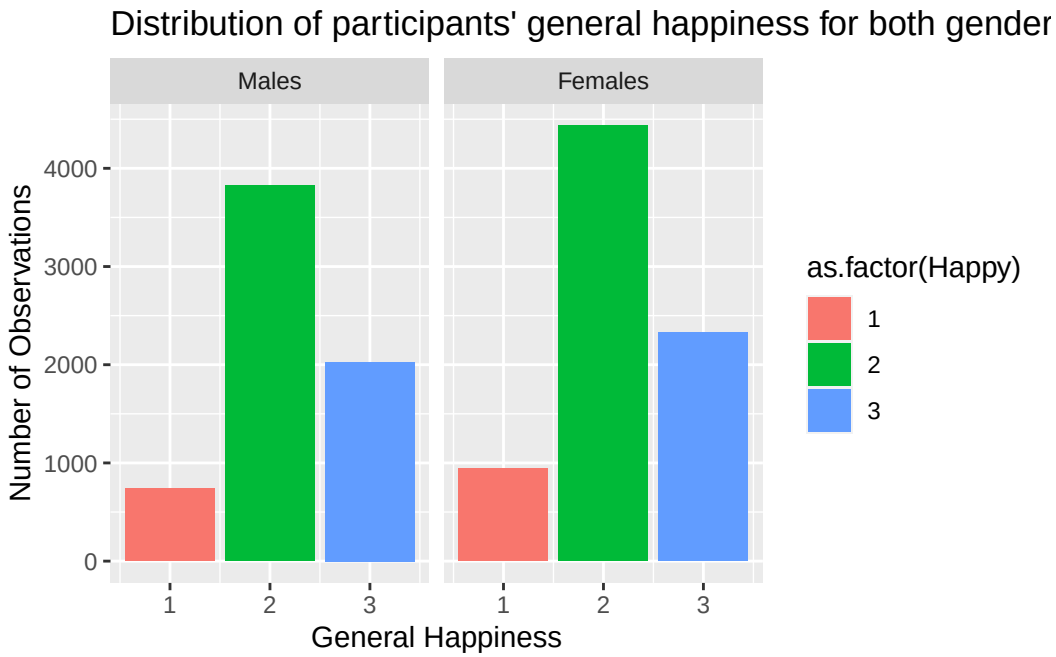
Tina Qian, Yanlin Li

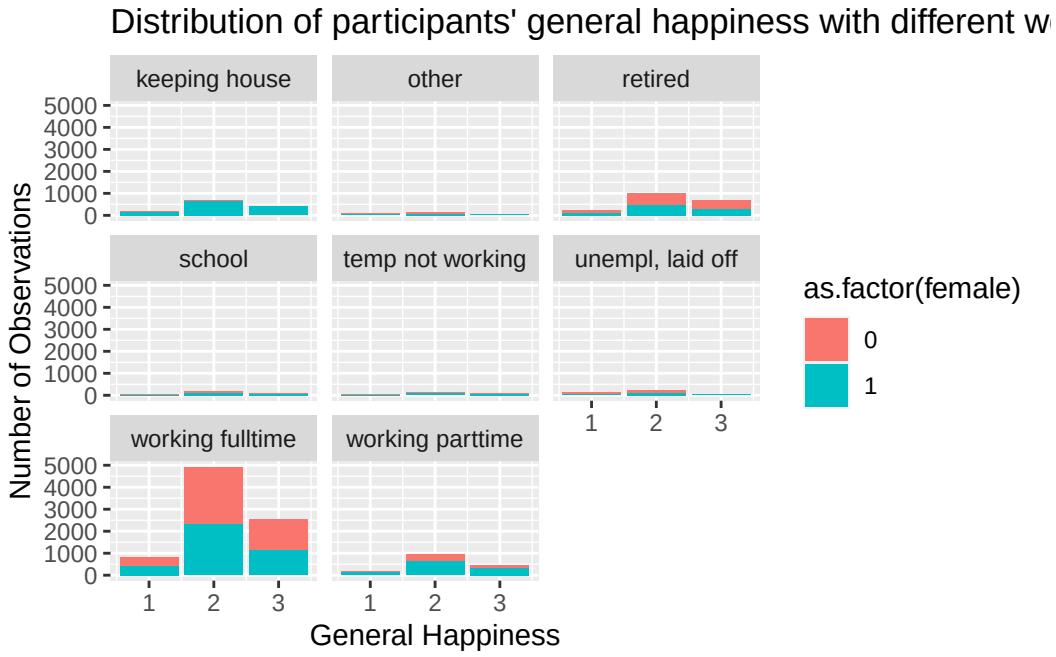
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
-- Attaching packages --------------------------------------- tidymodels 1.1.1 --

v broom        1.0.5     v rsample      1.2.0
v dials        1.2.0     v tune         1.1.2
v infer        1.0.5     v workflows    1.1.3
v modeldata    1.2.0     v workflowsets 1.0.1
v parsnip      1.1.1     v yardstick    1.2.0
v recipes      1.0.8

-- Conflicts ------------------------------------------ tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Learn how to get started at https://www.tidymodels.org/start/

Rows: 17137 Columns: 34
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (9): workstat, divorce, widowed, reg16, income, region, attend, happy, ...
dbl (25): rownames, year, prestige, educ, babies, preteen, teens, tvhours, v...
```

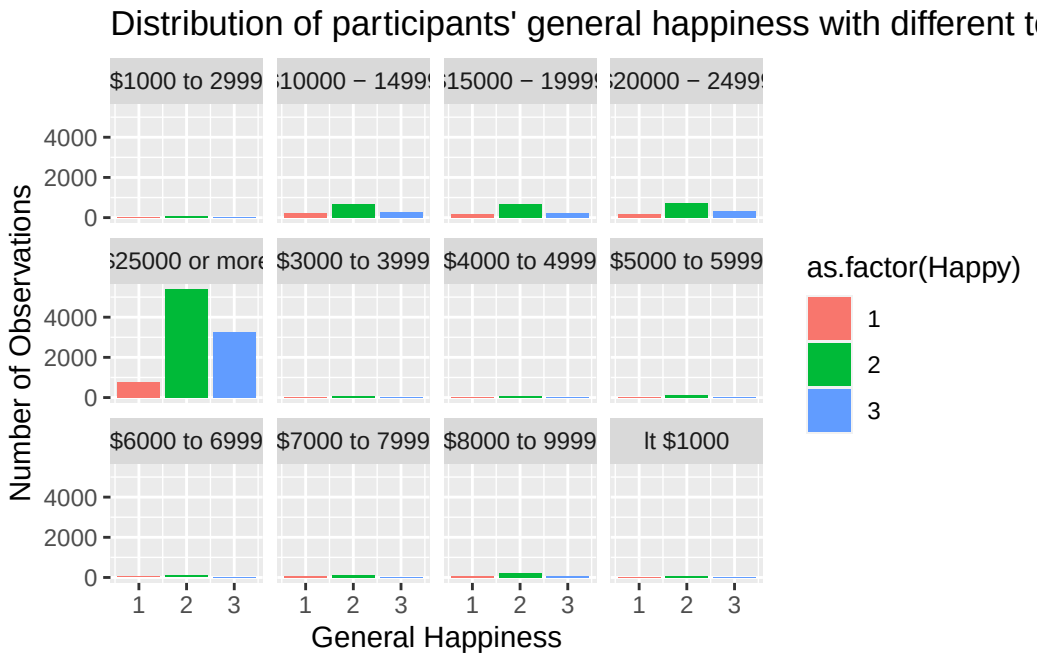
```
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Exploratory Analysis




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


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



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.

Since we focus on the effect of socioeconomic status on happiness in our research question,

we selected specific predictor variables: 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").

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 Multiple Imputation via Chained Equations (MICE) to perform multiple imputation to the data. We assume MAR is satisfied from our visual examination before.***

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.

Testing Interaction & Assumptions

Among these 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.

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 + as.factor(workstat) * female,
    data = Happiness)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4713 -0.2906 -0.1498  0.6804  1.4855

Coefficients:
                                          Estimate Std. Error t value
(Intercept)                               1.750413   0.089412  19.577
as.factor(workstat)other                 -0.154164   0.087481  -1.762
as.factor(workstat)retired                0.158361   0.073248   2.162
as.factor(workstat)school                 0.029637   0.089693   0.330
as.factor(workstat)temp not working      -0.080136   0.088377  -0.907
as.factor(workstat)unempl, laid off      -0.285579   0.081197  -3.517
as.factor(workstat)working fulltime       0.033895   0.071246   0.476
as.factor(workstat)working parttime      -0.067877   0.075716  -0.896
educ                                      0.013611   0.001862   7.310
young_members                             0.005022   0.004939   1.017
as.factor(income)$10000 - 14999           0.087776   0.054148   1.621
as.factor(income)$15000 - 19999           0.099341   0.054472   1.824
as.factor(income)$20000 - 24999           0.136333   0.054099   2.520
as.factor(income)$25000 or more           0.290342   0.051832   5.602
as.factor(income)$3000 to 3999           -0.015361   0.072947  -0.211
as.factor(income)$4000 to 4999           -0.021025   0.073850  -0.285
as.factor(income)$5000 to 5999            0.096634   0.068795   1.405
as.factor(income)$6000 to 6999           -0.081748   0.069326  -1.179
```

```
as.factor(income)$7000 to 7999             -0.047267   0.067359  -0.702
as.factor(income)$8000 to 9999              0.011261   0.060286   0.187
as.factor(income)lt $1000                   0.040015   0.071964   0.556
female                                      0.095102   0.072647   1.309
as.factor(workstat)other:female            -0.075533   0.102128  -0.740
as.factor(workstat)retired:female          -0.124099   0.077903  -1.593
as.factor(workstat)school:female           -0.021063   0.102659  -0.205
as.factor(workstat)temp not working:female -0.018963   0.101667  -0.187
as.factor(workstat)unempl, laid off:female  0.026220   0.094152   0.278
as.factor(workstat)working fulltime:female -0.124118   0.073889  -1.680
as.factor(workstat)working parttime:female  0.035318   0.079843   0.442
                                           Pr(>|t|)
(Intercept)                                 < 2e-16 ***
as.factor(workstat)other                   0.078048 .
as.factor(workstat)retired                 0.030637 *
as.factor(workstat)school                  0.741085
as.factor(workstat)temp not working        0.364550
as.factor(workstat)unempl, laid off        0.000438 ***
as.factor(workstat)working fulltime        0.634257
as.factor(workstat)working parttime        0.370020
educ                                       2.81e-13 ***
young_members                              0.309249
as.factor(income)$10000 - 14999            0.105030
as.factor(income)$15000 - 19999            0.068217 .
as.factor(income)$20000 - 24999            0.011744 *
as.factor(income)$25000 or more            2.16e-08 ***
as.factor(income)$3000 to 3999             0.833222
as.factor(income)$4000 to 4999             0.775875
as.factor(income)$5000 to 5999             0.160146
as.factor(income)$6000 to 6999             0.238342
as.factor(income)$7000 to 7999             0.482873
as.factor(income)$8000 to 9999             0.851823
as.factor(income)lt $1000                  0.578198
female                                     0.190521
as.factor(workstat)other:female            0.459562
as.factor(workstat)retired:female          0.111184
as.factor(workstat)school:female           0.837438
as.factor(workstat)temp not working:female 0.852037
as.factor(workstat)unempl, laid off:female 0.780643
as.factor(workstat)working fulltime:female 0.093020 .
as.factor(workstat)working parttime:female 0.658248
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.6062 on 14281 degrees of freedom
Multiple R-squared:  0.05366,   Adjusted R-squared:  0.05181
F-statistic: 28.92 on 28 and 14281 DF,  p-value: < 2.2e-16
```

interpretations:

The p value for interaction term between full-time working status and female is 0.09, which is higher than a significance level of 0.05 and not statistically significant. We do not have sufficient evidence to reject the null hypothesis, and thus the relationship between happiness and whether one is working full-time is not depended on whether the respondent is female.

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

Residuals:
    Min      1Q  Median      3Q     Max
-1.4508 -0.2844 -0.1569  0.6804  1.4327

Coefficients:
                                          Estimate Std. Error t value
(Intercept)                               1.876728   0.066522  28.212
as.factor(workstat)other                 -0.235609   0.040290  -5.848
as.factor(workstat)retired                0.053637   0.023173   2.315
as.factor(workstat)school                -0.015969   0.040208  -0.397
as.factor(workstat)temp not working      -0.130230   0.039535  -3.294
as.factor(workstat)unempl, laid off      -0.317619   0.034817  -9.123
as.factor(workstat)working fulltime      -0.070377   0.019364  -3.634
as.factor(workstat)working parttime      -0.068891   0.023174  -2.973
educ                                      0.013541   0.001863   7.270
young_members                            -0.048317   0.046285  -1.044
as.factor(income)$10000 - 14999           0.051439   0.063748   0.807
as.factor(income)$15000 - 19999           0.058373   0.064287   0.908
as.factor(income)$20000 - 24999           0.122706   0.063780   1.924
as.factor(income)$25000 or more           0.247903   0.061184   4.052
as.factor(income)$3000 to 3999           -0.062579   0.086003  -0.728
as.factor(income)$4000 to 4999           -0.011663   0.085663  -0.136
as.factor(income)$5000 to 5999            0.079082   0.079451   0.995
as.factor(income)$6000 to 6999           -0.073860   0.079500  -0.929
as.factor(income)$7000 to 7999           -0.065636   0.076536  -0.858
```

```
as.factor(income)$8000 to 9999                -0.035514   0.070139  -0.506
as.factor(income)lt $1000                     -0.013376   0.083586  -0.160
female                                         0.001681   0.010709   0.157
young_members:as.factor(income)$10000 - 14999  0.047881   0.049286   0.971
young_members:as.factor(income)$15000 - 19999  0.056663   0.049981   1.134
young_members:as.factor(income)$20000 - 24999  0.016747   0.048923   0.342
young_members:as.factor(income)$25000 or more  0.064563   0.046625   1.385
young_members:as.factor(income)$3000 to 3999   0.068412   0.070767   0.967
young_members:as.factor(income)$4000 to 4999  -0.029005   0.067653  -0.429
young_members:as.factor(income)$5000 to 5999   0.022889   0.066951   0.342
young_members:as.factor(income)$6000 to 6999  -0.052616   0.065432  -0.804
young_members:as.factor(income)$7000 to 7999  -0.024965   0.079233  -0.315
young_members:as.factor(income)$8000 to 9999   0.073173   0.056613   1.293
young_members:as.factor(income)lt $1000        0.072501   0.071885   1.009
                                              Pr(>|t|)
(Intercept)                                    < 2e-16 ***
as.factor(workstat)other                      5.09e-09 ***
as.factor(workstat)retired                     0.02064 *
as.factor(workstat)school                      0.69125
as.factor(workstat)temp not working            0.00099 ***
as.factor(workstat)unempl, laid off            < 2e-16 ***
as.factor(workstat)working fulltime            0.00028 ***
as.factor(workstat)working parttime            0.00296 **
educ                                          3.79e-13 ***
young_members                                  0.29655
as.factor(income)$10000 - 14999                0.41973
as.factor(income)$15000 - 19999                0.36389
as.factor(income)$20000 - 24999                0.05439 .
as.factor(income)$25000 or more               5.11e-05 ***
as.factor(income)$3000 to 3999                 0.46685
as.factor(income)$4000 to 4999                 0.89171
as.factor(income)$5000 to 5999                 0.31958
as.factor(income)$6000 to 6999                 0.35288
as.factor(income)$7000 to 7999                 0.39114
as.factor(income)$8000 to 9999                 0.61263
as.factor(income)lt $1000                      0.87286
female                                         0.87524
young_members:as.factor(income)$10000 - 14999  0.33133
young_members:as.factor(income)$15000 - 19999  0.25694
young_members:as.factor(income)$20000 - 24999  0.73212
young_members:as.factor(income)$25000 or more  0.16615
young_members:as.factor(income)$3000 to 3999   0.33370
young_members:as.factor(income)$4000 to 4999   0.66812
```

```
young_members:as.factor(income)$5000 to 5999   0.73244
young_members:as.factor(income)$6000 to 6999   0.42134
young_members:as.factor(income)$7000 to 7999   0.75271
young_members:as.factor(income)$8000 to 9999   0.19620
young_members:as.factor(income)lt $1000        0.31320
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6065 on 14277 degrees of freedom
Multiple R-squared:  0.05312,   Adjusted R-squared:  0.05099
F-statistic: 25.03 on 32 and 14277 DF,  p-value: < 2.2e-16
```

The p value for interaction term between household income and number of children across all levels are higher than 0.05, which is not statistically significant. We do not have sufficient evidence to reject the null hypothesis, and thus the relationship between score of general happiness and household income is not depended on the number of children in the household.

Ordinal Regression

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.