Mini-Project 2

Stat 212: Interim 2022

General Instructions

This assignment is due **Sunday, January 16 at 11:59 PM**. You will submit one pdf file for the whole group, knitted from an RMarkdown document, to Moodle with all contributing group member names. Your file name should include the last name of everyone in the group and the project number, `BagginsGamgeeBrandybuckTook_Project2.pdf`. You do NOT need to submit your RMarkdown .Rmd file, *however* it needs to be saved in your "Submit" folder in the R server so I can access it if I need to look at your code more closely.

You should create your own RMarkdown file and start from scratch. You may copy and paste the code from this document to read in the data. The formatting of this project report should be neater than the first. For example, I don't want to see any R code in the pdf, but I would like to see the R output for your simulated null distributions and p-values. You can do this by including `echo = FALSE` at the start of your R chunks[1]. You can also change the plot sizes to better fit the pages and not take up too much space, play around with changing the figure height and width values.

Exploring Binary Categorical Variables

We are still using the survey data from the NREEC. This time you will be using *categorical variables only*. Since none of the original variables have only two responses, we will need to transform them into binary variables. For all of the questions with likert (Strongly agree, Agree,...) scales, this is can easily be done by breaking the responses down into a general "Agree" category and "Other" or "Not agree" group. See the code chunk below for an example of how to do this in R. For the demographic variables, you can also create two groups by choosing your own criteria. For example, if you want to consider Age, you could either filter out two age groups to compare, or combine age groups to cover a wider range. The race/ethnicity variables are a little tricky because they are just indicators, meaning each column only has one type of response and an `NA` otherwise. For example, the variable `Black` has the "Black or African American" response, but we can't assume that every respondent with an `NA` is not Black (someone who identifies as Black may not have answered this question). So you will need to consider the other race/ethnicity columns to check whether the respondent selected another option. See the code below for an example of how I created the `IsWhite` variable.

> Note: For the examples below, you will need to combine several wrangling parts into one big chain. You might use Example 1B, 2A, and 3 all in one chain to create a dataset you'll use called `survey_binary`. DO NOT try to filter/mutate seperately and combine afterwards.

[1] You can also add a setup R chuck that includes `knitr::opts_chunk$set(echo = FALSE)` to suppress code from ALL chunks at once.

1. Identify Research Questions and Background Like the last part of Project #1, you will create your own research questions to explore using the data from the NREEC survey.

a. Clearly state two research questions using binary categorical variables. Since you can only use categorical variables, you will be limited to only asking certain kinds of questions at this point. Make sure you have at least one question that:

Addresses whether two categorical variables are independent (e.g. the Dolphins and Mythbusters examples in class).

> Question 1: *Is the proportion of residents aged 34 year or younger who think Northfield values culture the same as the proportion of residents over 34 years old sharing the same opinion?*

Examines a single categorical variable and compares the proportion of "success" to a fixed value (e.g. the ESP and Reese's examples in class).

> Question 2: *Is the proportion of residents who agree that there is access to housing for people of various races significantly larger than 50% (a significant majority)?*

b. Create a **variable codebook** for the data. The codebook should be similar to the one you made in Project #1, but I DO NOT need a row for all of the variables in NHANES, just the the 2-3 you've decided to focus on or changed above. I've included the start of a blank table in the .Rmd version of this document.

Variable name	Original name	Description	Type	Levels/Encoding
`Housing`	`Housing`	There is access to housing in Northfield for people of all races and ethnicities.	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
`Age`	`Age`	In which category is your age?	categorical (ordinal)	18-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, 65-74 years, 75 years or older
`Culture_valued`	`Culture_valued`	I believe my culture is valued in Northfield.	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree

c. Perform a short literature review for your research questions. This could require a little digging, but you should be able to find something. Feel free to reach out to the library to ask for help on where to start looking. Here is a link to some of their resources (link). Find 1-2 supporting, peer-reviewed papers that are related to your questions. Write a brief (one paragraph) summary for each of the papers. Some hints for finding and writing about other research:

- Societal Value Culture: Latent and Dynamic https://journals.sagepub.com/doi/pdf/10.1177/0022022113513404. Schwartz discusses the way perceived societal values complement the individuals values as influences on behavior. It focuses on several parts of culture such as ethnic, professional, religious and family and how that influence is more direct than the societal value of culture. It discusses the misunderstanding of culture in a societal value.
- Exploring Multiple Levels of Access to Rental Subsidies and Supportive Housing https://www-webofscience-com.ezproxy.stolaf.edu/wos/woscc/full-record/WOS:000334070400007.Quinn writes about the benefits of stable housing and the barriers for people of low-income and those that suffer homelessness have to deal with. It focuses on predicting greater or a more limited access to housing. It touches upon how some people have greater access to housing but for others their options are very limited.

2. Variable Exploration Perform some EDA on your chosen variables. Follow the same guidelines for Project #1 in what should be included. Briefly write about the trends you can or cannot see in the plots and tables.

Remember to filter out missing values and try to keep your plots clean with clear titles and good color choices.

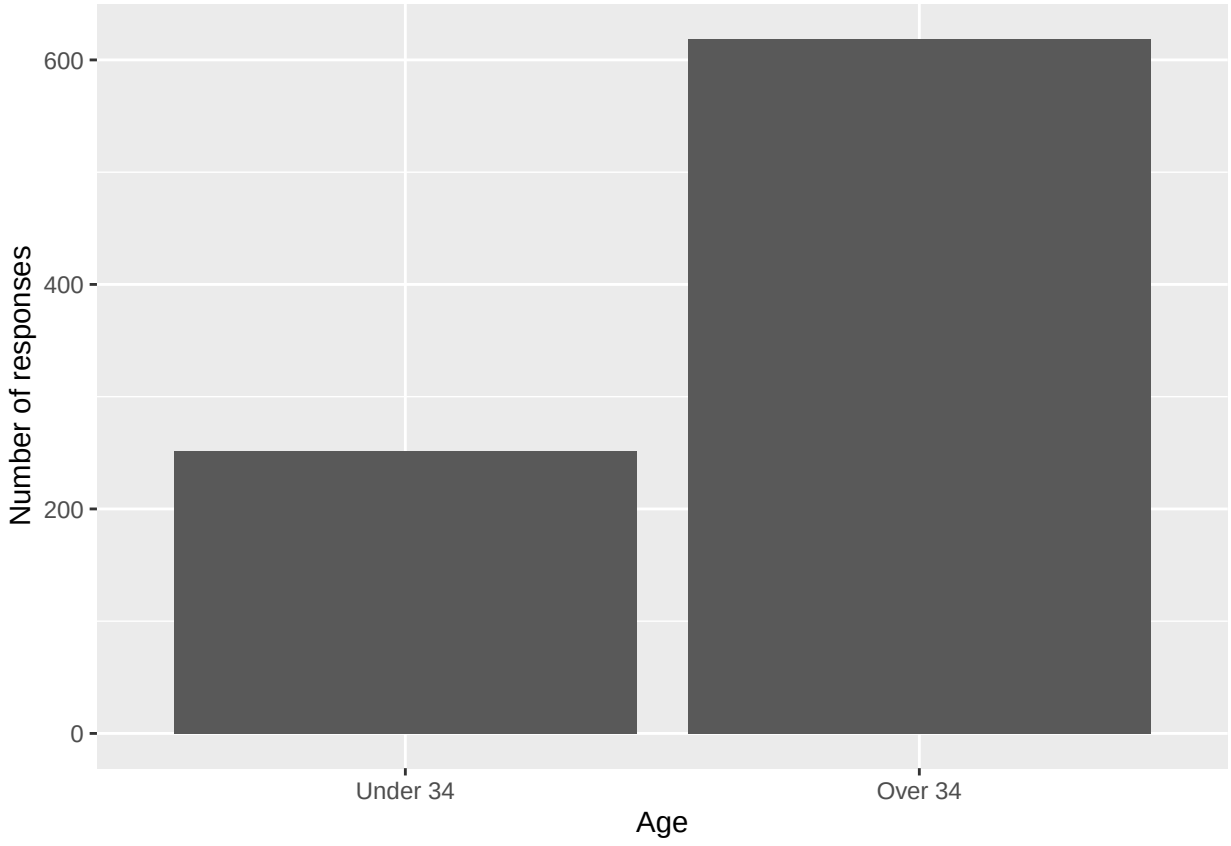
- Tables - both counts and proportions
- Visual plot of the individual variable
- Conditional tables - Does it make sense to create proportions conditional on your explanatory variable? (Probably...)
- Visual plots of the variable pairs (this is the most interesting part where we can look for relationships)

  Question 1: *Is the proportion of residents aged 34 year or younger who think Northfield values culture the same as the proportion of residents over 34 years old sharing the same opinion?*

EDA for Age_combine

```
##
## Under 34  Over 34
##      252      619
```

```
##
## Under 34  Over 34
##    0.289    0.711
```



EDA for Culture_valued_bin

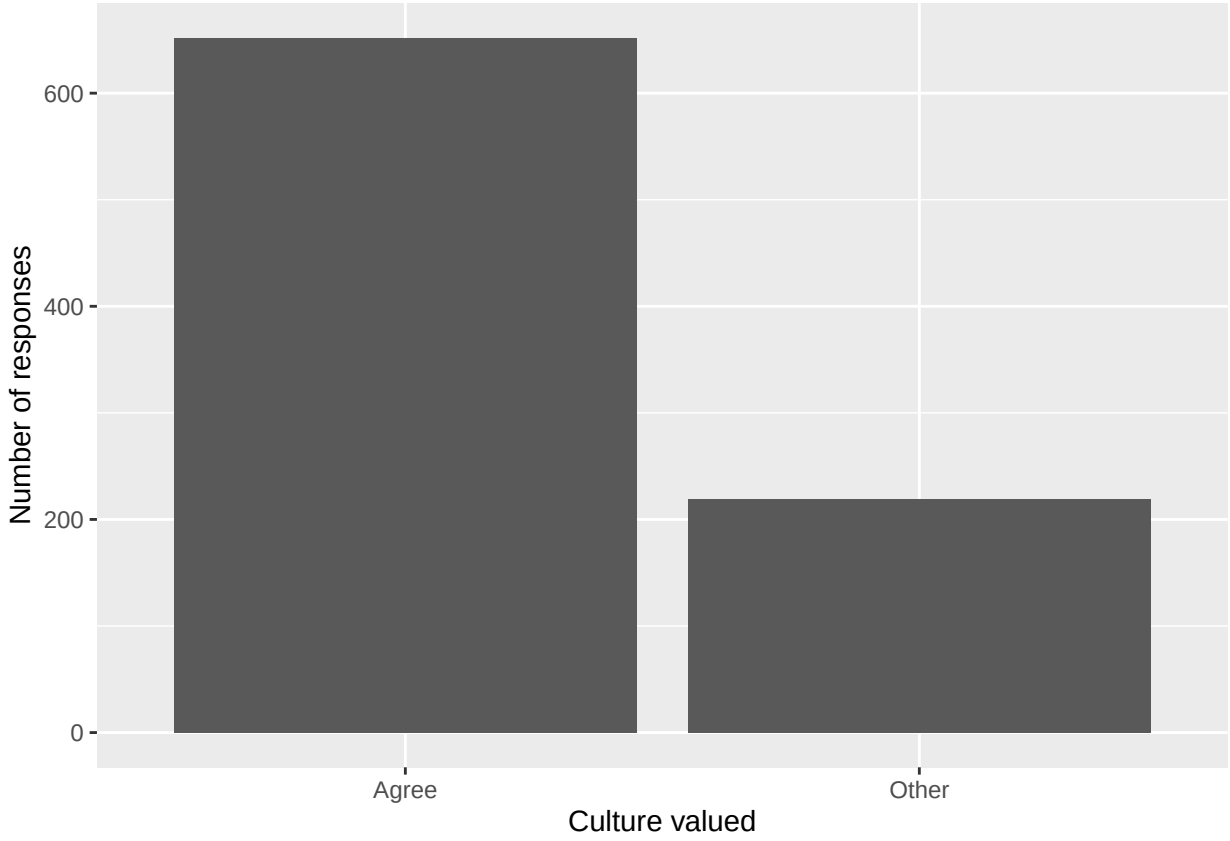
```
table2 <- table(survey_binary$Culture_valued_bin)
table2
```

```
##
## Agree Other
##   652   219
```

```
round(prop.table(table2), 3)
```

```
##
## Agree Other
## 0.749 0.251
```

```
ggplot(data = survey_binary) +
  geom_bar(aes(x = Culture_valued_bin)) +
  labs(x = 'Culture valued', y = 'Number of responses')
```



EDA for both variables

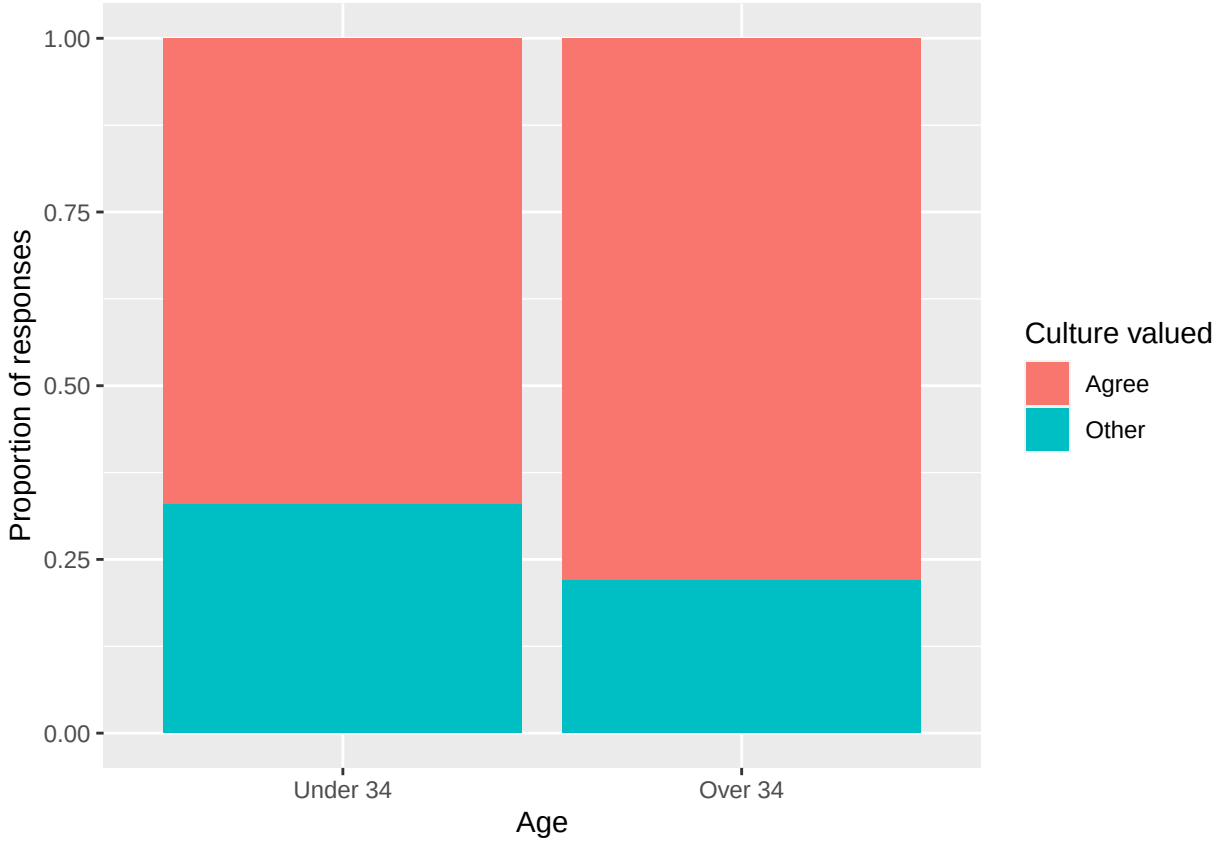
```
table3 <- table(survey_binary$Age_combine, survey_binary$Culture_valued_bin)
table3
```

```
##
##            Agree Other
##   Under 34   169    83
##   Over 34    483   136
```

```
round(prop.table(table3, 1),3)
```

```
##
##            Agree Other
##   Under 34 0.671 0.329
##   Over 34  0.780 0.220
```

```
ggplot(data = survey_binary) +
  geom_bar(mapping = aes(x = Age_combine, fill = Culture_valued_bin), position = "fill") +
  labs(x = 'Age', fill = 'Culture valued', y = "Proportion of responses")
```



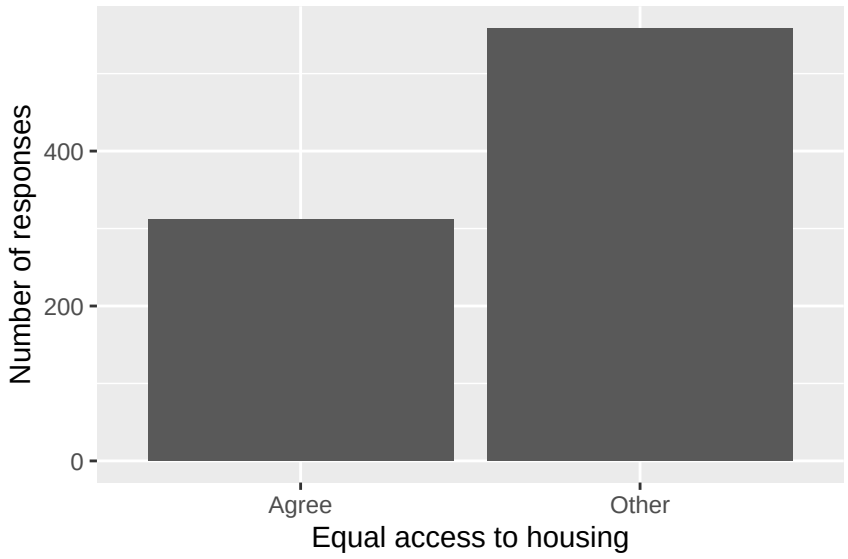
This data set shows the relationship of the age of Northfield residents and how much they feel culture is valued. From the bar plot we can see that the age group of 34 or older feel agree that Northfield values culture while those 34 or under disagree with it more. Overall there seems to also show that there are more 34 or older response which may show be a reason for the differences in the responses. But we can see that generally the age group of 34 or older and 34 or younger are not the same proportions and more 34 or younger residents think culture isn't valued while more 34 or older residents do feel like culture is valued.

> Question 2: *Is the proportion of residents who agree that there is access to housing for people of various races significantly larger than 50% (a significant majority)?*

EDA for Housing_bin

```
##
## Agree Other
##   312   559
```

```
##
##    Agree    Other
## 0.358209 0.641791
```



The bar plot shows us that more residents of Northfield feel that the housing offered in the area is not offered to various races. Though the survey originally has 5 categories “strongly disagree” “agree” “neutral” “disagree” and “strongly disagree” they’re split into 2 categories, of just “agree” and “disagree”, taking out the “neutral” answers. Through this bar plot and the data table we can conclude that there is a little over half of residents who don’t feel that the housing offered is open to vraious races but it isn’t a significant amount more who feel they disagree.

3. Hypothesis Testing Formally state the null and alternative hypotheses for testing your research questions in both words and symbols. Perform a randomized hypothesis test and simulate a null distribution to determine how unusual your samples are under the assumption the null hypothesis is correct. You may use the code from the Dolphin and ESP examples as a reference. Remember to include the following:

a. Clear statement of null and alternative hypothesis for both research questions in words and well-defined symbols.

- You may use LaTeX syntax in an RMarkdown document for symbols, or simplify things using underscores (e.g. p_d, p_c)

b. A histogram of your simulated null distribution with an indication of where your real observed sample falls.

c. A calculation of the p-value for your observed sample.

d. An interpretation of your p-value and a formal decision/conclusion of your hypothesis test, using the context of your original question.

Question 1: *Is the proportion of residents aged 34 year or younger who think Northfield values culture the same as the proportion of residents over 34 years old sharing the same opinion?*

a.

Null hypothesis: The proportion of 34 yr olds who think Northfield values culture is the same as the proportion of over 34 years olds.

Alternative hypothesis: The proportion of 34 yr olds who think Northfield values culture is not the same as the proportion of over 34 years olds.

$H_0 : \hat{p}_{under34} = \hat{p}_{over34}$, $H_A : \hat{p}_{under34} \neq \hat{p}_{over34}$

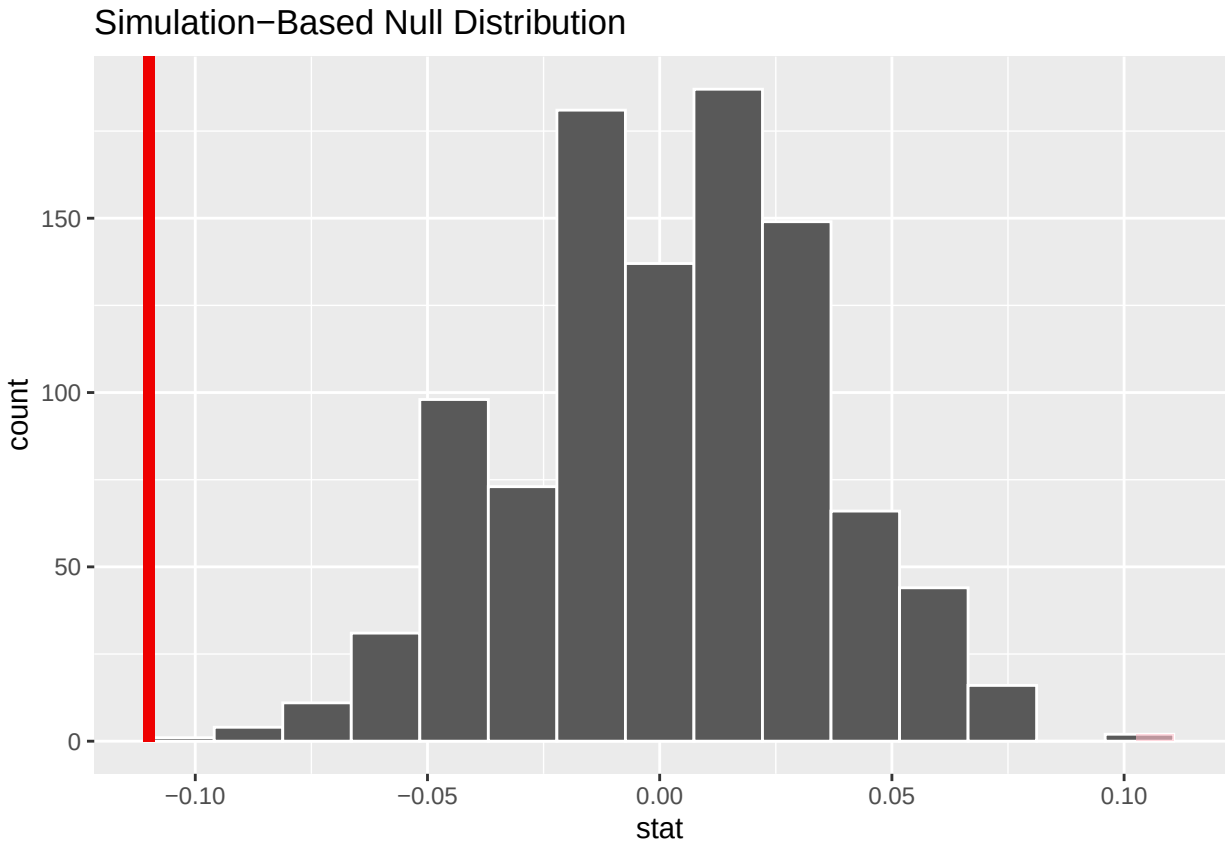
b.

```
## # A tibble: 871 x 2
##    Age_combine Culture_valued_bin
##    <fct>       <fct>
##  1 Over 34     Agree
##  2 Under 34    Other
##  3 Under 34    Other
##  4 Under 34    Other
##  5 Under 34    Agree
##  6 Under 34    Agree
##  7 Under 34    Agree
##  8 Under 34    Other
##  9 Over 34     Agree
## 10 Over 34     Agree
## # ... with 861 more rows
```

```
## `summarise()` has grouped output by 'Age_combine'. You can override using the `.groups` argument.
```

```
## # A tibble: 4 x 3
## # Groups:   Age_combine [2]
##   Age_combine Culture_valued_bin     n
##   <fct>       <fct>              <int>
## 1 Under 34    Agree                169
## 2 Under 34    Other                 83
## 3 Over 34     Agree                483
## 4 Over 34     Other                136
```

Proportion agreeing under 34: 170 / 253 = 0.672, Proportion agreeing 34 or older: 487 / 623 = 0.782, Difference = -0.110,



c.

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step. See
## `?get_p_value()` for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

d.

p value of 0 indicates that there is enough evidence to reject the null hypothesis. We can conclude that the proportion of 34 years old or under who think Northfield values culture is less then the proportion of people aged over 34 years.

> Question 2: *Is the proportion of residents who agree that there is access to housing for people of various races significantly larger than 50% (a significant majority)?*

a.

Null hypothesis: The proportion of residents who agree that there is access to housing for people of various races is not larger then 50%.

Alternative hypothesis: The proportion of residents who agree that there is access to housing for people of various races is significantly larger then 50%.

$H_0 : \hat{p} <= 0.5$, $H_A : \hat{p} > 0.5$

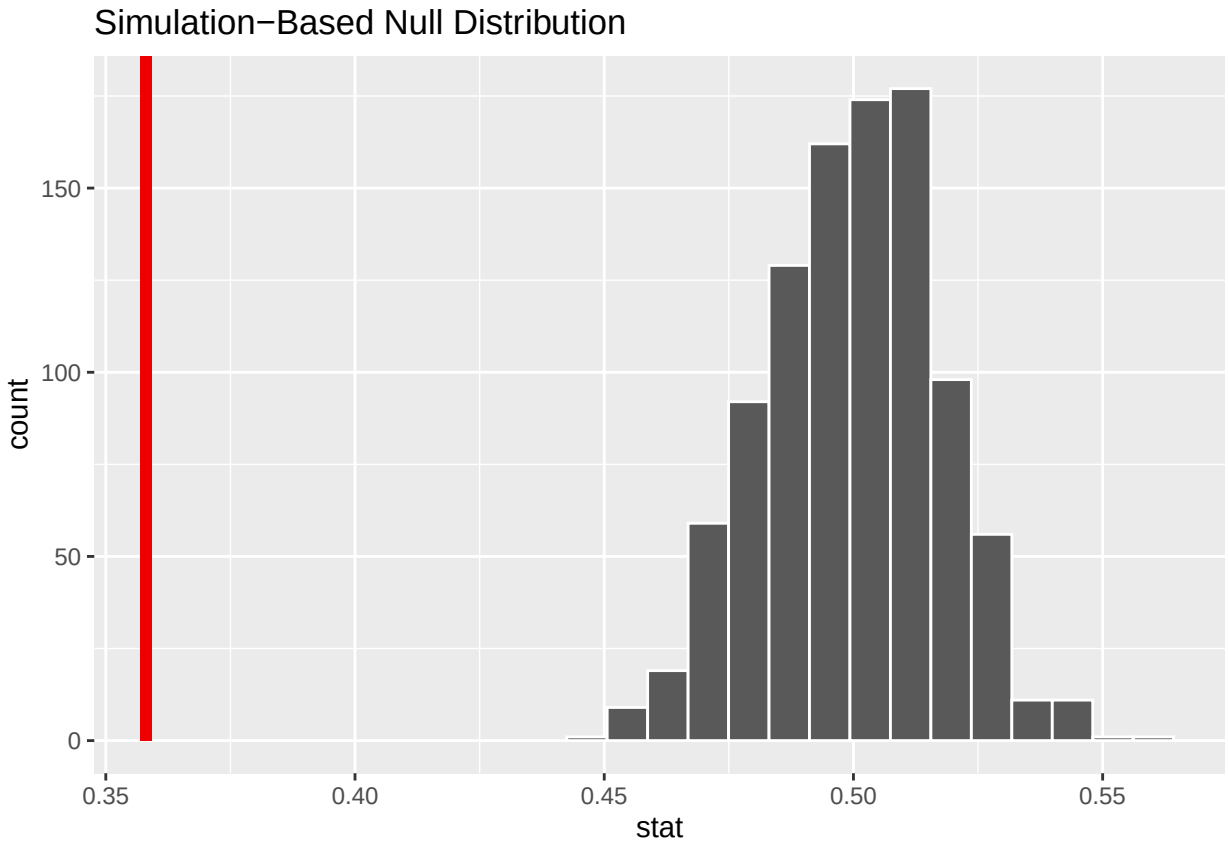
b.

```
## # A tibble: 871 x 1
##    Housing_bin
##    <fct>
##  1 Other
##  2 Other
##  3 Agree
##  4 Other
##  5 Agree
##  6 Agree
##  7 Agree
##  8 Other
##  9 Other
## 10 Other
## # ... with 861 more rows
```

```
## # A tibble: 2 x 2
##   Housing_bin     n
##   <fct>       <int>
## 1 Agree         312
## 2 Other         559
```

Proportion = 0.358

```
## [1] 0.358
```



c.

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
```

```
## approximation based on the number of `reps` chosen in the `generate()` step. See
## `?get_p_value()` for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

d.

For p value of 0 and significance level of 0.05 we can conclude that there is enough evidence to reject the null hypothesis. Thus, we conclude that The proportion of residents who agree that there is access to housing for people of various races is significantly larger then 50%.