MATH 254 - Predicting Marriage Happiness

Sophie Maniscalco Rachel O'Brien Tural Sadigov

October 17, 2022

Due to the high rate of divorce in the United States, a topic of interest for many investigators is relationship happiness in marriages. This project aimed to construct a model predicting relationship happiness using numerous relationship variables. Permutation tests were used to select variables that were significantly correlated to relationship happiness – results indicated that only three out of five potential relationship variables were significantly related to relationship happiness. Seven linear models were constructed using the three variables and the best model based on the training-validation slipt and selecting the maximized r-squared value was selected to predict relationship happiness – specifically, an interaction between hours of sleep per night and argument frequency best predicted relationship happiness. Hours of sleep appeared to serve as a protective factor for couples with higher argument frequency, effectively reducing the negative effect of argument frequency on relationship happiness.

Table of contents

1 Background and Significance **2**

2 Methods **3**

3 Results **5**

3.1 Descriptive statistics 5

3.1.1 Descriptive statistics for relationship happiness, our response variable. . 5

3.1.2 Descriptive statistics for years married. 6

3.1.3 Descriptive statistics for cohabitation before marriage. 6

3.1.4 Descriptive Statistics for hours of sleep. 7

3.1.5 Descriptive Statistics for argument frequency. 8

3.1.6 Descriptive Statistics for days in the past three months that spouse consumed more than 4 drinks 8

3.2 Hypothesis Tests 9
3.2.1 Permutation test for relation between years married and relationship happiness. 9
3.2.2 Preliminary result of relationship between months cohabiting before marriage and relationship happiness. 11
3.2.3 Permutation test for relation between cohabitation before marriage and relationship happiness. 12
3.2.4 Permutation Test for hours of sleep per night and relationship happiness. 14
3.2.5 Permutation test for independence between days partner consumed more than 4 drinks in the last 90 days and relationship happiness: 17
3.2.6 Permutation test for independence between argument frequency and relationship happiness: 20
3.2.7 New variable ARGUEFRQ1 23
3.3 Linear Model 24

4 Discussion/Conclusions 27

5 Appendix 28
5.1 Loading Libraries and Data File. 28
5.2 Visualizing cohabitation before marriage. 28
5.3 Visualizing months cohabiting before marriage. 29
5.4 Visualizing years married. 30
5.5 Visualizing stress levels. 31
5.6 Visualizing relationship happiness, our response variable. 32
5.7 Visualizing relationship warmness. 33
5.8 Visualizing how critical one's spouse is. 34
5.9 Visualizing argument frequency. 35
5.10 Visualizing depression symptoms. 36
5.11 Visualizing days in which spouse has had more than four drinks in one day in the last three months 37
5.12 Visualizing hours of sleep. 38

References 39

1 Background and Significance

In 2019, there were 2,015,603 marriages in the United States ("Marriage and Divorce." 2022). This means that marriage is a crucial part of life for a lot of American people. According to marriage.com, marriage "acts as a social and legal contract that gives a partner someone to rely on, brings a greater degree of intimacy and emotional security" Smith (2022). At least that this is the ideal goal of marriage. However, also according to the CDC, in 2019 there were 746,971 divorces in the United States, so this idealistic goal of marriage is not reached in every relationship.

Given this grim outlook on marriage, we were interested as to what components of a marriage make it a successful one. Previous studies have shown that cohabitation before marriage is bad for marriage, with studies finding on average that couples that cohabited before marriage had a 33% higher rate of divorce than couples that did not cohabit before marriage Fox (2014). In addition, marriages where at least one member is depressed are less positive than marriages where neither member is depressed Johnson and Jacob (1997). Additionally, when looking at sleep and marriage, The American Academy of Sleep Medicine states that "being stably married or gaining a partner is associated with better sleep in women than being unmarried or losing a partner" Wagner (2009). With so many factors that can affect a relationship, how can we predict happiness in a marriage?

Our main analysis looks at how eleven different variables influence relationship happiness. We hypothesize that years married, cohabiting before marriage, hours of sleep a night, days binge drinking, and argument frequency all impact the happiness of a marriage. Our goal is to see if it is possible to predict happiness in a marriage through a combination of these variables.

2 Methods

a. *Data collection.* The data, which was collected between 2014 and 2015 only includes couples that "were legally married and had been living together for a minimum of three years at the time of the study." The sample was collected using different methods. About 70% of the same-sex couples in the sample were selected from the Massachusetts Registry of Vital Records and participants received invitations to participate in the mail. About 40% of the different-sex couples in the sample were selected from public records and then asked to participate. The rest of the participants were recruited through referrals. After all that, the couples were then asked to refer any other couples that met the study's requirements. The data itself was collected through a series of 10-day surveys with questions about their relationship which were referred to as diaries. The couples were asked to complete these questionnaires separately. "90% of participants completed all 10 days". All of this information was provided by icpsr.umich.edu under Health and Relationships Project, United States, 2014-2015 (ICPSR 37404).

b. *Variable creation.*There are twelve variables we looked at in our analysis. RELHAPPY, which is our primary variable of interest, is a discrete numerical variable with range 1 to 7 assessing the respondent's degree of happiness in the relationship, with higher scores indicating greater relationship happiness. COHB4MAR is a normal categorical variable indicating whether or not a couple cohabitated before marriage. MNCOHB4M is a continuous numerical variable assessing the number of months that the couple spent cohabiting before marriage if they did cohabit before marriage, with range 1 to 180. YRSMAR is a continuous numerical variable assessing how many years a couple has been married for, with range 0 to 13 years. RAGE is a continuous numerical variable with range 35 to 65 indicating the respondent's age at the time of the survey. STRSCALE is a discrete numerical variable with range 9 to 41 indicating how stressful the marriage is; it is an aggregate measure of multiple variables that were calculated by adding up the scores (from 1 to 7) of the other items. RELREWAR is a discrete numerical variable with range 1 to 6 representing how rewarding the relationship is, with higher scores indicating greater feelings that the relationship with their spouse is rewarding. SPCRITIC is a discrete numerical variable with range 1 to 5 representing how critical their spouse is, with higher scores indicating a more critical spouse. ARGUEFRQ is a discrete numerical variable with range 1 to 5 assessing how often the respondent argues with their spouse, with higher scores indicating a higher frequency of arguing. DEPSYM is a discrete numerical variable with range 11 to 38 assessing depressive symptoms in the respondent; this is an aggregate measure calculated by adding up the positive variable items in a reverse-coded fashion. CDDAYSBD is a discrete numerical variable with range 0 to 90 indicating how many days out of the last three months that the respondent's spouse had four or more drinks. Finally, HRSLEEP is a continuous numerical variable with range 2 to 10, indicating how many hours of sleep the respondent gets on an average night. All variables were taken from Health and Relationships Project, United States, 2014-2015 and, unless otherwise stated, were direct measures.

RELHAPPY1 was created by us in order to better use the data provided. Originally the data was in string format and we converted it into numeric. We extracted the numbers provided in the original strings to create RELHAPPY1.

ARGUEFRQ1 was created by us in order to better use the data provided. Originally the data was in string format and we converted it into numeric. We extracted the numbers provided in the original strings to create ARGUFRQ1.

c. *Analytic Methods.* A bar graph was used to illustrate the distribution of respondents who cohabited and did not cohabit before marriage. A histogram was used to illustrate the distribution of daily hours of sleep received by respondents at night. Five initial permutation tests were conducted to assess the relationship between the outcome variable, relationship happiness, and the potential predictor variables, years married, nightly hours of sleep, number of days in a three-month period where the partner had more than three drinks, and argument frequency. Two new variables were constructed for relationship happiness and argument frequency by extracting the numerical value from the string data. The relationship between relationship happiness and cohabitation before marriage was assessed using side-by-side boxplots.

Once we had initial results for our five variables of interest in relation to relationship happiness, we chose three variables to do linear regression models with. We chose these variables based on if they were significantly correlated to relationship happiness or not. Through our hypothesis testing, we found that two variables, years married and argument frequency had a significant non-zero correlation coefficient with relationship happiness. We also chose a third variable, hours of sleep per night to use in our modeling because it had the lowest p-value of the variables with a non-significant relationship with relationship happiness. Then we chose 7 different possible combinations of these variables to test to see if a linear regression would be a good model for our data which were, argument frequency, argument frequency added to years married, argument frequency added to years married, and hours of sleep per night, argument frequency multiplied by hours of sleep per night, square root of years married, a degree two polynomial of argument frequency, and a degree three polynomial of argument frequency. We split our data into 3 subsections: testing data, training data, and validation data. Then we used our training data subgroup to create linear regressions based on our 7 models. Finally, we selected our best model based on the maximized r-squared value for the validation data. We ended up selecting our fourth model which was the model that contained interaction between hours of sleep and argument frequency. Then we used our testing data to test this model.

3 Results

In the results sections we used libraries from the following sources: Couch et al. (2021), Henry and Wickham (2022), Ushey et al. (2022), and Wickham et al. (2019).

~~~
```{r}
#| message: false
library(tidyverse)
library(infer)
da37404.0001 <- read_csv('https://raw.githubusercontent.com/turalsadigov/MATH_254/main/Datase
#load(file = '37404-0001-Data.rda')
```
~~~

3.1 Descriptive statistics

3.1.1 Descriptive statistics for relationship happiness, our response variable.

Relationship happiness - Descriptive statistics First, we dropped 7 NAs in the data for a total of 831 observations. From the resulting subset, the mean relationship happiness was 5.19, with range 1 to 7.

~~~
```{r}
data <- da37404.0001 %>%
  mutate(RELHAPPY1 = str_sub(RELHAPPY, 2, 2)) %>%
  mutate(RELHAPPY1 = as.numeric(RELHAPPY))

data %>%
  select(RELHAPPY1) %>%
  drop_na() %>%
  summarise(mean = mean(RELHAPPY1),
            min = min(RELHAPPY1),
~~~

~~~
            max = max(RELHAPPY1))
```
~~~

```
      mean min max
1 5.190132   1   7
```

3.1.2 Descriptive statistics for years married.

Years Married - Descriptive statistics No NAs were dropped. The mean number of years married was 8.69, with range from 0 to 42.5.

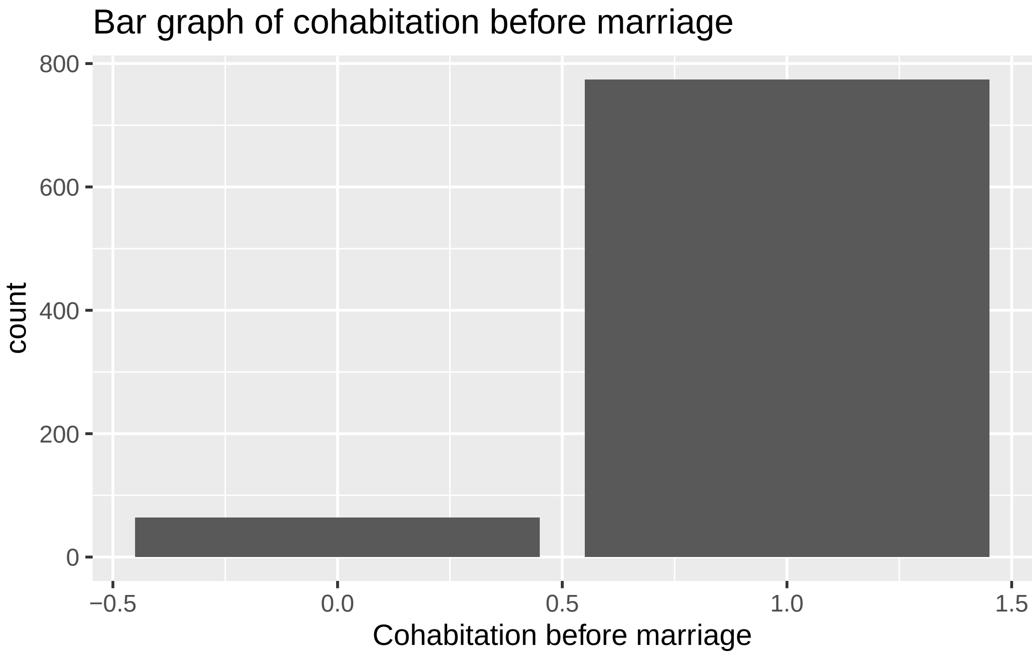
~~~
```{r}
data %>%
  select(YRSMAR) %>%
  summarise(mean = mean(YRSMAR), min = min(YRSMAR), max = max(YRSMAR))
```
~~~

```
      mean min  max
1 8.690036   0 42.5
```

3.1.3 Descriptive statistics for cohabitation before marriage.

Cohabitation before marriage - Descriptive statistics No NAs were dropped. The majority of respondents cohabited before marriage, as depicted in the bar graph below.

~~~
```{r}
data %>%
  select(COHB4MAR) %>%
  mutate(COHB4MAR1 = str_sub(COHB4MAR, 2, 2)) %>%
  mutate(COHB4MAR1 = as.numeric(COHB4MAR1)) %>%
  ggplot(aes(COHB4MAR1)) +
  geom_bar() +
  ggtitle('Bar graph of cohabitation before marriage') +
  xlab('Cohabitation before marriage')
```
~~~

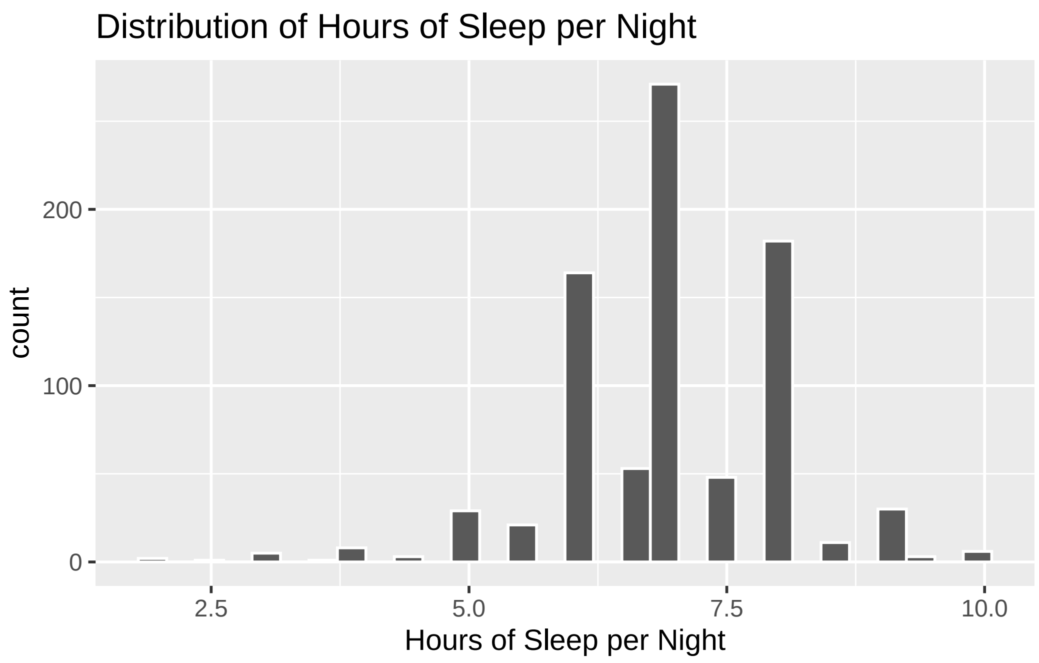


~~~
```{r}
data %>%
  select(HRSLEEP) %>%
  summarize(mean = mean(HRSLEEP),
            minimum = min(HRSLEEP),
            maximum = max(HRSLEEP))
```
~~~

```
      mean minimum maximum
1 6.950477       2      10
```

3.1.4 Descriptive Statistics for hours of sleep.

~~~
```{r}
data %>%
  select(HRSLEEP) %>%
  ggplot(aes(x = HRSLEEP)) +
  geom_histogram(col = 'white') +
  ggtitle('Distribution of Hours of Sleep per Night') +
  xlab('Hours of Sleep per Night')
```
~~~



3.1.5 Descriptive Statistics for argument frequency.

~~~
```{r}
data <- data %>%
  mutate(ARGUEFRQ1 = str_sub(ARGUEFRQ, 2, 2)) %>%
  mutate(ARGUEFRQ1 = as.numeric(ARGUEFRQ))

data %>%
  summarise(mean = mean(ARGUEFRQ1),
            minimum = min(ARGUEFRQ1),
            maximum = max(ARGUEFRQ1))
```
~~~

```
      mean minimum maximum
1 2.353222       1       5
```

3.1.6 Descriptive Statistics for days in the past three months that spouse consumed more than 4 drinks

~~~
```{r}
data %>%
~~~

~~~r
  select(CDDAYSBD) %>%
  drop_na() %>%
  summarize(mean = mean(CDDAYSBD),
            minimum = min(CDDAYSBD),
            maximum = max(CDDAYSBD))

data %>%
  select(CDDAYSBD) %>%
  drop_na() %>%
  count()
```
~~~

```
      mean minimum maximum
1 4.223058       0      90
    n
1 798
```

3.2 Hypothesis Tests

For the hypothesis tests, we will assess using a significance level of 0.05. Given that we are doing multiple tests and models, we will divide the significance level by the number of tests and coefficients in our models to avoid p-hacking. So, for each individual test, the significance level is 0.0025.

3.2.1 Permutation test for relation between years married and relationship happiness.

H_o: Relationship happiness is not related to years married; the correlation coefficient is 0.

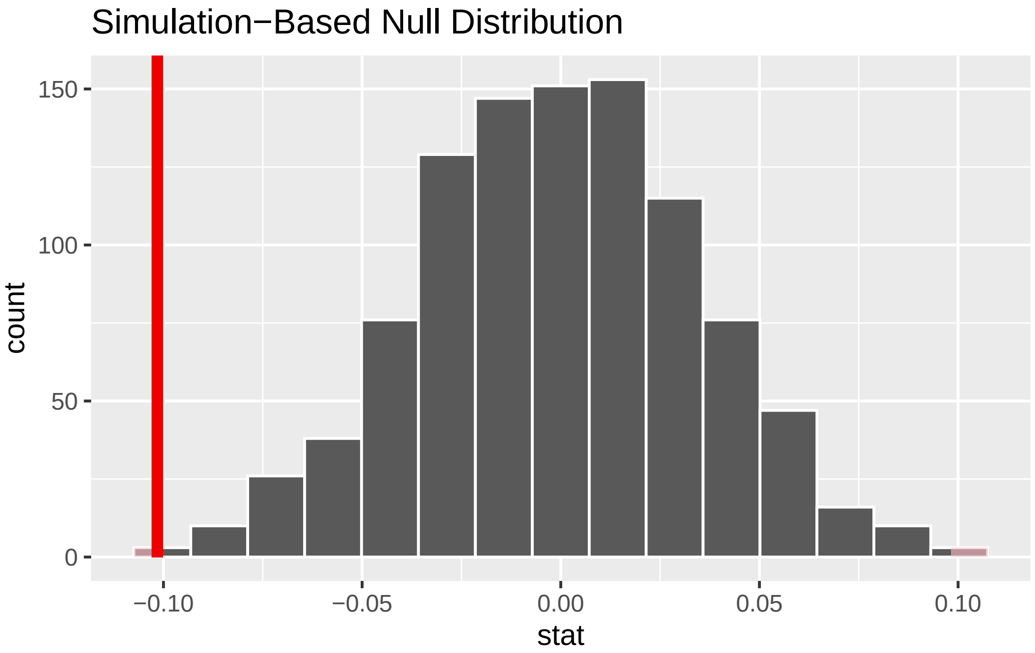
H_a : Relationship happiness is negatively related to years married: the correlation coefficient is negative.

Years married vs. relationship happiness from a sub-sample of 831 respondents (after NAs were dropped), the data revealed a significant negative correlation, based on our p-value, between years married and relationship happiness, such that as years married increases, relationship happiness decreases.

~~~r
```{r}
correlation_hat <-
  data %>%
  specify(RELHAPPY1 ~ YRSMAR) %>%
  calculate(stat = 'correlation')
~~~

~~~
correlation_hat

set.seed(2023)
null_dist <-
  data %>%
  specify(RELHAPPY1 ~ YRSMAR) %>%
  hypothesize(null = 'independence') %>%
  generate(reps = 1000, type = 'permute') %>%
  calculate(stat = 'correlation')

null_dist %>%
  visualize() +
  shade_p_value(obs_stat = correlation_hat,
                direction = "two-sided")

null_dist %>%
  get_p_value(obs_stat = correlation_hat,
              direction = 'two-sided')
```
~~~

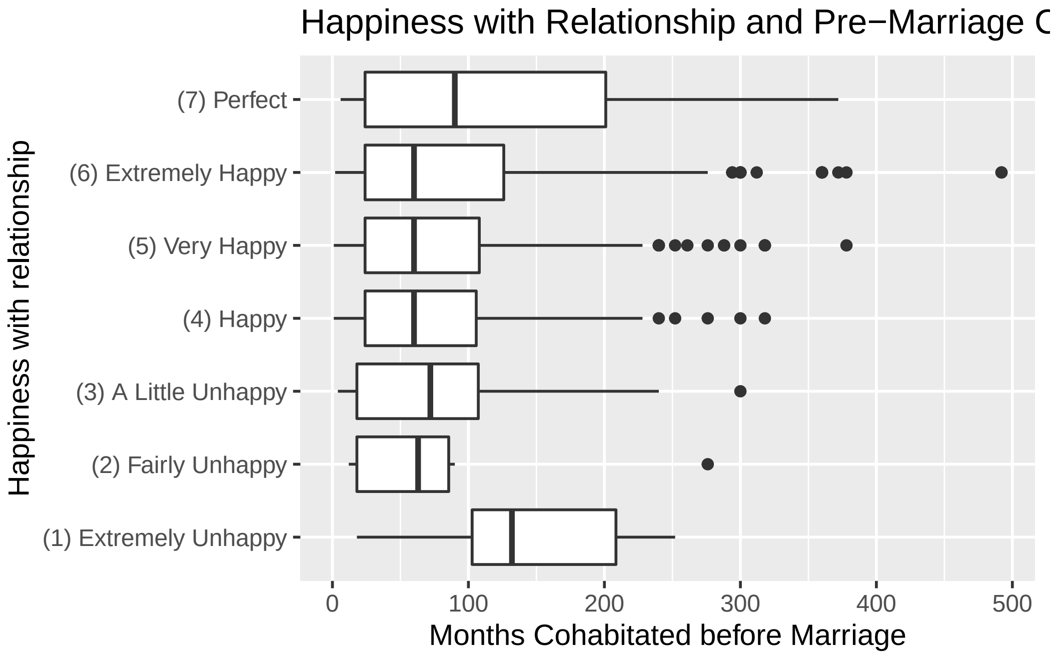
```
Response: RELHAPPY1 (numeric)
Explanatory: YRSMAR (numeric)
# A tibble: 1 x 1
    stat
   <dbl>
1 -0.102
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.002
```



3.2.2 Preliminary result of relationship between months cohabiting before marriage and relationship happiness.

These boxplots provide preliminary results that the tail ends of happiness (i.e., perfect relationship happiness and extreme unhappiness) correspond to longer cohabitation periods before marriage, as indicated by the IQR of the boxplots.

~~~
```{r}
da37404.0001 %>%
  select(MNCOHB4M, RELHAPPY) %>%
  drop_na() %>%
  ggplot(aes(x = MNCOHB4M, y = RELHAPPY)) +
  geom_boxplot() +
  ggtitle('Happiness with Relationship and Pre-Marriage Cohabitation') +
  xlab('Months Cohabitated before Marriage') +
  ylab("Happiness with relationship")
```
~~~



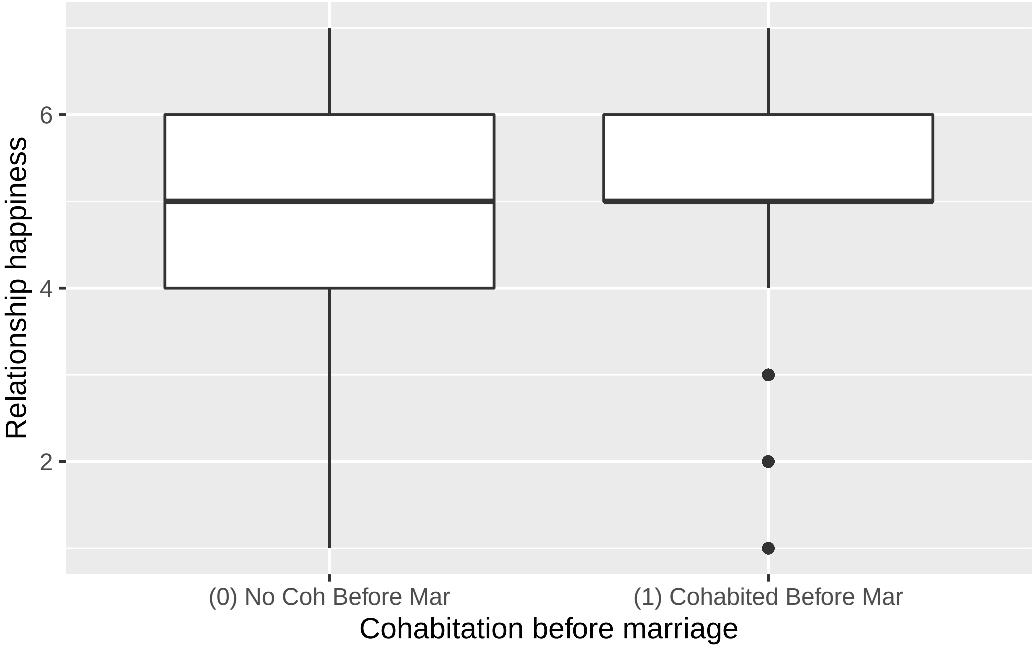
3.2.3 Permutation test for relation between cohabitation before marriage and relationship happiness.

H_o: Relationship happiness is not related to cohabitation before marriage; the correlation coefficient is 0.

H_a : Relationship happiness is negatively related to cohabitation before marriage; the correlation coefficient is negative.

Cohabitation before marriage vs. relationship happiness from a sub-sample of 831 respondents (after NAs were dropped). The boxplot for respondents that cohabited before marriage has a smaller inter-quartile range than the boxplot for respondents who did not cohabit before marriage, indicating that couples that cohabited before marriage are generally happier with their relationship. This finding is contrary to previous findings that suggest cohabiting before marriage is bad for the relationship. However, follow-up hypothesis testing using a permutation test for difference in means revealed that there was a non-significant difference in relationship happiness between couples that cohabited and those that did not cohabit before marriage.

~~~
```{r}
data %>%
  select(RELHAPPY1, COHB4MAR) %>%
  drop_na() %>%
  ggplot(aes(COHB4MAR, RELHAPPY1)) +
~~~

```
  geom_boxplot() +
  xlab('Cohabitation before marriage') +
  ylab('Relationship happiness')
```
```


6
4
2
Relationship happiness
(0) No Coh Before Mar
(1) Cohabited Before Mar
Cohabitation before marriage

```
```{r}
obs_stat <-
  data %>%
  specify(RELHAPPY1 ~ COHB4MAR) %>%
  calculate(stat = 'diff in means')
obs_stat

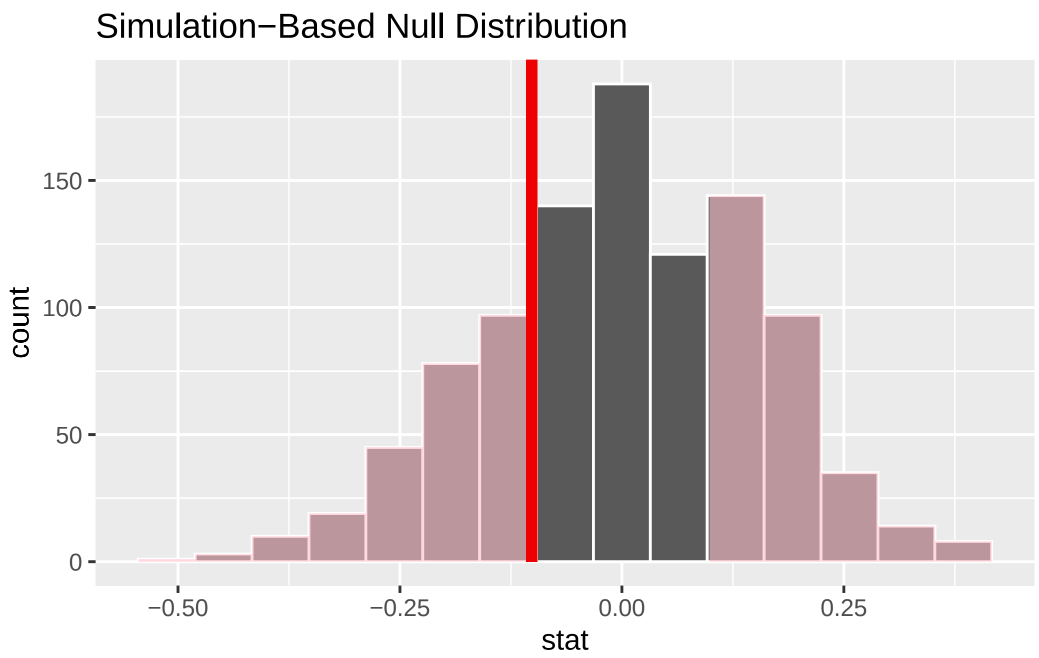
set.seed(2023)
null_dist <-
  data %>%
  specify(RELHAPPY1 ~ COHB4MAR) %>%
  hypothesize(null = 'independence') %>%
  generate(reps = 1000, type = 'permute') %>%
  calculate(stat = 'diff in means')

null_dist %>%
  visualize() +
  shade_p_value(obs_stat = correlation_hat,
```

~~~~
                direction = "two-sided")

null_dist %>%
  get_p_value(obs_stat = correlation_hat,
              direction = 'two-sided')
```
~~~~

```
Response: RELHAPPY1 (numeric)
Explanatory: COHB4MAR (factor)
# A tibble: 1 x 1
    stat
   <dbl>
1 -0.375
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.506
```

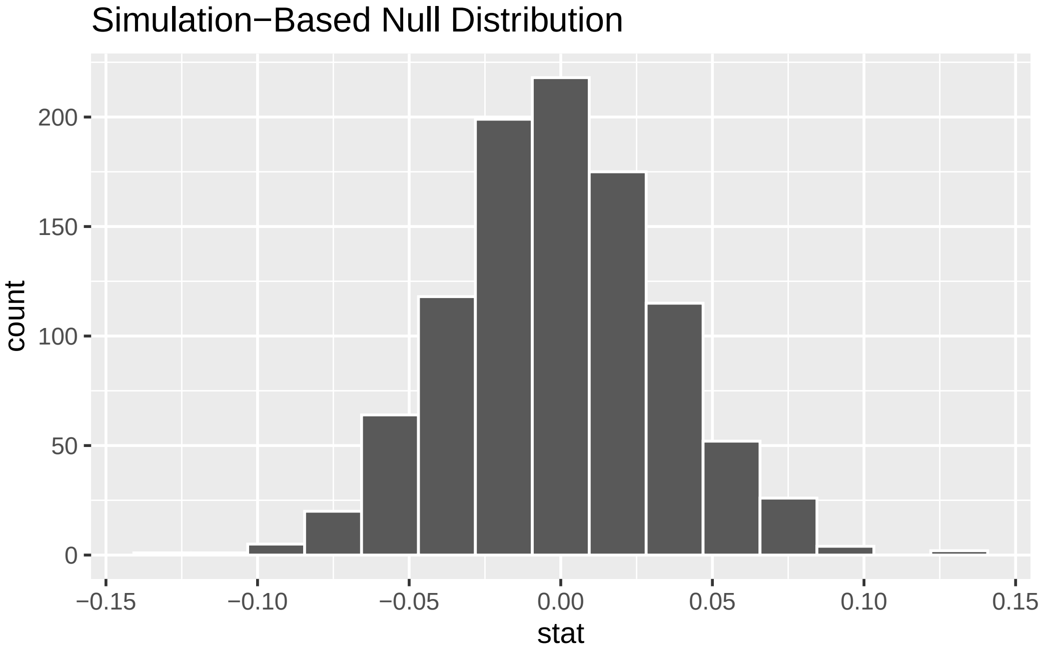
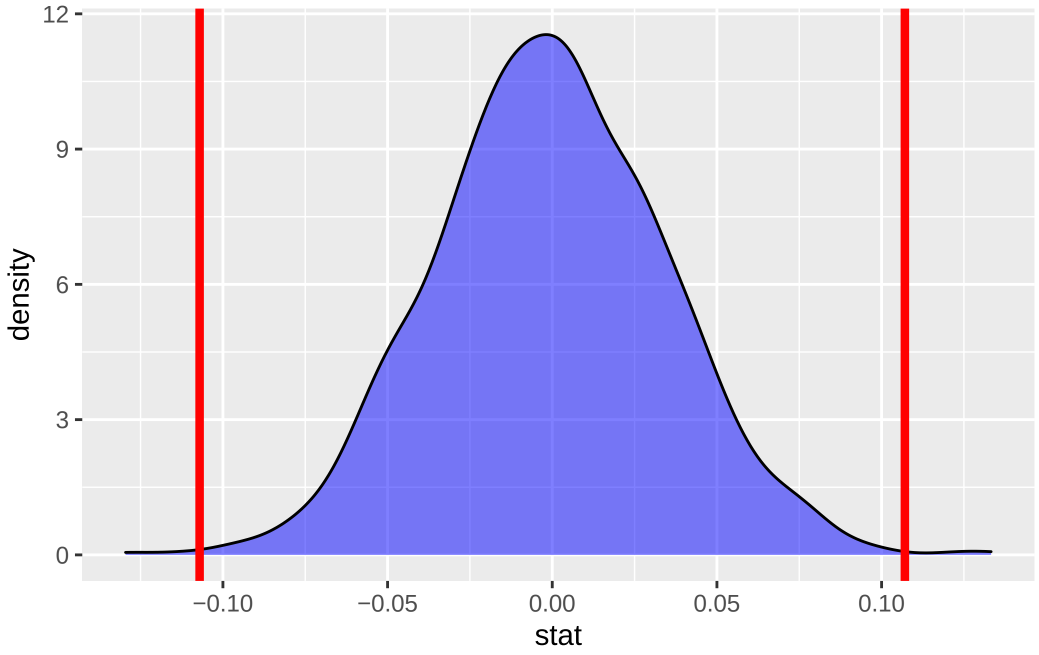


3.2.4 Permutation Test for hours of sleep per night and relationship happiness.

Hypothesis Testing for independence between hours of sleep per night and relationship happiness

Null Hypothesis: hours of sleep per night and relationship happiness are independent. The correlation coefficient is 0.

Alternative Hypothesis: hours of sleep per night and relationship happiness are not independent. The correlation coefficient is not 0.

~~~
```{r}
#| message: false
#| warning: false
obs_stat <-
  data %>%
  specify(RELHAPPY1 ~ HRSLEEP) %>%
  calculate(stat = 'correlation')

set.seed(2022)
null_dist <-
  data %>%
  specify(RELHAPPY1 ~ HRSLEEP) %>%
  hypothesize(null = 'independence') %>%
  generate(reps = 1000, type = 'permute') %>%
  calculate(stat = 'correlation')

null_dist %>%
  visualize()

p_value <-
  null_dist %>%
  get_p_value(obs_stat = obs_stat, direction = 'both')
p_value

null_dist %>%
  ggplot(aes(stat)) +
  geom_density(fill = 'blue', alpha = 0.5)  +
  geom_vline(xintercept = obs_stat$stat,
             color = "red",
             size=1.5) +
  geom_vline(xintercept = - obs_stat$stat,
             color = "red",
             size=1.5)
obs_stat
```
~~~

```
# A tibble: 1 x 1
```

```
  p_value
    <dbl>
1   0.004
Response: RELHAPPY1 (numeric)
Explanatory: HRSLEEP (numeric)
# A tibble: 1 x 1
   stat
  <dbl>
1 0.107
```





This permutation test results in a p-value of 0.004, which is slightly larger than our significance level of 0.0025. Therefore we fail to reject the null-hypothesis that the correlation coefficient between relationship happiness and hours of sleep per night is 0.

3.2.5 Permutation test for independence between days partner consumed more than 4 drinks in the last 90 days and relationship happiness:

Null Hypothesis: Days partner consumed more than 4 drinks in the last 40 days and relationship happiness are independent. The correlation coefficient is 0.

Alternative Hypothesis: Days partner consumed more than 4 drinks in the last 40 days are not independent. The correlation coefficient is not 0.

~~~
```{r}
obs_stat <-
  data %>%
  specify(RELHAPPY1 ~ CDDAYSBD) %>%
  calculate(stat = 'correlation')

set.seed(2022)
null_dist <-
  data %>%
  specify(RELHAPPY1 ~ CDDAYSBD) %>%
~~~

```
  hypothesize(null = 'independence') %>%
  generate(reps = 1000, type = 'permute') %>%
  calculate(stat = 'correlation')

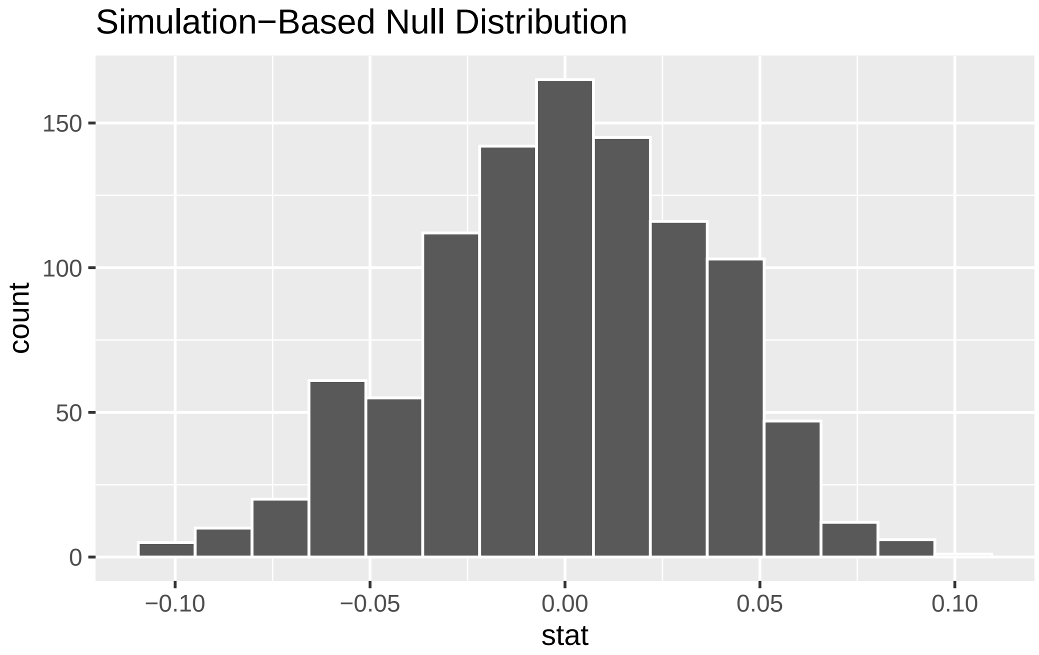
null_dist %>%
  visualize()

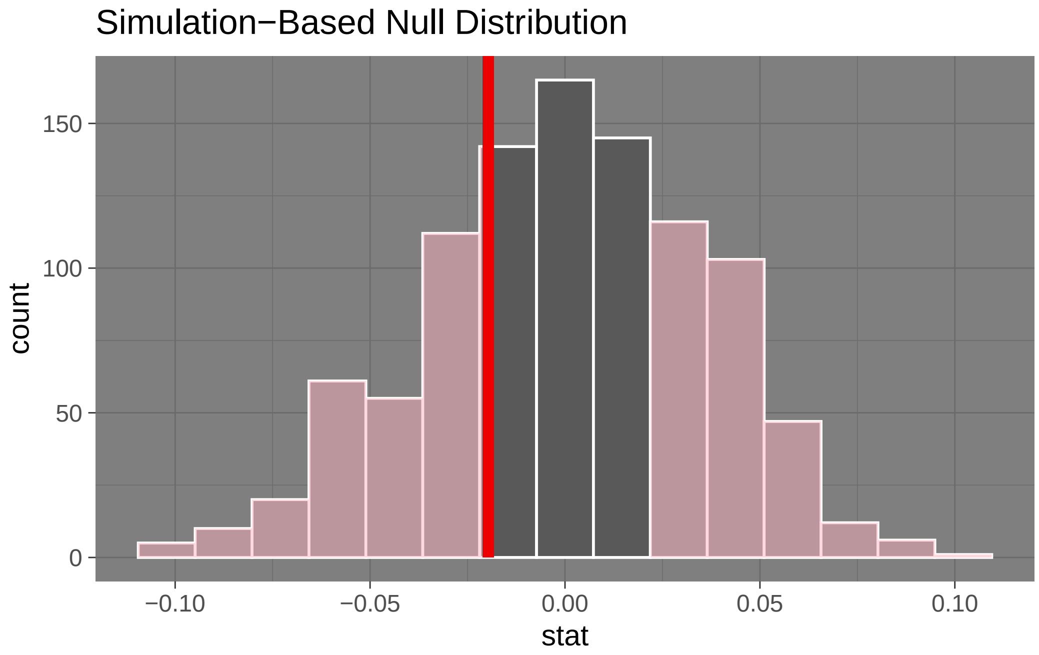
p_value <-
  null_dist %>%
  get_p_value(obs_stat = obs_stat, direction = 'both')
p_value

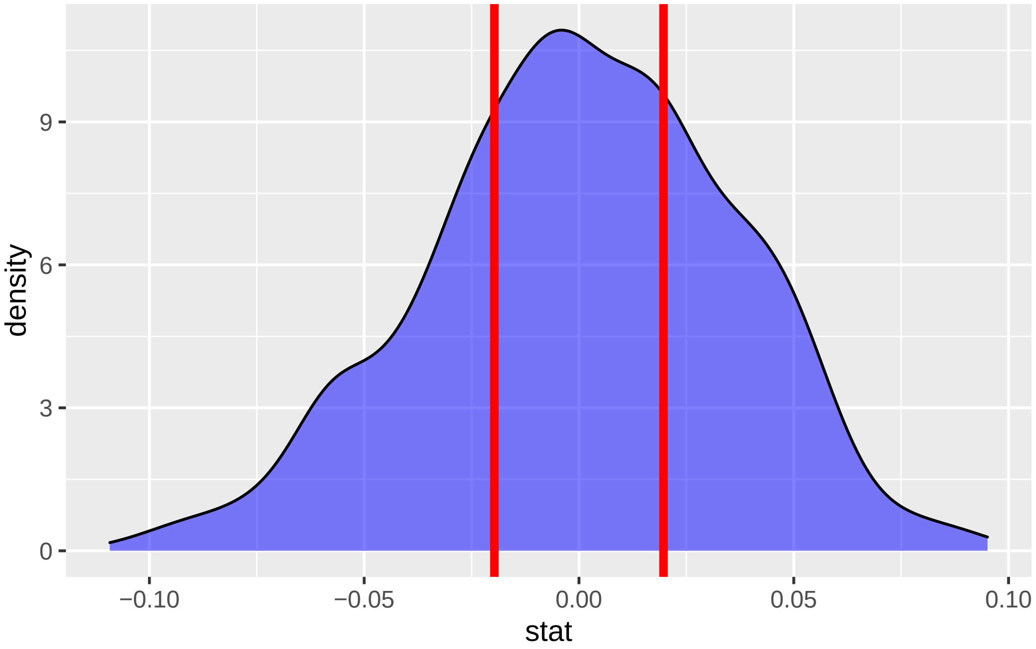
null_dist %>%
  visualize() +
  theme_dark() +
  shade_p_value(obs_stat = obs_stat, direction = 'both')

null_dist %>%
  ggplot(aes(stat)) +
  geom_density(fill = 'blue', alpha = 0.5)  +
  geom_vline(xintercept = obs_stat$stat,
             color = "red",
             size=1.5) +
  geom_vline(xintercept = - obs_stat$stat,
             color = "red",
             size=1.5)
obs_stat
```
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.566
Response: RELHAPPY1 (numeric)
Explanatory: CDDAYSBD (numeric)
# A tibble: 1 x 1
     stat
    <dbl>
1 -0.0197
```
```


Simulation−Based Null Distribution
150
100
50
0
count
−0.10
−0.05
0.00
0.05
0.10
stat


Simulation−Based Null Distribution
150
100
50
0
count
−0.10
−0.05
0.00
0.05
0.10
stat

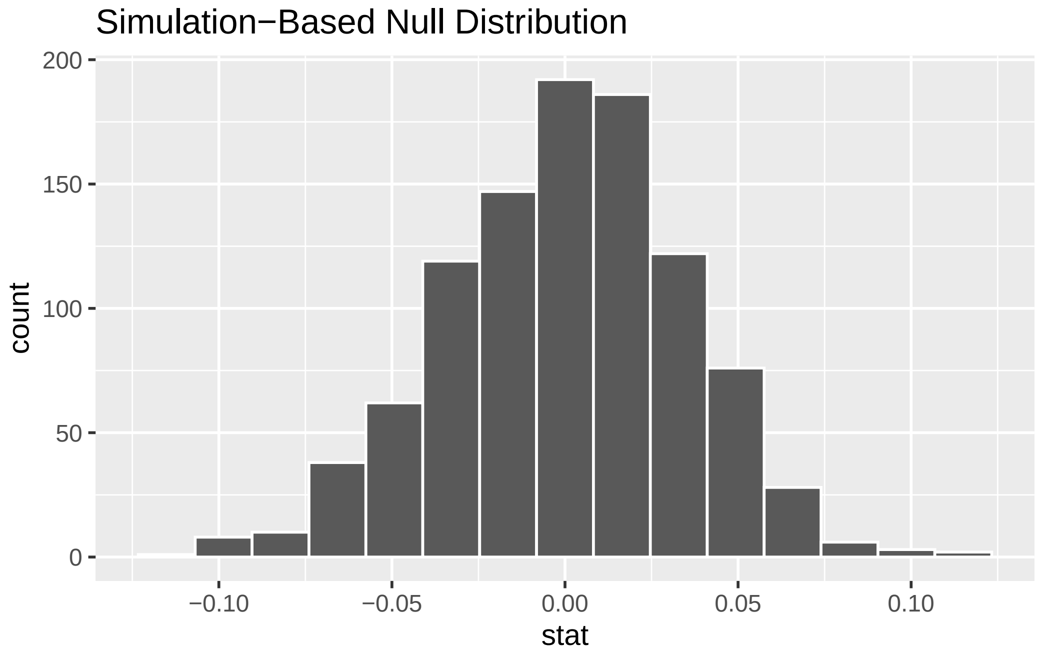
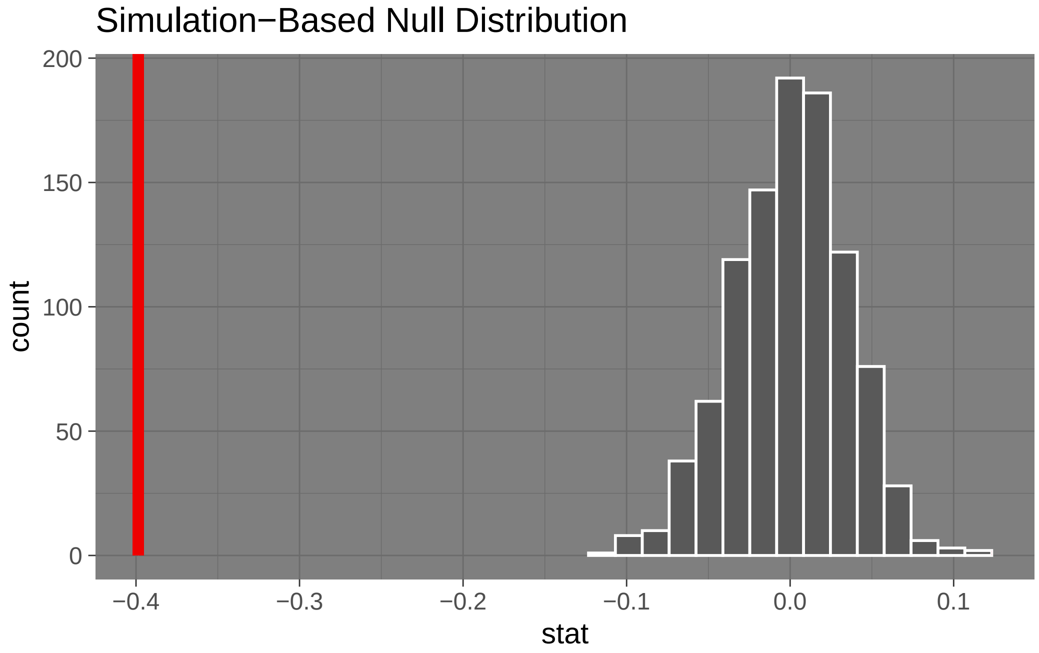
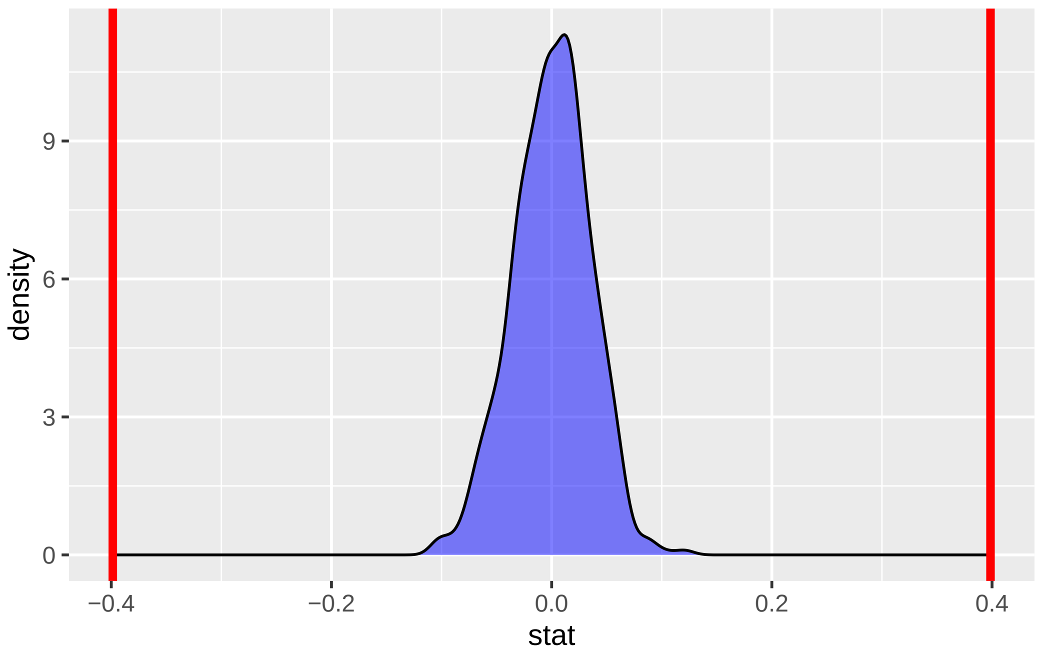


This permutation test results in a p-value of 0.566 which is greater than our significance level of 0.0025, so we fail to reject the null hypothesis that days partner consumed more than 4 drinks in the last 40 days is independent from relationship happiness.

### 3.2.6 Permutation test for independence between argument frequency and relationship happiness:

Null Hypothesis: Argument frequency and relationship happiness are independent. The correlation coefficient is 0.

Alternative Hypothesis: argument frequency and relationship happiness are not independent. The correlation coefficient is not 0.

~~~
```{r}
#| warning: false

obs_stat <-
  data %>%
  specify(RELHAPPY1 ~ ARGUEFRQ1) %>%
  calculate(stat = 'correlation')

set.seed(2022)
null_dist <-
~~~

~~~
  data %>%
  specify(RELHAPPY1 ~ ARGUEFRQ1) %>%
  hypothesize(null = 'independence') %>%
  generate(reps = 1000, type = 'permute') %>%
  calculate(stat = 'correlation')

null_dist %>%
  visualize()

p_value <-
  null_dist %>%
  get_p_value(obs_stat = obs_stat, direction = 'both')
p_value

null_dist %>%
  visualize() +
  theme_dark() +
  shade_p_value(obs_stat = obs_stat, direction = 'both')

null_dist %>%
  ggplot(aes(stat)) +
  geom_density(fill = 'blue', alpha = 0.5)  +
  geom_vline(xintercept = obs_stat$stat,
             color = "red",
             size=1.5) +
  geom_vline(xintercept = - obs_stat$stat,
             color = "red",
             size=1.5)
obs_stat
```
~~~

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
Response: RELHAPPY1 (numeric)
Explanatory: ARGUEFRQ1 (numeric)
# A tibble: 1 x 1
    stat
   <dbl>
1 -0.399
```


Simulation−Based Null Distribution
count
200
150
100
50
0
−0.10
−0.05
0.00
0.05
0.10
stat


Simulation−Based Null Distribution
count
200
150
100
50
0
−0.4
−0.3
−0.2
−0.1
0.0
0.1
stat



This permutation test results in a p-value of 0 which is less than our significance level of 0.0025 which leads us to reject the null hypothesis that the correlation coefficient between argument frequency and relationship happiness is 0. So there is a significant correlation between the argument frequency and relationship happiness.

3.2.7 New variable ARGUEFRQ1

The variable ARGUEFRQ1 was constructed by taking the second character in the ARGUEFRQ strings in the data and mutating them into a new column that is a numeric variable so that we could more easily use the data for testing.

~~~
```{r}
data %>%
  select(ARGUEFRQ1) %>%
  summarize(mean = mean(ARGUEFRQ1))
```
~~~

```
      mean
1 2.353222
```

3.3 Linear Model

For our linear model, we used libraries from the following sources: Kuhn and Wickham (2020), Wickham et al. (2019), Tang, Horikoshi, and Li (2016), and Greenwell and Boehmke (2020).

~~~~
```{r}
library(tidymodels)
library(tidyverse)
library(ggfortify)
library(vip)

df <- data %>%
  select(RELHAPPY1, ARGUEFRQ1, HRSLEEP, YRSMAR) %>%
  drop_na()

set.seed(1800)
df_split <- initial_split(data = df, prop = 0.80, strata = RELHAPPY1)
df_split

df_not_testing <- training(df_split)
df_testing <- testing(df_split)


set.seed(123)
df_split_2 <- initial_split(data = df_not_testing,
                            prop = 0.75,
                            strata = RELHAPPY1)

df_training <- training(df_split_2)
df_validation <-testing(df_split_2)

model_specs <-
  linear_reg() %>%
  set_engine('lm') %>%
  set_mode('regression')

model_1 <-
  model_specs %>%
  fit(RELHAPPY1 ~ ARGUEFRQ1, data = df_training)

model_2 <-
  model_specs %>%
  fit(RELHAPPY1 ~ ARGUEFRQ1 + YRSMAR, data = df_training)
~~~~

```r
model_3 <-
  model_specs %>%
  fit(RELHAPPY1 ~ ARGUEFRQ1 + YRSMAR + HRSLEEP, data = df_training)

model_4 <-
  model_specs %>%
  fit(RELHAPPY1 ~ ARGUEFRQ1*HRSLEEP, data = df_training)

model_5 <-
  model_specs %>%
  fit(RELHAPPY1 ~ sqrt(YRSMAR), data = df_training)

model_6 <-
  model_specs %>%
  fit(RELHAPPY1 ~ poly(ARGUEFRQ1, degree = 2, raw = TRUE), data = df_training)

model_7 <-
  model_specs %>%
  fit(RELHAPPY1 ~ poly(ARGUEFRQ1, degree = 2, raw = TRUE) + YRSMAR, data = df_training)

fitted_models <- list(model_1, model_2, model_3,
                      model_4, model_5, model_6,
                      model_7)

# function for validation performance
validation_performance <- function(model){
  model_results <-
    df_validation %>%
    bind_cols(predict(model, df_validation)) %>%
    select(RELHAPPY1, .pred)
  my_metrics <- metric_set(rmse, rsq, mae)
  output <-
    model_results %>%
    my_metrics(RELHAPPY1, .pred) %>%
    select(-.estimator) %>%
    pivot_wider(names_from = .metric, values_from = .estimate)
  return(output)
}

validation_results <- tibble()
for(model in fitted_models){
  validation_results <-
```

~~~
    validation_results %>%
    bind_rows(validation_performance(model))
}
validation_results

# CHOOSE A MODEL and do a last fit!

model_4_updated <-
  model_specs %>%
  fit(RELHAPPY1 ~ ARGUEFRQ1*HRSLEEP, data = df_not_testing)

tidy(model_4_updated)

# report the following test result

test_results <-
  df_testing %>%
  bind_cols(predict(model_4_updated, new_data = df_testing)) %>%
  select(RELHAPPY1, .pred)

my_metrics <- metric_set(rmse, rsq, mae)

test_results %>%
  my_metrics(RELHAPPY1, .pred)
```
~~~

```
<Training/Testing/Total>
<663/168/831>
# A tibble: 7 x 3
   rmse    rsq   mae
  <dbl>  <dbl> <dbl>
1  1.04 0.132  0.821
2  1.03 0.135  0.813
3  1.03 0.139  0.809
4  1.03 0.144  0.813
5  1.10 0.0152 0.871
6  1.03 0.141  0.821
7  1.03 0.144  0.813
# A tibble: 4 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)          7.29      0.820      8.90 5.54e-18
```

```
2 ARGUEFRQ1          -1.02      0.318      -3.21 1.40e- 3
3 HRSLEEP            -0.140     0.116      -1.22 2.25e- 1
4 ARGUEFRQ1:HRSLEEP   0.0788    0.0452      1.74 8.20e- 2
# A tibble: 3 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       1.11
2 rsq     standard       0.193
3 mae     standard       0.842
```

4 Discussion/Conclusions

Our objective in this project was to find a way to predict happiness within a marriage using a combination of factors including, hours of sleep per night, years married, cohabitation before marriage, argument frequency, and drinking habits. The best model that we found is $$ RELHAPPY1 =

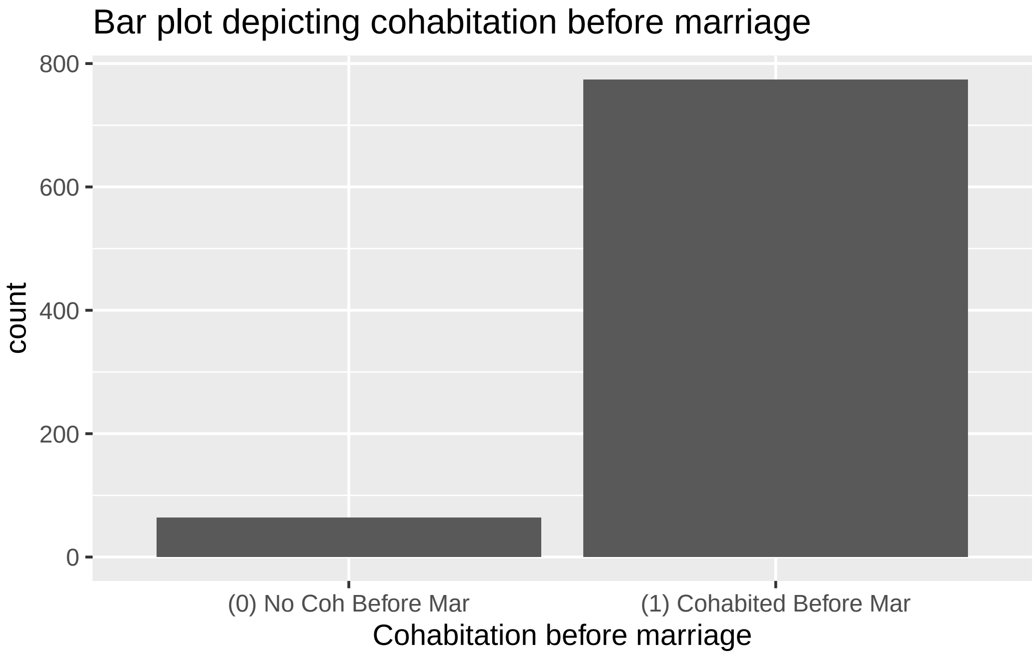
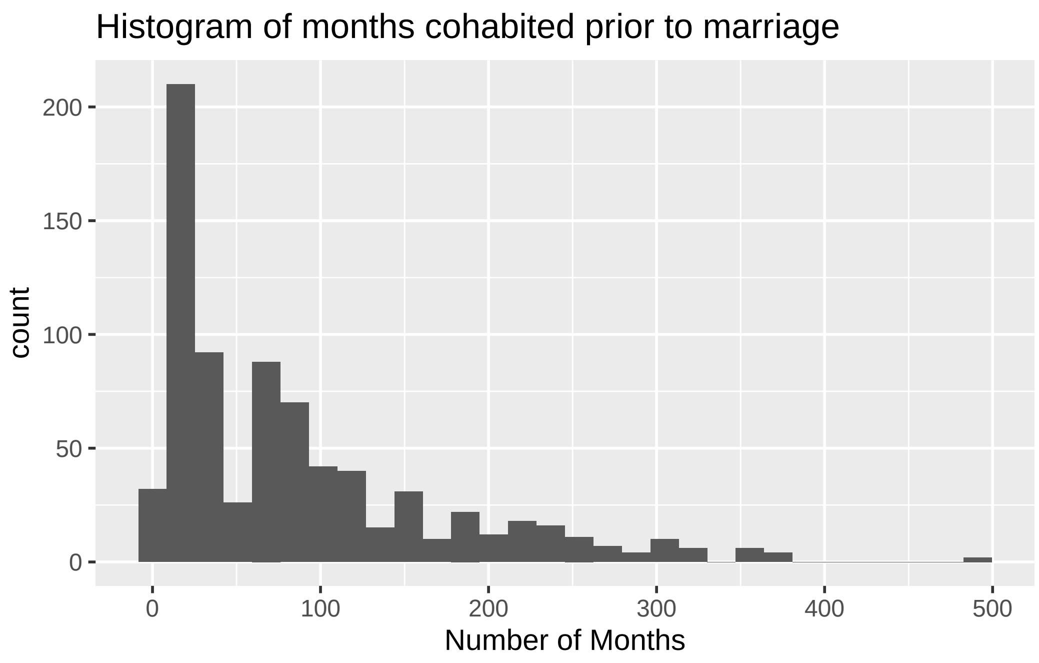
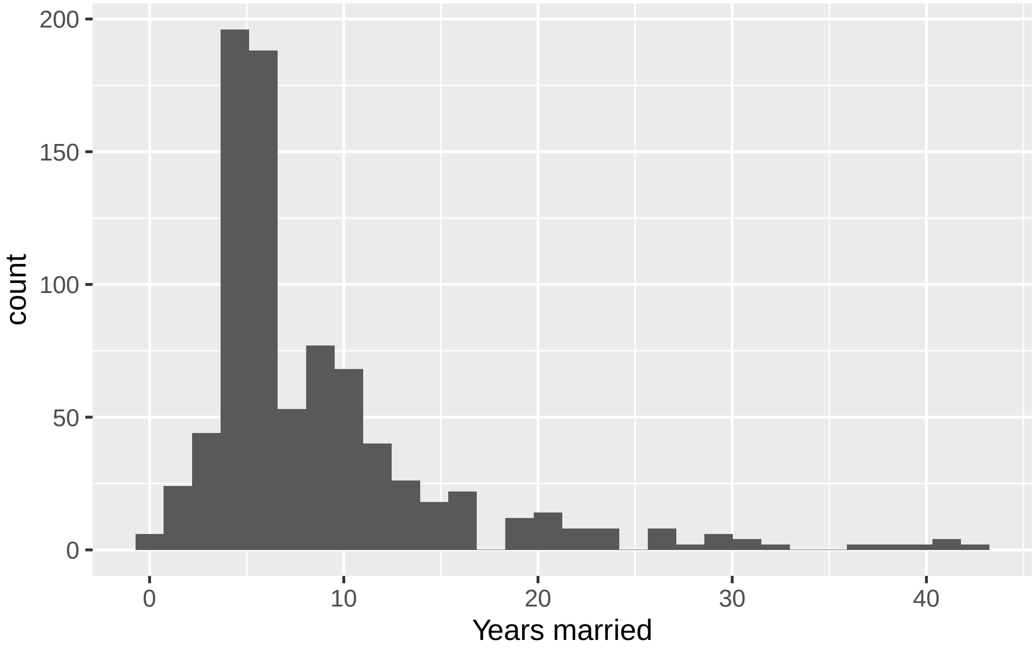
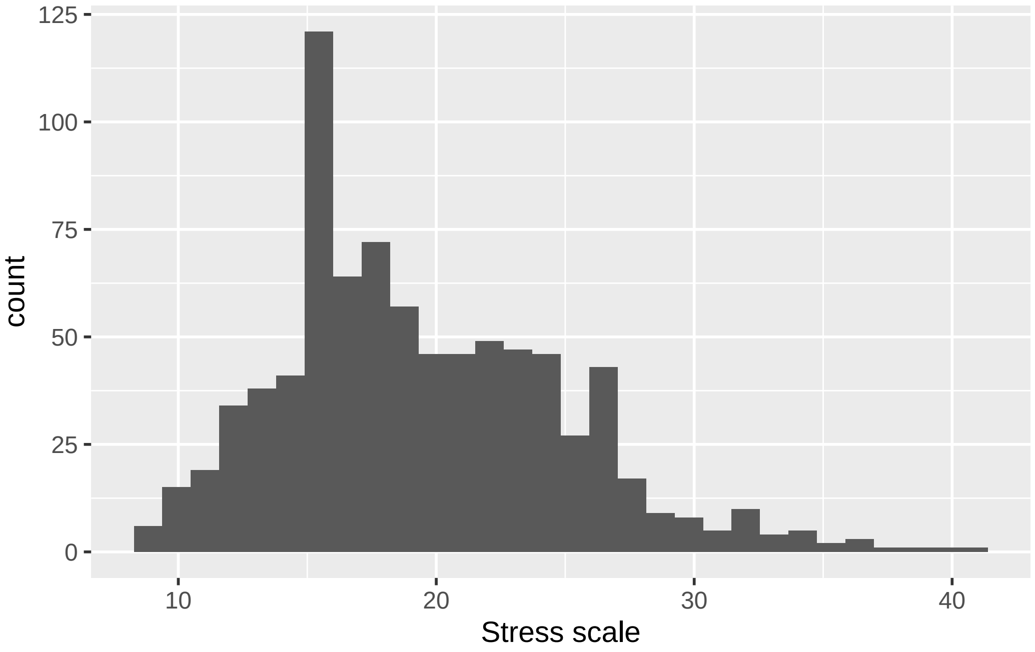
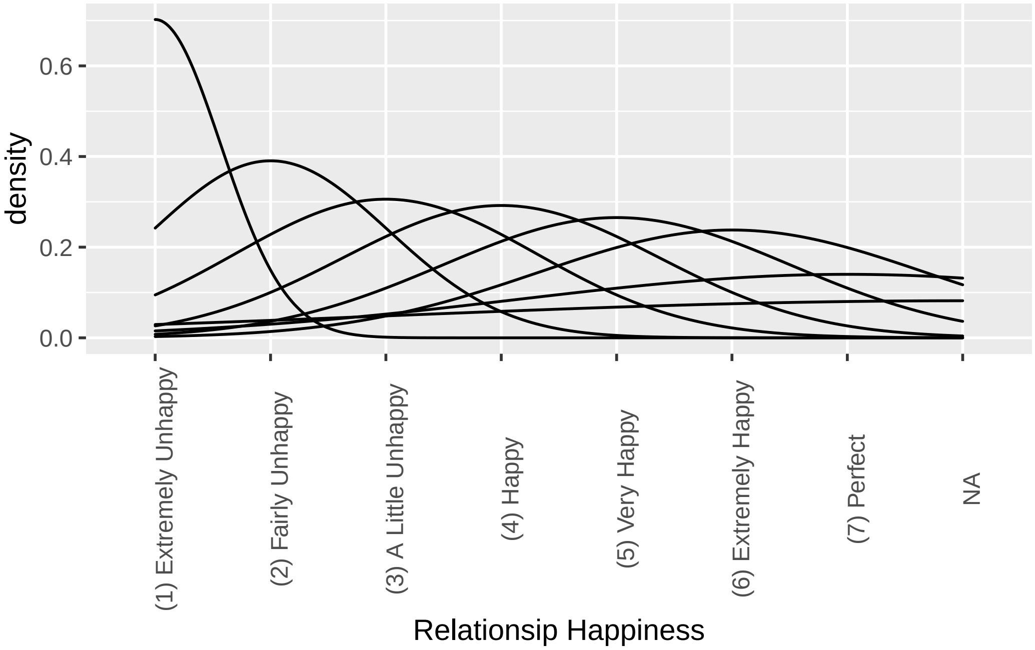
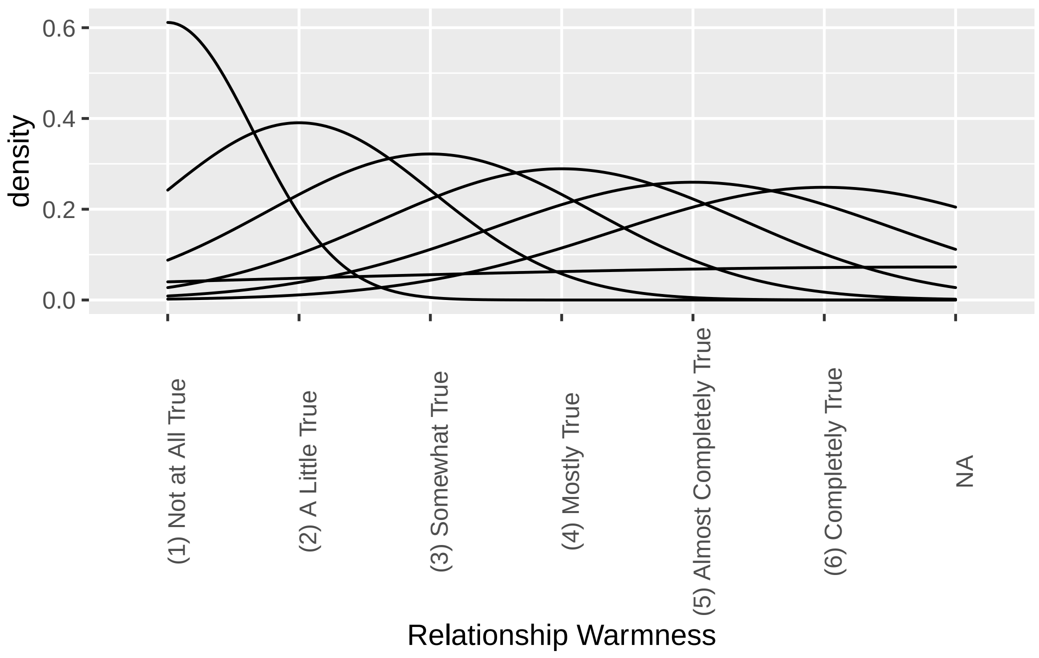
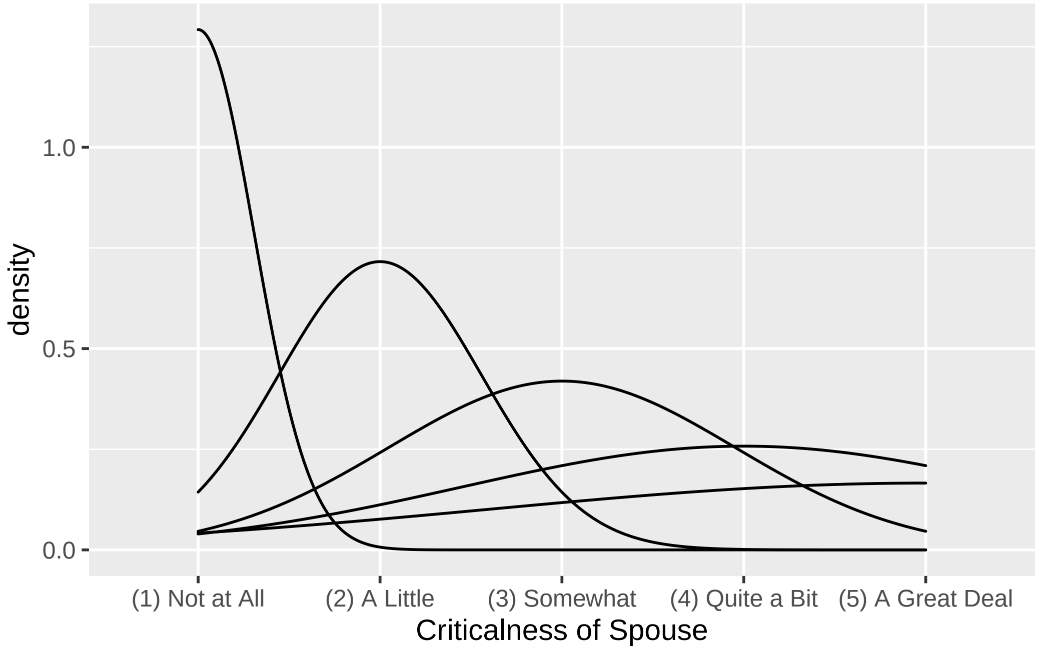
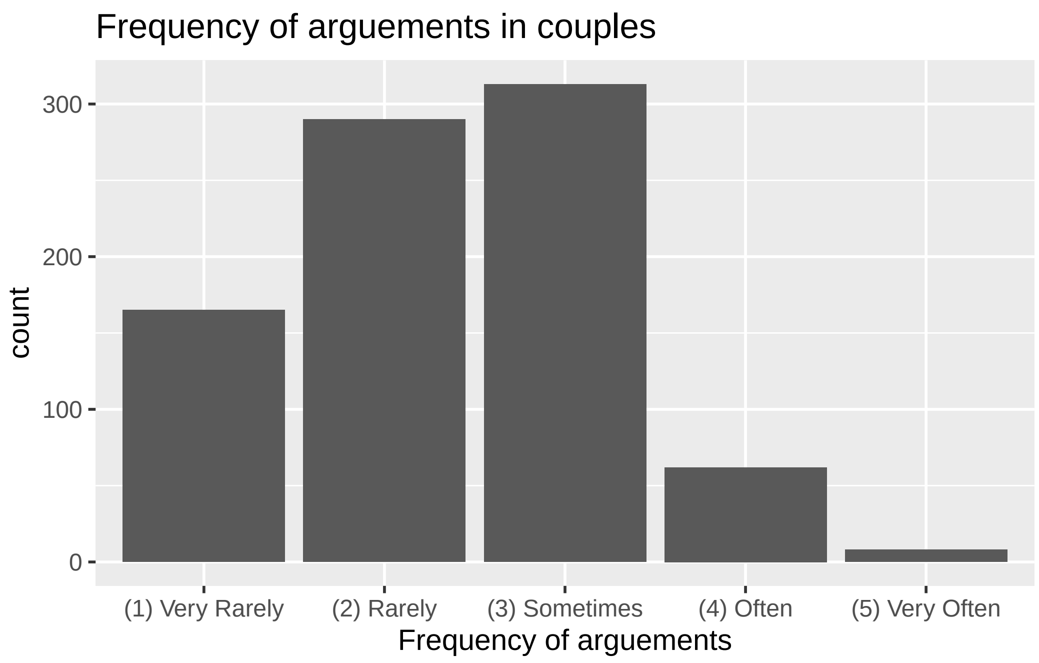
7.29 - 1.01ARGUEFRQ1 - 0.1404HRSLEEP + 0.0787ARGUEFRQ1*HRSLEEP. $$ The R^2 value for this model is .193. This means that around 19% of

the variance in marriage relationship happiness can be explained by our model. Our model suggests that a higher frequency of arguments predicts a lower rate of happiness in a relationship. However, if the couple sleeps more hours per night, the argument frequency has less of a negative effect on happiness. Previous studies have shown that cohabitation before marriage is bad for marriage Fox (2014), however, we did not find a significant relationship between the relationship happiness and cohabitation before marriage. Additionally, we found that hours of sleep has a positive effect on a marriage which aligns with The American Academy of Sleep Medicine's study on sleep and relationships Wagner (2009). Happiness is hard to find and define in real life, and it is also hard to predict, so there are some limitations we have to consider. Our data set contained around 500 variables, and we did not have the capacity to use all of them. Additionally, many of the variables we did not choose had a high population of N/A inputs which is one reason they did not end up on our final list of variables. Since we only selected five possible explanatory variables, there were a lot of other factors that could lead to relationship happiness including, age, political beliefs, religion, and other variables. If possible future research could look at more models with different variables by using imputation techniques to use variables with a lot of missing values.

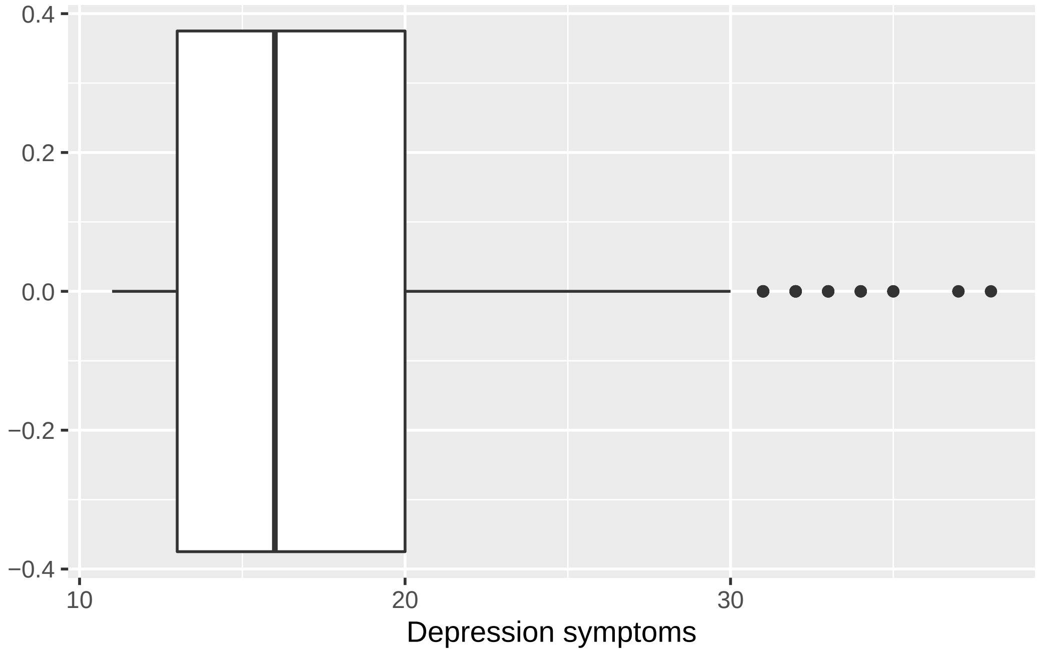
5 Appendix

In the appedix we used the libraries Wickham et al. (2019) and Couch et al. (2021).

5.1 Loading Libraries and Data File.

~~~
```{r}
library(tidyverse)
library(infer)
load(file = '37404-0001-Data.rda')
```
~~~

5.2 Visualizing cohabitation before marriage.

Visualizing breakdown of cohabitation before marriage - COHAB4MAR This variable is a non-ordinal categorical variable with two levels that tells whether or not the couple cohabited before marriage. Seems like the vast majority of couples cohabit before marriage.

~~~
```{r}
da37404.0001 %>%
  ggplot(aes(x = COHB4MAR)) +
  geom_bar() +
  ggtitle('Bar plot depicting cohabitation before marriage') +
  xlab('Cohabitation before marriage')
```
~~~



5.3 Visualizing months cohabiting before marriage.

Visualizing number of months cohabiting before marriage - MNCOHB4M This is a continuous numerical variable with range 0 to infinity that tells us the number of months that the couple spent cohabiting before marriage. Problem: if couple did not cohabit before marriage, data was entered as NA rather than 0. Will have to mutate the variable.

~~~
```{r}
da37404.0001 %>%
  ggplot(aes(x = MNCOHB4M)) +
  geom_histogram() +
  ggtitle('Histogram of months cohabited prior to marriage') +
  xlab('Number of Months')
```
~~~



5.4 Visualizing years married.

Visualizing years married among couples in this data set - YRSMAR. This is a continuous numerical variable with range 0 to infinity that tells us how long a couple has been married at the time of their interview.

~~~
```{r}
da37404.0001 %>%
  select(YRSMAR) %>%
  ggplot(aes(x = YRSMAR)) +
  geom_histogram() +
  xlab('Years married')
```
~~~



5.5 Visualizing stress levels.

Visualizing stress scores - STRSCALE This is a continuous numerical variable with range 0 to 41, indicating the sum of stress related question scores. Higher numbers correspond to higher stress levels.

~~~
```{r}
da37404.0001 %>%
  select(STRSCALE) %>%
  ggplot(aes(STRSCALE))+
  geom_histogram() +
  xlab('Stress scale')

da37404.0001 %>%
  summarise(range = range(STRSCALE))
```
~~~

```
  range
1     9
2    41
```



5.6 Visualizing relationship happiness, our response variable.

Visualizing relationship happiness - RELHAPPY This is a ordinal categorical variable with range 1 to 7.

~~~
```{r}
da37404.0001 %>%
  select(RELHAPPY) %>%
  ggplot(aes(RELHAPPY)) +
  geom_density() +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab('Relationsip Happiness')
```
~~~



5.7 Visualizing relationship warmness.

Visualing relationship warmness - RELRWAR This is a ordinal categorical variable with range 1 to 7.

~~~
```{r}
da37404.0001 %>%
  select(RELWARM) %>%
  ggplot(aes(RELWARM)) +
  geom_density() +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab('Relationship Warmness')
```
~~~



5.8 Visualizing how critical one's spouse is.

Visualizing how critical spouse's partner is - SPCRICTIC This is a ordinal categorical variable with range 1 to 5.

~~~
```{r}
da37404.0001 %>%
  select(SPCRITIC) %>%
  ggplot(aes(SPCRITIC)) +
  geom_density() +
  xlab('Criticalness of Spouse')
```
~~~



5.9 Visualizing argument frequency.

ARGUEFRQ: An ordinal categorical variable with how often the couple argues with 5 levels, (1) Very Rarely, (2) Rarely, (3) Sometimes, (4) Often, (5) Very Often. This variable gives us information about how frequently the couple argues.

~~~
```{r}
da37404.0001 %>%
  select(ARGUEFRQ) %>%
  ggplot(aes(x = ARGUEFRQ)) +
  geom_bar() +
  ggtitle("Frequency of arguements in couples") +
  xlab("Frequency of arguements")
```
~~~



5.10 Visualizing depression symptoms.

DEPSYM: A discrete numerical variable range 11 to 38 that tells us about how many depressive symptoms the respondant faces.

Note: This variable was created by adding up the ranking (1-5) of other categories to determine the overall depressive symptoms numerical value.

~~~
```{r}
da37404.0001 %>%
  select(DEPSYM) %>%
  summarise(range = range(DEPSYM))
```
~~~

```
  range
1    11
2    38
```

~~~
```{r}
da37404.0001 %>%
  select(DEPSYM) %>%
  ggplot(aes(x = DEPSYM)) +
  geom_boxplot() +
~~~

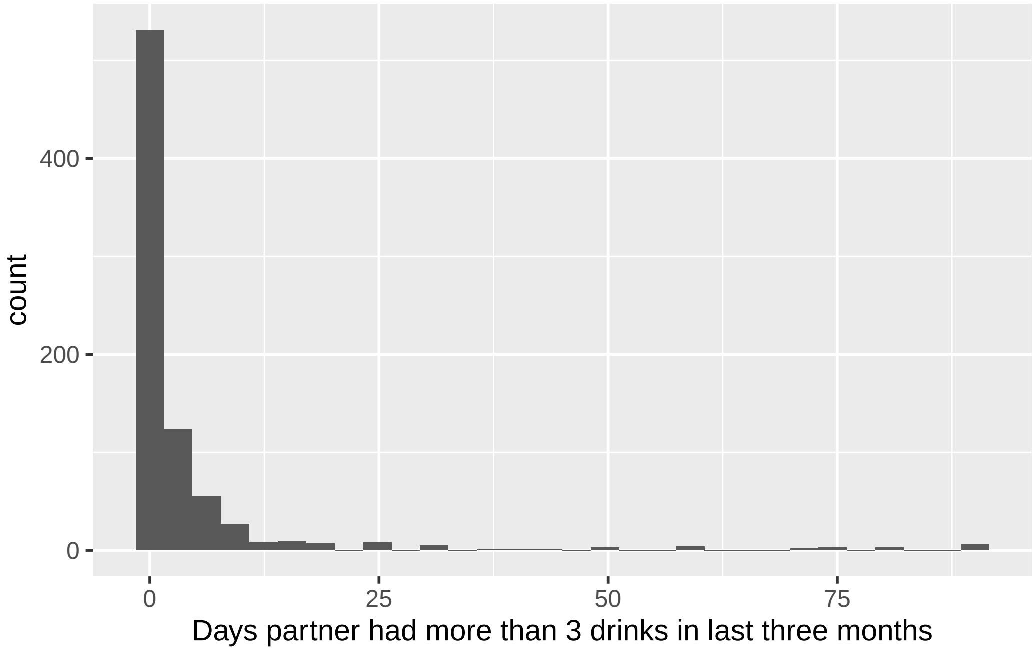
```
  xlab('Depression symptoms')
```



5.11 Visualizing days in which spouse has had more than four drinks in one day in the last three months

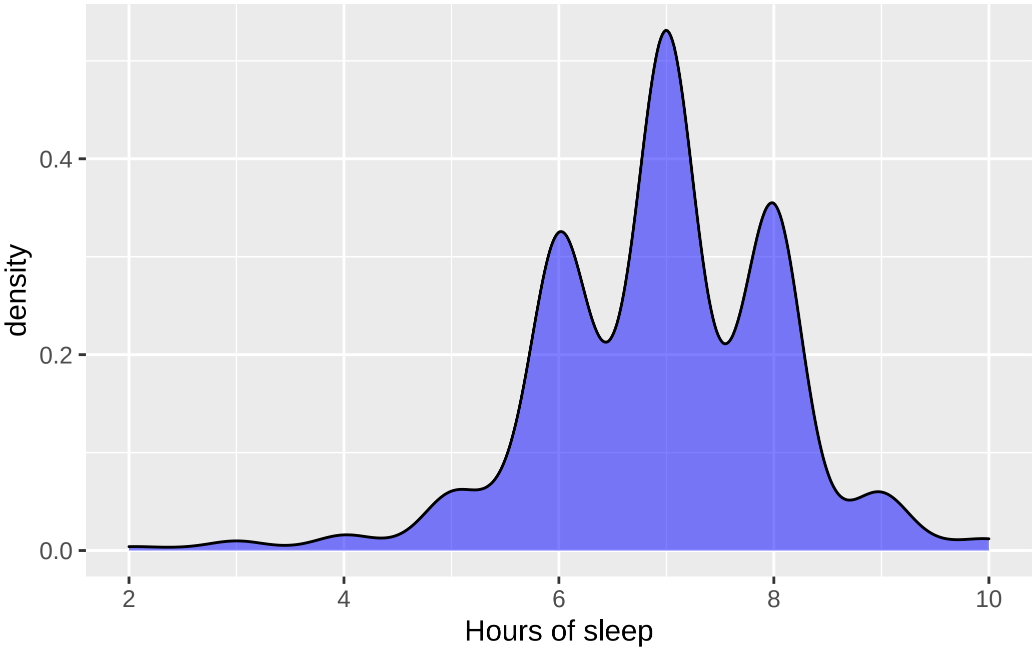
CDDAYSBD: A continuous numerical variable with range 0 to 90 which tells us how many days has the spouse had 4+ drinks in the last 3 months.

Problem: There are some N/As.

~~~
```{r}
da37404.0001 %>%
  select(CDDAYSBD) %>%
  drop_na() %>%
  summarise(range = range(CDDAYSBD))
```
~~~

```
  range
1     0
2    90
```

~~~
```{r}
da37404.0001 %>%
  select(CDDAYSBD) %>%
  drop_na() %>%
  ggplot(aes(x = CDDAYSBD)) +
  geom_histogram() +
  xlab('Days partner had more than 3 drinks in last three months')
```
~~~



5.12 Visualizing hours of sleep.

HRSLEEP: A continuous numerical variable with range 2 to 10 that tells us how many hours of sleep per/day the respondent gets.

~~~
```{r}
da37404.0001 %>%
  select(HRSLEEP) %>%
  summarise(range = range(HRSLEEP))
```
~~~

```
  range
1     2
2    10
```

```{r}
da37404.0001 %>%
  select(HRSLEEP) %>%
  drop_na() %>%
  ggplot(aes(x = HRSLEEP)) +
  geom_density(fill = 'blue', alpha = 0.5) +
  xlab('Hours of sleep')
```



References

Couch, Simon P., Andrew P. Bray, Chester Ismay, Evgeni Chasnovski, Benjamin S. Baumer, and Mine Çetinkaya-Rundel. 2021. "{Infer}: An {r} Package for Tidyverse-Friendly Statistical Inference" 6: 3661. https://doi.org/10.21105/joss.03661.

Fox, Lauren. 2014. "The Science of Cohabitation: A Step Toward Marriage, Not a Rebellion." *The Atlantic*. https://www.theatlantic.com/health/archive/2014/03/the-science-of-cohabitation-a-step-toward-marriage-not-a-rebellion/284512/.

Greenwell, Brandon M., and Bradley C. Boehmke. 2020. "Variable Importance Plots—an Introduction to the Vip Package" 12. https://doi.org/10.32614/RJ-2020-013.

Henry, Lionel, and Hadley Wickham. 2022. "Rlang: Functions for Base Types and Core r and 'Tidyverse' Features." https://CRAN.R-project.org/package=rlang.

Johnson, Sheri L., and Theodore Jacob. 1997. "Marital Interactions of Depressed Men and

Women." *Journal of Consulting and Clinical Psychology* 65 (1): 15–23. https://doi.org/10.1037/0022-006x.65.1.15.

Kuhn, Max, and Hadley Wickham. 2020. "Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles." https://www.tidymodels.org.

"Marriage and Divorce." 2022. *Centers for Disease Control and Prevention*. cdc.gov/nchs/fastats/marriage-divorce.htm.

Smith, Sylvia. 2022. "What Is Marriage?" *Marriage.com*. https://www.marriage.com/advice/relationship/five-facets-of-the-true-meaning-of-marriage/.

Tang, Yuan, Masaaki Horikoshi, and Wenxuan Li. 2016. "Ggfortify: Unified Interface to Visualize Statistical Result of Popular r Packages" 8. https://doi.org/10.32614/RJ-2016-060.

Ushey, Kevin, JJ Allaire, Hadley Wickham, and Gary Ritchie. 2022. "Rstudioapi: Safely Access the RStudio API." https://CRAN.R-project.org/package=rstudioapi.

Wagner, Kelly. 2009. "Stable Marriage Is Linked with Better Sleep In Women." *American Academy of Sleep Medicine - Association for Sleep Clinicians and Researchers*. https://aasm.org/stable-marriage-is-linked-with-better-sleep-in-women/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the {Tidyverse}" 4: 1686. https://doi.org/10.21105/joss.01686.