NREEC Projects

Stat 212 Interim 2022

Background

This term, we will be consulting for the city of Northfield and the Northfield Racial and Ethnic Equity Collaborative (NREEC). Last fall the NREEC implemented a survey to the community to collect data on residents' perceptions and experiences. They have provided us with the results of that survey to summarize and analyze. The survey was administered through the following outlets:

- City of Northfield
- Northfield Public Schools
- Healthy Community Initiative
- Carleton College
- St. Olaf College
- Northfield Hospital + Clinics
- Rice County
- FiftyNorth
- Northfield Area Family YMCA
- Community Action Center
- Adult Basic Education
- Northfield Area Interfaith Association
- Northfield Shares
- Rice County Neighbors United
- Northfield Union of Youth
- Three Rivers Community Action
- Northfield Area Chamber of Commerce & Tourism
- Rice County Area United Way
- Allina Health
- Minnesota Humanities Center

Each organization was asked to send it out to their employees and clients. In addition there were posters put up downtown that had a QR code which people could use to access the survey. There were also several local churches that put it out via their bulletins/newsletters and had paper copies for parishioners to fill out. The Adult Basic Education department of the School District pushed it out to their students and there were a couple of high school classes that used it as part of their curriculum this fall. Information was also pushed out via social media posts - primarily by the city and Healthy Community Initiative.

Through four mini-projects, we will explore the data and try to illuminate any interesting patterns and trends while also running statistical analysis to determine which patterns may be representative of the entire population of Northfield, rather than just a product of sampling.

Mini-Projects

In small teams, you will be asked to address a different part of the statistical analysis each week. We will discuss some of the data together in class, and we will check in with our NREEC community partner several times to make sure we are on the right track. Here is the outline of our projects.

Week	Mini-Project Topic	Due Date
1/3-1/7	Exploratory Data Analysis	Friday 1/9
1/10-1/14	Tests for categorical variables	Friday 1/14
1/17-1/21	Tests for numeric variables	Friday 1/21
1/24-1/28	Regression and final report	Friday 1/28

Mini-Project 1 - Exploratory Data Analysis

1. The Data and Variables Before we jump into any visualization or summaries of the data, we need to understand the data as a whole. Knowing the intended population, sample size, and variables will help us to better formulate research questions we can answer statistically and make sure we know exactly to whom we can generalize our results. Use the information above, supplemental information provided on Moodle, and the NREEC survey data file and that is read in using the code below to answer each question. (Hint: The functions `dim()` and `name()`, among others may be useful here.)

```
## Use this blank R chunk to write any code you may need to answer questions below.
## You may also create new R chunks by pressing Ctrl+Alt+I, new chunks help to organize your analysis
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggmosaic)
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following objects are masked from 'package:dplyr':
##
```

```
##     count, do, tally

## The following object is masked from 'package:purrr':
##
##     cross

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
survey <- read_csv("~/Stats 212 I22/Class/Data/NREEC Equity Survey.csv")
```

```
## Rows: 1112 Columns: 26

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (25): Status, NF_Welcomes, NF_Attracts, NF_Respects, EEO, Housing, Welco...
## dbl  (1): ResponseID
##
## 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.
```

```
survey <- survey %>%
  mutate_if(is.character, as.factor)  # This code is important to change the categorical variable type
```

a. Understanding the sample and population.

   - What does each observation represent in this data? Each observation would be another response taken for the survey. Each of them has its own unique response ID.

   - What do you think the intended population is for this sample? We believe that the inteded population would be the entire population of the city of Northfield.

   - Do you think the sample is representative of the intended population? In what ways might it be biased or misrepresentative? This is a difficult question to answer as I believe it can go both ways, in terms of demographics they have a good representation of the population but looking at the locations that these surveys were implemented could lead to bias because they are all places that are known to have good expieriences.

b. What are the *dimensions* of this dataset? In other words, how many observations (rows) and how many variables (columns) have been collected.

```
dim(survey)
```

```
## [1] 1112   26
```

According to the code above the dataset has 26 variables and 1112 rows.

c. Create a **variable codebook** for the data. This is a summary of all variable names (as they appear in the raw data), what they represent, how they are measured, and what the units of measurement are. I've started the table below, add in additional lines for each remaining variable. Check the original survey on Moodle as a reference.

Variable name	Description	Type	Levels/Encoding
ResponseID	Unique survey ID	numeric	identifier
Status	Completion status	categorical	Partial/Complete
NF_welcomes	Northfield creates a welcoming environment for people of racial and ethnically diverse backgrounds.	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
NF_Attracts	Northfield attracts people of racial and ethnically diverse backgrounds	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
NF_Respects	Northfield residents demonstrate respect for residents of racial and ethnically diverse backgrounds	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
EEO	There are equal employment opportunities in Northfield for people of all raves and ethnicities	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
Housing	There is acces to housing in northfield for people of all raves and ethnicities	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
Welcome_important	It is important for people of all racial and ethnic backgrounds to feel welcome in Northfield	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
Feel_Welcome	I feel welcome in Northfield	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
Feel_connected	I feel connected with other residents in the Northfield community	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
Feel_safe	I feel safe in Northfield	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
Culture_valued	I belive my culture is values in Northfield	categorical (ordianl)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
Welcomed_business	I feel welcome in business establishments in Northfield	categorical (ordinal)	Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree
Exp_discrim_race	Have you experienced discrimination based on your race or ethnicity in Northfield in the past 12 months?	categorical	yes/no
Exp_discrim_language	Have you experienced discrimination based on your language or country of origin in Nortfield in the past 12 months?	ordinal	yes/no
Age	In which category is your age?	categorical (ordinal)	Under 18, 18-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, 65-74 years, 75 years or older
AmericanIndian_AlaskaNative	What is your race and/or ethnicity?	categorical	yes/no
AAPI	What is your race and/or ethnicity?	categorical	yes/no
Black	What is your race and/or ethnicity?	categorical	yes/no
White	What is your race and/or ethnicity?	categorical	yes/no
TwoMore	What is your race and/or ethnicity?	categorical	yes/no
Other	What is your race and/or ethnicity?	categorical	yes/no

Variable name	Description	Type	Levels/Encoding
CountryOrigin	Which best describes your country of origin, regardless of current nationality?	categorical	My country of origin is the USA/I am originally from another country
AnnualIncome	What is the current annual income in your houshold?	categorical (ordinal)	Under (25,000), (25,000-49,999), (50,000-74,999), (75,000-99,999), (100,000 or more)
IsWhite	What is your race and/or ethnicity?	categorical	yes/no

2. Variable Exploration All statistical projects start with an exploration of the data. We typically need to do this because we don't go into a project knowing exactly what patterns and trends to test for. And even if we have a good idea of what we want to test for, we could be missing some interesting relationships if we don't poke around the data first. Use the exploratory data analysis skills we've practiced to look for patterns/trends/relationships in individual variables as well as combinations of variables. Good EDA should always include the following:

- Numeric summaries or tables (depending on variable type) of the individual variables of interest
- Visual plot of the individual variable
- Numeric summaries or two-way tables of variable pairs (this is the most interesting part where we can look for relationships)
- Visual plots of the variable pairs (this is the most interesting part where we can look for relationships)

Never leave the summaries and plots by themselves. Include a brief description of the variable (remember to CUSS). Consider whether you see any interesting similarities or differences across groups. You do not need to perform any formal statistical analysis on this data. Your statements can rely on visual or casual inspection of the data, in other words, just things you notice about the plots or summaries. Provide some EDA for the following variables from this data:

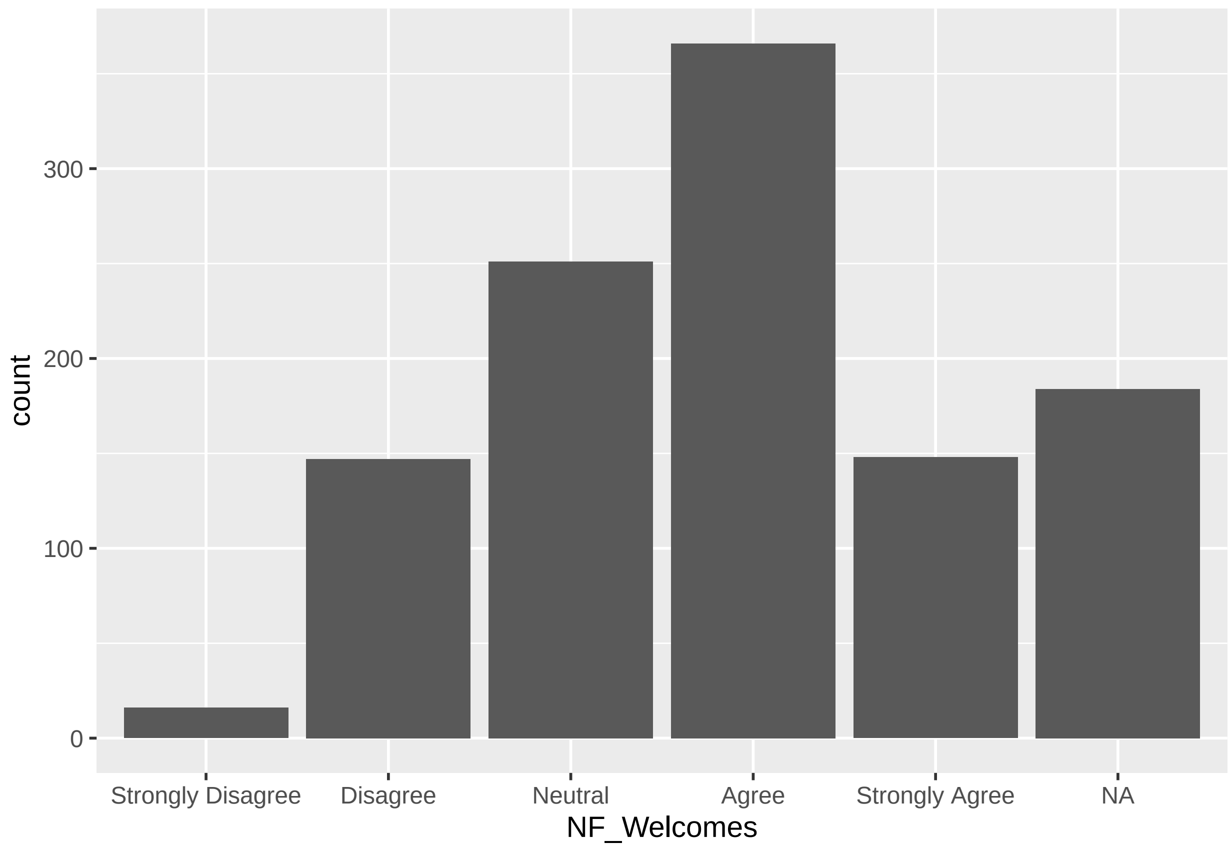
a. Summary of overall responses for how welcoming residents view Northfield (`NF_welcomes`). Summary of the ages of respondents (`Age`), how many fall into each age group, what are the percentages? Combine the two and provide a breakdown of `NF_welcomes` by Age groups.

```r
survey_fct <- survey %>%
  mutate(NF_Welcomes = fct_relevel(NF_Welcomes, "Strongly Disagree", "Disagree",
                                   "Neutral", "Agree", "Strongly Agree"),
         Age = fct_relevel(Age, "Under 18", "18-24 years", "25-34 years",
                           "35-44 years", "45-54 years", "55-64 years",
                           "65-74 years", "75 years or older"))

table(survey_fct$NF_Welcomes)
```

```
##
## Strongly Disagree          Disagree           Neutral             Agree
##                16               147               251               366
##    Strongly Agree
##               148
```

```r
ggplot(data = survey_fct, aes(x = NF_Welcomes)) +
  geom_histogram(stat = 'count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

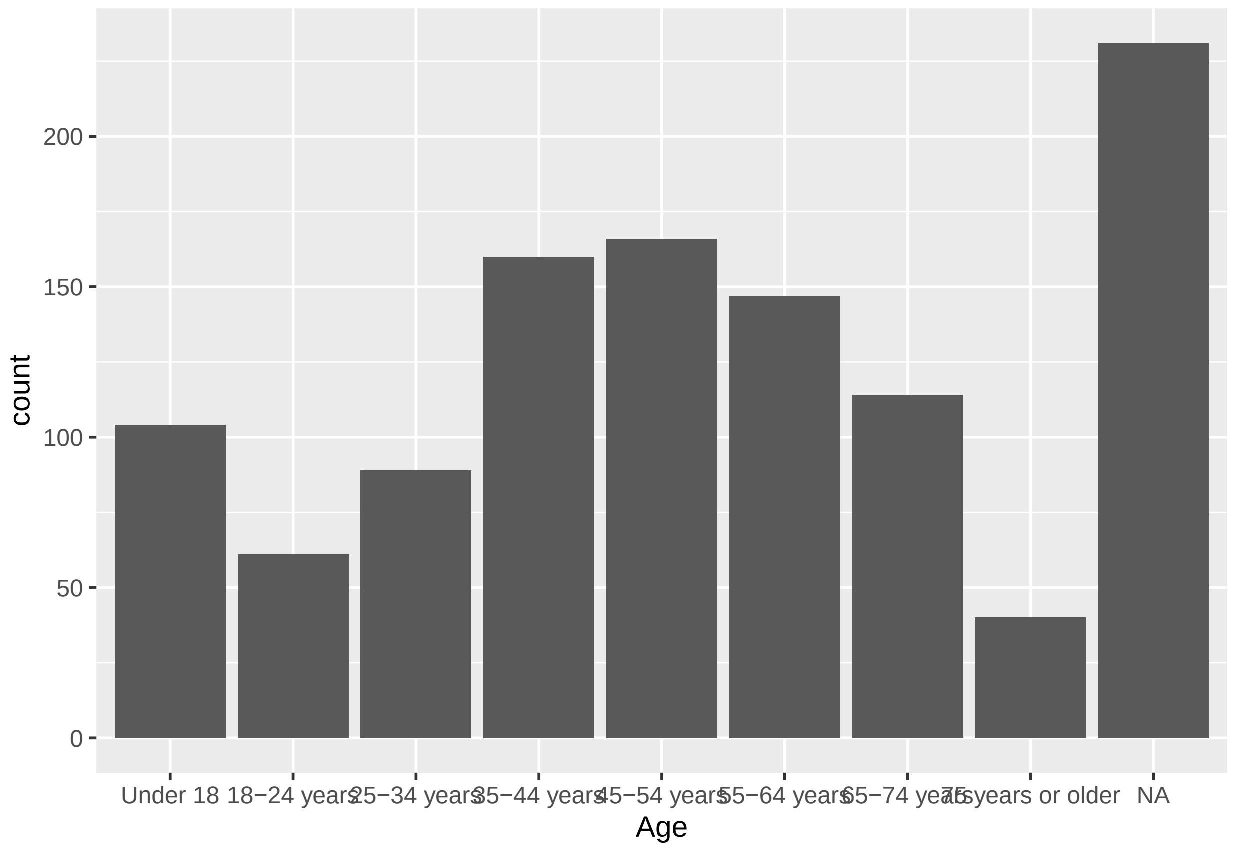


```
survey_fct %>%
  group_by(Age) %>%
  summarise(n = n(), perc = 100 * n / nrow(survey))
```

```
## # A tibble: 9 x 3
##   Age                    n  perc
##   <fct>              <int> <dbl>
## 1 Under 18             104  9.35
## 2 18-24 years           61  5.49
## 3 25-34 years           89  8.00
## 4 35-44 years          160 14.4
## 5 45-54 years          166 14.9
## 6 55-64 years          147 13.2
## 7 65-74 years          114 10.3
## 8 75 years or older     40  3.60
## 9 <NA>                 231 20.8
```

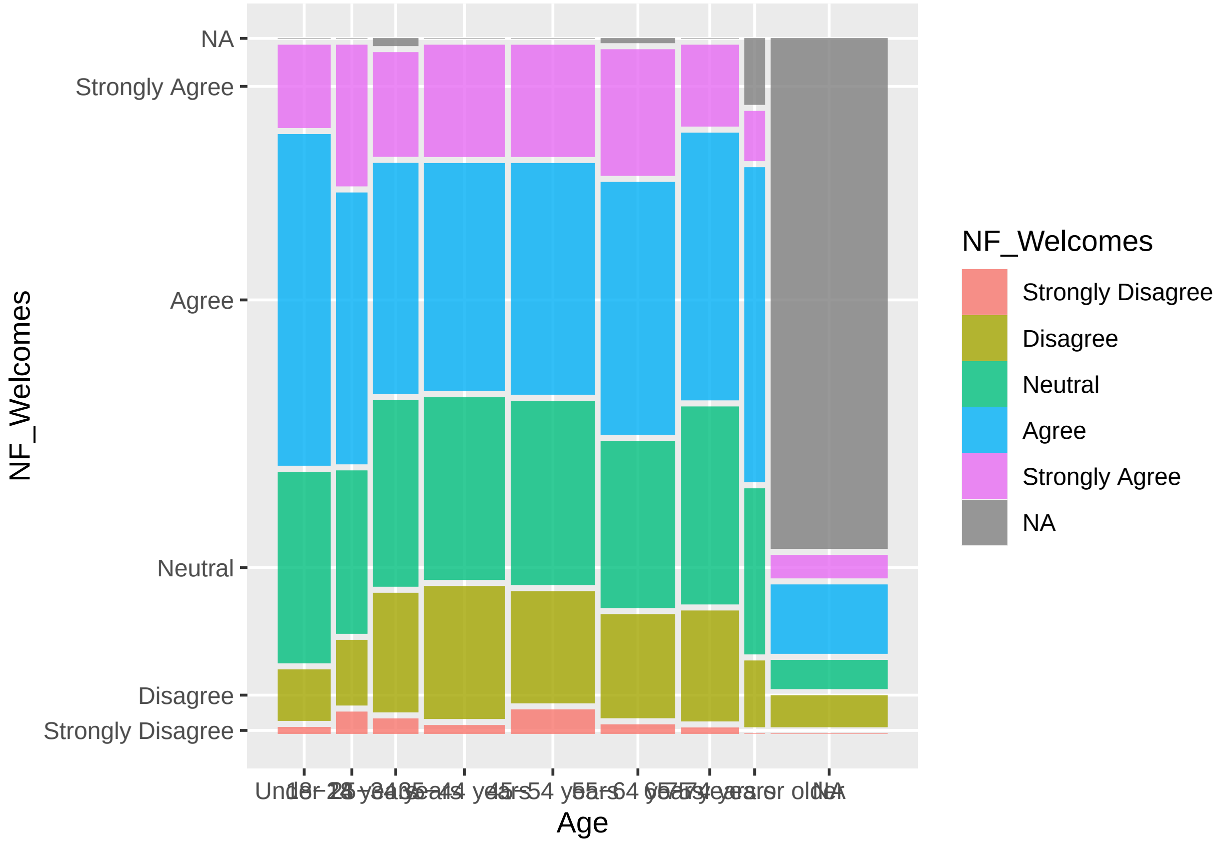
```
ggplot(data = survey_fct, aes(x = Age)) +
  geom_histogram(stat = 'count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```
survey_fct %>%
  select(ResponseID, NF_Welcomes, Age) %>%
  ggplot() +
  geom_mosaic(aes(x = product(Age), fill = NF_Welcomes))
```

```
## Warning: `unite_()` was deprecated in tidyr 1.2.0.
## Please use `unite()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```



As it can be seen above the histogram for the NF_Welcomes suggests that "Agree" has the most responses. We can also say that the responses were mostly well spread, the only exception is the "Strongly disagree" which has very few responses. Overall it can be noted that the responders mostly believe that Northfield is welcoming for people of different backgrounds, however that position is not that strong and it mostly varies between "Agree" and "Neutral". Also it should be noted that there were several missing answers for that question. We can see that the histogram of the age of responders is bimodal and has peaks at "under 18 years" and "45-54 years" age groups. It can be said that the age of the responders is quite well spread with the exception of people aged "18-24 years" and "75 years or older". However it is hard to make proper conclusions as there were a lot of missing responses for this question. The number of missing responses is actually higher than the number of people in any age group. The mosaic plot shows that the answer to the NF_Welcome did not differ that much across age groups. The only notable difference is that more people aged 25-74 tended to choose Disagree compared to other age groups. Some of the people who did not mention their age replied to the NA_Welcomes questions however for them the answers were equaly spread so we can say that even if we knew the ages of those people the mosaic plot is not that likely to have a big change.

b. Summary of overall responses for how connected residents feel (`Feel_connected`). Summary of white/non-white respondents (`IsWhite`), how many people identify as each, what are the percentages? Combine the two and provide a breakdown of `Feel_connected` by white/non-white residents.

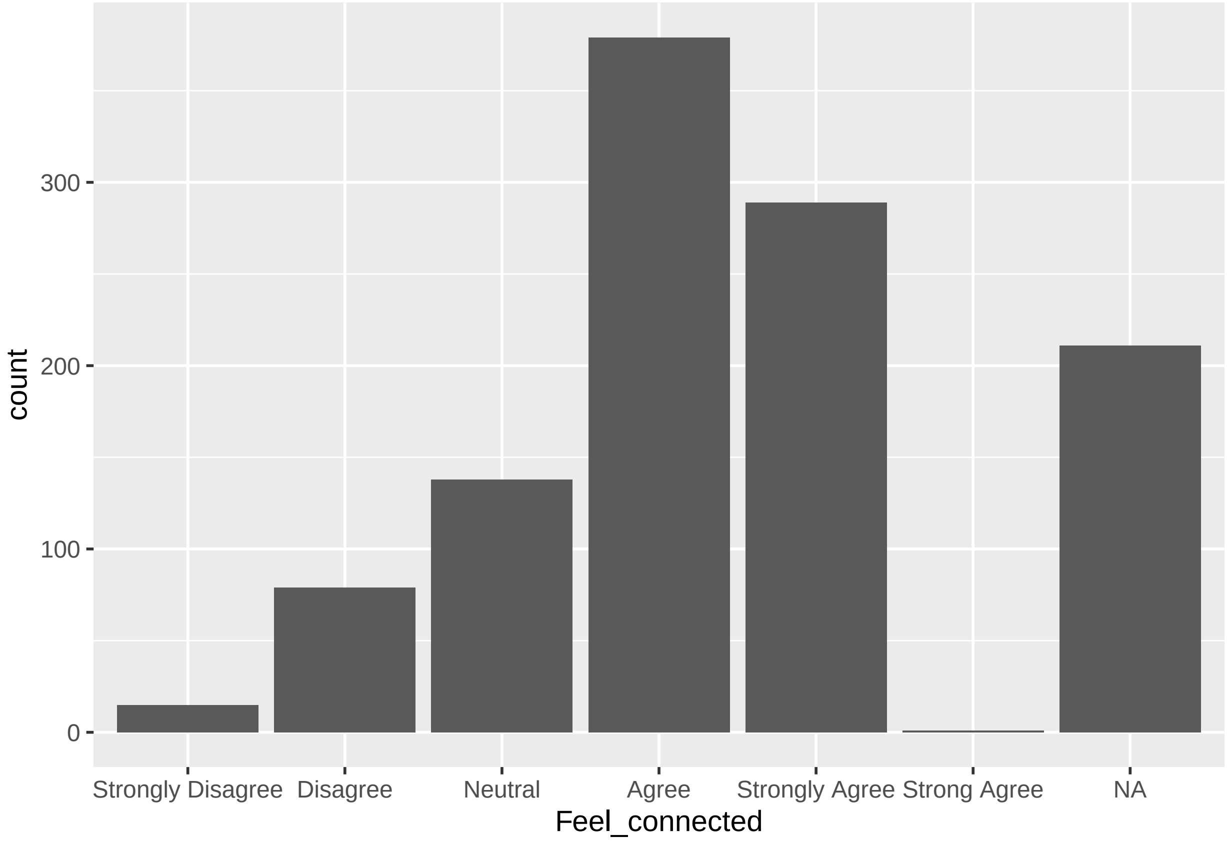
```
survey_fct <- survey %>%
  mutate(Feel_connected = fct_relevel(Feel_connected, "Strongly Disagree", "Disagree",
                                      "Neutral", "Agree", "Strongly Agree"),
         IsWhite = fct_relevel(IsWhite, "Non-white", "White"))

table(survey_fct$Feel_connected)
```

```
##
```

```
## Strongly Disagree           Disagree            Neutral              Agree
##                 15                 79                138                379
##     Strongly Agree      Strong Agree
##                289                  1
```

```
ggplot(data = survey_fct, aes(x = Feel_connected)) +
  geom_histogram(stat = 'count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

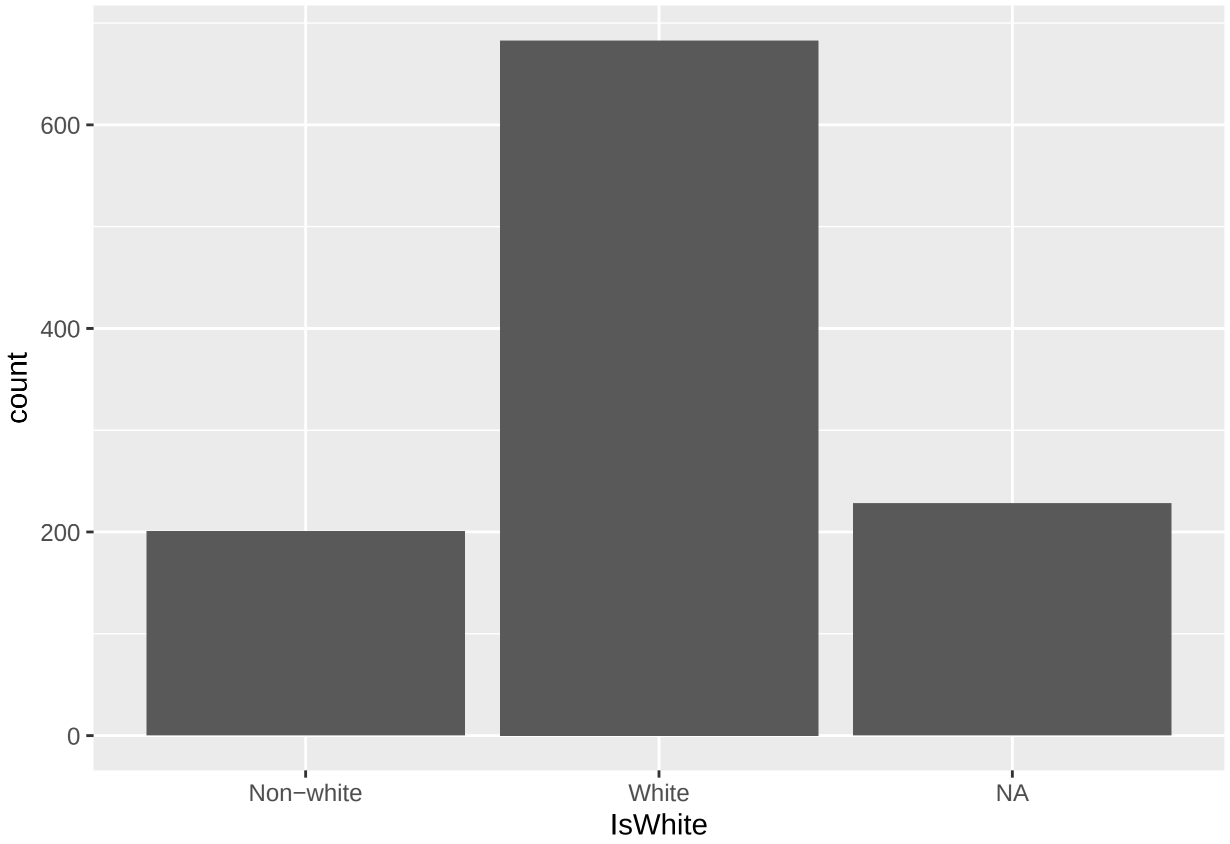


```
survey_fct %>%
  group_by(IsWhite) %>%
  summarise(n = n(), perc = 100 * n / nrow(survey))
```

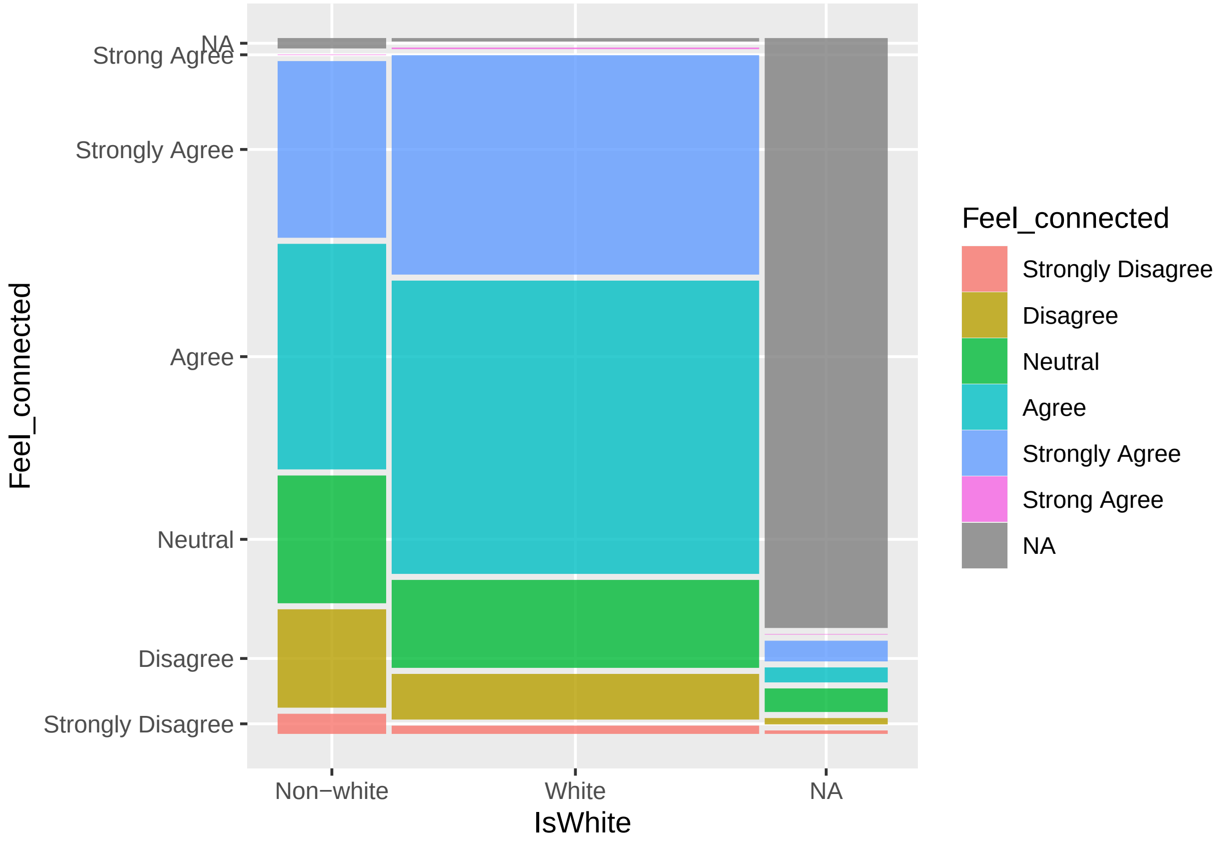
```
## # A tibble: 3 x 3
##   IsWhite       n  perc
##   <fct>     <int> <dbl>
## 1 Non-white   201  18.1
## 2 White       683  61.4
## 3 <NA>        228  20.5
```

```
ggplot(data = survey_fct, aes(x = IsWhite)) +
  geom_histogram(stat = 'count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```
survey_fct %>%
  select(ResponseID, Feel_connected, IsWhite) %>%
  ggplot() +
  geom_mosaic(aes(x = product(IsWhite), fill = Feel_connected))
```



As it can be seen from the histogram for the Feel_connected most of the responders choose “Agree”. The responses were not that spread. Majority of the responses choose “Agree” or “Strongly Agree”. There was one response of “Strong Agree” which is just a typing error and should be viewed as a “Strongly Agree” response. All of this suggests that mostly the responders feel quite connected. The answer to the question IsWhite has just two possible options which means it does not really make sense to discuss it as a histogram. We can only say that there are more than 3 times more people who identify as white compared to people who do not. We can also say that there are around 200 people who did not answer this question. Even if all of them are not white, the majority of the responders would still turn out to be white. Looking at the mosaic plot we can see that the response to the question Feel_connected actually differs for white and non-white responders. Non-white responders selected answer “Strongly disagree”, “Disagree” and “Neutral” motically more compared to white responders, while white responders had higher proportion of answers as “Agree” and “Strongly Agree”. Overall we can conclude that white residents feel slightly more connected to other residents compared to non-white residents.

c. Use the code below to “wrangle” the survey responses from ordinal to numeric. Then summarize the breakdowns for both parts above (welcomes by age and connected by white/non-white) using the survey question as a numeric variable with appropriate techniques.

```
# An example of wrangling with the NF_Respects variable
survey_subset <- survey %>%      # Never change the original data, create new versions to edit
  drop_na(NF_Welcomes, Feel_connected) %>%       # Removes missing responses
  mutate(NF_Welcomes = fct_relevel(NF_Welcomes, "Strongly Disagree", "Disagree",
                                  "Neutral", "Agree", "Strongly Agree"),   # Reorders the responses
         NF_Welcomes_num = recode(NF_Welcomes,
                                 "Strongly Disagree" = -2,
                                 "Disagree" = -1,
                                 "Neutral" = 0,
                                 "Agree" = 1,
                                 "Strongly Agree" = 2), # Create new numeric var from categorical

         Feel_connected = fct_relevel(Feel_connected, "Strongly Disagree", "Disagree",
                                  "Neutral", "Agree", "Strongly Agree"), # Reorders the responses
         Feel_connected_num = recode(Feel_connected,
                                 "Strongly Disagree" = -2,
                                 "Disagree" = -1,
                                 "Neutral" = 0,
                                 "Agree" = 1,
                                 "Strongly Agree" = 2), # Create new numeric var from categorical
         Age = fct_relevel(Age, "Under 18", "18-24 years", "25-34 years",
                           "35-44 years", "45-54 years", "55-64 years",
                           "65-74 years", "75 years or older"),
         IsWhite = fct_relevel(IsWhite, "Non-white", "White")
         )
```

```
favstats(survey_subset$NF_Welcomes_num)
```

```
##  min Q1 median Q3 max     mean        sd   n missing
##   -2  0      1  1   2 0.522905 0.9951113 895       0
```

```
favstats(~NF_Welcomes_num | Age, data = survey_subset) %>%
  mutate(perc = 100 * n / nrow(survey_subset))
```

```
##                 Age min Q1 median Q3 max      mean        sd   n missing
## 1          Under 18  -2  0      1  1   2 0.6538462 0.8333956 104       0
## 2       18-24 years  -2  0      1  1   2 0.6721311 1.0282884  61       0
## 3       25-34 years  -2  0      1  1   2 0.4431818 1.0378797  88       0
## 4       35-44 years  -2  0      1  1   2 0.4562500 1.0330810 160       0
## 5       45-54 years  -2  0      1  1   2 0.4457831 1.0702932 166       0
## 6       55-64 years  -2  0      1  1   2 0.5902778 1.0132907 144       0
## 7       65-74 years  -2  0      1  1   2 0.4649123 0.9424110 114       0
## 8 75 years or older  -1  0      1  1   2 0.5588235 0.8235612  34       0
##        perc
## 1 11.620112
## 2  6.815642
## 3  9.832402
## 4 17.877095
## 5 18.547486
## 6 16.089385
## 7 12.737430
## 8  3.798883
```

```
favstats(survey_subset$Feel_connected_num)
```

```
##  min Q1 median Q3 max      mean        sd   n missing
##   -2  0      1  2   2 0.9451902 0.9838075 894       1
```

```
favstats(~ Feel_connected_num | IsWhite, data = survey_subset) %>%
  mutate(perc = 100 * n / nrow(survey_subset))
```

```
##     IsWhite min Q1 median Q3 max      mean       sd   n missing     perc
## 1 Non-white  -2  0      1  2   2 0.6818182 1.119632 198       0 22.12291
## 2     White  -2  1      1  2   2 1.0340741 0.918178 675       1 75.41899
```

As it can be seen in the code above the answers to this questions changed from categorical to numerical which means we can now use their centers. For NF_Welcomes we can see that the mean of the responses is 0.52 and the median is 1, is suggests that the data is not symmetrical. That actually matches with the conclusion that we gave before. The average of the responds conclude that people have a somewhat positive feeling weather Northfield welcomes residents with different backgrounds, however it is not that strong and just slightly better then neutral. From the table with age groups we can say that younger responders (24 and under) felt that Northfield is welcoming to other residents compared to older responders. We can also see that our conclusion that residents aged 25-74 do not feel that Northfield is welcoming city was true as well.

We can see that residents responded way more positively for the Feel_connected questions. The mean is 0.95 and the median is 1 which suggests that the answers to this question were actually very symmetrical. We can also say that the residents mostly feel connected to other residents which we already concluded before. We can also see that the median for non-white and white responders was 1 and the difference in mean is not significant (3.68 and 4.04) which suggests that our conclusion was again correct and the slight difference in responses exists: White residents feel a bit more connected to other compared to non-white residents.

3. Create your own research question One important skill for any researcher/statistician is the ability to create a *useful* research question that can be answered with the available data. For example, "Does a feeling of connection differ among white and non-white residents in Northfield?" is an okay question to ask with this data. A better question is, "*How much* does the feeling of connection differ among white and non-white residents in Northfield and is that difference significant among the population?" Try to be specific, and try to quantify (if possible) what you would like to answer. Good research questions usually include more than one variable as well, comparing groups or looking for patterns as values change. "What's the average income in Northfield?" is not a great research question.

a. Create two research questions based on variable combinations not used in part 2. One should focus on a question in the **Views** section of the survey and one should focus on a question in the **Experiences** section of the survey (original survey posted on Moodle). A good idea is to compare responses by one of the demographic variables, but this is not necessary. For example you could summarize responses to an experiences question broken down by a views question. Be sure to:

   - Clearly state your specific question
   - Provide some summary statistics for the variables you are using
   - Plot the variables and comment on any patterns (or lack thereof) you notice

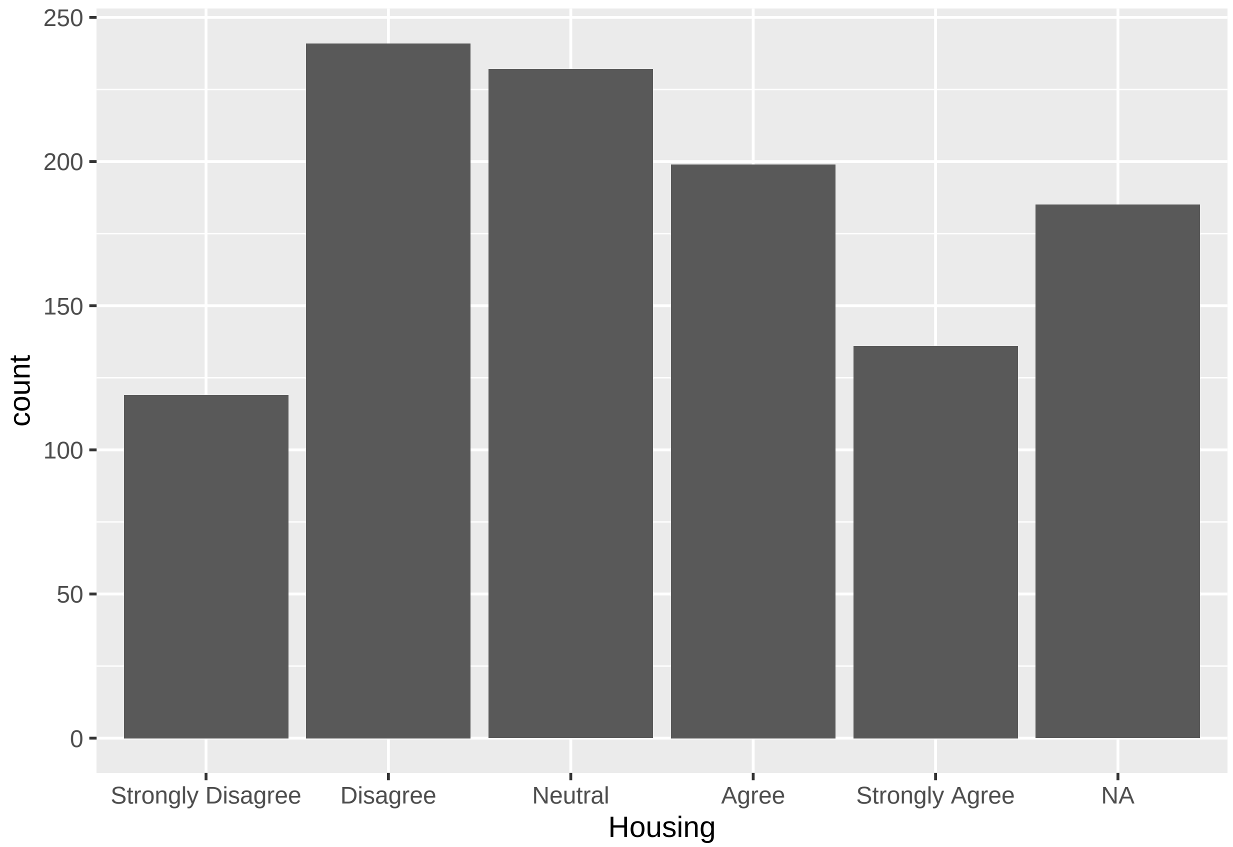
#1) How much does the view that there are equal howsing opportunities for people of all races and ethnicities differ between white and non-white residents.

```
survey_fct <- survey %>%
  mutate(Housing = fct_relevel(Housing, "Strongly Disagree", "Disagree",
                                   "Neutral", "Agree", "Strongly Agree"),
         IsWhite = fct_relevel(IsWhite, "Non-white", "White"))


table(survey_fct$Housing)
```

```
##
## Strongly Disagree          Disagree           Neutral             Agree
##               119               241               232               199
##    Strongly Agree
##               136
```

```
ggplot(data = survey_fct, aes(x = Housing)) +
  geom_histogram(stat = 'count')
```

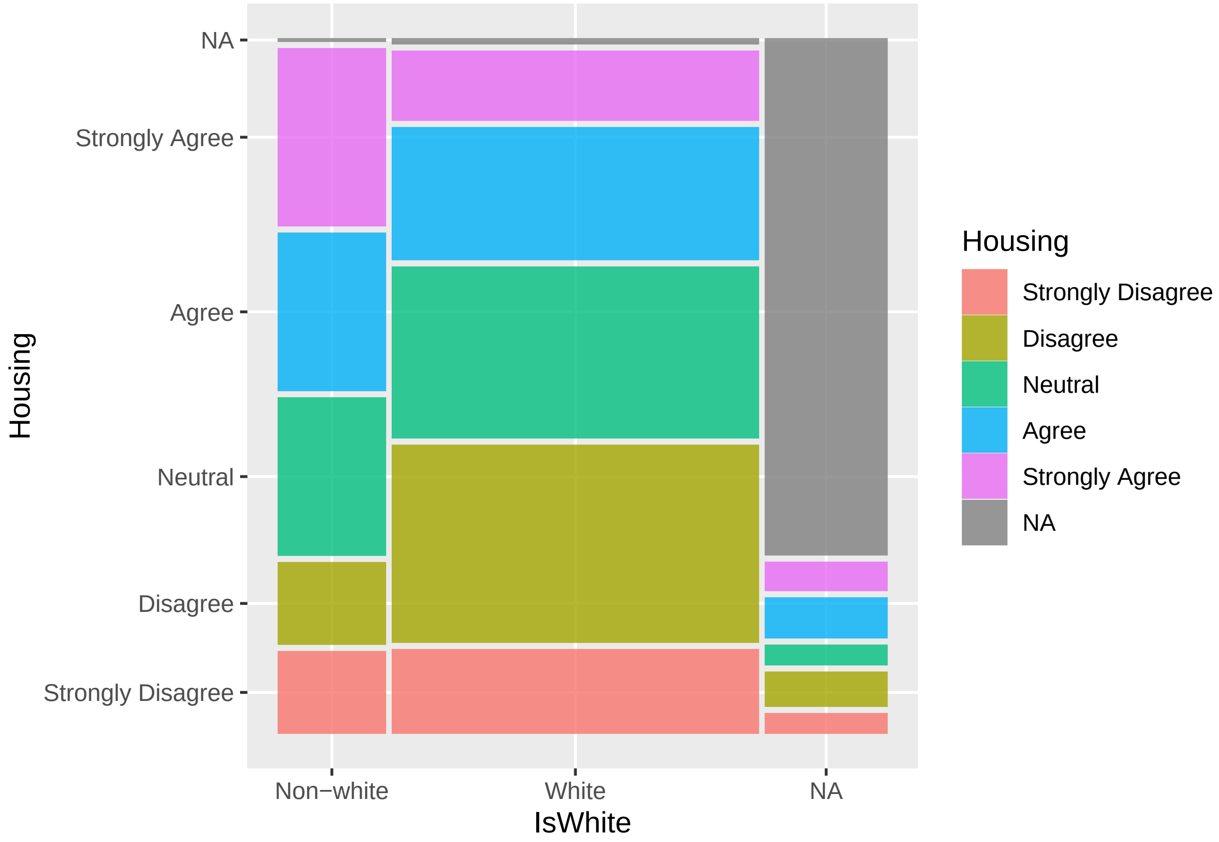
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```
survey_fct %>%
  group_by(IsWhite) %>%
  summarise(n = n(), perc = 100 * n / nrow(survey))
```

```
## # A tibble: 3 x 3
##   IsWhite       n  perc
##   <fct>     <int> <dbl>
## 1 Non-white   201  18.1
## 2 White       683  61.4
## 3 <NA>        228  20.5
```

```
survey_fct %>%
  select(ResponseID, Housing, IsWhite) %>%
  ggplot() +
  geom_mosaic(aes(x = product(IsWhite), fill = Housing))
```



We can see that the responses for question about equal housing is quite well spread except the “Strongly disagree” option, however it seems like the center of the responses is between “Disagree” and “Neutral”. There are several missing answers for this question which means that we can not say much about the true story of how residents perceive the equality of housing. We can however see in the mosaic plot that white residents disagree almost twice more (proportion-wise) compared to non-white residents. We can conclude that non-white residents perceive housing in Northfield to be quite equal in comparison to white residents.

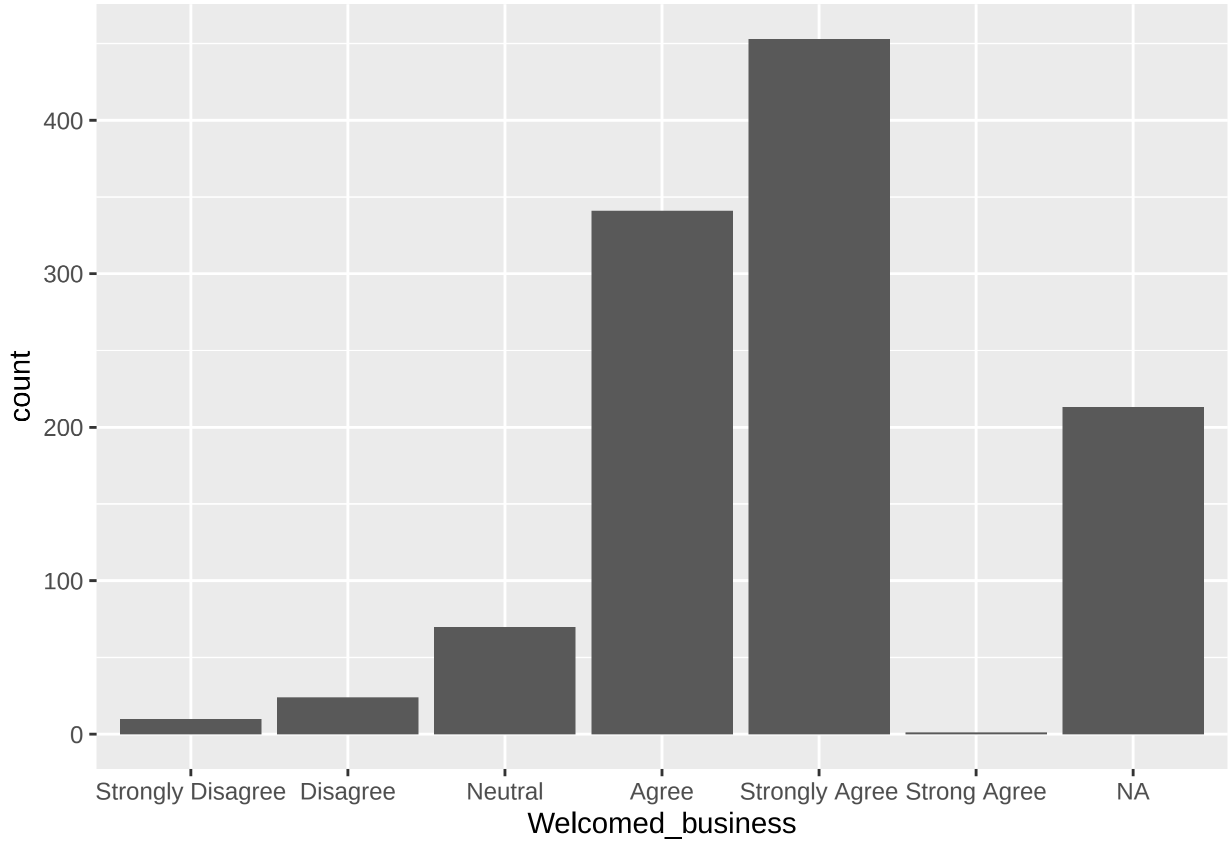
#2) Does the experience with feeling welcomed in business establishemnts differ among people with different socioeconomic status (annual income of family).

```
survey_fct <- survey %>%
  mutate(Welcomed_business = fct_relevel(Welcomed_business, "Strongly Disagree", "Disagree",
                                  "Neutral", "Agree", "Strongly Agree"),
         AnnualIncome = fct_relevel(AnnualIncome, "Under $25,000",
                                   "$25,000-$49,999", "$50,000-$74,999",
                                   "$75,000-$99,999", "$100,000 or more"))


table(survey_fct$Welcomed_business)
```

```
##
## Strongly Disagree          Disagree           Neutral             Agree
##                10                24                70               341
##    Strongly Agree     Strong Agree
##               453                 1
```

```
ggplot(data = survey_fct, aes(x = Welcomed_business)) +
  geom_histogram(stat = 'count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

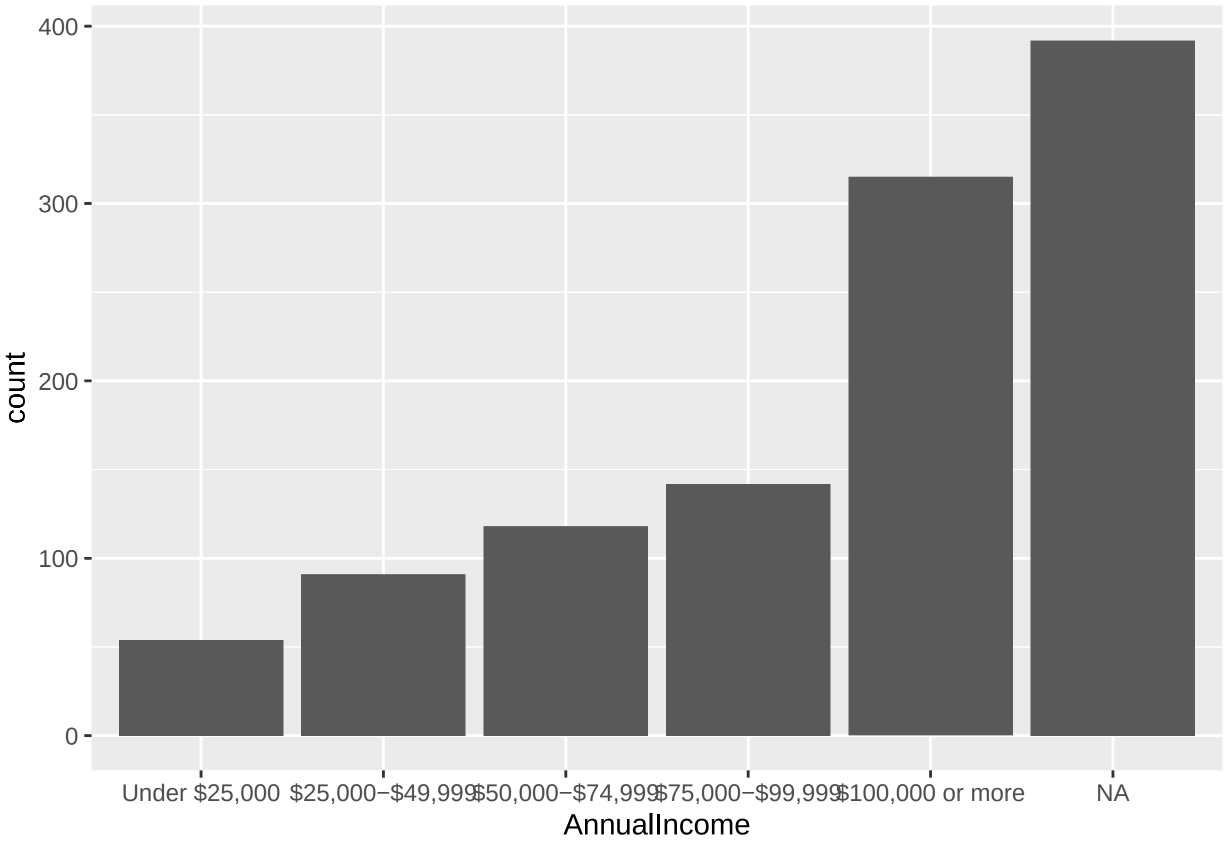


```
survey_fct %>%
  group_by(AnnualIncome) %>%
  summarise(n = n(), perc = 100 * n / nrow(survey))
```

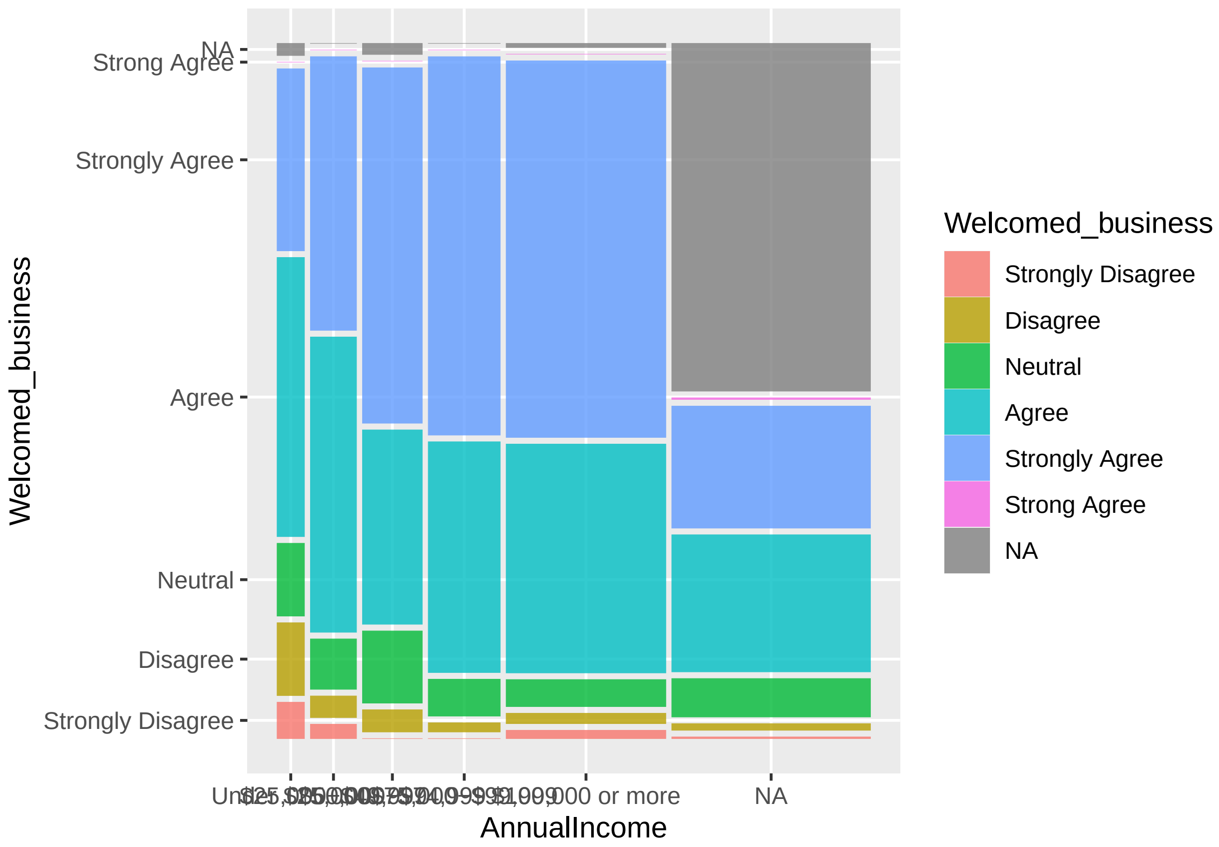
```
## # A tibble: 6 x 3
##   AnnualIncome         n  perc
##   <fct>            <int> <dbl>
## 1 Under $25,000       54  4.86
## 2 $25,000-$49,999     91  8.18
## 3 $50,000-$74,999    118 10.6
## 4 $75,000-$99,999    142 12.8
## 5 $100,000 or more   315 28.3
## 6 <NA>               392 35.3
```

```
ggplot(data = survey_fct, aes(x = AnnualIncome)) +
  geom_histogram(stat = 'count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```
survey_fct %>%
  select(ResponseID, Welcomed_business, AnnualIncome) %>%
  ggplot() +
  geom_mosaic(aes(x = product(AnnualIncome), fill = Welcomed_business))
```



We can see in the annual income for the majority of responders is “$100,000 or more” and the responses are not spread at all. Also we can see that the vast majority of responders feel very welcomed in business. There are arlmost no strongly disagree, disagree and neutral responses compared to agree and strongly agree. Enen though in the mosaic plot we can see that there is a strong relationship between family income and the feeling of being welcomed in business facilities in Northfield. The number of storngly disagree and disagree responses decrease as the family inclomde dicreases while the agree and strongly agree responses increase.

b. Identify any potential limitations in the data for answering your research question. Is there any additional data you would like to see collected in the future to better explain/explore possible trends?

1) It could be helpful to get another survey question that would better breakdown what type of housing these people are in but obviously not too personal of a question. Also the answer to this question in survey provides only the opinion of different residents about the overall situation in Northfield. It would be helpful to get data about the actual situation of housing depending on the race and ethnicity of residents which would show the real picture weather there are equal housing opportunities in Northfield.

2) A limitation could be the very few number of responses that consider that the responders do not feel welcomed in business facilities. Also there are several residents who did not report their annual income. The situation can have a change if we actually know the family income of all residents who choose not to answer the question.

Keeping track of your work

All of your work, code, and written summaries should be completed in an R Markdown file. Your group should create a folder with your names in the Project > Mini Project 1 folder that contains your work. Keep track of any filtering, variable mutating/creation, and other wrangling that you perform. I will have access to this folder so if you have questions on code, be sure to 1) tell me/screenshot the exact error you are receiving and 2) tell me where the issue is occurring in your code (the line number is usually fine). Create clear sections in your R Markdown document for each task you are performing and use the # symbol in your code chunks to comment on what your code is intended to do. Make sure that you explain enough so that anyone who has taken Stat 212 but knows nothing about the project could open your file and understand what you are doing.