Week 3: Practical Exercise - Model Answers

Calum Webb

27/09/2022

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

In R Markdown you need to style your writing in plain text - it will then read this plain text and format it however you want it to be formatted. For example, headings always begin with a # sign like above. The more # signs, the smaller the heading. **Bold** text is text with two asterixes or **underscores** around it and *italicised* text is text with only one asterix or *underscore* around it.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## Warning: package 'tidyr' was built under R version 4.0.5
## Warning: package 'readr' was built under R version 4.0.5
## Warning: package 'dplyr' was built under R version 4.0.5
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(rcompanion)
library(janitor)
```

```
##
## Attaching package: 'janitor'
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
ssa <- read_rds("ssa_tidied.rds")
ssa
```

```
## # A tibble: 1,409 x 25
##    pserial rsex    rage incsour  leftr~1 libauth emplo~2 emplo~3 party~4 gov_t~5
##      <dbl> <fct>  <dbl> <fct>      <dbl>   <dbl> <fct>   <fct>   <chr>   <ord>
##  1  151343 Male      31 Wages/p~     1.2    2.83 3. Emp~ 1. Emp~ Green ~ Never ~
```

```
##  2  151856 Female    41 Wages/p~     1.8    3.5  3. Emp~ 1. Emp~ Green ~ Someti~
##  3  151537 Male      53 Wages/p~     2      4    3. Emp~ 1. Emp~ Scotti~ Never ~
##  4  151369 Male      39 State B~     1.8    4.67 5. Per~ <NA>    Green ~ Never ~
##  5  152010 Female    43 State B~     2.4    3.83 7. Hom~ <NA>    <NA>    Never ~
##  6  151793 Female    60 Wages/p~     1.8    3.83 3. Emp~ 1. Emp~ <NA>    <NA>
##  7  151589 Female    86 State B~     1      3.67 6. Ret~ <NA>    Scotti~ Never ~
##  8  151792 Male      49 Wages/p~     2.4    4.33 3. Emp~ 1. Emp~ Scotti~ Someti~
##  9  152332 Male      43 State B~     2      4.17 4. Une~ 0. Une~ <NA>    Never ~
## 10  152417 Female    53 Wages/p~     2.4    3.67 3. Emp~ 1. Emp~ <NA>    Never ~
## # ... with 1,399 more rows, 15 more variables: tax_view <ord>, eu_policy <ord>,
## #   ev_cameron <dbl>, ev_salmond <dbl>, knowind <ord>, likely_vote <dbl>,
## #   referend_vote <fct>, union_benef <chr>, refvote_dum <fct>,
## #   uk_scot_spend_fair <ord>, scot_identity <dbl>, highest_quals <fct>,
## #   dole_fct <ord>, imig_fct <ord>, imig_fct_3 <ord>, and abbreviated variable
## #   names 1: leftrigh, 2: employment, 3: employmentdum, 4: party_allg,
## #   5: gov_trust
```

You can run the code in the chunks just like you would in a normal R script: just highlight it and click run or press control/command and enter while your typing cursor is on it.

Introduction

This week we are going to be taking a break from writing `R` code itself and will be focusing on interpreting the code that is output.

Below, you will find snippets of `R` code that have been written by a researcher studying the associations between age, education, views on immigration, and Scottish identity.

Only one problem — the researcher (like your module leader) — has totally neglected to add comments about what they have done and what the interpretation of their results should be! Wow, they are so lazy. It is *definitely* not something I would do and leave as a problem for future me.

It is your job to run and interpret the code left behind and write below what the research results are and their interpretation. You should write a brief description of everything for a general audience, even if it is something that should feel self-explanatory like a data visualisation. Not only is it generally good practice to put your interpretation of a graph into words, it makes your research more accessible for people with visual impairments.

Before you are given each of the steps, below is a description of the data being used and the variables in the data.

Description of data

This data is an extract taken from the 2014 Scottish Social Attitudes Survey (https://natcen.ac.uk/our-research/research/scottish-social-attitudes/) conducted by NatCen. It includes variables about attitudes towards welfare, Scottish Independence, Scottish identity, and political affiliation. This extract is taken from the materials in Fogarty (2019). It includes a somewhat small but representative sample of between around 1,000 and 2,000 people aged over 16 and living in Scotland.

Description of variables

The following variables are included in the data extract. You don't need to commit these to memory, but you may need to come back and refer to this list to understand and interpret the results below.

- **pserial** = respondent serial number (identifier)
- **rsex** = respondent's sex
- **rage** = respondent's age

- **incsour** = respondent's income source
- **leftrigh** = respondent's position on a left-right political ideology scale
  - 1 = Most Left Leaning, 5 = More Right Leaning
- **libauth** = respondent's position on a libertarian-authoritarian political ideology scale
  - 1 = Most Libertarian Leaning, 5 = Most Authoritarian Leaning
- **employment** = respondent's employment status
- **employmentdum** = simplified employment status (employed or unemployed)
- **party_allg** = respondent's political party allegiance
- **gov_trust** = the extent to which the respondent says they trust the **British** government
- **tax_view** = respondent's views on taxation and taxes
- **eu_policy** = respondent's views on the UK's relationship with the EU
- **ev_cameron** = respondent's evaluation of then Prime Minister David Cameron
  - 0 to 10 with 0 = "very bad" and 10 = "very good"
- **ev_salmond** = respondent's evaluation of then First Minister Alex Salmond
  - 0 to 10 with 0 = "very bad" and 10 = "very good"
- **knowind** = how much the respondent reports they feel they know about the Scottish independence vote
- **likely_vote** = respondent's self reported likelihood of voting in the independence referendum
  - 0 to 10 with 0 = "very unlikely" and 10 = "very likely"
- **referend_vote** = respondent's expected vote in the independence referendum
- **union_benef** = respondent's view on who benefits from the Union (between Scotland and England)
- **refvote_dum** = simplified expected referendum vote (those who answered either Yes or No)
- **uk_scot_spend_fair** = respondent's view on the share of UK spending that Scotland receives
- **scot_identity** = respondent's reported strength of Scottish identity
  - 1 to 7 with 1 = "very weak" and 7 = "very strong"
- **highest_quals** = highest education qualification attained
- **dole_fct** = respondent's view on whether benefits are too high, too low, or neither too high or too low
- **imig_fct** = respondent's view on how immigration to Scotland should change in the future
- **imig_fct_3** = simplified view on how immigration should change in the future

Aim

Identify the associations between age (`rage`), education (`highest_quals`), views on immigration (`imig_fct_3`), and strength of Scottish identity (`scot_identity`)?

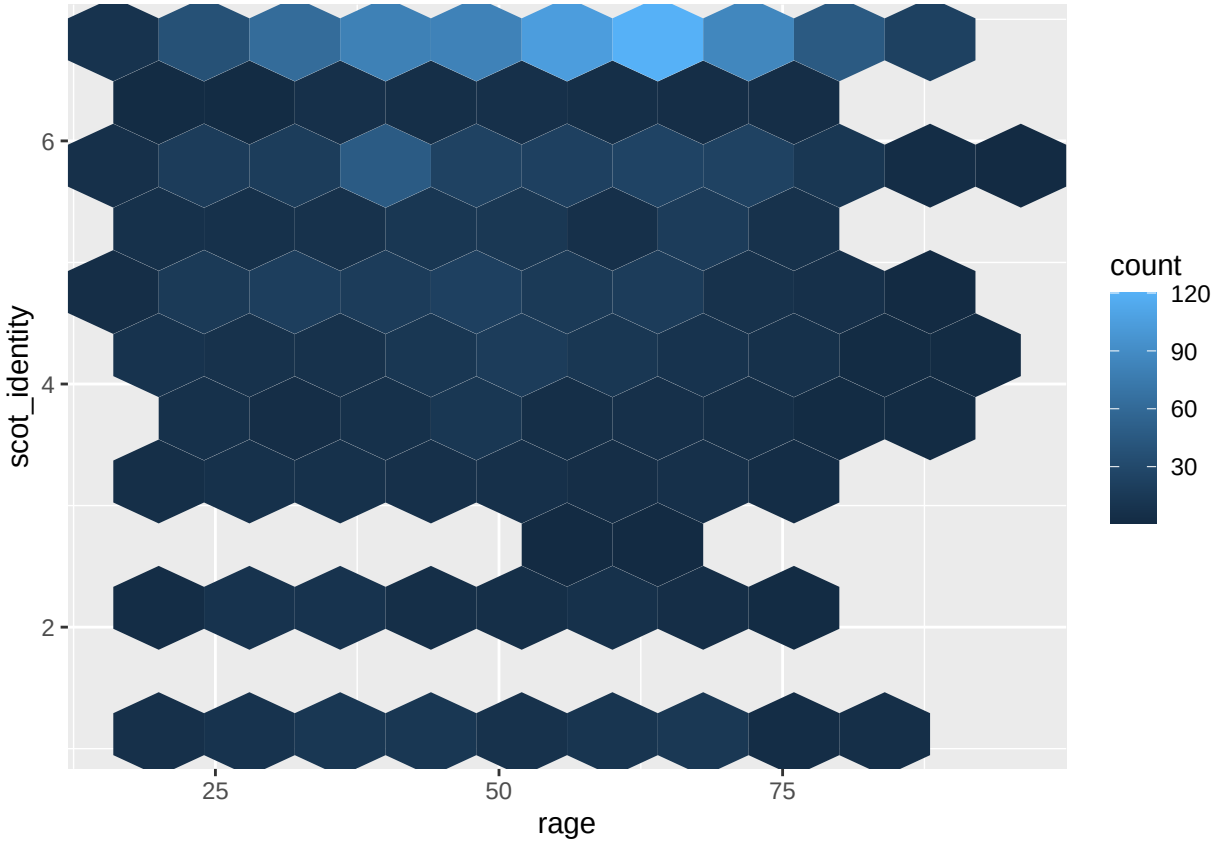
Research questions

- Does strength of Scottish identity vary by age?
- Are people with different levels of qualifications more or less likely to identify strongly with their nationality?
- Are people with lower educational qualifications more likely to hold the view that immigration should be reduced?

Research Question 1: Does strength of Scottish identity vary by age?

```
ssa %>%
  ggplot() +
    geom_hex(
      aes(x = rage, y = scot_identity), bins = 10
    )
```

```
## Warning: Removed 3 rows containing non-finite values (stat_binhex).
```




```
cor(ssa$rage, ssa$scot_identity,
    use = "complete.obs",
    method = "pearson")
```

```
## [1] 0.1301157
```

```
cor(ssa$rage, ssa$scot_identity,
    use = "complete.obs",
    method = "spearman")
```

```
## [1] 0.1504902
```

Describe how the association between strength of Scottish national identity and age was visualised, and what the visualisation suggests.

The association between strength of Scottish national identity and age was visualised using a hexbin plot. The hexbin plot suggests that, overall, the majority of respondents in the sample had strong (greater than 4 out of 7) affiliation with their Scottish identity.

This affiliation appeared to be stronger in older respondents, with the largest single group of respondents with strong Scottish identity being aged around 60-65 years old.

Describe the strength of the association in terms of correlation - state whether you would describe the association as very weak, weak, moderate, strong, or very strong and explain the direction of the association (e.g. as X increases, Y increases/decreases).

Both descriptive statistics of correlation suggest that as age increases, strength of Scottish identity also tends to increase, however this correlation would be best described as weak or negligible (Pearson's R =

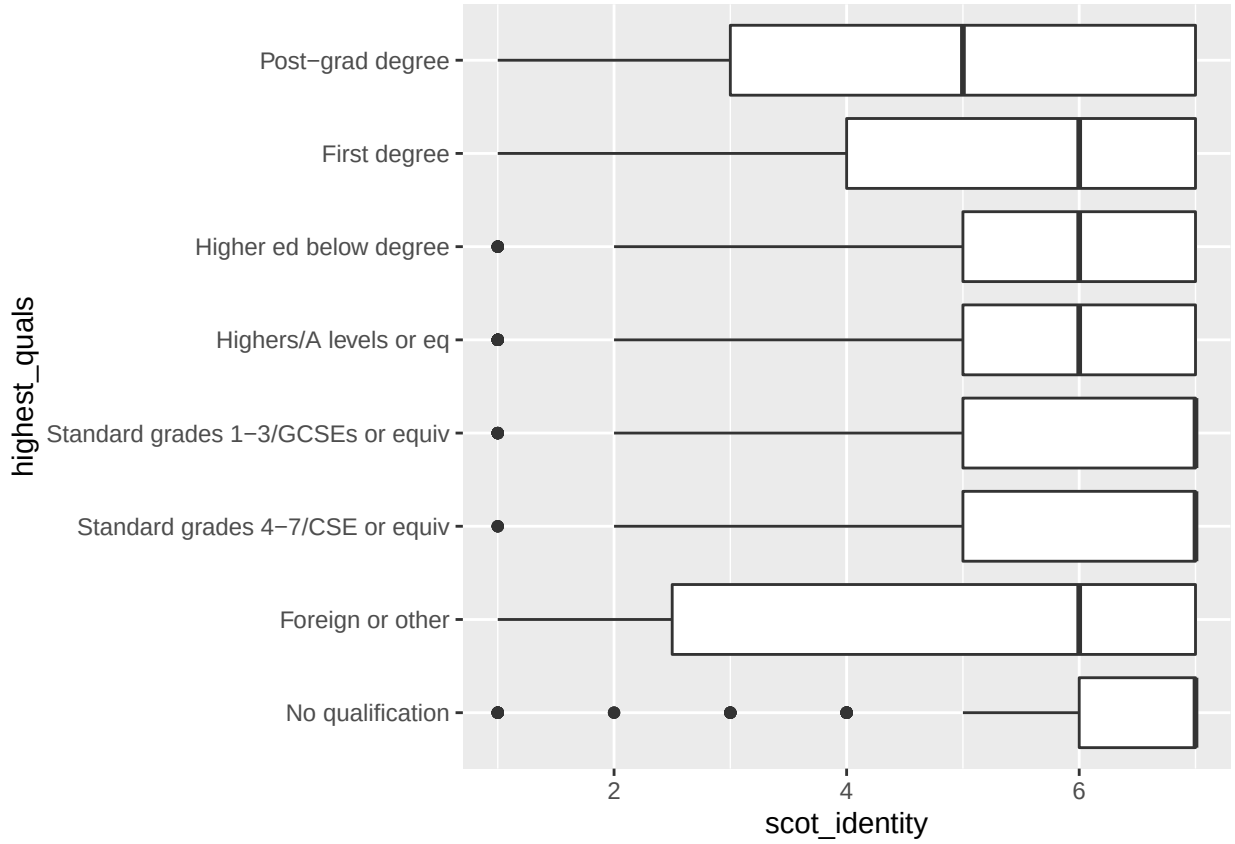
0.13, Spearman's Rho = 0.15).

As the strength of Scottish identity is an ordinal variable rather than a continuous variable, a more appropriate description of the association may be a contingency table or an analysis of the mean and median differences in age.

Research Question 2: Are people with different levels of qualifications more or less likely to identify strongly with their nationality?

```
ssa %>%
  filter(!is.na(highest_quals)) %>% # Remove missing responses
  ggplot() +
    geom_boxplot(
      aes(x = highest_quals, y = scot_identity)
    ) +
  coord_flip() # Flip the coordinates to make labels easier to read
```

```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
```




```
library(modeest)
```

```
## Registered S3 method overwritten by 'rmutil':
##   method         from
##   print.response httr
```

```
ssa %>%
  filter(!is.na(highest_quals)) %>% # Remove missing responses
```

```
  group_by(highest_quals) %>%
  summarise(
    mean_scotid = mean(scot_identity, na.rm = TRUE),
    median_scotid = median(scot_identity, na.rm = TRUE),
    mfv_scotid = mfv(scot_identity, na_rm = TRUE),
    sample_size = n()
  )
```

```
## # A tibble: 8 x 5
##   highest_quals                       mean_scotid median_scotid mfv_sco~1 sampl~2
##   <fct>                                     <dbl>         <dbl>     <dbl>   <int>
## 1 No qualification                           6.11             7         7     265
## 2 Foreign or other                           4.82             6         7      22
## 3 Standard grades 4-7/CSE or equiv           5.86             7         7      91
## 4 Standard grades 1-3/GCSEs or equiv         5.78             7         7     228
## 5 Highers/A levels or eq                     5.60             6         7     268
## 6 Higher ed below degree                     5.51             6         7     209
## 7 First degree                               5.18             6         7     238
## 8 Post-grad degree                           4.72             5         7      83
## # ... with abbreviated variable names 1: mfv_scotid, 2: sample_size
```

Describe how strength of Scottish identity varies by highest qualification held. Make reference to descriptive statistics.

The boxplot shows that there is some degree of variation in strength of Scottish identity associated with the highest qualification the respondent holds. The greatest amount of variation was found in the post-grad degree holders and in those holding foreign or other qualifications. The least amount of variation was found in respondents who hold no formal qualifications.

Those with no formal qualifications had the highest mean Scottish identity score (6.11), with the second and third highest mean scores being among those whose highest qualifications were Standard grades 1-3 (5.78) or 4-7 (5.86). These three groups also had the highest medians of 7.

Those with post-graduate degrees or with foreign or other qualifications had the lowest mean Scottish identity scores of 4.72 and 4.82 respectively. Further, the median score for post-graduate degree holders was 5 (out of a maximum of seven).

Despite these differences respondents, on average, had a strong sense of Scottish identity, with all groups having a mean or median score over four (on a range between 1 and 7). Further, the most common (modal) response for all groups was 7, indicating very a very strong affiliation with their Scottish national identity.

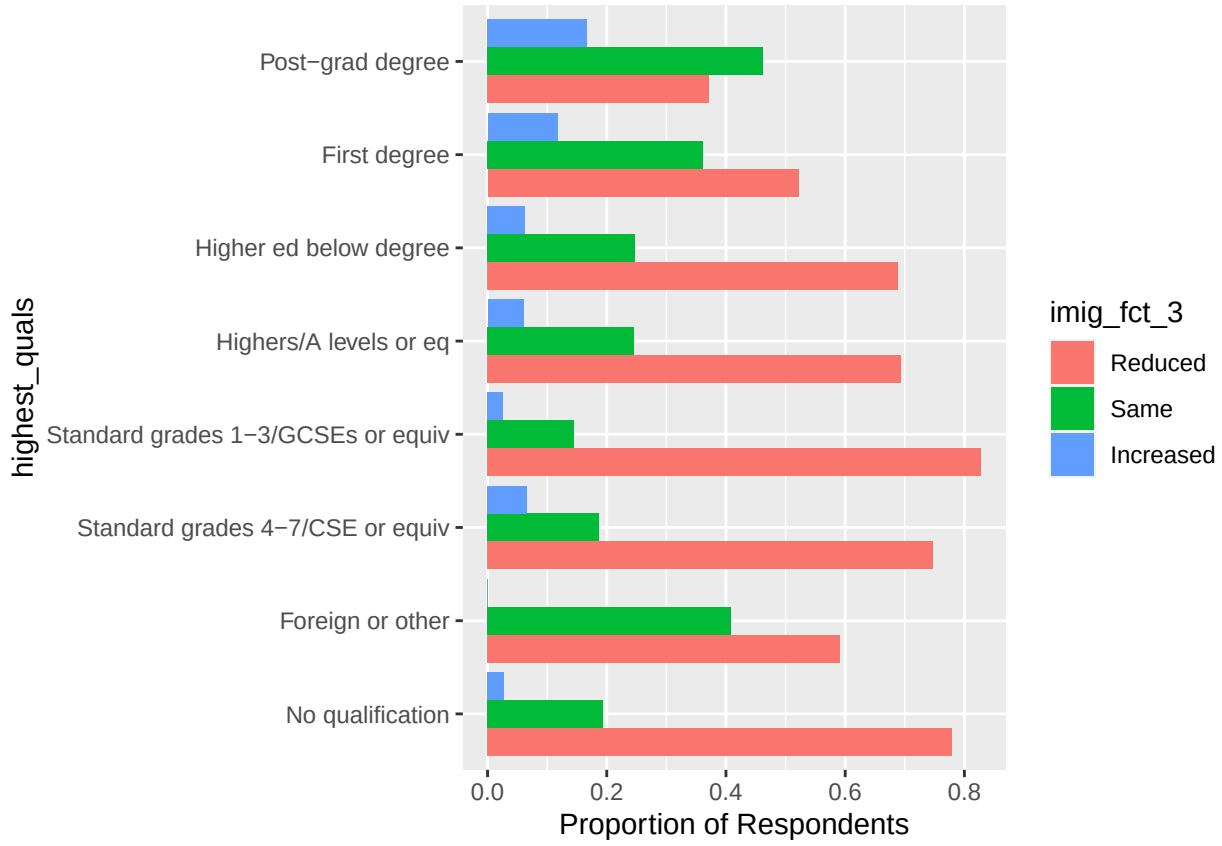
What other type of visualisation could the researcher have used here?

The researcher could also have used a Ridgeplot to visualise the results.

Research Question 3: Are people with lower educational qualifications more likely to hold the view that immigration should be reduced?

```
ssa %>%
  filter(!is.na(highest_quals) & !is.na(imig_fct_3)) %>% # Remove missing responses
  tabyl(highest_quals, imig_fct_3) %>% # create a contingency table
  adorn_percentages("row") %>% # convert to proportions within education categories
  pivot_longer(-1, names_to = "imig_fct_3") %>% # convert to long format
  mutate(imig_fct_3 = factor(imig_fct_3, levels = c("Reduced", "Same", "Increased"))) %>% # add order t
  ggplot() +
```

```
    geom_col(
      aes(x = highest_quals, fill = imig_fct_3, y = value),
      position = "dodge" # creates side-by-side bar chart
    ) +
  coord_flip() + # Flip the coordinates to make labels easier to read
  ylab("Proportion of Respondents")
```




```
library(janitor)

ssa %>%
  tabyl(highest_quals, imig_fct_3)
```

```
##                      highest_quals Reduced Same Increased NA_
##                   No qualification     202   50         7   6
##                   Foreign or other      13    9         0   0
##    Standard grades 4-7/CSE or equiv      68   17         6   0
##  Standard grades 1-3/GCSEs or equiv     188   33         6   1
##               Highers/A levels or eq     184   65        16   3
##               Higher ed below degree     142   51        13   3
##                         First degree     120   83        27   8
##                   Post-grad degree      29   36        13   5
##                                <NA>       2    2         1   0
```

```
library(janitor)

ssa %>%
  filter(!is.na(highest_quals) & !is.na(imig_fct_3)) %>% # remove missing
```

```
  tabyl(highest_quals, imig_fct_3) %>% # create cross table
  adorn_totals(where = "col") %>% # add totals
  adorn_percentages(denominator = "row") %>% # create percentages
  adorn_pct_formatting() # format percentages
```

```
##                       highest_quals Reduced  Same Increased  Total
##                    No qualification   78.0% 19.3%      2.7% 100.0%
##                    Foreign or other   59.1% 40.9%      0.0% 100.0%
##    Standard grades 4-7/CSE or equiv   74.7% 18.7%      6.6% 100.0%
##  Standard grades 1-3/GCSEs or equiv   82.8% 14.5%      2.6% 100.0%
##              Highers/A levels or eq   69.4% 24.5%      6.0% 100.0%
##              Higher ed below degree   68.9% 24.8%      6.3% 100.0%
##                         First degree   52.2% 36.1%     11.7% 100.0%
##                    Post-grad degree   37.2% 46.2%     16.7% 100.0%
```

Describe how attiudes towards immigration differ based on educational qualifications held; which types of qualifications are most strongly associated with views that immigration should be reduced and which are most strongly associated with views that it should be increased? Remember to use both percentages and frequency counts (raw numbers of people in the sample) in your description.

Respondents with Scottish Standard grades 1-3 (82.8%, N = 188), Standard grades 4-7 (74.7%, N = 68), and no qualifications (78%, N = 202), were the most likely to hold the view that immigration to Scotland should be reduced. Respondents holding post-graduate degrees and first degrees were the most likely to respond that immigration should be increased, with 16.7% (N = 13) and 11.7% (N = 27) reporting they held this view. Similarly, post-graduate degree holders (46.2%, N = 36), foreign or other qualifications holders (40.9%, N = 9), and first degree holders (36.1%, N = 83) were the most likely to respond that they felt immigration should remain the same as it currently is.

```
cramerV(ssa$highest_quals, ssa$imig_fct_3)
```

```
## Cramer V
##   0.1603
```

Interpret the Cramer's V statistic: does it suggest the association between education and attitudes towards immigration is strong, moderate, weak, or negligible and why?

The Cramer's V statistic was 0.1603; the contingency table from Cohen (1988) indicates that 2 degrees of freedom is appropriate for assessing the strength of the effect. Because V was between 0.07 and 0.21 the effect would generally be classed as weak under these guidelines. This statistic does not tell us the direction of the association, just that attitudes towards immigration are not equally distributed across education levels.

What other descriptive statistic could the researchers have used to describe the relationship between the two variables? Would there be any benefits to using the alternative statistic you suggested compared to Cramer's V, and if so, what are they?

The researchers could have used Spearman's Rho (Rank Order Correlation Coefficient) to describe the strength of the relationship between these two variables because both variables can be considered ordinal. The benefit of using Spearman's Rho would be that the researchers could infer a direction to the association - they would be able to say how strongly pro-immigration attitudes increased or decreased as educational qualifications increased.

Week 3 Challenge

Come up with two additional research questions (they can be about anything, they don't have to just be about Scottish identity or immigration) you could investigate using this dataset. Produce relevant visualisations

and descriptive statistics for exploring the associations that can answer your research questions. Remember to describe your findings!